

# Benchmarking EngGPT2-16B-A3B against Comparable Italian & International Open-source LLMs

Andrea Sassella<sup>2\*</sup>, Andrea Chizzola<sup>1\*</sup>, Tommaso Bianchi<sup>1\*</sup>, Luca Alessandrelli<sup>1\*</sup>, Mark J. Carman<sup>2</sup>

<sup>1</sup>AIRIC, Politecnico di Milano

<sup>2</sup>DEIB, Politecnico di Milano

andrea.sassella@polimi.it, andrea.chizzola@polimi.it, tommaso.bianchi@polimi.it,  
luca.alessandrelli@polimi.it, mark.carman@polimi.it

## Abstract

This report benchmarks the performance of ENGINEERING Ingegneria Informatica S.p.A.’s EngGPT2MoE-16B-A3B Large Language Model (LLM), a 16 billion parameter Mixture of Experts (MoE) model with 3 billion active parameters. Performance is investigated across a wide variety of representative LLM benchmarks<sup>1</sup>, including both international and Italian datasets, and is compared against comparably-sized open-source MoE and dense models.

In comparison with popular Italian models, namely FastwebMIIA-7B, Minerva-7B-instruct-v1.0, Velvet-14B, and LLaMAntino-3-ANITA-8B, the new EngGPT2MoE-16B-A3B model performs as well or better on the international benchmarks: ARC-Challenge, GSM8K, AIME2024, AIME2025, MMLU, and HumanEval. It achieves the best performance for the longest context setting (32k) of the RULER benchmark. On the Italian benchmark dataset ITALIC, the model performs as well or better than the other models except for Velvet-14B, which outperforms it.

Compared with popular Mixture-of-Experts models of comparable size, the new model reports higher values than DeepSeek-MoE-16B-Chat on all considered benchmarks. It has higher values than Moonlight-16B-A3B-Instruct on HumanEval, MMLU, AIME-2024, AIME-2025, GSM8K, and the 32k RULER setting, but lower on the function calling benchmark BFCL and some ARC-Challenge and ITALIC settings. Finally it has lower values than GPT-OSS-20B on most benchmarks, including HumanEval, MMLU, AIME 2024, AIME 2025, GSM8K, ARC, BFCL, and the RULER 32k context setting. When compared with popular dense models, EngGPT2MoE-16B-A3B reports higher values on AIME-2024 and AIME-2025 than Llama-3.1-8B-Instruct, Gemma-3-12b-it, and Ministral-3-8B-Instruct-2512-BF16, but lower values on ITALIC, BFCL, and RULER with a 32k context.

When performance is aggregated in a dataset-wise fashion across all benchmark metrics (in terms of either score or rank), EngGPT2MoE-16B-A3B shows higher performance than the Italian models under evaluation - namely FastwebMIIA-7B, Minerva-7B-instruct-v1.0, Velvet-14B, and LLaMAntino-3-ANITA-8B - while achieving lower results than some of the most performant international models, in particular GPT-5 nano and Qwen3-8B. Taken together, our findings find the new model to be a step forward for native Italian Large Language Models.

## 1 Introduction

Over the last few years, Large Language Modelling (LLM) technology [1] based on generative pretrained Transformer models [2] have demonstrated remarkable capabilities across all types of Natural Language Processing tasks from machine translation [3]-[4] to question answering [5], text generation [6], etc.

LLM technology has evolved over time into two branches, large proprietary models, that can only be accessed via paid programatic interfaces (APIs) and smaller open-source models that can be deployed locally with the advantage of increased control and reduced risk of data loss [7]-[8]. The amount of training tokens used to pre-train these models has steadily increased over time and along with it the performance of the models, particularly the smaller open-source ones [9].

Moreover, the development of instruction tuning techniques to improve the model’s ability to follow instructions [10], chain of thought prompting to improve the model’s ability to generate and leverage

\*These authors contributed equally to this work.

<sup>1</sup>Complete code used for this project is available at <https://github.com/AIRIC-Polimi/EngGPT-evaluation>

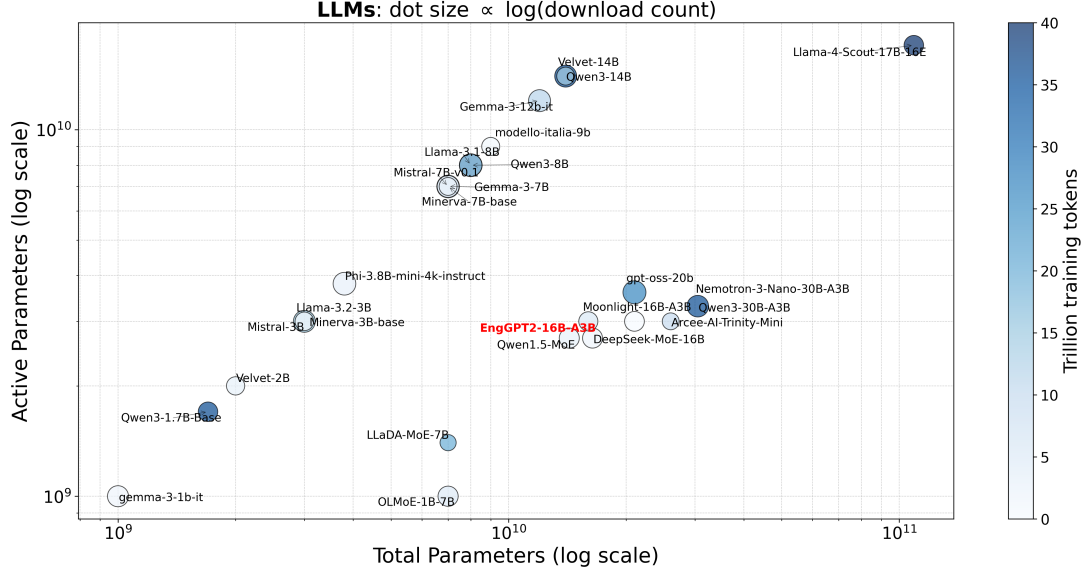


Figure 1: Comparable open-source models in April 2026. The horizontal axis shows the total parameter count, and the vertical axis is the active parameter count, both on a logarithmic scale. The size of each circle shows the popularity of the model as measured by the number of downloads from the Huggingface repository, while the colour indicates the size of the training corpus used to pretrain the model (with darker indicating more tokens). We see the set of dense models (where active parameters equals the total parameter count) along a diagonal on the top left of the graph, and a cluster of Mixture of Expert (MoE) models in the centre of the graph.

intermediate results [11], and reinforcement learning for improving the reasoning capabilities of the models [12], have all resulted in performance improvements across difficult benchmarks [13]. Finally, specialized training has enabled the models to be used for tool calling [14] and skill acquisition [15].

On the architecture side, there have been a steady series of improvements to increase the [16] the performance of the models, in terms of throughput [17], answer quality, and ability to work with large inputs [16]. Sliding window attention [18] with key-value caches have allowed the models to work with very long context lengths, greatly reducing the limitations of the early models to process only relatively short documents. In order to further increase inference speed, Mixture of Experts (MoE) models [19, 20] allow for a subset of the parameters of the model to be used during forward passes, thereby speeding up model inference, while also reducing training time [21].

EngGPT2-16B-A3B<sup>2</sup> [22] is a 16 billion parameter Mixture of Experts (MoE) model with 3 billion active parameters, developed and trained by ENGINEERING Ingegneria Informatica S.p.A. for the Italian context. In this paper, we evaluate the performance of the model across a wide variety of representative LLM benchmark datasets, including general international benchmarks which are primarily in English as well as more specific benchmarks for Italian. Performance is compared with comparable sized state-of-the-art open-source models (both MoE and dense) from leading international research teams as well as competitive LLMs that have been developed specifically for the Italian market.

## 2 Evaluation Methodology

We now describe the methodology adopted for selecting appropriately dimensioned comparison models and benchmarks for evaluating them.

### 2.1 Selection of Models for Comparison

In order to select an appropriate set of models for comparison, we first visualise popular models in Figure 1 in terms of their (i) total parameter count, (ii) active parameter count (parameters used during a forward inference pass), (iii) popularity (as estimated by the number of downloads from the HuggingFace repository), and (iv) amount of training data (number of tokens) used to pre-train the model. A set of

<sup>2</sup>Huggingface link: <https://huggingface.co/engineering-group/EngGPT2-16B-A3B>

| Model                      | Date Released | # Parameters | # Active Parameters | Context Length | Thinking | Tool Calling |
|----------------------------|---------------|--------------|---------------------|----------------|----------|--------------|
| EngGPT2-MoE-16B            | Mar 30, 2025  | 16b          | 3b                  | 32k            | Yes      | Yes          |
| FastwebMIIA-7B             | May 29, 2025  | 7.4b         | 7.4b                | 16k            | No       | No           |
| Minerva-7B-instruct-v1.0   | Sep 18, 2024  | 7b           | 7b                  | 4k             | No       | No           |
| Velvet-14B                 | Jan 31, 2025  | 14b          | 14b                 | 128k           | No       | Yes          |
| LLaMAntino-3-ANITA-8B      | May 10, 2024  | 8b           | 8b                  | 8k             | No       | Yes          |
| Qwen3-4B                   | Apr 28, 2025  | 4b           | 4b                  | 32k            | Yes      | Yes          |
| Qwen3-8B                   | Apr 28, 2025  | 8b           | 8b                  | 32k            | Yes      | Yes          |
| Llama-3.2-3B-Instruct      | Sep 18, 2024  | 3b           | 3b                  | 128k           | No       | Yes          |
| Llama-3.1-8B-Instruct      | Jul 15, 2024  | 8b           | 8b                  | 128k           | No       | Yes          |
| gemma-3-4b-it              | Mar 12, 2025  | 4b           | 4b                  | 128k           | No       | Yes          |
| gemma-3-12b-it             | Mar 12, 2025  | 12b          | 12b                 | 128k           | No       | Yes          |
| Ministral-3-8B             | Jan 13, 2026  | 8b           | 8b                  | 256k           | No       | Yes          |
| GPT-5 nano                 | Mar 12, 2025  | N.A.         | N.A.                | 400k           | Yes      | Yes          |
| deepseek-moe-16b-chat      | Jan 9, 2024   | 16.4b        | 2.7b                | 4k             | No       | No           |
| gpt-oss-20b                | Aug 5, 2025   | 21b          | 3.6b                | 128k           | Yes      | Yes          |
| Moonlight-16B-A3B-Instruct | Feb 24, 2025  | 12b          | 3b                  | 128k           | No       | No           |

Table 1: Models chosen for comparison with their relevant characteristics.

dense models (where all parameters are active) are seen along a diagonal on the top left of the graph, while a cluster of Mixture of Expert (MoE) models (containing also EngGPT2-16B-A3B) appears in the centre of the graph. Our aim is to select the most representative and popular models from both clusters.

For the analysis, we will compare the EngGPT2MoE-16B-A3 model with the models listed in Table 1, specifically:

- **Dense Italian models** that have been developed specifically for the Italian context, namely *FastwebMIIA-7B*<sup>3</sup> and *Minerva-7B-instruct-v1.0* [23], both leveraging seven billion parameters, *Velvet-14B*<sup>4</sup> with 14 billion parameters, and *LLaMAntino-3-ANITA-8B-Inst-DPO-ITA*<sup>5</sup> [24], a model with 8 billion parameters that has been fine-tuned from a LLaMA3 model on Italian datasets;
- **Dense International models** that are extremely popular, including two models from Alibaba, *Qwen3-4B* and *Qwen3-8B* [25] with four and eight billion parameters respectively, two models from Meta, *Llama-3.2-3B-Instruct* and the older *Llama-3.1-8B-Instruct* [26] with 3 and 8 billion parameters respectively, a model from mistral *Ministral-3-8B-Base-2512*<sup>6</sup> [27] and finally two models from Google, *gemma-3-4b-it* and *gemma-3-12b-it* [28] with 4 and 12 billion parameters respectively. We additionally include a proprietary model with undisclosed weights, *GPT-5 nano*, to establish a baseline using commercially available APIs and to better understand the gap between open-weight and closed models.
- **Mixture of Expert (MoE) models** of comparable size: *deepseek-moe-16b-chat* [29] from Deepseek with 16 billion parameters of which 2.7 billion are active, *gpt-oss-20b* [30] from OpenAI with 20 billion parameters and 3.6 billion active, and *Moonlight-16B-A3B-Instruct* [31], with the most similar amount of total and active parameters with respect to EngGPT2-MoE-16B.

We note that the chosen models in Table 1 differ not only in terms of number of total and active parameters, but also in their context length (affecting the maximum size of conversations they can deal with), whether or not they have been trained as "thinking models" [32, 33] to output their reasoning in a structured format when responding, and whether they have been trained for tool/function calling [14, 34, 35]. These architectural and training differences are highly consequential and directly dictate performance across specific benchmarks; for instance, native tool-calling capabilities are decisive for success on the BFCL benchmark, while structured chain-of-thought training is critical for achieving high accuracy on complex, multi-step problem-solving tasks such as AIME, MMLU, and GSM8K.

## 2.2 Selection of Benchmarks

In order to identify a comprehensive set of general benchmarks that can be used to evaluate the chosen models across a representative set of tasks, we start from the articles describing recent and popular LLMs:

<sup>3</sup><https://huggingface.co/Fastweb/FastwebMIIA-7B>

<sup>4</sup><https://huggingface.co/Almawave/Velvet-14B>

<sup>5</sup>From here on *LLaMAntino-3-ANITA-8B-Inst-DPO-ITA* and *LLaMAntino-3-ANITA-8B* will be used interchangeably

<sup>6</sup>From here on *Ministral-3-8B-Base-2512* and *Ministral-3-8B* will be used interchangeably

| Dataset                   | Llama3 | Qwen3 | gpt_oss | DeepSeek MoE |
|---------------------------|--------|-------|---------|--------------|
| AIME [36]-[37]            |        | ✓     | ✓       |              |
| AlignBench v1.1 [38]      |        | ✓     |         |              |
| API-Bank [39]             | ✓      |       |         |              |
| ARC-Challenge [40]        | ✓      |       |         | ✓            |
| Arena-Hard [41]           |        | ✓     |         |              |
| AutoLogi [42]             |        | ✓     |         |              |
| BFCL [43]                 | ✓      | ✓     |         |              |
| C-Eval [44]               |        | ✓     |         | ✓            |
| CHID [45]                 |        |       |         | ✓            |
| CLUEWSC [46]              |        |       |         | ✓            |
| CMMLU [47]                |        |       |         | ✓            |
| CodeElo [48]              |        | ✓     | ✓       |              |
| Creative Writing V3       |        | ✓     |         |              |
| DROP [49]                 |        |       |         | ✓            |
| Gorilla API-Bench [50]    | ✓      |       |         |              |
| GPQA [51]                 | ✓      |       | ✓       |              |
| GPQA-Diamond [51]         |        | ✓     |         |              |
| GSM8K [52]                | ✓      |       |         | ✓            |
| HealthBench [53]          |        |       | ✓       |              |
| HellaSwag [54]            |        |       |         | ✓            |
| HLE [55]                  |        |       | ✓       |              |
| HumanEval [56]            | ✓      |       |         |              |
| IFEval [57]               | ✓      | ✓     |         |              |
| INCLUDE [58]              |        | ✓     |         |              |
| InfiniteBench [59]        | ✓      |       |         |              |
| LiveBench [60]            |        | ✓     |         |              |
| LiveCodeBench [61]        |        | ✓     |         |              |
| MATH [62]                 | ✓      |       |         | ✓            |
| MATH-500 [63]             |        | ✓     |         |              |
| MBPP EvalPlus [64]        | ✓      |       |         |              |
| MGSM [65]                 | ✓      |       |         |              |
| MlogiQA [66]              |        | ✓     |         |              |
| MMMLU [67]                |        | ✓     | ✓       |              |
| MMLU [67]                 | ✓      |       | ✓       | ✓            |
| MMLU-Pro [68]             | ✓      |       |         |              |
| MMLU-Redux [69]           |        | ✓     |         |              |
| Multi-IF [70]             |        | ✓     |         |              |
| MultiPL-E [71]            | ✓      |       |         |              |
| Needle-in-a-Haystack [72] | ✓      |       |         |              |
| Nexus [73]                | ✓      |       |         |              |
| Pile [74]                 |        |       |         | ✓            |
| PolyMath [75]             |        | ✓     |         |              |
| RULER [76]                |        | ✓     |         |              |
| TruthfulQA [77]           |        |       |         | ✓            |
| WinoGrande [78]           |        |       |         | ✓            |
| WritingBench [79]         |        | ✓     |         |              |
| ZebraLogic [80]           |        | ✓     |         |              |
| ZeroSCROLLS [81]          | ✓      |       |         |              |
| $\tau$ -Bench Retail [82] |        |       | ✓       |              |

Table 2: Datasets used for benchmarking in the papers introducing four popular state-of-the-art post-trained LLMs: *Llama3*, *Qwen3*, *GPT-oss*, and *DeepSeekMoE*. We highlight the ones that were used in this work to perform the benchmarking.

*Llama3* [26], *Qwen3* [25], *GPT-Oss* [30], and *DeepSeekMoE* [29], and investigate the set of datasets that were used to evaluate them. Table 2 lists all the datasets used to evaluate the capabilities of these state-of-the-art *post-trained* LLMs. We note that there is significant variation in the set of benchmarks used in the evaluation across the different models, an issue that we aim to address in this work by adopting a standardized, representative and reproducible set of benchmarks.

We investigate the properties of the benchmark datasets in Figure 2, where we categorise them in terms of their category (general knowledge vs coding, etc.), their language (English, Multilingual or Italian), size (total number of samples), popularity (downloads from Huggingface) and whether they are generative tasks or multiple choice tasks. We split the dataset into 7 categories, and for each of them we identify the selected benchmarks, which are summarized in Table 3. Based on the visualisation, we then decide to start our preliminary evaluation using the most popular of the datasets within their category: *MMLU* [67, 69] for general tasks, *HumanEval* [56] for coding, *ARC-Challenge* [40] for reasoning, *GSM8K* [52] and *AIME 2024 & 2025* [36, 37] for math reasoning, *RULER* [76] for Long context, and *BFCL* [43] for Tool Use.

Since we aim at evaluating the model also on the Italian language, we consider including at least one native Italian dataset. Work by NLP researchers in Italy has resulted in the creation of two standard set of benchmark datasets in Italian, called *Evalita-LLM* [83] (extending a previous ITA-bench benchmark [84]) and *ITALIC* [85]. Both *Evalita-LLM* and *ITALIC* contain a large amount of native Italian text, distinguishing it from other Italian datasets which tended to be created through a process of machine translation. For the sake of our evaluation, we consider the *ITALIC* benchmark, as it is more aligned with current benchmarks adopted to assess LLMs capabilities.

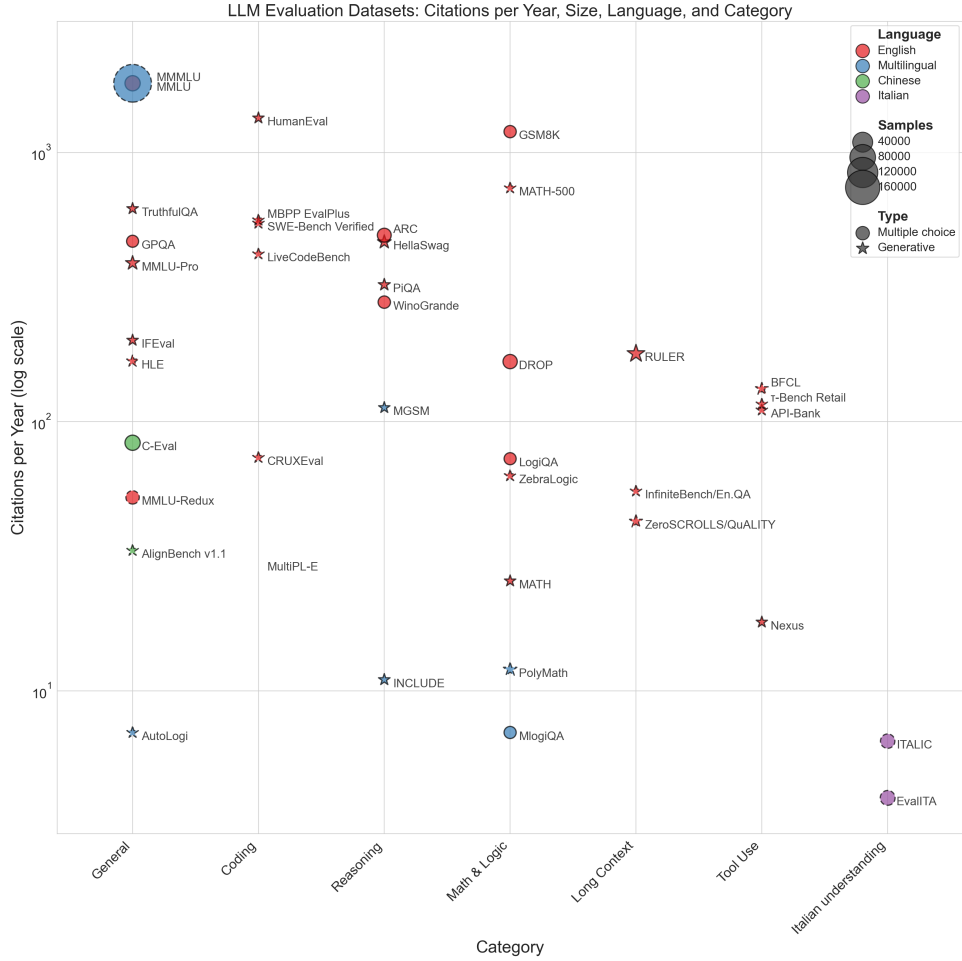


Figure 2: Overview of popular potential datasets, including their popularity (measured by yearly downloads), category, size (in terms of the number of samples), language, and type (multiple choice vs generative).

## 2.3 Implementation Details

To conduct a rigorous and reproducible evaluation across the selected benchmarks, we utilized a combination of standardized frameworks and custom integrations tailored to the specific requirements of each task. The majority of the evaluations were executed using EleutherAI’s `lm-evaluation-harness` Python package [86]. Within this framework, all associated tasks were configured using the `generate_until` output type. This setting instructs the model to perform open-ended text generation until a predefined stopping criterion (such as a specific token or maximum length limit) is met, thereby capturing complete reasoning trajectories rather than merely calculating log-likelihoods for predefined multiple-choice tokens. To automatically evaluate these free-form responses, the harness employs regular expression (regex) functions to parse and isolate the final answer from the surrounding generated text. For our evaluations, we utilized both *strict* and *flexible* extraction methodologies. Strict extraction enforces rigid formatting constraints (e.g., requiring the answer to appear at the very end of the text or within specific delimiters), penalizing models that fail to follow precise formatting instructions. Conversely, flexible extraction uses broader regex patterns to identify the correct answer anywhere within the output. Employing both methods is crucial, as it allows us to effectively decouple a model’s underlying reasoning and factual capabilities from its strict formatting compliance.

To maximize inference throughput and computational efficiency, the evaluated models were predominantly served utilizing the vLLM backend [87]. There were, however, two necessary deviations from this default serving configuration. First, the evaluation of the *GPT-5-nano* model was conducted via API calls using the `azure-openai-chat-completions` model option of the harness. During this process, we encountered instances where specific task prompts inadvertently triggered Azure’s built-in safety and content filters, resulting in API exceptions that halted the execution. To ensure an uninter-

| Task category          | Selected dataset            |
|------------------------|-----------------------------|
| General tasks          | MMLU                        |
| Coding                 | HumanEval                   |
| Reasoning              | ARC-Challenge               |
| Math reasoning         | GSM8K, AIME-2024, AIME-2025 |
| Long context           | RULER                       |
| Tool use               | BFCL                        |
| Language understanding | ITALIC                      |

Table 3: Datasets selected for the evaluation, chosen as the most popular benchmark within each task category.

rupted evaluation pipeline, we modified the underlying `lm-evaluation-harness` source code to include an exception-handling mechanism; this allowed the system to catch safety-related errors, bypass the blocked queries, and automatically proceed to the subsequent test sample. Second, for the evaluation of *Ministral-3-8B-Instruct-2512-BF16*, we were required to use the `hf-mistral3` model option of the harness, which uses Huggingface’s `transformers` library instead of `vLLM` as a backend. This fallback was necessitated by compatibility constraints between the model’s underlying architecture and the Hugging Face version currently supported by the `lm-evaluation-harness`. Consequently, this compatibility workaround resulted in significantly slower inference and overall evaluation times for this specific model compared to the other ones.

For benchmarks outside the scope of the primary evaluation *harness*, we employed alternative software solutions. The BFCL evaluation was carried out using the `evalscope` Python package [88], since tool calling is currently not compatible with the architecture of EleutherAI’s `lm-evaluation-harness`. Furthermore, since the *ITALIC* benchmark is not natively supported by these standard evaluation libraries, we conducted its evaluation by directly importing the original task repository into our custom codebase, in addition to developing custom tasks for integrating it with the `lm-evaluation-harness` framework.

For reproducibility purposes, comprehensive technical specifications, including the exact software versions and environment utilized during all experiments, are documented in Appendix E.

### 3 Benchmarking Datasets

This section presents the benchmarking datasets selected for our evaluation. For each benchmark, we briefly describe its main characteristics, we provide an illustrative example to clarify the type of task it contains and we describe the selected evaluation setting. For additional technical details on how each evaluation task was configured (including custom modifications) starting from the selected benchmarking datasets, see Appendix D.

#### 3.1 Reasoning-Intensive Benchmark - ARC-Challenge

ARC-Challenge (AI2 Reasoning Challenge) [40] is a benchmark composed of grade-school-level multiple-choice science questions designed to evaluate commonsense and scientific reasoning abilities (see Table 4). The dataset focuses on questions that are difficult for simple retrieval-based and word co-occurrence methods, making it a more challenging subset of the broader ARC benchmark. In particular, we consider the Challenge split, which is composed of 2.6k examples (1,119 train, 299 validation, and 1,172 test). To assess the impact of intermediate reasoning on model performance, we evaluate three distinct configurations of this task:

- `arc_challenge_chat`, which evaluates direct answering *without enabling reasoning*;
- `arc_challenge_cot_custom`, which we introduced as an extension of the previous one to *explicitly prompt the model to reason* (see Appendix D.2 for implementation details);
- `arc_challenge_cot_it_custom`, which is an *Italian machine-translated version* of the previous one (see Appendix D.2 for implementation details).

This dual approach allows for a comprehensive comparison, ensuring that models natively supporting reasoning paradigms can be fairly evaluated alongside those that do not. For all the configurations, the evaluation is formulated as an open-ended generation task utilizing the `generate_until` output type,

| Example Questions                                                                                                                                                                                                                                                                                                                                                                                             | Desired Answers                                                                                                                                                                                                                                                                                                                                 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>ARC-Challenge:</b><br>George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?<br>A. dry palms<br>B. wet palms<br>C. palms covered with oil<br>D. palms covered with lotion                                                                                                                                                                                  | A. dry palms                                                                                                                                                                                                                                                                                                                                    |
| <b>GSM8K:</b><br>Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?                                                                                                                                                                                                                                  | Natalia sold $48/2 = \llbracket 48/2=24 \rrbracket 24$ clips in May. Natalia sold $48+24 = \llbracket 48+24=72 \rrbracket 72$ clips altogether in April and May. ##### 72                                                                                                                                                                       |
| <b>AIME-2024:</b><br>Let $x, y$ and $z$ be positive real numbers that satisfy the following system of equations: $\log_2 \left( \frac{x}{yz} \right) = \frac{1}{2}$ $\log_2 \left( \frac{y}{xz} \right) = \frac{1}{3}$ $\log_2 \left( \frac{z}{xy} \right) = \frac{1}{4}$ Then the value of $ \log_2(x^4 y^3 z^2) $ is $\frac{m}{n}$ where $m$ and $n$ are relatively prime positive integers. Find $m + n$ . | Denote $\log_2(x) = a$ , $\log_2(y) = b$ , and $\log_2(z) = c$ . Then, we have: $a - b - c = \frac{1}{2}$ , $-a + b - c = \frac{1}{3}$ , $-a - b + c = \frac{1}{4}$ . Now, we can solve to get $a = \frac{-7}{24}$ , $b = \frac{-9}{24}$ , $c = \frac{-5}{12}$ . Plugging these values in, we obtain $ 4a + 3b + 2c  = \frac{25}{8} = 0.3125$ . |
| <b>MMLU:</b><br>Find the degree for the given field extension $Q(\sqrt{2}, \sqrt{3}, \sqrt{18})$ over $Q$ .<br>A. 0<br>B. 4<br>C. 2<br>D. 6<br>Which of these principles is not an element of the responsibility to protect?<br>A. The responsibility to prevent<br>B. The responsibility to react<br>C. The responsibility to remain sovereign<br>D. The responsibility to rebuild                           | B<br><br><br><br><br><br><br>C                                                                                                                                                                                                                                                                                                                  |

Table 4: Illustrative examples of questions and correct answers from the *ARC-Challenge*, *GSM8K*, *AIME-2024*, and *MMLU* benchmarks.

wherein the entirety of the model’s generated text is considered. Model performance is quantified using the mean exact match score. Finally, since the reasoning-enabled `arc_challenge_cot_custom` variant produces verbose outputs, we use both flexible and strict extraction functions to accurately parse the final answer from the preceding reasoning traces. The same applies to the Italian version. The exact code implementation for these extraction methodologies is detailed in Appendix D.2.

### 3.2 Math - GSM8K

GSM8K (Grade School Math 8K) [52] is a benchmark composed of grade-school level mathematical word problems designed to evaluate numerical and multi-step reasoning abilities. The dataset focuses on problems expressed in natural language, where solving the task typically requires performing a sequence of arithmetic operations rather than retrieving a direct fact. The dataset is composed by 7.5k examples (see Table 4 for an example).

For our evaluation, we specifically selected the `gsm8k_llama` variant of the task, as it incorporates superior parsing rules for extracting mathematical answers from complex model outputs. The task is structured using the `generate_until` output type, meaning that the entirety of the model’s generated text is considered during evaluation. Performance is measured using the mean exact match score. To reliably isolate the final numerical answer from the generated reasoning traces, we applied both flexible and strict extraction methodologies using their default configurations.

### 3.3 Hard Math - AIME-2024 & AIME-2025

AIME-2024 [36] and AIME-2025 [37] are benchmarks derived from the 2024 and 2025 editions of the American Invitational Mathematics Examination (AIME), a competition aimed at evaluating advanced problem-solving skills in pre-college mathematics. Compared with datasets such as GSM8K, AIME problems are significantly more challenging (see Table 4), often requiring advanced reasoning, algebraic manipulation, combinatorics, number theory, or geometry. This benchmark is designed to assess high-level mathematical reasoning on competition-style questions, where success depends not only on

| Prompt                                                                                                                                                                                                                                                    | Canonical Solution                                   | Test                                                                                                                                                                                                                              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <pre>def greatest_common_divisor(a: int, b: int) -&gt; int:     """Return a greatest common divisor     of two integers a and b     &gt;&gt;&gt; greatest_common_divisor(3, 5)     1     &gt;&gt;&gt; greatest_common_divisor(25, 15)     5     """</pre> | <pre>while b:     a, b = b, a % b     return a</pre> | <pre>METADATA = {     'author': 'jt',     'dataset': 'test' } def check(candidate):     assert candidate(3, 7) == 1     assert candidate(10, 15) == 5     assert candidate(49, 14) == 7     assert candidate(144, 60) == 12</pre> |

Table 5: Illustrative example from the HumanEval benchmark style.

obtaining the correct final answer but also on correctly identifying the sequence of intermediate logical and computational steps needed to solve the problem.

Consistent with our other evaluations, this task utilizes the `generate_until` output type, meaning the entirety of the model’s generated text is considered to capture the complete reasoning trajectory. For the standard evaluation runs, overall performance is measured using the mean exact match score. In addition to the standard evaluations, we performed custom runs to compute the pass@k and avg@k metrics, specifically for k=1 and k=8. These custom runs are identical to the standard configurations in all respects except for the evaluation metrics applied; the exact code used to compute these custom metrics is provided in Appendix D.1. Finally, for models evaluated via API calls (e.g., OpenAI’s GPT-5-nano), we modified the task prompt to explicitly instruct the model to enclose its final answer using the `\boxed{}` syntax, ensuring reliable extraction. The exact prompt template utilized for these API-based evaluations is detailed in Appendix D.1.

### 3.4 General Knowledge Benchmarks - MMLU & MMLU Redux

MMLU (Massive Multitask Language Understanding) [67] is a benchmark composed of multiple-choice questions drawn from diverse branches of knowledge, designed to evaluate broad world knowledge and problem-solving abilities (see example in Table 4). The benchmark spans subjects in the humanities, social sciences, hard sciences, and other domains, covering 57 tasks. The dataset includes 99.8k examples in the auxiliary training split, 285 examples in the dev split, 1.5k examples in the validation split, and 14k examples in the test split. For our evaluation, we selected two distinct variants of this benchmark. The first is *MMLU\_cot\_llama*, a chain-of-thought version of MMLU designed to elicit step-by-step reasoning. To effectively handle the outputs of this variant, we integrated enhanced parsing rules and created a variant of the benchmark named *MMLU\_cot\_llama\_custom*, the details of which are described in Appendix D.3. The second variant is *mmlu\_redux* [69], a carefully annotated and corrected version of the original dataset that provides a more accurate and reliable benchmark for evaluating language model performance. For both configurations, the task is structured using the `generate_until` output type, meaning that the entirety of the model’s generated text is considered during evaluation. Overall, performance across both variants is quantified using the mean exact match score. Additionally, because the *MMLU\_cot\_llama* variant involves verbose reasoning traces, we applied both flexible and strict extraction methodologies to accurately isolate the final multiple-choice answer from the generated text.

### 3.5 Coding - HumanEval

HumanEval is a benchmark composed of handwritten Python programming problems designed to evaluate code generation and functional reasoning abilities [56]. The dataset focuses on tasks where a model must generate a correct function implementation from a natural-language specification and function signature, and correctness is assessed through unit tests. Each example includes a prompt, a canonical solution, a test suite, and an entry point. The dataset is composed of 164 examples.

For our evaluation, we utilized the default variant of the task, enhanced with custom parsing rules. Specifically, we observed that several models generated functionally correct code that inadvertently included unsupported Unicode characters, which caused Python syntax errors during execution. To address this, we implemented an additional sanitization step to remove these characters prior to running the unit tests. The generation phase was configured with the `generate_until` output type, meaning that the entirety of the model’s generated text is considered. Finally, to rigorously evaluate code generation performance, we measured the pass@k and avg@k metrics for  $k = 1$  and  $k = 8$ .

| Overall Score                                          |                                                  |                                           |                                                           |
|--------------------------------------------------------|--------------------------------------------------|-------------------------------------------|-----------------------------------------------------------|
| Unweighted Average<br>(4423)                           |                                                  |                                           |                                                           |
| Non-Live (Single-Turn)<br>Unweighted Average<br>(1150) | Live (Single-Turn)<br>Weighted Average<br>(1351) | Multi-Turn<br>Unweighted Average<br>(800) | Hallucination Measurement<br>Unweighted Average<br>(1122) |

Table 6: Structure of the BFCL dataset used for the evaluations in terms of task groups. The number in parentheses indicate the number of samples for each task and task group, while the averaging rules indicate how scores were grouped together for each task group and then globally. Further details on the specific tasks for each category can be found in Table 26 in the Appendix.

### 3.6 Tool Calling - BFCL

*BFCL* (Berkeley Function Calling Leaderboard) [43] is a benchmark dataset for evaluating large language models on function/tool calling. It pairs user requests with function specifications and measures whether a model selects the right tool, formats the call correctly, fills arguments accurately, abstains when no tool is appropriate, and, in newer versions, handles multi-turn and multi-step agentic interactions. *BFCL* covers multiple tool-use settings such as single, multiple, and parallel calls, and spans several programming-language/tool formats. In particular, we consider the version v3 of the benchmark, which includes four main groups of tasks that we scored separately:

- **Non-live:** contains expert-curated samples spanning the Python, Java and JavaScript programming languages in which each one describes a single function that needs to be called correctly, a multiple category in which the model is given more than one tool and needs to call the correct one, a parallel category in which the model needs to call more than one tool, and a parallel multiple category which is the most complex and require selection of  $m$  correct tools out of  $N$  provided.
- **Live:** contains user-contributed samples, spanning the same categories as the Non-live group.
- **Hallucination Measurement:** contains both expert-curated and user-contributed samples where the model is given non relevant tools and should correctly recognize that it does not need to use them to answer. This group measures how well the model behaves in the presence of irrelevant data.
- **Multi-Turn:** contains samples in which multiple tool-based interactions are needed to complete the task, such as listing the files in a folder and then reading the correct one. This group contains a base category, a missing functions category in which some tools are not available in the first turn of conversation but are revealed later, a missing parameters category in which some key information is held out from the original user message and needs to be requested during the conversation, and a long context category in which large volumes of extraneous data (e.g., long files, long directory listings) are added to measure how well the model can extract important information out of noisy data.

Table 6 summarizes the structure of the BFCL dataset, including the number of samples for each task group and the aggregation logic used to report the overall scores. Further details about each specific task can be found in Appendix C.

The metric used to evaluate model responses depends on the task: for the live and non-live tasks, the generated tool calls are parsed into their Abstract Syntax Tree (AST), which is in turn used for determining whether the functions and their respective parameters and types are correct; for multi-turn tasks, at each turn both the state (e.g., a virtual filesystem) and the response of the models are checked against a ground truth, and both checks must succeed for all turns for the entry to be marked as a success; for hallucination measurement, we simply check that the model outputs no tool calls at all (which may cause non-responsive models to inadvertently perform well on this task).

This dataset required ad hoc treatment due to the complexity and many nuances of function calling. In particular, since function calling in general is not supported by EleutherAI’s LM Evaluation Harness suite due to architectural limitations, we adopted the Evalscope evaluation library [88], which in turn wraps the original implementation of the BFCL evaluation benchmark. In addition, we served each model using the vLLM [89] inference server, which enabled us to make use of their built-in parsers for tool calling for the models that support it. More details on which parser was used for each model can be found in Table 7.

| Model                                                                                                                                                        | FC mode                              | Tool call parser                                                                                                                          | Reasoning parser                                                               |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| EngGPT2-MoE-16B<br>FastwebMIIA-7B<br>Minerva-7B-instruct-v1.0<br>Velvet-14B<br>LLaMAntino-3-ANITA-8B                                                         | ✓                                    | hermes<br><br>unknown <sup>1</sup>                                                                                                        | qwen3                                                                          |
| Qwen3-4B<br>Qwen3-8B<br>Llama-3.2-3B-Instruct<br>Llama-3.1-8B-Instruct<br>gemma-3-4b-it<br>gemma-3-12b-it<br>Ministral-3-8B-Instruct-2512-BF16<br>GPT-5 nano | ✓<br>✓<br>✓<br>✓<br>✓<br>✓<br>✓<br>✓ | hermes<br>hermes<br>llama3_json, pythonic<br>llama3_json<br>pythonic <sup>2</sup><br>pythonic <sup>2</sup><br>mistral<br>N/A <sup>3</sup> | qwen, deepseek_r1<br>qwen, deepseek_r1<br><br><br><br><br><br>N/A <sup>3</sup> |
| deepseek-moe-16b-chat<br>gpt-oss-20b<br>Moonlight-16B-A3B-Instruct                                                                                           | ✓                                    | openai                                                                                                                                    | auto-detected by vLLM <sup>4</sup>                                             |

<sup>1</sup> Velvet models have reportedly been trained for tool calling, but we were unable to find details regarding the parsing style used. Tests with plausible parsers (mistral, granite) proved ineffective.

<sup>2</sup> During evaluation the pythonic tool call parser for gemma was unable to properly parse model output.

<sup>3</sup> The GPT5 model family was accessed through the Azure OpenAI Cloud service, so we did not have explicit control over tool and reasoning parsing, though it was in effect server-side.

<sup>4</sup> The gpt-oss family of models uses Harmony tokenizer response format, which is natively auto-detected and parsed by the vLLM engine without specific configuration.

Table 7: vLLM tool call parsers used for each model in the BFCL evaluation.

It is worth noting that the evaluation logic supports both injecting tool descriptions in the model input as raw strings and then parsing them from the raw textual output (so-called non-FC or PROMPT mode), as well as using dedicated fields in the Request and Response objects of the OpenAI standard API format and leaving serialization and deserialization to and from the model to the inference server (so-called native function calling or FC mode). For models that support both, we tested them and reported the best results; for the other ones we used PROMPT mode.

| Messages                                                                                                                                                                       | Options                                                                        | Expected Answer |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|-----------------|
| La frase "Secondo recenti sondaggi il candidato repubblicano gode di scarsissima popolarità: la possibilità di una sua vittoria si allontana sempre di più" contiene un verbo: | [ "riflessivo diretto", "impersonale", "passivo", "intransitivo pronominale" ] | D               |

Table 8: Illustrative example from the ITALIC benchmark.

### 3.7 Italian Benchmark - ITALIC

ITALIC is a large-scale benchmark introduced to assess the ability of large language models to understand Italian language, culture, and commonsense knowledge. The dataset comprises 10,000 multiple-choice questions drawn from authentic public examination materials and covers 12 domains across two broad dimensions: *Culture and Commonsense* and *Language Capability*. By focusing on culturally grounded and linguistically rich Italian content, ITALIC provides a more representative evaluation setting for Italian than benchmarks obtained through simple translation from English[85].

While ITALIC is not natively supported by the `lm_evaluation_harness` library, we employed a dual-evaluation strategy to ensure a robust and fair assessment of the models. The original ITALIC evaluation suite structures prompts as a multi-turn conversation, injecting few-shot examples and the target question as a history of prior chat messages. However, to decouple intrinsic factual knowledge from chat-formatting adherence, we evaluated the benchmark using both the original repository’s chat-based format (manually integrated into our codebase) and custom task configurations designed for native compatibility with the `lm_evaluation_harness`.

For the original chat-formatted evaluation, we maintained the two distinct prompting strategies available in the original implementation: a "fast" variant, which strictly instructs the model to output only the correct letter, and a "chain-of-thought" (*CoT*) variant, which prompts the model for a brief reasoning trace. Both utilize a 5-shot conversational setting. While we adopted the original implementation, we introduced specific enhancements to ensure reliable extraction. For the *CoT* variant, we implemented more

robust parsing rules to mitigate false negatives caused by functionally correct answers being misclassified. Furthermore, we refined the few-shot formatting to improve structural clarity, explicitly formulating the example answers as 'Risposta: A' rather than a standalone 'A' (see Appendix D.4 for exact changes).

In our custom `lm_evaluation_harness` integration, questions are provided directly to the models, stripping away the conversational overhead. All evaluations within this harness were conducted in a 5-shot setting. To systematically isolate the effect of the prompt structure, we implemented this integration in two phases. First, we evaluated the models using the exact same prompt wording as the original chat-based version, simply adapted for direct input; we designated these variants as `italic_fast` and `italic_cot` (see Appendix D.4). Second, we explored the impact of more highly structured prompting by adapting templates from established benchmarks. This resulted in a standard non-*CoT* variant (`italic_fast_custom`) and a *CoT* variant (`italic_cot_custom`), whose prompt templates were directly translated from the `arc_challenge_chat` and `mmlu_cot_llama` tasks, respectively (see Appendix D.4).

Adapting the *CoT* variants (`italic_cot` and `italic_cot_custom`) for the direct-prompting harness required addressing a limitation in the source dataset: the standard few-shot examples provide only the final answer without underlying reasoning traces. Consequently, we could not rely on the standard `num_fewshot` parameter for dynamic prompt construction. Instead, we utilized Gemini 3.1 Pro to synthetically generate accurate, high-quality explanations for the five baseline few-shot examples. These generated reasoning traces were statically embedded into the prompt templates of both *CoT* configurations to properly condition the models for deliberative generation.

As with previous tasks, the entirety of the model’s generated text is considered, and overall performance is quantified using accuracy as the primary metric. Analyzing the results across both evaluation formats (chat-based versus direct) and prompting modes (fast versus *CoT*) provides a highly nuanced perspective on model capabilities. This multi-faceted approach allows us to accurately determine whether a model’s performance on Italian-centric tasks genuinely benefits from deliberative processing, or if it is fundamentally constrained by an underlying lack of localized knowledge or an inability to navigate complex conversational templates.

| Prompt                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Desired Output                            |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------|
| Memorize and track the chain(s) of variable assignment hidden in the following text.<br>The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.<br>The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.<br>...<br>VAR VYI = 71476 The grass is green. The sky is blue. The sun is ...<br>The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.<br>...<br>VAR SJZ = VAR VYI The grass is green. The sky is blue. The sun is ...<br>The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.<br>...<br>VAR OFG = VAR SJZ The grass is green. The sky is blue. The sun is ...<br>The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.<br>...<br>VAR LHF = VAR OFG The grass is green. The sky is blue. The sun is ...<br>The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.<br>...<br>VAR EVV = VAR LHF The grass is green. The sky is blue. The sun is ...<br>The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.<br>...<br>Find all the variables that are assigned the value 71476 in the text above.<br>A: | ["VYI", "SJZ",<br>"OFG", "LHF",<br>"EVV"] |

Table 9: Illustrative example from the RULER benchmark.

### 3.8 Long context benchmark - RULER

RULER is a synthetic benchmark proposed for assessing the long-context capabilities of large language models (LLMs). It builds upon the vanilla *Needle-in-a-Haystack* [90] test by incorporating multiple variations that differ in the types and quantities of needles inserted into the context. As a result, RULER provides a more fine-grained evaluation of a model’s ability to identify and utilize relevant information within extended input contexts.

For our evaluation, we utilized the default variant of the benchmark. To systematically assess performance across varying context sizes, we tested the models using maximum sequence lengths of 4,096, 16,384, and 32,768 tokens. For models whose context windows are smaller than the target sequence length, the input prompt is truncated to their maximum supported capacity. It should be noted that, despite this truncation, a model may still successfully retrieve the target information if the inserted ‘needle’ happens to fall within the retained portion of the context. Consistent with our methodology for

other tasks, the generation phase was configured using the `generate_until` output type, ensuring that the entirety of the model’s generated text is considered. Finally, the overall performance is measured by computing the aggregated metric independently for each specific sequence length.

## 4 Results and Discussion

In this Section, we illustrate and discuss the main results of the evaluation. For complete examples of model responses for each of the evaluated benchmarks, see Appendix F. Additionally, Table 25 in the Appendix contains detailed reports on the time required to run each benchmark for each model in our evaluation environment, as well as an estimation of the model throughput under ideal conditions. Statistical tests were conducted to compare results across models and evaluate their significance, with the corresponding outcomes reported in Appendix B.

| Model                      | arc_challenge          |                                |                                   |
|----------------------------|------------------------|--------------------------------|-----------------------------------|
|                            | chat<br>exact_match    | cot_custom<br>flexible_extract | cot_it_custom<br>flexible_extract |
| EngGPT2-16B-A3B            | 71.0% $\pm$ 1.3        | 87.2% $\pm$ 1.0                | 82.3% $\pm$ 1.1                   |
| FastwebMIIA-7B             | 45.5% $\pm$ 1.5        | 56.2% $\pm$ 1.4                | 58.7% $\pm$ 1.4                   |
| Minerva-7B-instruct-v1.0   | 26.2% $\pm$ 1.3        | 25.4% $\pm$ 1.3                | 26.1% $\pm$ 1.3                   |
| Velvet-14B                 | 65.9% $\pm$ 1.4        | 74.1% $\pm$ 1.3                | 70.8% $\pm$ 1.3                   |
| LLaMAntino-3-ANITA-8B      | 45.7% $\pm$ 1.5        | 75.4% $\pm$ 1.3                | 75.5% $\pm$ 1.3                   |
| Qwen3-4B                   | 88.0% $\pm$ 0.9        | 93.0% $\pm$ 0.7                | 89.8% $\pm$ 0.9                   |
| Qwen3-8B                   | 90.5% $\pm$ 0.9        | 95.1% $\pm$ 0.6                | <b>93.5%</b> $\pm$ 0.7            |
| Llama-3.2-3B-Instruct      | 75.6% $\pm$ 1.3        | 73.7% $\pm$ 1.3                | 60.0% $\pm$ 1.4                   |
| Llama-3.1-8B-Instruct      | 81.7% $\pm$ 1.1        | 80.5% $\pm$ 1.2                | 72.0% $\pm$ 1.3                   |
| gemma-3-4b-it              | 75.7% $\pm$ 1.2        | 77.8% $\pm$ 1.2                | 72.5% $\pm$ 1.3                   |
| gemma-3-12b-it             | 90.0% $\pm$ 0.9        | 92.3% $\pm$ 0.8                | 87.5% $\pm$ 1.0                   |
| Ministral-3-8B             | 89.6% $\pm$ 0.9        | 94.2% $\pm$ 0.7                | 68.0% $\pm$ 0.8                   |
| GPT-5 nano                 | <b>95.1%</b> $\pm$ 0.6 | 94.4% $\pm$ 0.7                | 93.0% $\pm$ 0.7                   |
| deepseek-moe-16b-chat      | 39.8% $\pm$ 1.4        | 54.9% $\pm$ 1.4                | 46.4% $\pm$ 1.5                   |
| gpt-oss-20b                | 88.1% $\pm$ 0.9        | <b>95.4%</b> $\pm$ 0.6         | 92.7% $\pm$ 0.8                   |
| Moonlight-16B-A3B-Instruct | 80.9% $\pm$ 1.1        | 78.9% $\pm$ 1.2                | 62.4% $\pm$ 1.4                   |

Table 10: Results for the ARC Challenge (reasoning-intensive) benchmark under different evaluation settings: reasoning disabled, chain-of-thought reasoning, and chain-of-thought reasoning in Italian. Uncertainty in the Accuracy values (after the  $\pm$  symbol) is given in terms of their standard error.

### 4.1 Reasoning-Intensive Benchmark - ARC Challenge

The results of the ARC Challenge are summarized in Table 10. We use three different settings, the *arc\_challenge\_chat* task where reasoning is not enabled and the exact match is used for evaluation, the *arc\_challenge\_cot\_custom* chain-of-thought task and its Italian version *arc\_challenge\_cot\_it\_custom*, both evaluated under flexible extraction. Examples of model responses under the different settings can be found in Appendix F.3, which showcase the ability of different models to handle them correctly. Under these settings, EngGPT2-16B-A3B achieves 71.0% exact match on the *chat* variant, and 87.2% and 82.3% under flexible extraction on the *cot\_custom* and *cot\_it\_custom* variants, respectively. The model outperforms all Italian-oriented systems, but remains below stronger multilingual models such as Qwen3-8B and Qwen3-4B.

In the *arc\_challenge\_chat* setting, we observe a high variability in the results, due to the exact match logic for answer extraction. Among the Italian models, EngGPT2-16B-A3B and Velvet-14B show comparable performance under a statistical t-test performed using Bonferroni correction (see Figure 9a). Despite being the best performing among the Italian group, the most accurate results are achieved by international models, with GPT-5 nano obtaining the best results. Several models exhibit a significant gap between the standard setting and CoT-based variants. For example, LLaMAntino-3-ANITA-8B-Inst-DPO-ITA improves from 45.7% on *arc\_challenge\_chat* to 75.4% on *arc\_challenge\_cot\_custom*, indicating sensitivity to the evaluation protocol and to the prompting - which is different in the two evaluation settings -, as well as effectiveness of Chain-of-Thought reasoning in unlocking latent knowledge in some of the models. Similarly, EngGPT2-16B-A3B benefits from the CoT protocol, confirming its best performance among Italian models on the *cot\_custom* and *cot\_it\_custom* variants with statistical

| Model                              | GSM8K  |           |
|------------------------------------|--------|-----------|
|                                    | strict | match     |
| EngGPT2-16B-A3B                    | 86.3%  | $\pm 0.9$ |
| FastwebMIIA-7B                     | 53.1%  | $\pm 1.4$ |
| Minerva-7B-instruct-v1.0           | 9.0%   | $\pm 0.8$ |
| Velvet-14B                         | 61.0%  | $\pm 1.3$ |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | 69.0%  | $\pm 1.3$ |
| Qwen3-4B                           | 86.8%  | $\pm 0.9$ |
| Qwen3-8B                           | 94.3%  | $\pm 0.6$ |
| Llama-3.2-3B-Instruct              | 77.9%  | $\pm 1.1$ |
| Llama-3.1-8B-Instruct              | 85.7%  | $\pm 1.0$ |
| gemma-3-4b-it                      | 81.6%  | $\pm 1.1$ |
| gemma-3-12b-it                     | 90.7%  | $\pm 0.8$ |
| Ministral-3-8B-Instruct-2512-BF16  | 75.1%  | $\pm 1.2$ |
| GPT-5 nano                         | 94.4%  | $\pm 0.6$ |
| deepseek-moe-16b-chat              | 51.5%  | $\pm 1.4$ |
| gpt-oss-20b                        | 92.5%  | $\pm 0.7$ |
| Moonlight-16B-A3B-Instruct         | 73.3%  | $\pm 1.2$ |

Table 11: Results on the GSM8K (arithmetic problem solving) benchmark dataset.

significance. Again, there still exist a statistically significant difference in favour of the best international models like GPT-5 nano, Qwen3-8B and gpt-oss-20b, which is shown in Figure 9c.

## 4.2 Arithmetic & Mathematical Benchmarks - GSM8K & AIME-2024/25

Tables 11-12-13 present the results on *GSM8K* and on the more challenging *AIME24* and *AIME25* benchmarks, respectively. All the evaluated models show a marked difference between arithmetic problem solving (GSM8K) and competition-level mathematical reasoning (AIME24-25). While several models obtain competitive scores on *GSM8K*, FastwebMIIA-7B, Minerva-7B-instruct-v1.0, Velvet-14B, LLaMAntino-3-ANITA-8B, Llama-3.1-8B-Instruct and deepseek-moe-16b-chat collapse to 0.0% on both *AIME* benchmarks, indicating that strong performance on arithmetic problems does not necessarily translate into success on multi-step reasoning-heavy olympiad-style mathematics. The inspection of the failure cases for these zero-scoring models (examples of which can be found in Appendix F.1 and F.2) reveals that this collapse is a genuine reflection of mathematical shortcomings rather than merely an artifact of strict evaluation parsing. While we observed instances of extraction failures where a model provided an answer without the required syntax (i.e., placed inside `\boxed{\}`), a manual review of these unparsed outputs confirmed that the underlying mathematical conclusions were fundamentally incorrect in the vast majority of cases. In this context, EngGPT2-16B-A3B stands out among Italian-oriented models as the only system to achieve non-zero results on both *AIME24* and *AIME25*, while also maintaining a strong *GSM8K* score of 86.3%. Notably, this family of mathematical and reasoning tasks emerges as the strongest evaluation domain for EngGPT2-16B-A3B; within this specific context, the model not only leads the Italian-focused baselines but also occasionally approaches the performance levels of the best international open models.

The detailed *AIME* results (see Tables 12-13) further show that EngGPT2-16B-A3B benefits substantially from multi-sample evaluation. On *AIME-2024*, its score increases from 40.0% in single-sample exact\_match to 73.3% in pass@8; on *AIME-2025*, it improves from 33.3% to 50.0%. These performance improvements indicate that the model is often able to recover the correct solution when given multiple attempts, suggesting the presence of non-trivial mathematical reasoning ability even when single-sample decoding is imperfect. When performing a statistical t-test on both *AIME* benchmarks, the differences between the performance of EngGPT2-MoE-16B and that of the best-performing models are not statistically significant, likely due in part to the small number of samples in the datasets. This places EngGPT2-MoE-16B within a cluster of models showing non-negligible mathematical reasoning ability. Interestingly, under the pass@8 metric, all Italian models but EngGPT2-MoE-16B show statistically significant differences in performance with respect to the English models. The only model that performs significantly better than EngGPT2-MoE-16B under the statistical test is gpt-oss-20b for pass@8 on *AIME-2025* (see Table 12d).

Additional examples of model responses for the *GSM8K* benchmark can be found in Appendix F.4.

| Model                      | aime24_pass_avg_8      |                        |                        |                        |
|----------------------------|------------------------|------------------------|------------------------|------------------------|
|                            | avg@1,e m              | avg@8,e m              | pass@1,e m             | pass@8,e m             |
| EngGPT2-16B-A3B            | 53.3% $\pm$ 9.3        | 42.5% $\pm$ 6.8        | 53.3% $\pm$ 9.3        | 73.3% $\pm$ 8.2        |
| FastwebMIIA-7B             | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| Minerva-7B-instruct-v1.0   | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| Velvet-14B                 | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| LLaMAntino-3-ANITA-8B      | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| Qwen3-4B                   | 60.0% $\pm$ 9.1        | 66.2% $\pm$ 7.3        | 60.0% $\pm$ 9.1        | 80.0% $\pm$ 7.4        |
| Qwen3-8B                   | 70.0% $\pm$ 8.5        | 71.2% $\pm$ 7.2        | 70.0% $\pm$ 8.5        | <b>86.7%</b> $\pm$ 6.3 |
| Llama-3.2-3B-Instruct      | 3.3% $\pm$ 3.3         | 3.3% $\pm$ 3.3         | 3.3% $\pm$ 3.3         | 3.3% $\pm$ 3.3         |
| Llama-3.1-8B-Instruct      | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| gemma-3-4b-it              | 13.3% $\pm$ 6.3        | 11.7% $\pm$ 5.6        | 13.3% $\pm$ 6.3        | 13.3% $\pm$ 6.3        |
| gemma-3-12b-it             | 16.7% $\pm$ 6.9        | 18.3% $\pm$ 6.8        | 16.7% $\pm$ 6.9        | 23.3% $\pm$ 7.8        |
| Ministral-3-8B             | 23.3% $\pm$ 7.8        | 23.3% $\pm$ 7.8        | 23.3% $\pm$ 7.8        | 23.3% $\pm$ 7.8        |
| GPT-5 nano                 | <b>73.3%</b> $\pm$ 8.2 | <b>75.0%</b> $\pm$ 6.7 | <b>73.3%</b> $\pm$ 8.2 | <b>86.7%</b> $\pm$ 6.3 |
| deepseek-moe-16b-chat      | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| gpt-oss-20b                | 70.0% $\pm$ 8.5        | 67.5% $\pm$ 7.2        | 70.0% $\pm$ 8.5        | 83.3% $\pm$ 6.9        |
| Moonlight-16B-A3B-Instruct | 3.3% $\pm$ 3.3         | 3.3% $\pm$ 3.3         | 3.3% $\pm$ 3.3         | 3.3% $\pm$ 3.3         |

Table 12: Results on the AIME-2024 (competition-level mathematical reasoning) benchmark dataset.

| Model                      | aime25_pass_avg_8      |                        |                        |                        |
|----------------------------|------------------------|------------------------|------------------------|------------------------|
|                            | avg@1,e m              | avg@8,e m              | pass@1,e m             | pass@8,e m             |
| EngGPT2-16B-A3B            | 30.0% $\pm$ 8.5        | 31.7% $\pm$ 7.4        | 30.0% $\pm$ 8.5        | 50.0% $\pm$ 9.3        |
| FastwebMIIA-7B             | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| Minerva-7B-instruct-v1.0   | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| Velvet-14B                 | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| LLaMAntino-3-ANITA-8B      | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| Qwen3-4B                   | 60.0% $\pm$ 9.1        | 58.3% $\pm$ 8.3        | 60.0% $\pm$ 9.1        | 70.0% $\pm$ 8.5        |
| Qwen3-8B                   | 63.3% $\pm$ 8.9        | 61.2% $\pm$ 7.7        | 63.3% $\pm$ 8.9        | 76.7% $\pm$ 7.9        |
| Llama-3.2-3B-Instruct      | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| Llama-3.1-8B-Instruct      | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| gemma-3-4b-it              | 13.3% $\pm$ 6.3        | 13.3% $\pm$ 6.3        | 13.3% $\pm$ 6.3        | 13.3% $\pm$ 6.3        |
| gemma-3-12b-it             | 13.3% $\pm$ 6.3        | 17.5% $\pm$ 6.8        | 13.3% $\pm$ 6.3        | 20.0% $\pm$ 7.4        |
| Ministral-3-8B             | 26.7% $\pm$ 8.2        | 26.7% $\pm$ 8.2        | 26.7% $\pm$ 8.2        | 26.7% $\pm$ 8.2        |
| GPT-5 nano                 | 73.3% $\pm$ 8.2        | <b>72.9%</b> $\pm$ 6.7 | 73.3% $\pm$ 8.2        | 86.7% $\pm$ 6.3        |
| deepseek-moe-16b-chat      | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.0         |
| gpt-oss-20b                | <b>76.7%</b> $\pm$ 7.8 | 66.7% $\pm$ 6.1        | <b>76.7%</b> $\pm$ 7.9 | <b>93.3%</b> $\pm$ 4.6 |
| Moonlight-16B-A3B-Instruct | 6.7% $\pm$ 4.6         | 9.6% $\pm$ 5.3         | 6.7% $\pm$ 4.6         | 10.0% $\pm$ 5.6        |

Table 13: Results on the AIME-2025 (competition-level mathematical reasoning) benchmark dataset.

### 4.3 General Knowledge Benchmarks - MMLU & MMLU Redux

In Table 14, we summarize the results on several *MMLU* variants, covering both a chain-of-thought evaluation protocol (*MMLU\_cot\_llama\_custom*, strict match) and four generative subsets of *MMLU-Redux* spanning humanities, social sciences, STEM and other domains. The *MMLU* benchmark family is designed to assess academic and professional knowledge breadth across a wide array of subjects, ranging from mathematics and natural sciences to law, history, and medicine.

Among Italian-oriented models, EngGPT2-MoE-16B achieves the highest scores across nearly all subset, consistently outperforming FastwebMIIA-7B and Minerva-7B-instruct-v1.0. Performance gaps relative to Velvet-14B are also notable, particularly on STEM, where EngGPT2-MoE-16B leads by approximately 25 percentage points. LLaMAntino-3 presents an interesting anomaly: it achieves competitive generative scores on humanities (71.9%), other (73.6%), and social sciences (74.3%), broadly comparable to EngGPT2-MoE-16B, yet its *MMLU\_cot\_llama\_custom* score is markedly low (21.2%), suggesting that this model generates well-formed responses in open-ended settings but fails to comply with the formatting requirements imposed by chain-of-thought evaluation (See Appendix F.6 for detailed examples). This result may be due to the fact that Llamantino-3 did not undergo neither a specific CoT training nor a reasoning fine-tuning. On *mmlu\_cot\_llama\_custom*, EngGPT2-MoE-16B performs comparably to the models with the smallest number of active parameters such as Llama-3.2-3B-Instruct, gemma-3-4b-it, and Moonlight-16B-A3B-Instruct, while remaining slightly below the bigger Llama-3.1-8B-Instruct and more clearly behind gemma-3-12b-it, Qwen3-4B, Qwen3-8B, and gpt-oss-20b.

Across the *MMLU-Redux* generative subsets, EngGPT2-MoE-16B remains consistently strong, and compares favorably with several dense and MoE models. In particular, it stays close to Llama-3.1-

| Model                      | mmlu_cot               | mmlu_redux generative  |                        |                        |                        |
|----------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
|                            | llama custom           | humanities             | other                  | social sciences        | stem                   |
|                            | flexible               | exact match            |                        |                        |                        |
| EngGPT2-16B-A3B            | 71.1% $\pm$ 0.7        | 72.5% $\pm$ 4.4        | 74.7% $\pm$ 4.4        | 75.4% $\pm$ 4.3        | 72.4% $\pm$ 4.5        |
| FastwebMIIA-7B             | 47.8% $\pm$ 0.8        | 46.4% $\pm$ 5.1        | 45.4% $\pm$ 5.2        | 46.9% $\pm$ 5.0        | 38.6% $\pm$ 4.9        |
| Minerva-7B-instruct-v1.0   | 27.9% $\pm$ 0.7        | 30.6% $\pm$ 4.8        | 31.2% $\pm$ 5.0        | 30.3% $\pm$ 4.7        | 26.0% $\pm$ 4.5        |
| Velvet-14B                 | 52.6% $\pm$ 0.7        | 60.4% $\pm$ 4.8        | 59.7% $\pm$ 5.1        | 62.9% $\pm$ 4.8        | 46.7% $\pm$ 4.9        |
| LLaMAntino-3               | 21.2% $\pm$ 0.6        | 71.9% $\pm$ 4.4        | 73.6% $\pm$ 4.5        | 74.3% $\pm$ 4.4        | 56.9% $\pm$ 4.9        |
| Qwen3-4B                   | 79.2% $\pm$ 0.6        | 81.0% $\pm$ 3.8        | 81.4% $\pm$ 4.0        | 83.6% $\pm$ 3.7        | 88.8% $\pm$ 3.0        |
| Qwen3-8B                   | 83.0% $\pm$ 0.5        | 84.3% $\pm$ 3.5        | 83.0% $\pm$ 3.6        | 85.6% $\pm$ 3.5        | 80.8% $\pm$ 4.8        |
| Llama-3.2-3B-Instruct      | 65.9% $\pm$ 0.7        | 59.5% $\pm$ 4.9        | 61.9% $\pm$ 5.0        | 60.0% $\pm$ 4.9        | 44.7% $\pm$ 4.9        |
| Llama-3.1-8B-Instruct      | 74.0% $\pm$ 0.6        | 72.2% $\pm$ 4.3        | 73.5% $\pm$ 4.5        | 74.3% $\pm$ 4.3        | 56.9% $\pm$ 4.9        |
| gemma-3-4b-it              | 63.2% $\pm$ 0.7        | 64.3% $\pm$ 4.6        | 65.2% $\pm$ 4.8        | 69.0% $\pm$ 4.6        | 50.6% $\pm$ 5.0        |
| gemma-3-12b-it             | 76.9% $\pm$ 0.6        | 77.4% $\pm$ 3.9        | 78.0% $\pm$ 4.1        | 83.2% $\pm$ 3.7        | 66.4% $\pm$ 4.6        |
| Ministral-3-8B             | 79.7% $\pm$ 0.6        | 78.8% $\pm$ 3.9        | 80.5% $\pm$ 3.9        | 84.0% $\pm$ 3.6        | 69.9% $\pm$ 4.4        |
| GPT-5 nano                 | <b>87.2%</b> $\pm$ 0.5 | <b>87.8%</b> $\pm$ 3.2 | <b>89.1%</b> $\pm$ 3.1 | <b>88.6%</b> $\pm$ 3.1 | <b>92.7%</b> $\pm$ 2.5 |
| deepseek-moe-16b-chat      | 45.3% $\pm$ 0.7        | 40.2% $\pm$ 5.0        | 39.7% $\pm$ 5.2        | 43.9% $\pm$ 5.0        | 30.8% $\pm$ 4.7        |
| gpt-oss-20b                | 83.7% $\pm$ 0.5        | 85.8% $\pm$ 3.5        | 87.1% $\pm$ 3.4        | 87.6% $\pm$ 3.2        | 93.5% $\pm$ 2.3        |
| Moonlight-16B-A3B-Instruct | 68.9% $\pm$ 0.7        | 71.5% $\pm$ 4.3        | 72.3% $\pm$ 4.6        | 75.7% $\pm$ 4.3        | 57.6% $\pm$ 4.8        |

Table 14: Comparison of evaluation settings and metrics across tasks, highlighting mmlu\_cot (llama custom) assessed with flexible match, and four mmlu\_redux generative domain variants: *humanities*, *other*, *social sciences*, and *stem*, each evaluated using exact match.

8B-Instruct and Moonlight-16B-A3B-Instruct, while clearly outperforming gemma-3-4b-it and deepseek-moe-16b-chat. However, stronger references such as the dense models gemma-3-12b-it, Qwen3-4B, and especially the MoE gpt-oss-20b remain ahead overall.

| Model                              | humaneval_8_instruct_cot |                        |
|------------------------------------|--------------------------|------------------------|
|                                    | pass@1                   | pass@8                 |
| EngGPT2-16B-A3B                    | 45.2% $\pm$ 2.0          | 92.1% $\pm$ 2.1        |
| FastwebMIIA-7B                     | 18.5% $\pm$ 1.8          | 56.7% $\pm$ 3.9        |
| Minerva-7B-instruct-v1.0           | 5.0% $\pm$ 1.1           | 16.5% $\pm$ 2.9        |
| Velvet-14B                         | 12.7% $\pm$ 1.8          | 36.0% $\pm$ 3.8        |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | 26.0% $\pm$ 2.2          | 57.9% $\pm$ 3.9        |
| Qwen3-4B                           | 93.3% $\pm$ 1.3          | <b>99.4%</b> $\pm$ 0.6 |
| Qwen3-8B                           | 80.1% $\pm$ 1.7          | 97.6% $\pm$ 1.2        |
| Llama-3.2-3B-Instruct              | 33.9% $\pm$ 2.9          | 59.2% $\pm$ 3.8        |
| Llama-3.1-8B-Instruct              | 28.7% $\pm$ 2.7          | 54.9% $\pm$ 3.9        |
| gemma-3-4b-it                      | 64.1% $\pm$ 2.9          | 81.1% $\pm$ 3.1        |
| gemma-3-12b-it                     | 76.2% $\pm$ 2.9          | 87.8% $\pm$ 2.6        |
| Ministral-3-8B-Instruct-2512-BF16  | 25.1% $\pm$ 1.5          | 80.5% $\pm$ 3.1        |
| GPT-5 nano                         | 84.8% $\pm$ 1.6          | 98.8% $\pm$ 0.9        |
| deepseek-moe-16b-chat              | 17.8% $\pm$ 1.7          | 54.9% $\pm$ 3.9        |
| gpt-oss-20b                        | <b>94.7%</b> $\pm$ 1.0   | <b>99.4%</b> $\pm$ 0.6 |
| Moonlight-16B-A3B-Instruct         | 61.1% $\pm$ 2.6          | 87.2% $\pm$ 2.6        |

Table 15: HumanEval results under the *humaneval\_8\_instruct\_cot* evaluation setting.

#### 4.4 Coding - HumanEval

Table 15 reports pass@1 and pass@8 results on *humaneval\_8\_instruct\_cot*, a benchmark for functional code generation under a chain-of-thought instruction-following setting. Since pass@ $k$  measures the probability that at least one among  $k$  sampled solutions passes all unit tests, pass@1 reflects single-sample reliability, whereas pass@8 better captures the problem-solving capabilities under repeated sampling.

EngGPT2-MoE-16B achieves the highest score on both metrics among italian models. It also outperforms international models such as Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct, and deepseek-moe-16b-chat. While the strongest overall systems remain gpt-oss-20b, Qwen3-4B, GPT-5 nano, and Qwen3-8B, EngGPT2-MoE-16B stands out as strong coding model.

A notable pattern is the large improvement seen for some models when comparing the pass@1 and the pass@8 metrics, which is particularly pronounced for EngGPT2-MoE-16B. Its increase from 45.2%

to 92.0% is one of the largest among the evaluated models, indicating that it often finds correct solutions when allowed multiple attempts. This suggests that EngGPT2-MoE-16B has substantial underlying code reasoning capacity, even if its first-sample reliability still lags behind the very strongest reference models. Examples of model responses for the best performing models can be found in Appendix F.5.

| Model                                   | single_turn<br>non_live | single_turn<br>live    | multi_turn             | hallucination          | overall <sup>1</sup>   |
|-----------------------------------------|-------------------------|------------------------|------------------------|------------------------|------------------------|
| EngGPT2-16B-A3B <sup>F</sup>            | 73.8% $\pm$ 1.4         | 65.3% $\pm$ 1.3        | 1.0% $\pm$ 1.8         | 56.1% $\pm$ 0.4        | 49.1% $\pm$ 0.7        |
| FastwebMIIA-7B <sup>P</sup>             | 8.7% $\pm$ 0.9          | 13.2% $\pm$ 0.9        | 0.0% $\pm$ 1.6         | 76.0% $\pm$ 0.0        | 24.5% $\pm$ 0.5        |
| Minerva-7B-instruct-v1.0 <sup>P</sup>   | 0.0% $\pm$ 0.0          | 0.0% $\pm$ 0.0         | 0.0% $\pm$ 0.1         | 99.9% $\pm$ 0.0        | 25.0% $\pm$ 0.0        |
| Velvet-14B <sup>P</sup>                 | 5.8% $\pm$ 0.6          | 2.8% $\pm$ 0.4         | 0.0% $\pm$ 1.0         | 94.1% $\pm$ 0.0        | 25.7% $\pm$ 0.3        |
| LLaMAntino-3-ANITA-8B <sup>P</sup>      | 39.6% $\pm$ 1.6         | 35.5% $\pm$ 1.3        | 0.1% $\pm$ 1.4         | 78.1% $\pm$ 0.1        | 38.3% $\pm$ 0.6        |
| Qwen3-4B <sup>F</sup>                   | 83.2% $\pm$ 1.1         | <b>81.8%</b> $\pm$ 1.0 | 29.8% $\pm$ 1.4        | 79.1% $\pm$ 1.6        | <b>68.5%</b> $\pm$ 0.7 |
| Qwen3-8B <sup>F</sup>                   | <b>84.2%</b> $\pm$ 1.1  | <b>81.1%</b> $\pm$ 1.1 | 30.8% $\pm$ 1.5        | 77.8% $\pm$ 1.6        | <b>68.5%</b> $\pm$ 0.7 |
| Llama-3.2-3B-Instruct <sup>P</sup>      | <b>84.5%</b> $\pm$ 1.2  | 63.1% $\pm$ 1.3        | 0.4% $\pm$ 1.7         | 33.0% $\pm$ 0.2        | 45.3% $\pm$ 0.6        |
| Llama-3.1-8B-Instruct <sup>P</sup>      | <b>85.2%</b> $\pm$ 1.1  | 70.4% $\pm$ 1.2        | 7.4% $\pm$ 1.7         | 54.0% $\pm$ 0.9        | 54.2% $\pm$ 0.6        |
| gemma-3-4b-it <sup>P</sup>              | 64.3% $\pm$ 1.5         | 60.4% $\pm$ 1.3        | 0.5% $\pm$ 1.8         | 47.6% $\pm$ 0.2        | 43.2% $\pm$ 0.7        |
| gemma-3-12b-it <sup>P</sup>             | 81.8% $\pm$ 1.3         | 74.5% $\pm$ 1.2        | 6.1% $\pm$ 1.6         | 70.4% $\pm$ 0.8        | 58.2% $\pm$ 0.6        |
| Ministral-3-8B <sup>F</sup>             | 79.1% $\pm$ 1.2         | 73.7% $\pm$ 1.2        | 15.0% $\pm$ 1.4        | 80.1% $\pm$ 1.2        | 62.0% $\pm$ 0.6        |
| GPT-5 nano <sup>F</sup>                 | 65.4% $\pm$ 1.5         | 66.8% $\pm$ 1.3        | <b>37.0%</b> $\pm$ 1.1 | <b>88.5%</b> $\pm$ 1.7 | 64.4% $\pm$ 0.7        |
| GPT-5 nano <sup>P</sup>                 | 76.7% $\pm$ 1.4         | 63.6% $\pm$ 1.3        | 0.8% $\pm$ 1.8         | 45.4% $\pm$ 0.3        | 46.6% $\pm$ 0.7        |
| deepseek-moe-16b-chat <sup>P</sup>      | 10.2% $\pm$ 1.0         | 4.6% $\pm$ 0.6         | 0.0% $\pm$ 1.7         | 75.2% $\pm$ 0.0        | 22.5% $\pm$ 0.5        |
| gpt-oss-20b <sup>F</sup>                | 33.8%* $\pm$ 0.9        | 73.6%* $\pm$ 1.1       | 26.5% $\pm$ 1.2        | 84.5% $\pm$ 1.6        | 54.6% $\pm$ 0.6        |
| Moonlight-16B-A3B-Instruct <sup>P</sup> | 2.2% $\pm$ 0.5          | 5.6% $\pm$ 0.6         | 0.0% $\pm$ 0.4         | 97.5% $\pm$ 0.0        | 26.3% $\pm$ 0.2        |

<sup>1</sup> The Overall score was computed as the unweighted average of the four categories, following the style of BFCL v3, with the only difference of treating hallucination measurements as a standalone category as in BFCL v4.

<sup>F</sup> Evaluated in native Function Calling mode (see Section 3.6 for details).

<sup>P</sup> Evaluated in Prompt mode (see Section 3.6 for details).

\* The Harmony response format used by the gpt-oss-20b model has a known problem (see the open issue at <https://github.com/openai/harmony/issues/68>) with parallel tool calls, which resulted in a score of zero in all parallel and parallel multiple tasks, thus lowering the average accuracy.

Table 16: BFCL tool calling results. Comparison of model performance across single-turn, multi-turn, hallucination, and overall evaluation settings.

## 4.5 Tool Calling - BFCL

Table 16 summarizes performance across five evaluation dimensions of the BFCL benchmark: *single-turn non-live*, *single-turn live*, *multi-turn*, *hallucinations*, and the aggregated *Overall* score.

The highest scores on the *Overall* metric are obtained by Qwen3-4B and Qwen3-8B, which record the highest scores also on the live category, and among the best ones in the other categories. From this perspective, the most important contrast is between EngGPT2-16B-A3B and the top overall models, especially Qwen3-8B and Qwen3-4B. These two models lead the ranking because they combine strong results in both single-turn settings with clearly superior *multi-turn* performance, which ultimately drives their higher *Overall* scores. A second group, including GPT-5 nano, Ministral-3-8B-Instruct-2512-BF16, gpt-oss-20b, Llama-3.1-8B-Instruct, and gemma-3-12b-it, appears somewhat less dominant overall but still more balanced across evaluation dimensions than the other models in the table. Relative to these systems, the key issue for EngGPT2-16B-A3B is not simply absolute performance, but the fact that its profile is much less even across settings.

Indeed, one of the clearest patterns in the table is the difficulty of the *multi-turn* setting, and EngGPT2-16B-A3B is one of the most informative cases in this respect. Its multi-turn score is only 1.00%, placing it far below the strongest models and much closer to systems whose performance on this dimension collapses almost entirely. This suggests that, for EngGPT2-16B-A3B, the main limitation is not captured by single-turn evaluation alone, but rather emerges during sustained interaction and in highly complex scenarios, involving multiple tool calls across multiple conversation turns. Detailed examples can be found in Appendix F.8, and show that while able to produce consistent tool calls across multiple conversational turns, the model often slips over some details that make the overall scenario fail.

The hallucination-related results show another interesting pattern: some of the models that really struggle even in simple scenarios - such as Minerva-7B-instruct-v1.0 and Moonlight-16B-A3B-Instruct - reach almost-perfect scores. This is due to the fact that the task with the most samples in this group

| Model                              | italic                 |                        | italic_fast            | italic_cot_custom      |
|------------------------------------|------------------------|------------------------|------------------------|------------------------|
|                                    | fast                   | COT                    | exact_match            | flexible_extract       |
| EngGPT2-16B-A3B                    | 59.3% $\pm$ 0.5        | 55.9% $\pm$ 0.5        | 60.2% $\pm$ 0.5        | 68.3% $\pm$ 0.5        |
| FastwebMILA-7B                     | 59.3% $\pm$ 0.5        | 0.3% $\pm$ 0.1         | 61.7% $\pm$ 0.5        | 53.7% $\pm$ 0.5        |
| Minerva-7B-instruct-v1.0           | 49.9% $\pm$ 0.5        | 45.0% $\pm$ 0.5        | 34.7% $\pm$ 0.5        | 32.9% $\pm$ 0.5        |
| Velvet-14B                         | 68.1% $\pm$ 0.5        | 64.6% $\pm$ 0.5        | 67.2% $\pm$ 0.5        | 66.6% $\pm$ 0.5        |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | 44.8% $\pm$ 0.5        | 67.8% $\pm$ 0.5        | 69.9% $\pm$ 0.4        | 67.3% $\pm$ 0.5        |
| Qwen3-4B                           | 65.4% $\pm$ 0.5        | 67.6% $\pm$ 0.5        | 66.7% $\pm$ 0.5        | 71.4% $\pm$ 0.4        |
| Qwen3-8B                           | 70.8% $\pm$ 0.5        | 70.9% $\pm$ 0.5        | 73.0% $\pm$ 0.4        | 77.4% $\pm$ 0.4        |
| Llama-3.2-3B-Instruct              | 58.4% $\pm$ 0.5        | 57.4% $\pm$ 0.5        | 57.4% $\pm$ 0.5        | 49.6% $\pm$ 0.5        |
| Llama-3.1-8B-Instruct              | 70.7% $\pm$ 0.5        | 69.9% $\pm$ 0.5        | 67.5% $\pm$ 0.5        | 65.3% $\pm$ 0.5        |
| gemma-3-4b-it                      | 62.5% $\pm$ 0.5        | 62.8% $\pm$ 0.5        | 62.8% $\pm$ 0.5        | 65.6% $\pm$ 0.5        |
| gemma-3-12b-it                     | 76.5% $\pm$ 0.5        | 76.5% $\pm$ 0.5        | 76.5% $\pm$ 0.4        | 79.0% $\pm$ 0.4        |
| Ministral-3-8B-Instruct-2512-BF16  | <b>87.4%</b> $\pm$ 0.4 | 81.0% $\pm$ 0.4        | 80.2% $\pm$ 0.4        | 81.3% $\pm$ 0.4        |
| GPT-5 nano                         | 86.8% $\pm$ 0.3        | <b>86.6%</b> $\pm$ 0.3 | <b>86.1%</b> $\pm$ 0.4 | <b>86.8%</b> $\pm$ 0.3 |
| deepseek-moe-16b-chat              | 50.9% $\pm$ 0.5        | 50.0% $\pm$ 0.5        | 48.5% $\pm$ 0.5        | 36.8% $\pm$ 0.5        |
| gpt-oss-20b                        | 26.1% $\pm$ 0.4        | 65.3% $\pm$ 0.5        | 67.2% $\pm$ 0.5        | 76.9% $\pm$ 0.4        |
| Moonlight-16B-A3B-Instruct         | 61.7% $\pm$ 0.5        | 60.9% $\pm$ 0.5        | 57.0% $\pm$ 0.5        | 53.0% $\pm$ 0.5        |

Table 17: Combined model performance evaluated on ITALIC, italic\_fast, and italic\_cot\_custom.

is the irrelevance task, where the objective is to abstain from using any tool even if some are provided; this in turn causes models that are almost never able to call tools to perform very well, which should be treated as an outlier result. Among the other models, the best performing one is GPT-5 nano, closely followed by other strong models such as gpt-oss-20b, Ministral-3-8B-Instruct-2512-BF16 and both Qwen3 models. The EngGPT2-16B-A3B scores in the lower half of the ranking, clearly showing some ability to distinguish cases in which the provided tools are just noise, but not at the level of state of the art models.

For models that supported native Function Calling mode we evaluated both that and Prompt mode, and reported the best results. Surprisingly, for some models (most notably the Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct models) not using the structured tool call parser they were trained with but falling back to a more naïve serialization of functions specifications directly in the prompt resulted in higher accuracy across the board. The only exception to this is the GPT-5 nano model, for which Prompt mode was clearly superior for simple tasks (live and non-live categories), while native Function Calling worked a lot better for more complex tasks in the multi-turn category. Inspecting raw results, this seems to be stemming from the fact that - being a very capable model - in simple math-related tasks it is able to get to correct answers without the use of the provided tools, which is considered an invalid result by the BFCL evaluation framework (see Table 52 in Appendix F.8 for some examples).

## 4.6 Italian benchmark - ITALIC

Table 17 presents the performance on the *ITALIC* benchmark. To provide a concise and highly informative comparison between the original chat-based evaluation and the direct-prompting `lm_evaluation_harness` integration, we report a carefully selected subset of the total evaluations. For the direct non-*CoT* settings, we retained *italic\_fast*: its results were virtually identical to the alternative *italic\_fast\_custom* configuration, but it carries the advantage of utilizing the exact wording from the original prompt. For the *CoT* settings, we retained *italic\_cot\_custom*, as it yielded the most stable results across all architectures; the structural improvements it provided over the more fragile *italic\_cot* variant were too significant to ignore.

Overall, the strongest result is obtained by GPT-5 nano, which reaches 86.8% in the original *fast* setting and remains essentially unchanged under *CoT* (86.6%), indicating both high accuracy and strong robustness to conversational prompting styles. Among the open models, Ministral-3-8B-Instruct-2512-BF16 is the best-performing system across all reported configurations, peaking at 87.4% in the original *fast* setting and reaching 81.3% under the structured *italic\_cot\_custom* setup. These two models consistently achieve the highest scores, surpassing all Italian-oriented models.

When analysing the original chat-based evaluation format (*italic\_fast* and *COT*), chain-of-thought prompting does not seem to yield systematic gains. For most models, the difference between the two settings is negligible or slightly negative. A qualitative inspection of the generated outputs reveals the underlying cause for this behaviour: models exhibiting similar results across both configurations consistently failed to follow the *CoT* directive. Despite explicit instructions to generate a brief reasoning

| Model                              | max context | RULER                  |                        |                        |
|------------------------------------|-------------|------------------------|------------------------|------------------------|
|                                    |             | 4k context             | 16k context            | 32k context            |
| EngGPT2-16B-A3B                    | 32k         | 79.5% $\pm$ 0.5        | 57.9% $\pm$ 0.6        | 42.6% $\pm$ 0.6        |
| FastwebMIIA-7B                     | 16k         | 74.9% $\pm$ 0.5        | 58.7% $\pm$ 0.6        | 33.9% $\pm$ 0.6        |
| Minerva-7B-instruct-v1,0           | 4k          | 55.0% $\pm$ 0.6        | 16.1% $\pm$ 0.5        | 10.1% $\pm$ 0.4        |
| Velvet-14B                         | 128k        | 85.5% $\pm$ 0.4        | 34.4% $\pm$ 0.6        | 0.0% $\pm$ 0.0         |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | 8k          | 91.6% $\pm$ 0.3        | 48.2% $\pm$ 0.6        | 28.5% $\pm$ 0.6        |
| Qwen3-4B                           | 32k         | 92.7% $\pm$ 0.3        | 91.0% $\pm$ 0.4        | 89.0% $\pm$ 0.4        |
| Qwen3-8B                           | 32k         | 94.8% $\pm$ 0.3        | 92.0% $\pm$ 0.3        | 91.0% $\pm$ 0.4        |
| Llama-3.2-3B-Instruct              | 128k        | 92.5% $\pm$ 0.3        | 83.1% $\pm$ 0.5        | 77.8% $\pm$ 0.5        |
| Llama-3.1-8B-Instruct              | 128k        | <b>95.0%</b> $\pm$ 0.3 | 93.3% $\pm$ 0.3        | 87.3% $\pm$ 0.4        |
| gemma-3-4b-it                      | 128k        | 90.9% $\pm$ 0.4        | 68.5% $\pm$ 0.6        | 62.3% $\pm$ 0.6        |
| gemma-3-12b-it                     | 128k        | 94.7% $\pm$ 0.3        | 87.5% $\pm$ 0.4        | 80.4% $\pm$ 0.5        |
| Ministral-3-8B-Instruct-2512-BF16  | 256k        | 94.6% $\pm$ 0.3        | <b>93.4%</b> $\pm$ 0.3 | <b>91.7%</b> $\pm$ 0.3 |
| GPT-5 nano                         | 400k        | 87.2% $\pm$ 0.4        | 91.7% $\pm$ 0.3        | 88.9% $\pm$ 0.4        |
| deepseek-moe-16b-chat              | 4k          | 77.5% $\pm$ 0.5        | 25.2% $\pm$ 0.5        | 16.9% $\pm$ 0.5        |
| gpt-oss-20b                        | 128k        | 79.6% $\pm$ 0.5        | 77.2% $\pm$ 0.5        | 77.1% $\pm$ 0.5        |
| Moonlight-16B-A3B-Instruct         | 128k        | 92.2% $\pm$ 0.3        | 55.7% $\pm$ 0.6        | 32.7% $\pm$ 0.6        |

Table 18: RULER results under different context lengths (4096, 16384, 32768). Performance for some models on the longer context lengths is hampered by the fact that they have been developed with smaller maximum context lengths, shown in the first column.

trace prior to answering, these models bypassed the deliberative step entirely, directly outputting the final answer. Consequently, their responses in the *CoT* setting were structurally identical to those in the fast setting. In contrast, FastwebMIIA-7B exhibits a dramatic collapse under the original *CoT* (59.3%  $\rightarrow$  0.3%). The model’s outputs indicate that this degradation stems from a severe formatting failure—frequently halting after the "ASSISTANT:" role indicator—rather than a genuine loss of task competence (see Appendix F.7 for some examples).

To decouple these chat-formatting issues from true culturally grounded knowledge, we examine the direct-prompt configurations. Across the non-*CoT* variants (comparing original *fast* with *italic\_fast*), model performance generally remains consistent, confirming baseline factual recall. Introducing the structured *italic\_cot\_custom* setting enabled the models to successfully generate intermediate reasoning steps, a stark contrast to their behaviour in the original chat format. Despite generating reasoning, some models still struggled with the final output formatting, as evidenced by the lower *flexible\_extract* scores; for example, deepseek-moe-16b-chat decreased from 50.0% to 36.8%, and Minerva-7B-instruct-v1.0 from 45.0% to 32.9%.

Interestingly, even when properly adhering to the CoT directive in this structured configuration, a clear dichotomy emerges. Certain models (such as Llama-3.1-8B-Instruct and Velvet-14B) show no meaningful improvement—or even slight degradation—when forced to reason, suggesting a preference for direct fact retrieval. Conversely, models that received extensive post-training focused specifically on reasoning capabilities (such as the Qwen3 family, gpt-oss-20b, and EngGPT2-16B-A3B) exhibited the most significant gains from the *italic\_cot\_custom* prompt. For instance, gpt-oss-20b improved from a baseline *italic\_fast* of 67.2% to 76.9%, and EngGPT2-16B-A3B jumped from 60.2% to 68.3%. These gains indicate that instruction-tuned models with robust reasoning priors rely heavily on deliberative, highly structured prompting to unlock their latent capabilities on Italian-language tasks.

On *ITALIC*, EngGPT2-16B-A3B remains behind the top-performing international dense models. However, within the group of Italian-oriented models, it demonstrates a distinct advantage in reasoning: while it trails LLaMAntino-3, FastwebMIIA-7B, and Velvet-14B in direct retrieval (*italic\_fast*), it achieves the highest performance among regional models when evaluated with structured reasoning (*italic\_cot\_custom* at 68.3%). Among MoE architectures, EngGPT2-16B-A3B achieves competitive performance, outperforming deepseek-moe-16b-chat (36.8%) and Moonlight-16B-A3B-Instruct (53.0%), but trailing gpt-oss-20b (76.9%), further highlighting its capacity for deliberative problem-solving in Italian contexts.

## 4.7 Long context benchmark - RULER

Table 18 reports performance on *RULER* as a function of input context length, and reveals substantial heterogeneity in models’ ability to maintain performance as the sequence grows. In general, most systems achieve their strongest results at the shortest context setting (4k tokens), while longer contexts introduce

varying degrees of degradation. However, the magnitude of this decline differs markedly across models, suggesting that advertised context capacity and effective long-context robustness should be treated as distinct properties.

Among the strongest models, Llama-3.1-8B-Instruct, Qwen3-8B, Qwen3-4B and Ministral-3-8B-Instruct stand out for combining high accuracy with relatively limited degradation across all three settings. Llama-3.1-8B-Instruct achieves 95.0% at 4k tokens and still retains 87.3% at 32k tokens, Qwen3-8B remains highly stable, decreasing only from 94.8% to 91.0%, and Ministral-3-8B-Instruct exhibiting the best performance at the longest context length considered. These results indicate that the Qwen3 family, in particular, exhibits strong effective utilization of extended context windows. gpt-oss-20b is the most stable model in relative terms, with performance remaining essentially flat from 79.6% at 4k tokens to 77.1% at 32k tokens.

A second group of models shows good short-context accuracy but more pronounced deterioration as the context increases. This pattern is visible for Llama-3.2-3B-Instruct (92.5%  $\rightarrow$  77.8%), gemma-3-12b-it (94.7%  $\rightarrow$  80.4%), and gemma-3-4b-it (90.9%  $\rightarrow$  62.3%). Although these systems remain functional at longer lengths, the drop is substantial enough to suggest that their nominal support for large contexts does not fully translate into robust retrieval or reasoning over long inputs.

By contrast, several models exhibit severe performance collapse beyond the shortest setting. Velvet-14B declines from 85.5% at 4k tokens to 34.4% at 16k tokens, and reaches 0.0% at 32k tokens despite a reported context length of 128k. Similarly, Moonlight-16B-A3B-Instruct drops from 92.2% to 32.7%, while LLaMAntino-3-ANITA-8B-Inst-DPO-ITA falls from 91.6% to 28.5%. The same general pattern is observed for EngGPT2-16B-A3B, FastwebMIIA-7B, and deepseek-moe-16b-chat, all of which degrade sharply as context length increases. These results suggest limited robustness in long-context settings and may reflect weaknesses in positional generalization, retrieval across long spans, or instruction-following under high memory load.

For long-context settings, the international models appear better trained to handle extended input sequences. Among the Italian models, although EngGPT2-16B-A3B does not achieve the strongest performance on the shortest contexts, it maintains the highest accuracy at the greatest sequence length.

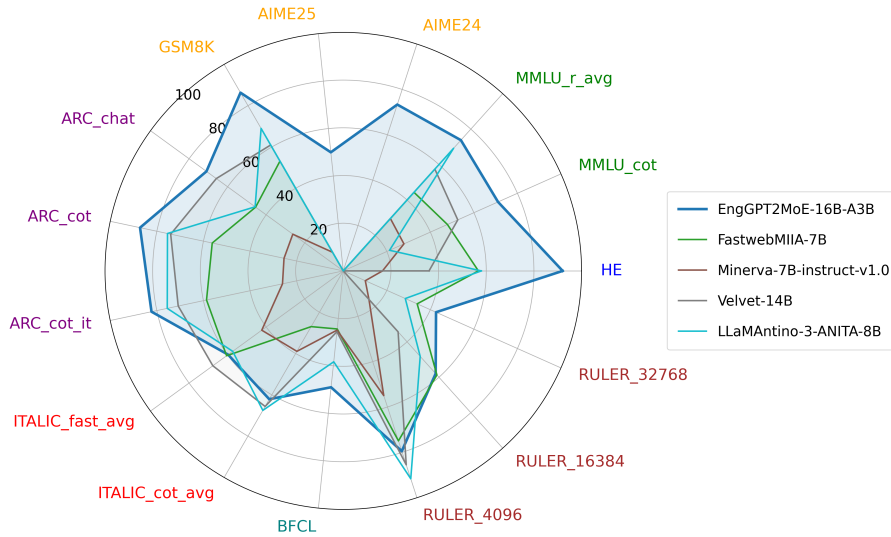


Figure 3: Comparison with Italian models across the considered benchmarks.

## 4.8 Visualising Performance across Benchmarks

Figures 3 to 6 provide a visual comparison of the performance across the selected benchmarks for various subsets of the models under consideration. To highlight differences in performance, we focus on a set of representative indicators. In particular, the results on *MMLU-Redux* are aggregated into a single average score. Similarly, for a clearer view, we also aggregate the results of the ITALIC dataset computing the mean of the fast and cot settings. The AIME24 and AIME25 scores shown below are the *pass@8*, in order to reflect the potential of the model in solving complex mathematical problems. The different models are divided into four subgroups, to better understand the overall capabilities of similar models: italian-based models, MoE models, larger and smaller dense models.

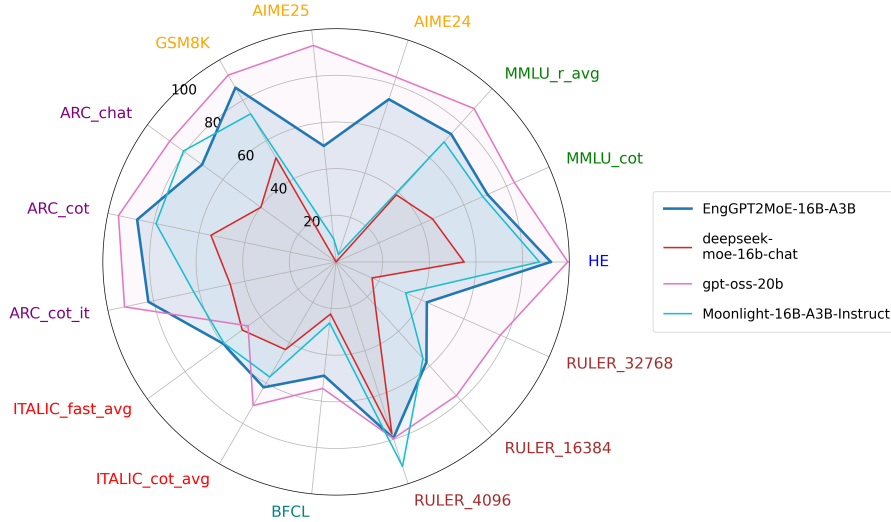


Figure 4: Comparison with Mixture-of-Experts models across the considered benchmarks.

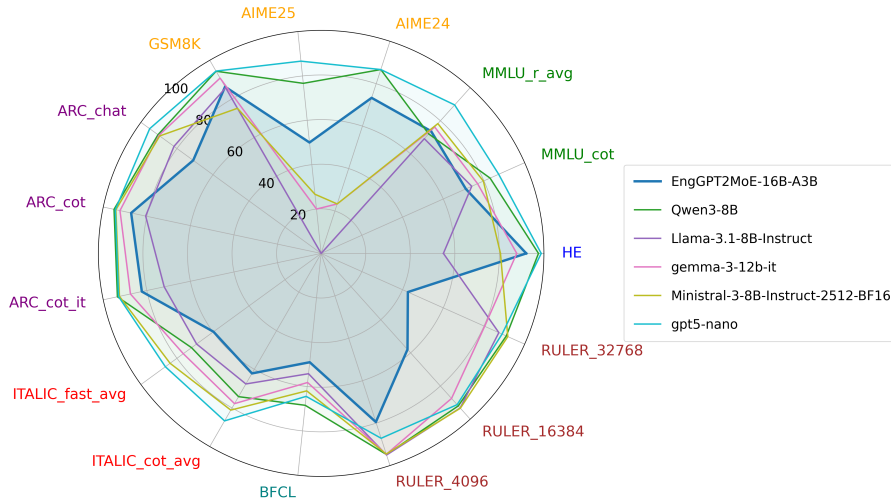


Figure 5: Comparison with the larger dense models across the considered benchmarks.

In Figure 3, we can observe that EngGPT2-MoE-16B outperforms the competitor Italian models on most of the evaluated benchmarks. It stands out particularly for its advanced reasoning capabilities, coding and tool calling. Figure 4 shows the comparison with Mixture-of-Experts models. EngGPT2-MoE-16B is trailing behind gpt-oss-20B on most benchmarks. On the other hand, it shows comparable performance with respect to Moonlight-16B-A3B, while clearly outperforming Deepseek-MoE-16B. Interestingly, when compared with Moonlight-16B-A3B-Instruct, which has the most similar parameters distribution (16 billion total, 3 billion active), EngGPT2-MoE-16B appears to achieve comparable-to-better performance, especially leading in mathematical reasoning and tool use, suggesting that it makes more effective use of its parameters.

Concerning larger dense models (with more than 5B parameters), Figure 5 shows that Qwen3-8B and GPT5-nano achieve the most balanced and strongest overall performance across the evaluated benchmarks. In contrast, EngGPT2-MoE-16B exhibits a less uniform performance profile. Its results are comparatively solid on challenging mathematical reasoning tasks, particularly the AIME benchmarks, and remain competitive on other reasoning and mathematical benchmarks (ARC, GSM8K). However, the model lags behind the strongest baselines on long-context benchmarks, especially at extended context lengths, and also shows weaker results on Italian and tool-call-related evaluations. Overall, this pattern suggests that EngGPT2-MoE-16B is relatively effective on high-difficulty reasoning tasks, but still requires improvements in Italian language, function-calling, and long-context robustness.

Finally, Figure 6 compares EngGPT2-MoE-16B against smaller dense models with a more similar

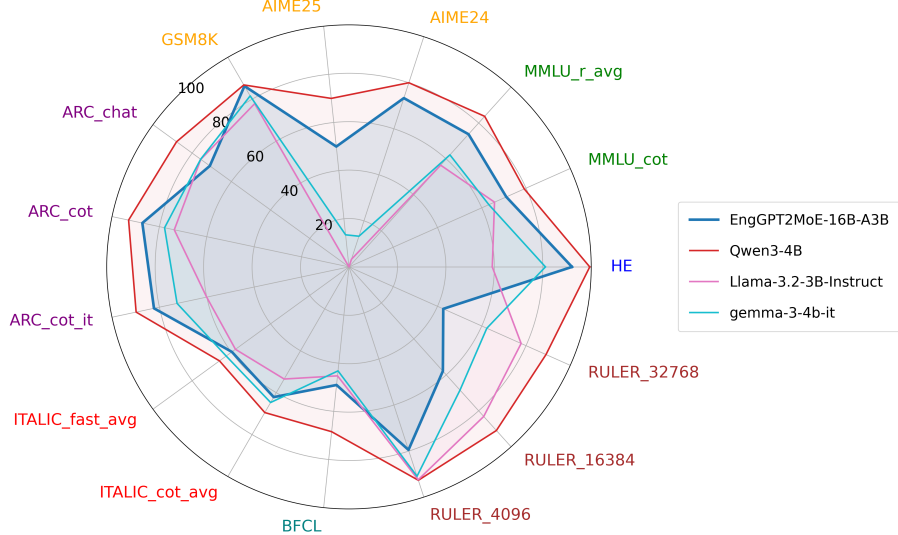


Figure 6: Comparison with the smaller dense models across the considered benchmarks.

number of active parameters. Qwen3-4B clearly achieves the strongest overall performance, showing the most consistent results across all benchmarks. EngGPT2-MoE-16B, however, remains competitive with the other baselines and often delivers comparable or better performance on several reasoning-oriented tasks. Its main weakness remain long-context evaluation, where performance degrades more visibly, especially on the RULER benchmarks at longer context lengths.

| Model             | AIME       | ARC        | GSM8K      | HE         | MMLU       | IT         | RU         | BFCL       | Average              |
|-------------------|------------|------------|------------|------------|------------|------------|------------|------------|----------------------|
| EngGPT2-16B-A3B   | 5.0        | 8.3        | 7.0        | 6.5        | 7.0        | 10.8       | 11.3       | 8.0        | 8.0 $\pm$ 0.8        |
| FastwebMIIA-7B    | 13.3       | 14.0       | 14.5       | 12.5       | 13.8       | 12.3       | 12.0       | 15.0       | 13.4 $\pm$ 0.4       |
| Minerva-7B        | 13.3       | 16.0       | 16.0       | 16.0       | 15.8       | 15.3       | 15.7       | 14.0       | 15.2 $\pm$ 0.4       |
| Velvet-14B        | 13.3       | 11.7       | 13.0       | 15.0       | 12.2       | 7.8        | 13.7       | 13.0       | 12.4 $\pm$ 0.8       |
| LLaMAntino-3-8B   | 13.3       | 10.7       | 12.0       | 11.0       | 10.4       | 8.5        | 11.3       | 11.0       | 11.0 $\pm$ 0.5       |
| Qwen3-4B          | 3.3        | 5.3        | 5.5        | 1.8        | 4.2        | 7.3        | 4.3        | 2.0        | 4.2 $\pm$ 0.7        |
| Qwen3-8B          | 2.5        | <b>1.7</b> | 2.0        | 4.0        | 3.2        | 4.0        | 2.3        | <b>1.0</b> | 2.6 $\pm$ 0.4        |
| Llama-3.2-3B      | 11.3       | 12.0       | 10.0       | 9.5        | 12.2       | 12.8       | 6.7        | 9.0        | 10.4 $\pm$ 0.7       |
| Llama-3.1-8B      | 13.3       | 8.3        | 7.0        | 11.8       | 8.6        | 6.8        | 2.7        | 7.0        | 8.2 $\pm$ 1.1        |
| gemma-3-4b-it     | 8.0        | 9.3        | 8.8        | 7.0        | 11.0       | 9.5        | 9.0        | 10.0       | 9.1 $\pm$ 0.4        |
| gemma-3-12b-it    | 6.9        | 5.0        | 4.5        | 5.5        | 6.2        | 3.0        | 5.0        | 5.0        | 5.1 $\pm$ 0.4        |
| Ministral-3-8B    | 6.1        | 4.0        | 7.0        | 10.5       | 4.8        | 1.8        | <b>2.0</b> | 4.0        | 5.0 $\pm$ 1.0        |
| GPT-5 nano        | <b>1.8</b> | 2.0        | <b>1.0</b> | 3.0        | <b>1.2</b> | <b>1.3</b> | 6.0        | 3.0        | <b>2.4</b> $\pm$ 0.6 |
| deepseek-moe-16b  | 13.3       | 15.0       | 14.5       | 13.8       | 14.8       | 14.3       | 14.3       | 16.0       | 14.5 $\pm$ 0.3       |
| gpt-oss-20b       | 2.5        | 3.0        | 3.0        | <b>1.3</b> | 1.8        | 9.3        | 9.3        | 6.0        | 4.5 $\pm$ 1.2        |
| Moonlight-16B-A3B | 9.3        | 9.7        | 10.3       | 7.0        | 8.8        | 11.8       | 10.3       | 12.0       | 9.9 $\pm$ 0.6        |

Table 19: Average ranking for each benchmark dataset, computed by averaging the ranks of the model over the metrics for the benchmark. The overall average rank across all benchmarks is shown on the right. Lower rank indicates better performance. (Some model names have been abbreviated as have the benchmark names (HE = HumanEval, IT = ITALIC, RU = Ruler)).

#### 4.9 Aggregating Results across Benchmark datasets

In order to provide a high level comparison across the models, we report aggregate measures of performance across benchmarks for each model in Tables 19 and 20. We caution the reader when interpreting aggregated results, since the ordering of systems depends heavily on the choice of datasets included in the survey.

For both tables, we first compute an average score for each benchmark over the individual metrics collected, and then average those results at the dataset level into a single aggregate score. This procedure weights each benchmark equally in terms of importance, regardless of the number of metrics taken. For most benchmarks we considered all reported metrics, with the exception of AIME, where we only

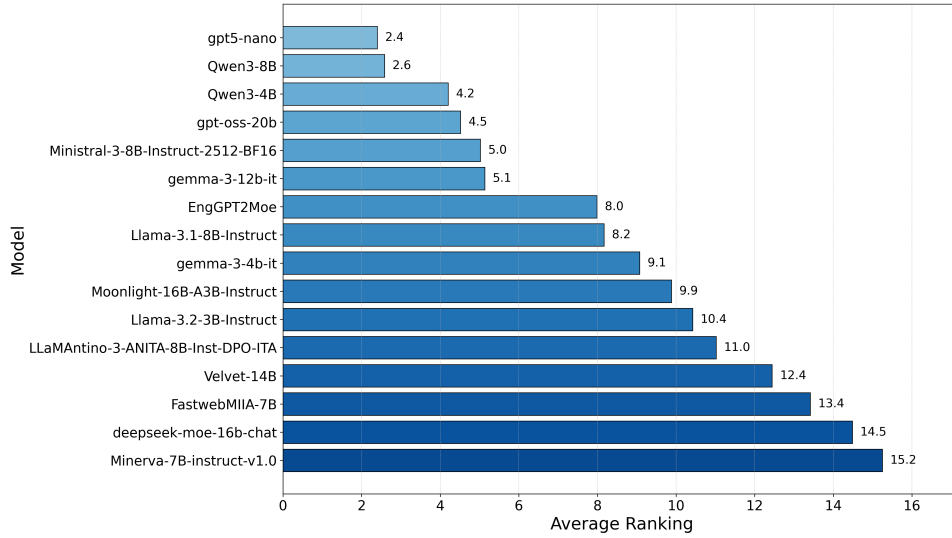


Figure 7: Average ranking of models across the various benchmark datasets. The ranking for each benchmark contributes equally to the overall ranking. For each benchmark, the rank of the model on each metric is first averaged, and then the average rankings for each benchmark are themselves aggregated.

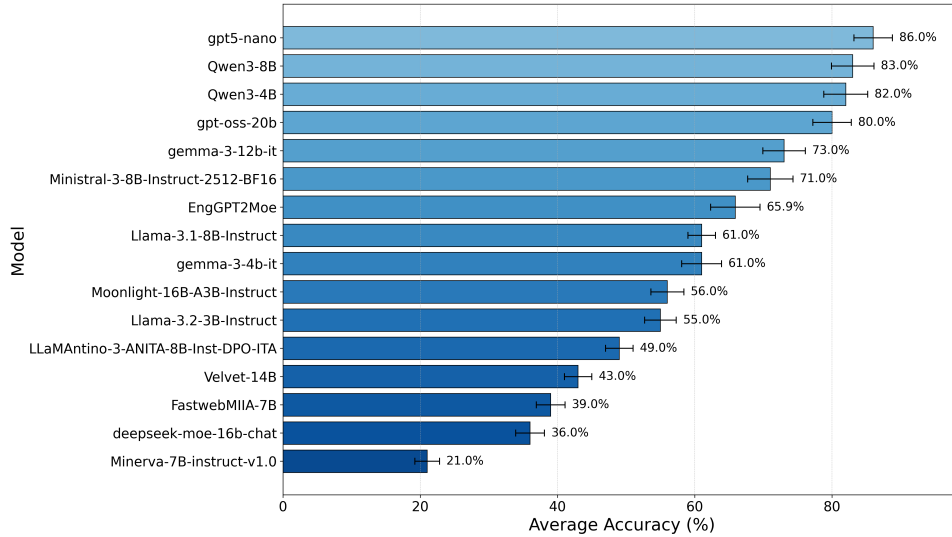


Figure 8: Average of the accuracy score of models across the various benchmarks datasets. Each benchmark contributes equally to the overall average. For each benchmark, the Accuracy score of the model on each metric is first averaged, and then the averages for each benchmark are themselves aggregated. Error bars show standard errors values, calculated by aggregating variances across the datasets from the standard error estimates for each metric value.

considered pass@1 and pass@8 metrics, for both 2024 and 2025 versions. Table 19 contains a ranking-based aggregate score: for each metric, the performance of each model in terms of accuracy is translated into a relative ranking with respect to other models, breaking ties taking the mid point. Table 20 contains an accuracy-based aggregate score: each metric accuracy score is used directly, and the standard error is computed considering the underlying standard error of each metric.

Figures 7-8 show the considered models ordered by ranking and average performance, respectively. The ranking-based aggregation clearly highlights several clusters of performance: GPT-5 nano and Qwen3-8B lead all other models, followed by a second group of similarly-performing models including Qwen3-4B, gemma-3-12b-it, Ministral-3-8B and gpt-oss-20b. Then, we find EngGPT2-16B-A3B and Llama-3.1-8B-Instruct, which are almost paired despite having a clearly distinct performance profile, the former leading in reasoning-heavy and coding tasks, the latter leading by a large margin in long-context

| Model             | AIME         | ARC          | GSM8K        | HE           | MMLU         | IT           | RU           | BFCL         | Average                 |
|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-------------------------|
| EngGPT2-16B-A3B   | 49.2%        | 80.2%        | 86.4%        | 68.6%        | 73.2%        | 60.9%        | 60.0%        | 49.1%        | 65.9% $\pm$ 3.6%        |
| FastwebMIIA-7B    | 0.0%         | 53.5%        | 53.1%        | 37.6%        | 45.0%        | 43.7%        | 55.9%        | 24.5%        | 39.0% $\pm$ 2.1%        |
| Minerva-7B        | 0.0%         | 25.9%        | 10.5%        | 10.7%        | 29.2%        | 40.6%        | 27.1%        | 25.0%        | 21.0% $\pm$ 1.8%        |
| Velvet-14B        | 0.0%         | 70.3%        | 61.0%        | 24.4%        | 56.5%        | 66.6%        | 40.0%        | 25.7%        | 43.0% $\pm$ 2.0%        |
| LLaMAntino-3-8B   | 0.0%         | 65.6%        | 69.0%        | 42.0%        | 59.6%        | 62.5%        | 56.1%        | 38.3%        | 49.0% $\pm$ 2.0%        |
| Qwen3-4B          | 70.8%        | 90.3%        | 87.5%        | 96.3%        | 82.8%        | 67.8%        | 90.9%        | <b>68.5%</b> | 82.0% $\pm$ 3.2%        |
| Qwen3-8B          | 74.2%        | 93.0%        | 94.3%        | 88.8%        | 83.3%        | 73.0%        | 92.6%        | <b>68.5%</b> | 83.0% $\pm$ 3.1%        |
| Llama-3.2-3B      | 1.7%         | 69.8%        | 78.1%        | 46.5%        | 58.4%        | 55.7%        | 84.5%        | 45.3%        | 55.0% $\pm$ 2.3%        |
| Llama-3.1-8B      | 0.0%         | 78.1%        | 86.1%        | 41.8%        | 70.2%        | 68.3%        | 91.9%        | 54.2%        | 61.0% $\pm$ 2.0%        |
| gemma-3-4b-it     | 12.5%        | 75.3%        | 81.5%        | 72.6%        | 62.5%        | 63.4%        | 73.9%        | 43.2%        | 61.0% $\pm$ 2.9%        |
| gemma-3-12b-it    | 20.0%        | 89.9%        | 90.7%        | 82.0%        | 76.4%        | 77.1%        | 87.5%        | 58.2%        | 73.0% $\pm$ 3.1%        |
| Ministral-3-8B    | 25.0%        | 92.1%        | 83.7%        | 52.8%        | 78.6%        | 82.5%        | <b>93.2%</b> | 62.0%        | 71.0% $\pm$ 3.3%        |
| GPT-5 nano        | <b>81.7%</b> | <b>94.2%</b> | <b>94.6%</b> | 91.8%        | <b>89.1%</b> | <b>86.6%</b> | 89.2%        | 64.4%        | <b>86.0%</b> $\pm$ 2.8% |
| deepseek-moe-16b  | 0.0%         | 47.0%        | 53.4%        | 36.4%        | 40.0%        | 46.6%        | 39.9%        | 22.5%        | 36.0% $\pm$ 2.1%        |
| gpt-oss-20b       | 79.2%        | 92.1%        | 92.9%        | <b>97.0%</b> | 87.5%        | 58.9%        | 78.0%        | 54.6%        | 80.0% $\pm$ 2.8%        |
| Moonlight-16B-A3B | 5.8%         | 74.1%        | 77.4%        | 74.2%        | 69.2%        | 58.2%        | 60.2%        | 26.3%        | 56.0% $\pm$ 2.4%        |

Table 20: Aggregated Accuracy values for each benchmark, computed by averaging accuracy values across metrics for the benchmark. The average across benchmarks is overall average across all categories; higher is better. Model names have been abbreviated to fit the table (full names can be found in Table 1). Some benchmark names have been abbreviated to fit the table (HE = HumanEval, IT = ITALIC, RU = Ruler).

understanding and less prominently in tool calling.

The more direct accuracy-based aggregation portrays a similar picture, although with minor numerical differences related to the different scale and the different importance given to the actual accuracy gap in each specific benchmark. Four models reach an average accuracy score of 80% or above: GPT-5 nano, both Qwen3 models and gpt-oss-20b. It is worth noting that, using both aggregation methods, the EngGPT2-16B-A3B largely leads all other Italian models, with a gap of more than 15 percentage points when aggregating the accuracy score.

Overall, the results portray EngGPT2MoE-16B-A3 as one of the most capable models across many of the various considered tasks. Compared with the other Italian-focused systems, it appears to be more robust and versatile. In particular, it is generally more reliable on reasoning-intensive evaluation, clearly stronger on mathematical problem solving, often superior on broad knowledge benchmarks, and solid in code generation. This makes it the only Italian model with competitive general-purpose capabilities. At the same time, the model does not yet reach the level of the strongest international dense models, which remain the reference point for overall performance. In particular, Qwen, GPT-5 nano, gpt-oss, and in some cases Ministral or the larger Gemma variant display a more mature overall profile, combining stronger first-try reliability, better compliance with complex evaluation protocols, stronger long-context behavior, and more effective multi-turn tool use.

## 5 Conclusions

In this work, we evaluated EngGPT2MoE-16B-A3 across a broad set of benchmarks covering reasoning, mathematics, coding, tool use, general knowledge, Italian-language understanding, and long-context capabilities, comparing it with a set of recent state-of-the-art models. The results show that the model achieves a strong and balanced performance profile, consistently ranking above all the other tested Italian-oriented models. At the same time, the evaluation highlights that EngGPT2MoE-16B-A3 remains below the strongest multilingual open models on the most challenging capabilities, especially in long context understanding, complex tool interaction, and high-precision coding tasks.

Overall, these findings position EngGPT2MoE-16B-A3 as one of the most complete Italian-focused models, combining local language specialization with credible general-purpose performance.

## 6 Acknowledgements

The research was carried out in the context of the European Union-funded “AVANT” project, with Engineering Ingegneria Informatica S.p.A. sponsorship, participating as an industrial research partner.

## References

- [1] Tom Brown et al. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.
- [2] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [3] Devalla Bhaskar Ganesh et al. “Advances in machine translation: A comprehensive survey of large language models”. In: *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE. 2025, pp. 1671–1675.
- [4] Zhaopeng Feng et al. “Tear: Improving llm-based machine translation with systematic self-refinement”. In: *Findings of the Association for Computational Linguistics: NAACL 2025*. 2025, pp. 3922–3938.
- [5] Murong Yue. “A survey of large language model agents for question answering”. In: *arXiv preprint arXiv:2503.19213* (2025).
- [6] Xun Liang et al. “Controllable text generation for large language models: A survey”. In: *arXiv preprint arXiv:2408.12599* (2024).
- [7] Yifan Yao et al. “A survey on large language model (llm) security and privacy: The good, the bad, and the ugly”. In: *High-Confidence Computing* 4.2 (2024), p. 100211.
- [8] Zhenyan Lu et al. “Small language models: Survey, measurements, and insights”. In: *arXiv preprint arXiv:2409.15790* (2024).
- [9] Jordan Hoffmann et al. “Training compute-optimal large language models”. In: *arXiv preprint arXiv:2203.15556* 10 (2022).
- [10] Jason Wei et al. “Finetuned language models are zero-shot learners”. In: *arXiv preprint arXiv:2109.01652* (2021).
- [11] Jason Wei et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems* 35 (2022).
- [12] Long Ouyang et al. “Training Language Models to Follow Instructions with Human Feedback”. In: *Advances in Neural Information Processing Systems* 35 (2022).
- [13] Hyung Won Chung et al. “Scaling instruction-finetuned language models”. In: *Journal of Machine Learning Research* 25.70 (2024), pp. 1–53.
- [14] Timo Schick et al. “Toolformer: Language models can teach themselves to use tools”. In: *Advances in neural information processing systems* 36 (2023), pp. 68539–68551.
- [15] Guanzhi Wang et al. “Voyager: An open-ended embodied agent with large language models”. In: *arXiv preprint arXiv:2305.16291* (2023).
- [16] Yiran Ding et al. “Longrope: Extending llm context window beyond 2 million tokens”. In: *arXiv preprint arXiv:2402.13753* (2024).
- [17] Joshua Ainslie et al. “Gqa: Training generalized multi-query transformer models from multi-head checkpoints”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 4895–4901.
- [18] Zichuan Fu et al. “Sliding window attention training for efficient large language models”. In: *arXiv preprint arXiv:2502.18845* (2025).
- [19] Noam Shazeer et al. “OUTRAGEOUSLY LARGE NEURAL NETWORKS: THE SPARSELY-GATED MIXTURE-OF-EXPERTS LAYER”. In: *International Conference on Learning Representations*. 2017. DOI: 1701.06538’.
- [20] William Fedus, Barret Zoph, and Noam Shazeer. “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity”. In: *Journal of Machine Learning Research* 23.120 (2022), pp. 1–39.
- [21] Dmitry Lepikhin et al. “GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding”. In: *arXiv preprint arXiv:2006.16668* (2020).
- [22] G. Ciarfaglia et al. *EngGPT2: Sovereign, Efficient and Open Intelligence*. 2026. arXiv: 2603.16430 [cs.CL]. URL: <https://arxiv.org/abs/2603.16430>.

- [23] Riccardo Orlando et al. “Minerva LLMs: The First Family of Large Language Models Trained from Scratch on Italian Data”. In: *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*. Ed. by Felice Dell’Orletta et al. Pisa, Italy: CEUR Workshop Proceedings, Dec. 2024, pp. 707–719. ISBN: 979-12-210-7060-6. URL: <https://aclanthology.org/2024.clicit-1.77/>.
- [24] Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. “Advanced natural-based interaction for the italian language: Llamantino-3-anita”. In: *Scientific Reports* (2026).
- [25] An Yang et al. “Qwen3 technical report”. In: *arXiv preprint arXiv:2505.09388* (2025).
- [26] Aaron Grattafiori et al. “The llama 3 herd of models”. In: *arXiv preprint arXiv:2407.21783* (2024).
- [27] Alexander H. Liu et al. *Minstral 3*. 2026. arXiv: 2601.08584 [cs.CL]. URL: <https://arxiv.org/abs/2601.08584>.
- [28] Gemma Team et al. *Gemma 3 Technical Report*. 2025. arXiv: 2503.19786 [cs.CL]. URL: <https://arxiv.org/abs/2503.19786>.
- [29] Damai Dai et al. “Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 1280–1297.
- [30] Sandhini Agarwal et al. “gpt-oss-120b & gpt-oss-20b model card”. In: *arXiv preprint arXiv:2508.10925* (2025).
- [31] Jingyuan Liu et al. “Muon is scalable for llm training”. In: *arXiv preprint arXiv:2502.16982* (2025).
- [32] Qiguang Chen et al. “Towards reasoning era: A survey of long chain-of-thought for reasoning large language models”. In: *arXiv preprint arXiv:2503.09567* (2025).
- [33] Fengli Xu et al. “Toward large reasoning models: A survey of reinforced reasoning with large language models”. In: *Patterns* 6.10 (2025).
- [34] Shishir G Patil et al. “Gorilla: Large language model connected with massive apis, 2023”. In: *URL https://arxiv.org/abs/2305.15334* (2023).
- [35] Yifan Song et al. “Restgpt: Connecting large language models with real-world restful apis”. In: *arXiv preprint arXiv:2306.06624* (2023).
- [36] AIME. *2024 AIME I*. Art of Problem Solving Wiki, accessed July 2025. 2024. URL: [https://artofproblemsolving.com/wiki/index.php/2024\\_AIME\\_I](https://artofproblemsolving.com/wiki/index.php/2024_AIME_I).
- [37] AIME. *2025 AIME I*. Art of Problem Solving Wiki. Held February 6, 2025. 2025. URL: [https://artofproblemsolving.com/wiki/index.php/2025\\_AIME\\_I](https://artofproblemsolving.com/wiki/index.php/2025_AIME_I).
- [38] Xiao Liu et al. “Alignbench: Benchmarking chinese alignment of large language models”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 11621–11640.
- [39] Minghao Li et al. “Api-bank: A comprehensive benchmark for tool-augmented llms”. In: *Proceedings of the 2023 conference on empirical methods in natural language processing*. 2023, pp. 3102–3116.
- [40] Peter Clark et al. “Think you have solved question answering? try arc, the ai2 reasoning challenge”. In: *arXiv preprint arXiv:1803.05457* (2018).
- [41] Tianle Li et al. “From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline”. In: *arXiv preprint arXiv:2406.11939* (2024).
- [42] Qin Zhu et al. “Autologi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models”. In: *arXiv preprint arXiv:2502.16906* (2025).
- [43] Shishir G Patil et al. “The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models”. In: *Forty-second International Conference on Machine Learning*. 2025.
- [44] Yuzhen Huang et al. “C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models”. In: *Advances in neural information processing systems* 36 (2023), pp. 62991–63010.
- [45] Chujie Zheng, Minlie Huang, and Aixin Sun. “ChID: A large-scale Chinese IDiom dataset for cloze test”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 778–787.

- [46] Liang Xu et al. “CLUE: A Chinese Language Understanding Evaluation Benchmark”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4762–4772. DOI: 10.18653/v1/2020.coling-main.419. URL: <https://aclanthology.org/2020.coling-main.419/>.
- [47] Haonan Li et al. “Cmmlu: Measuring massive multitask language understanding in chinese”. In: *Findings of the Association for Computational Linguistics: ACL 2024*. 2024, pp. 11260–11285.
- [48] Shanghaoran Quan et al. “Codeelo: Benchmarking competition-level code generation of llms with human-comparable elo ratings”. In: *arXiv preprint arXiv:2501.01257* (2025).
- [49] Dheeru Dua et al. “DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 2368–2378.
- [50] Shishir G Patil et al. “Gorilla: Large language model connected with massive apis”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 126544–126565.
- [51] David Rein et al. “Gpqa: A graduate-level google-proof q&a benchmark”. In: *First conference on language modeling*. 2024.
- [52] Karl Cobbe et al. “Training verifiers to solve math word problems”. In: *arXiv preprint arXiv:2110.14168* (2021).
- [53] Rahul K Arora et al. “Healthbench: Evaluating large language models towards improved human health”. In: *arXiv preprint arXiv:2505.08775* (2025).
- [54] Rowan Zellers et al. “Hellaswag: Can a machine really finish your sentence?” In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2019, pp. 4791–4800.
- [55] Long Phan et al. “Humanity’s last exam”. In: *arXiv preprint arXiv:2501.14249* (2025).
- [56] Mark Chen et al. “Evaluating large language models trained on code”. In: *arXiv preprint arXiv:2107.03374* (2021).
- [57] Jeffrey Zhou et al. “Instruction-following evaluation for large language models”. In: *arXiv preprint arXiv:2311.07911* (2023).
- [58] Angelika Romanou et al. “Include: Evaluating multilingual language understanding with regional knowledge”. In: *arXiv preprint arXiv:2411.19799* (2024).
- [59] Xinrong Zhang et al. “ $\infty$ Bench: Extending long context evaluation beyond 100K tokens”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 15262–15277.
- [60] Colin White et al. “Livebench: A challenging, contamination-free llm benchmark”. In: *arXiv preprint arXiv:2406.19314* 4 (2024), p. 2.
- [61] Naman Jain et al. “Livecodebench: Holistic and contamination free evaluation of large language models for code”. In: *arXiv preprint arXiv:2403.07974* (2024).
- [62] Dan Hendrycks et al. “Measuring mathematical problem solving with the math dataset”. In: *arXiv preprint arXiv:2103.03874* (2021).
- [63] Hunter Lightman et al. “Let’s verify step by step”. In: *The twelfth international conference on learning representations*. 2023.
- [64] Jiawei Liu et al. “Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation”. In: *Advances in neural information processing systems* 36 (2023), pp. 21558–21572.
- [65] Freda Shi et al. “Language models are multilingual chain-of-thought reasoners”. In: *arXiv preprint arXiv:2210.03057* (2022).
- [66] Yidan Zhang et al. “P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms”. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025, pp. 4809–4836.
- [67] Dan Hendrycks et al. “Measuring massive multitask language understanding”. In: *arXiv preprint arXiv:2009.03300* (2020).

- [68] Yubo Wang et al. “Mmlu-pro: A more robust and challenging multi-task language understanding benchmark”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 95266–95290.
- [69] Aryo Pradipta Gema et al. “Are we done with mmlu?” In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2025, pp. 5069–5096.
- [70] Yun He et al. “Multi-if: Benchmarking llms on multi-turn and multilingual instructions following”. In: *arXiv preprint arXiv:2410.15553* (2024).
- [71] Federico Cassano et al. “Multipl-e: A scalable and polyglot approach to benchmarking neural code generation”. In: *IEEE Transactions on Software Engineering* 49.7 (2023), pp. 3675–3691.
- [72] Gregory Kamradt. “Llmtest\_needleinahaystack”. In: *GitHub repository* (2023).
- [73] Venkat Krishna Srinivasan et al. “Nexusraven: a commercially-permissive language model for function calling”. In: *NeurIPS 2023 Foundation Models for Decision Making Workshop*. 2023.
- [74] Leo Gao et al. “The pile: An 800gb dataset of diverse text for language modeling”. In: *arXiv preprint arXiv:2101.00027* (2020).
- [75] Yiming Wang et al. “Polymath: Evaluating mathematical reasoning in multilingual contexts”. In: *arXiv preprint arXiv:2504.18428* (2025).
- [76] Cheng-Ping Hsieh et al. “RULER: What’s the real context size of your long-context language models?” In: *arXiv preprint arXiv:2404.06654* (2024).
- [77] Stephanie Lin, Jacob Hilton, and Owain Evans. “Truthfulqa: Measuring how models mimic human falsehoods”. In: *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*. 2022, pp. 3214–3252.
- [78] Keisuke Sakaguchi et al. “Winogrande: An adversarial winograd schema challenge at scale”. In: *Communications of the ACM* 64.9 (2021), pp. 99–106.
- [79] Yuning Wu et al. “Writingbench: A comprehensive benchmark for generative writing”. In: *arXiv preprint arXiv:2503.05244* (2025).
- [80] Bill Yuchen Lin et al. “ZebraLogic: On the scaling limits of llms for logical reasoning”. In: *arXiv preprint arXiv:2502.01100* (2025).
- [81] Uri Shaham et al. “ZeroSCROLLS: A zero-shot benchmark for long text understanding”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pp. 7977–7989.
- [82] Shunyu Yao et al. “ $\tau$ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains”. In: *arXiv preprint arXiv:2406.12045* (2024).
- [83] Bernardo Magnini et al. “Evalita-llm: Benchmarking large language models on italian”. In: *arXiv preprint arXiv:2502.02289* (2025).
- [84] Luca Moroni et al. “Towards a more comprehensive evaluation for Italian LLMs”. In: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*. 2024, pp. 584–599.
- [85] Andrea Seveso et al. “ITALIC: An Italian culture-aware natural language benchmark”. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2025, pp. 1469–1478.
- [86] Leo Gao et al. *The Language Model Evaluation Harness*. Version v0.4.3. July 2024. DOI: 10.5281/zenodo.12608602. URL: <https://zenodo.org/records/12608602>.
- [87] Woosuk Kwon. “vLLM: An Efficient Inference Engine for Large Language Models”. PhD thesis. UC Berkeley, 2025.
- [88] ModelScope Team. *EvalScope: Evaluation Framework for Large Models*. 2024. URL: <https://github.com/modelscope/evalscope>.
- [89] Woosuk Kwon et al. “Efficient Memory Management for Large Language Model Serving with PagedAttention”. In: *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*. 2023.
- [90] Elliot Nelson et al. “Needle in the haystack for memory based large language models”. In: *arXiv preprint arXiv:2407.01437* (2024).

## A Timings and Model Throughput

Tables 21, 22, 23, and 24 contain details regarding the wall-clock time taken to run each evaluation benchmark for each model, in the evaluation environment described in Appendix E, as well as the number of samples for each benchmarking dataset. It is worth noting that the per-sample time can vary quite dramatically between both different tasks and different models, mainly as a consequence of the number of generated tokens; for instance, tasks requiring long reasoning traces such as AIME24 and AIME25 typically consume two to three orders of magnitude more tokens than knowledge-based tasks such as MMLU, and models heavily trained to produce long and detailed reasoning traces (e.g., Qwen models) can easily generate hundreds of tokens more even for simple tasks than models that did not undergo specific reasoning training (e.g., Llama3 models).

Table 21: Evaluation Times (in seconds) for the AIME 24 & AIME 25 tasks

| Model \ Task (n samples)           | aime24              | aime24             | aime25              | aime25             |
|------------------------------------|---------------------|--------------------|---------------------|--------------------|
|                                    | single_pass<br>(30) | pass_avg_8<br>(30) | single_pass<br>(30) | pass_avg_8<br>(30) |
| EngGPT2-16B-A3B                    | 2379                | 16496              | 2634                | 16903              |
| FastwebMIIA-7B                     | 107                 | 192                | 111                 | 241                |
| Minerva-7B-instruct-v1.0           | 77                  | 128                | 82                  | 150                |
| Velvet-14B                         | 3633                | 12279              | 3877                | 10090              |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | 151                 | 198                | 153                 | 199                |
| Qwen3-4B                           | 718                 | 6592               | 1461                | 8522               |
| Qwen3-8B                           | 1312                | 8008               | 1919                | 13213              |
| Llama-3.2-3B-Instruct              | 925                 | 3809               | 881                 | 4232               |
| Llama-3.1-8B-Instruct              | 1895                | 6339               | 1988                | 8206               |
| gemma-3-4b-it                      | 635                 | 969                | 426                 | 497                |
| gemma-3-12b-it                     | 2735                | 10038              | 971                 | 4526               |
| deepseek-moe-16b-chat              | 136                 | 301                | 289                 | 350                |
| gpt-oss-20b                        | 723                 | 1686               | 744                 | 1733               |
| Moonlight-16B-A3B-Instruct         | 227                 | 376                | 187                 | 231                |
| Ministral-3-8B-Instruct-2512-BF16* | 14432               | 110570             | 16507               | 135549             |

\* Evaluation times computed serving the model via Huggingface (not vLLM).

Table 22: Evaluation Times (in seconds) for the ARC tasks

| Model \ Task (n samples)           | arc_challenge<br>chat<br>(1172) | arc_challenge<br>cot_custom<br>(1172) | arc_challenge<br>cot_it_custom<br>(1169) |
|------------------------------------|---------------------------------|---------------------------------------|------------------------------------------|
|                                    |                                 |                                       |                                          |
| EngGPT2-16B-A3B                    | 81                              | 605                                   | 1034                                     |
| FastwebMIIA-7B                     | 61                              | 248                                   | 206                                      |
| Minerva-7B-instruct-v1.0           | 61                              | 188                                   | 250                                      |
| Velvet-14B                         | 84                              | 280                                   | 354                                      |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | 135                             | 601                                   | 268                                      |
| Qwen3-4B                           | 162                             | 340                                   | 404                                      |
| Qwen3-8B                           | 69                              | 441                                   | 611                                      |
| Llama-3.2-3B-Instruct              | 125                             | 130                                   | 137                                      |
| Llama-3.1-8B-Instruct              | 216                             | 286                                   | 343                                      |
| gemma-3-4b-it                      | 209                             | 192                                   | 236                                      |
| gemma-3-12b-it                     | 462                             | 425                                   | 598                                      |
| deepseek-moe-16b-chat              | 427                             | 251                                   | 858                                      |
| gpt-oss-20b                        | 69                              | 197                                   | 183                                      |
| Moonlight-16B-A3B-Instruct         | 386                             | 202                                   | 179                                      |
| Ministral-3-8B-Instruct-2512-BF16* | 324                             | 12026                                 | 18766                                    |

\* Evaluation times computed serving the model via Huggingface (not vLLM).

To provide a more standardized perspective on the raw performance of each model (although specific performance testing remains outside the scope of this paper), Table 25 reports the models’ throughput measured as wall-clock speed over a full batch run of requests, using the *vLLM* serving framework. To isolate core model inference speed from tokenization overhead, we constructed a benchmark of 1,000 synthetic requests per model. Each prompt consisted of exactly 1,024 token IDs uniformly sampled from the model’s vocabulary.

Models were instantiated via the LLM class, utilizing a constrained context window (`max_model_len=4096`) to optimize KV-cache memory and prevent out-of-memory errors during batch execution. To ensure uniform, maximum-load evaluation conditions, we configured generation parameters using the `SamplingParams` class with `max_tokens=128` and `ignore_eos=True`. This forced every model to output exactly 128 tokens per prompt. Throughput metrics were derived by measuring the end-to-end execution time of `llm.generate()`, which processes the entire batch of requests.

Table 23: Evaluation Times (in seconds) for the GSM8K, HumanEval, and RULER tasks

| Model \ Task (n samples)           | gsm8k<br>llama<br>(1319) | humaneval_8<br>instruct_cot<br>(164) | ruler<br>(19500) |
|------------------------------------|--------------------------|--------------------------------------|------------------|
| EngGPT2-16B-A3B                    | 5729                     | 16434                                | 27890            |
| FastwebMIIA-7B                     | 168                      | 3045                                 | 18082            |
| Minerva-7B-instruct-v1.0           | 138                      | 3036                                 | 5429             |
| Velvet-14B                         | 164                      | 3067                                 | 65225            |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | 269                      | 3173                                 | 9003             |
| Qwen3-4B                           | 202                      | 5504                                 | 25965            |
| Qwen3-8B                           | 1040                     | 7027                                 | 33237            |
| Llama-3.2-3B-Instruct              | 98                       | 2811                                 | 17493            |
| Llama-3.1-8B-Instruct              | 131                      | 3037                                 | 30983            |
| gemma-3-4b-it                      | 189                      | 3091                                 | 15138            |
| gemma-3-12b-it                     | 274                      | 3076                                 | 50193            |
| deepseek-moe-16b-chat              | 211                      | 3303                                 | 6659             |
| gpt-oss-20b                        | 147                      | 3061                                 | 33640            |
| Moonlight-16B-A3B-Instruct         | 211                      | 3108                                 | 7850             |
| Ministral-3-8B-Instruct-2512-BF16* | 10203                    | 88368                                | 196186           |

\* Evaluation times computed serving the model via Huggingface (not vLLM).

Table 24: Evaluation Times (in seconds) for the MMLU tasks

| Model \ Task (n samples)           | mmlu_cot                   | mmlu_redux_generative |                 |                              |                |
|------------------------------------|----------------------------|-----------------------|-----------------|------------------------------|----------------|
|                                    | llama<br>custom<br>(14042) | humanities<br>(1208)  | other<br>(1162) | social<br>sciences<br>(1147) | stem<br>(1813) |
| EngGPT2-16B-A3B                    | 30517                      | 879                   | 692             | 718                          | 3039           |
| FastwebMIIA-7B                     | 1709                       | 83                    | 123             | 79                           | 102            |
| Minerva-7B-instruct-v1.0           | 942                        | 88                    | 82              | 84                           | 92             |
| Velvet-14B                         | 1449                       | 112                   | 95              | 93                           | 117            |
| LLaMAntino-3-ANITA-8B              | 6642                       | 187                   | 187             | 175                          | 240            |
| Qwen3-4B                           | 36693                      | 329                   | 299             | 327                          | 1108           |
| Qwen3-8B                           | 32258                      | 441                   | 427             | 323                          | 143            |
| Llama-3.2-3B-Instruct              | 587                        | 88                    | 68              | 64                           | 79             |
| Llama-3.1-8B-Instruct              | 943                        | 101                   | 73              | 88                           | 91             |
| gemma-3-4b-it                      | 649                        | 123                   | 104             | 100                          | 115            |
| gemma-3-12b-it                     | 2101                       | 149                   | 129             | 136                          | 142            |
| deepseek-moe-16b-chat              | 2703                       | 203                   | 146             | 142                          | 232            |
| gpt-oss-20b                        | 894                        | 140                   | 245             | 123                          | 216            |
| Moonlight-16B-A3B-Instruct         | 1138                       | 100                   | 108             | 106                          | 876            |
| Ministral-3-8B-Instruct-2512-BF16* | 159499                     | 160                   | 152             | 143                          | 212            |

\* Evaluation times computed serving the model via Huggingface (not vLLM).

Table 25: Model throughputs computed via vLLM.

| Model                             | Tokens per second | Requests per second |
|-----------------------------------|-------------------|---------------------|
| EngGPT2-MoE-16B                   | 1023              | 8.00                |
| FastwebMIIA-7B                    | 1265              | 9.89                |
| Minerva-7B-instruct-v1.0          | 1277              | 9.98                |
| Velvet-14B                        | 572               | 4.48                |
| LLaMAntino-3-ANITA-8B             | 1267              | 9.90                |
| Qwen3-4B                          | 1848              | 14.44               |
| Qwen3-8B                          | 1228              | 9.59                |
| Llama-3.2-3B-Instruct             | 2435              | 19.03               |
| Llama-3.1-8B-Instruct             | 1259              | 9.84                |
| gemma-3-4b-it                     | 2058              | 16.08               |
| gemma-3-12b-it                    | 442               | 3.46                |
| Ministral-3-8B-Instruct-2512-BF16 | 1159              | 9.06                |
| deepseek-moe-16b-chat             | 531               | 4.16                |
| gpt-oss-20b                       | 2057              | 16.07               |
| Moonlight-16B-A3B-Instruct        | 1531              | 11.96               |

## B Heatmaps for Statistical t-test with Bonferroni correction

This section reports the full set of pairwise significance heatmaps of every dataset. For each metric and for each benchmark, pairs of models were compared using a two-sided Welch test using the reported means and standard errors. Since many pairwise tests are performed for each metric,  $p$ -values were adjusted using a Bonferroni correction within each column. Each heatmap displays the Bonferroni-adjusted pairwise  $p$ -values for one benchmark+metric combination. Models are ordered by the mean score for that metric, from strongest to weakest. Green cells indicate statistically significant differences at  $\alpha = 0.05$ , yellow cells lie near the decision threshold, and red cells indicate that the comparison is not significant after correction. Because the figures show adjusted  $p$ -values rather than signed effect sizes, the color encodes confidence in a difference, while the axis ordering mimics the performance ranking.

### B.1 ARC-Challenge

ARC-Challenge is shown for three evaluation settings. These plots make it easy to see whether the same groups of models remain statistically distinguishable when the prompting or extraction setting is changed.

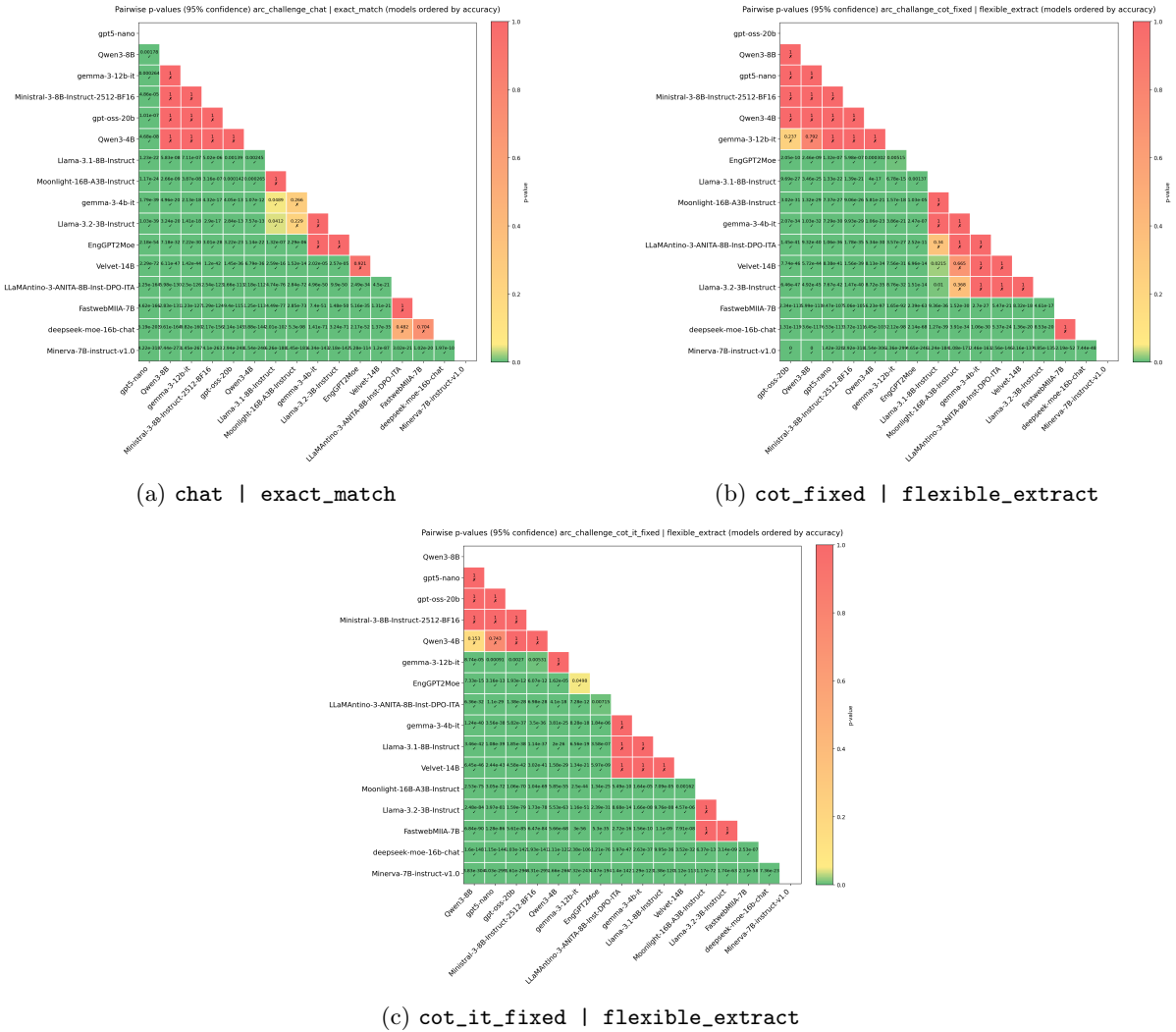


Figure 9: Bonferroni-adjusted pairwise significance heatmaps for ARC-Challenge under three settings: chat | exact\_match, cot\_fixed | flexible\_extract, and cot\_it\_fixed | flexible\_extract.

## B.2 GSM8K

For GSM8K, the plot below summarizes the pairwise evidence under the strict-match metric. This figure helps us identify whether models achieve statistically indistinguishable performance at the chosen threshold after correction.

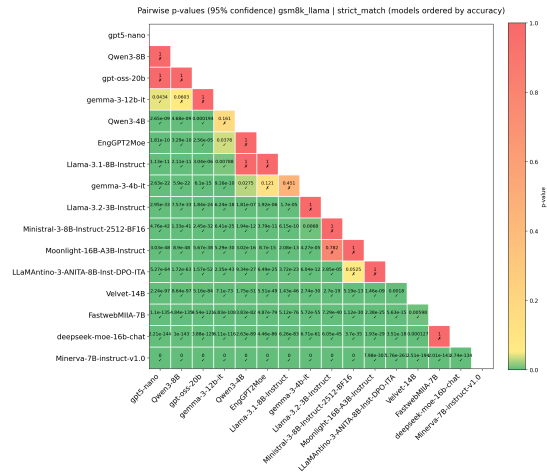


Figure 10: GSM8K, llama | strict\_match. Bonferroni-adjusted pairwise significance heatmap for GSM8K under the strict\_match metric.

### B.3 AIME-2024 & AIME-2025

We report the comparison of the considered models on AIME-2024 and AIME-2025 in Figures 11 and 12, respectively. The comparison is performed across all four metrics.

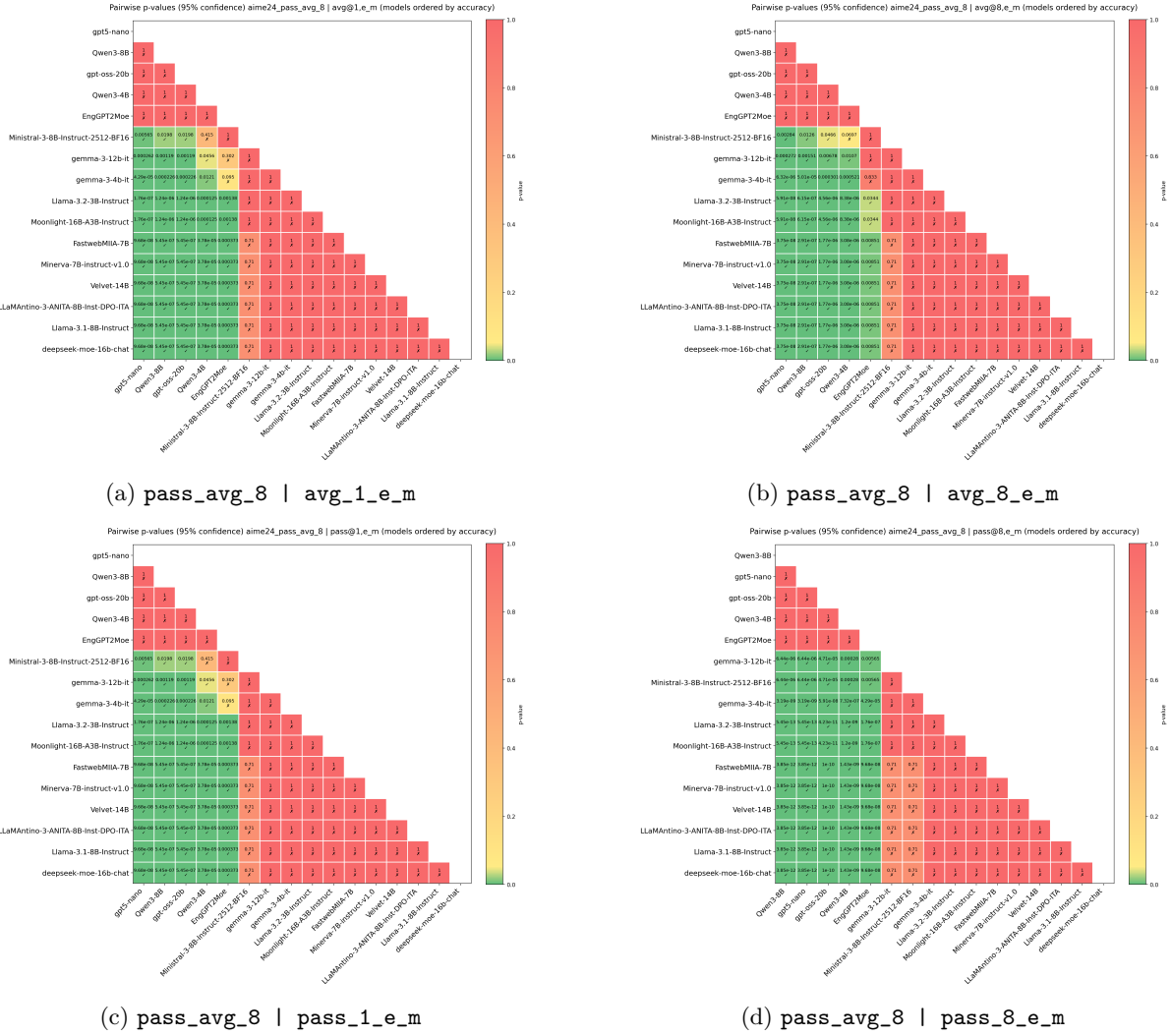
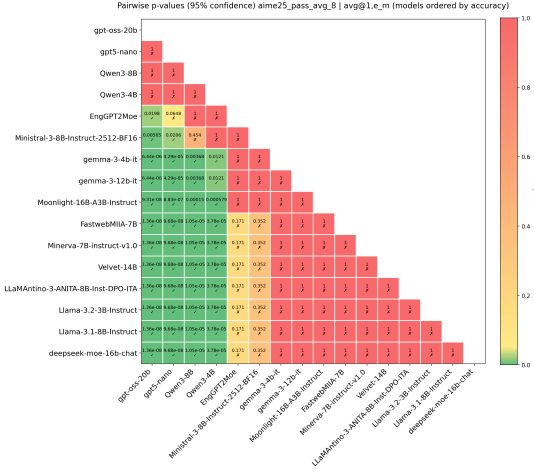
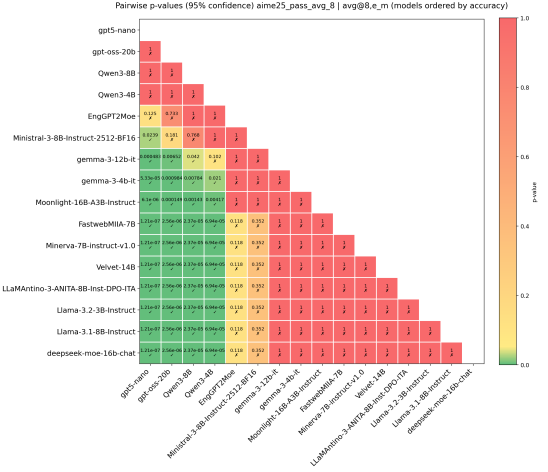


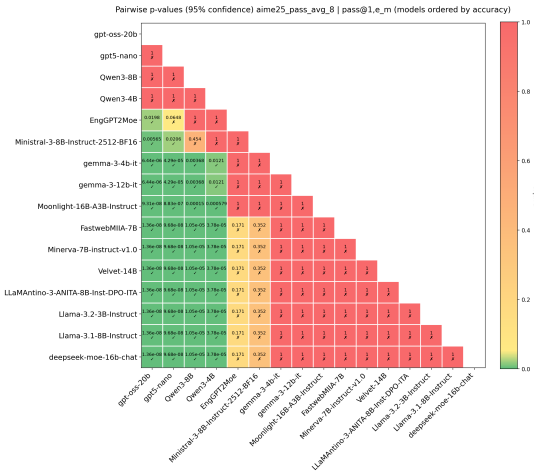
Figure 11: Bonferroni-adjusted pairwise significance heatmaps for AIME24 under the `pass_avg_8` setting, for the metrics `avg_1_e_m`, `avg_8_e_m`, `pass_1_e_m`, and `pass_8_e_m`.



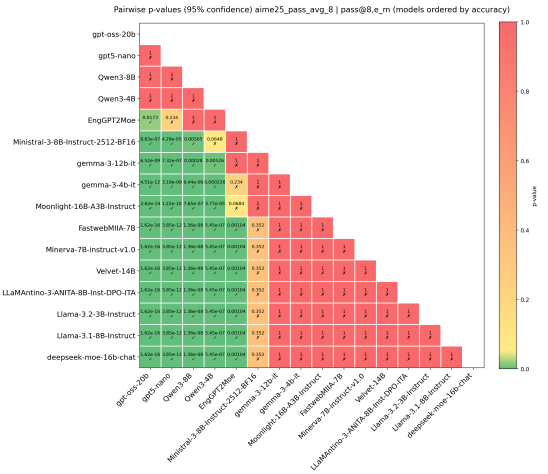
(a) pass\_avg\_8 | avg\_1\_e\_m



(b) pass\_avg\_8 | avg\_8\_e\_m



(c) pass\_avg\_8 | pass\_1\_e\_m



(d) pass\_avg\_8 | pass\_8\_e\_m

Figure 12: Bonferroni-adjusted pairwise significance heatmaps for AIME25 under the pass\_avg\_8 setting, for the metrics avg\_1\_e\_m, avg\_8\_e\_m, pass\_1\_e\_m, and pass\_8\_e\_m.

## B.4 MMLU & MMLU-Redux

Figure 13 reports the pairwise significance for the MMLU dataset. MMLU-Redux is decomposed into four domain groups as shown in Figure 14.

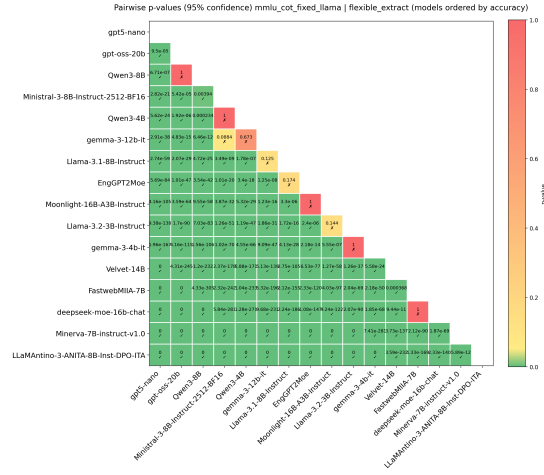
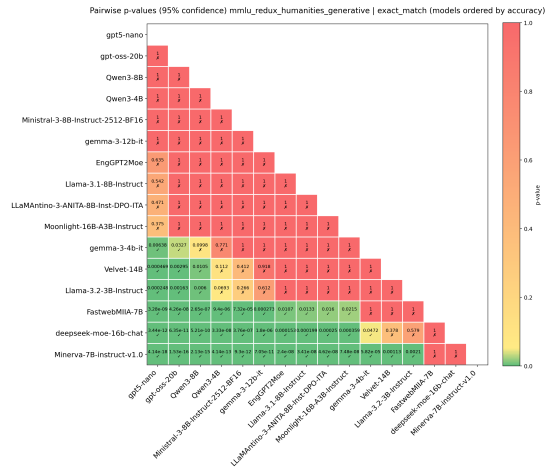
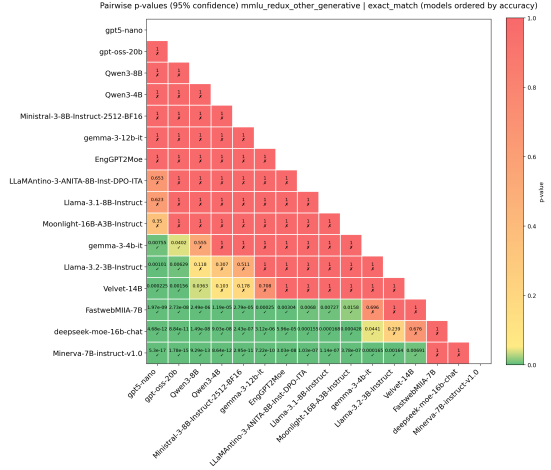


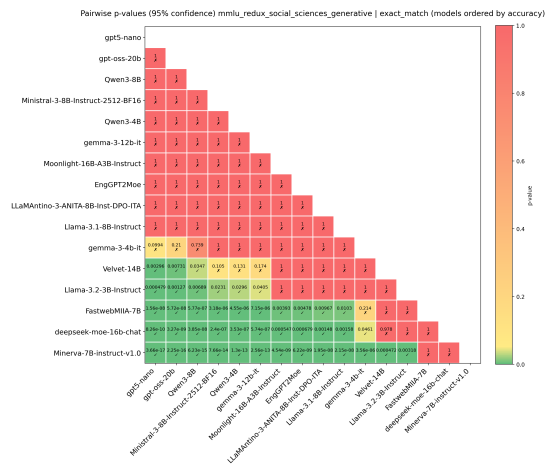
Figure 13: . Bonferroni-adjusted pairwise significance heatmap for MMLU under the `cot_fixed_llama` / `flexible_extract` setting.



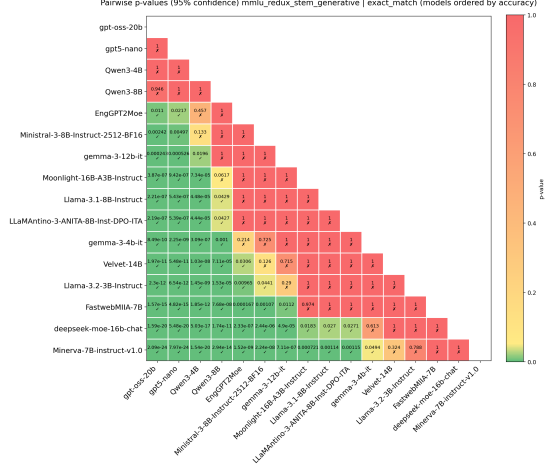
(a) Humanities, **generative | exact\_match**



(b) Other, **generative | exact\_match**



(c) Social Sciences, **generative | exact\_match**



(d) STEM, **generative | exact\_match**

Figure 14: Bonferroni-adjusted pairwise significance heatmaps for the MMLU-Redux subsets Humanities, Other, Social Sciences, and STEM under the **generative | exact\_match** setting.

## B.5 HumanEval

HumanEval is shown in Figure 15 for two `create_test` metrics.

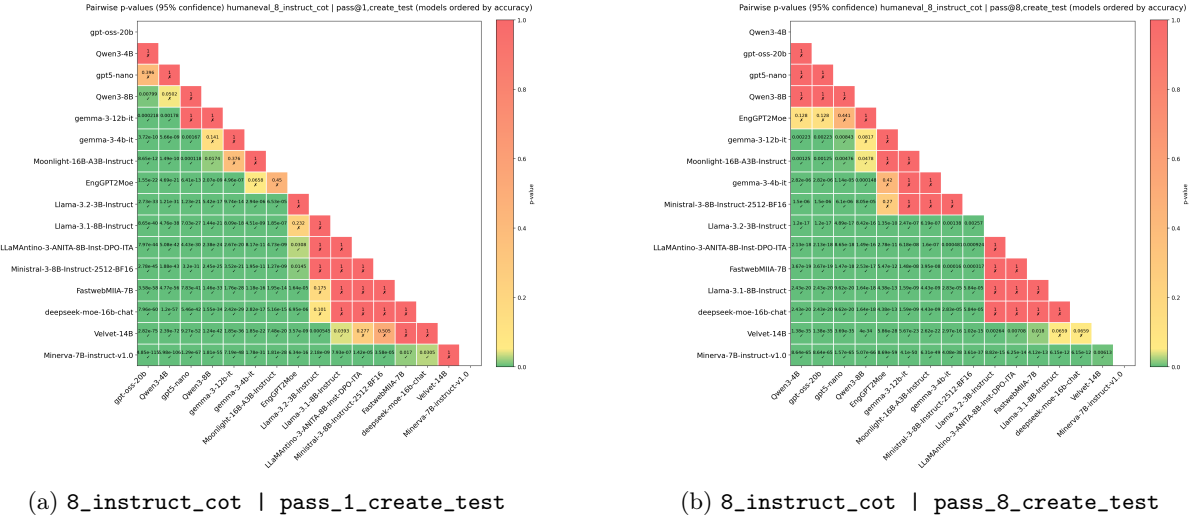


Figure 15: Bonferroni-adjusted pairwise significance heatmaps for HumanEval under the `8_instruct_cot` setting, for the metrics `pass_1_create_test` and `pass_8_create_test`.

## B.6 BFCL

Figure 16 shows the results of the statistical test on BFCL dataset.

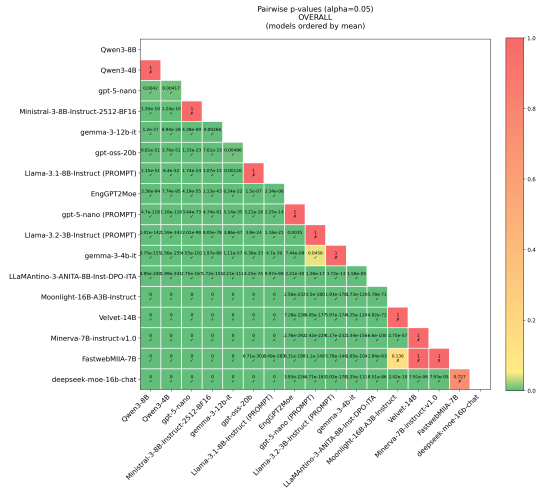


Figure 16: Bonferroni-adjusted pairwise significance heatmap for the BFCL Overall score.

## B.7 ITALIC

ITALIC is reported in Figure 17 for the four evaluation modes.

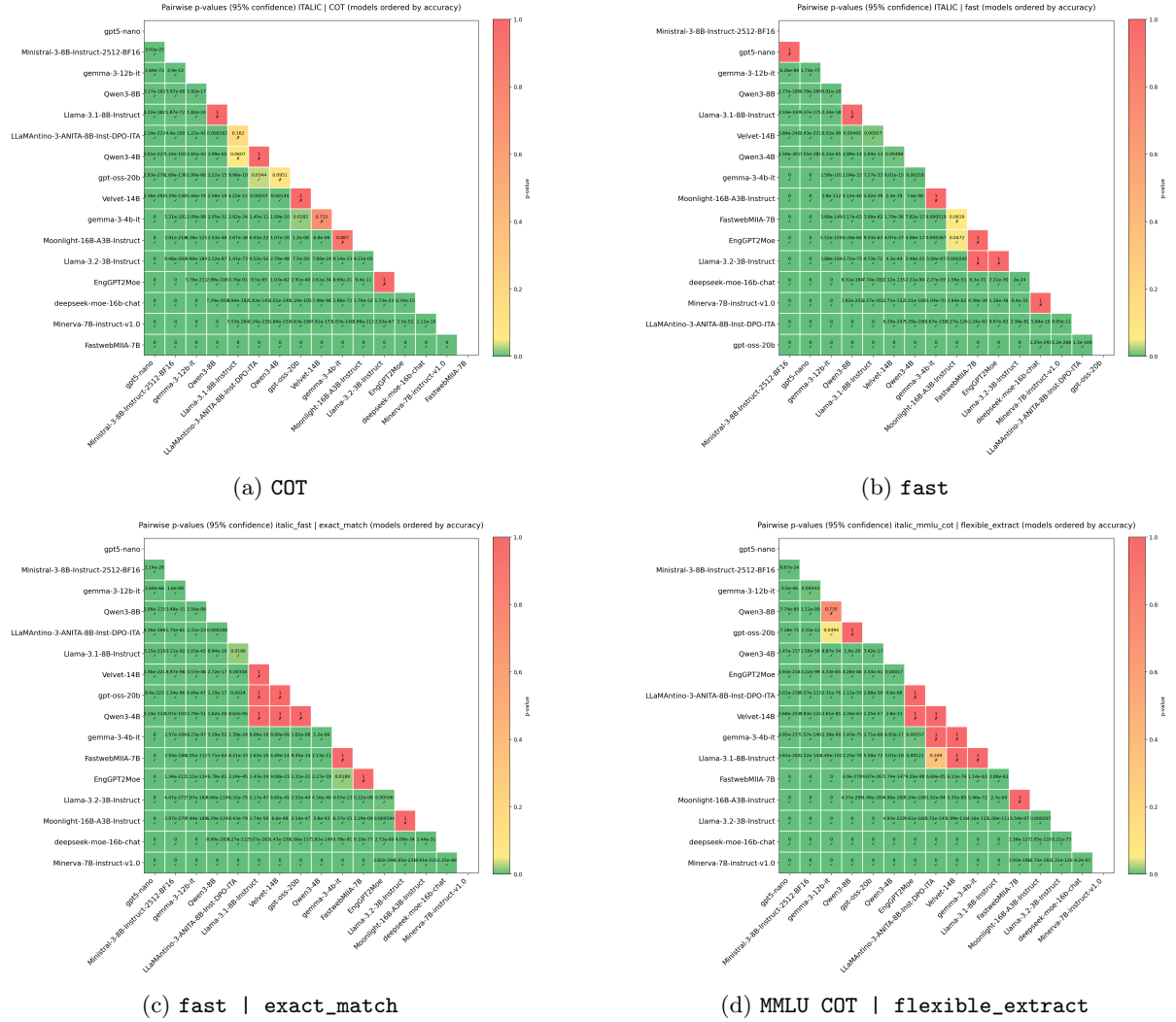


Figure 17: Bonferroni-adjusted pairwise significance heatmaps for ITALIC under the settings COT, fast, fast | exact\_match, and MMLU COT | flexible\_extract.

## B.8 RULER

The RULER benchmark is shown in Figure 18 at three different context lengths. These plots are useful for assessing that longer context length makes model differences become more pronounced.

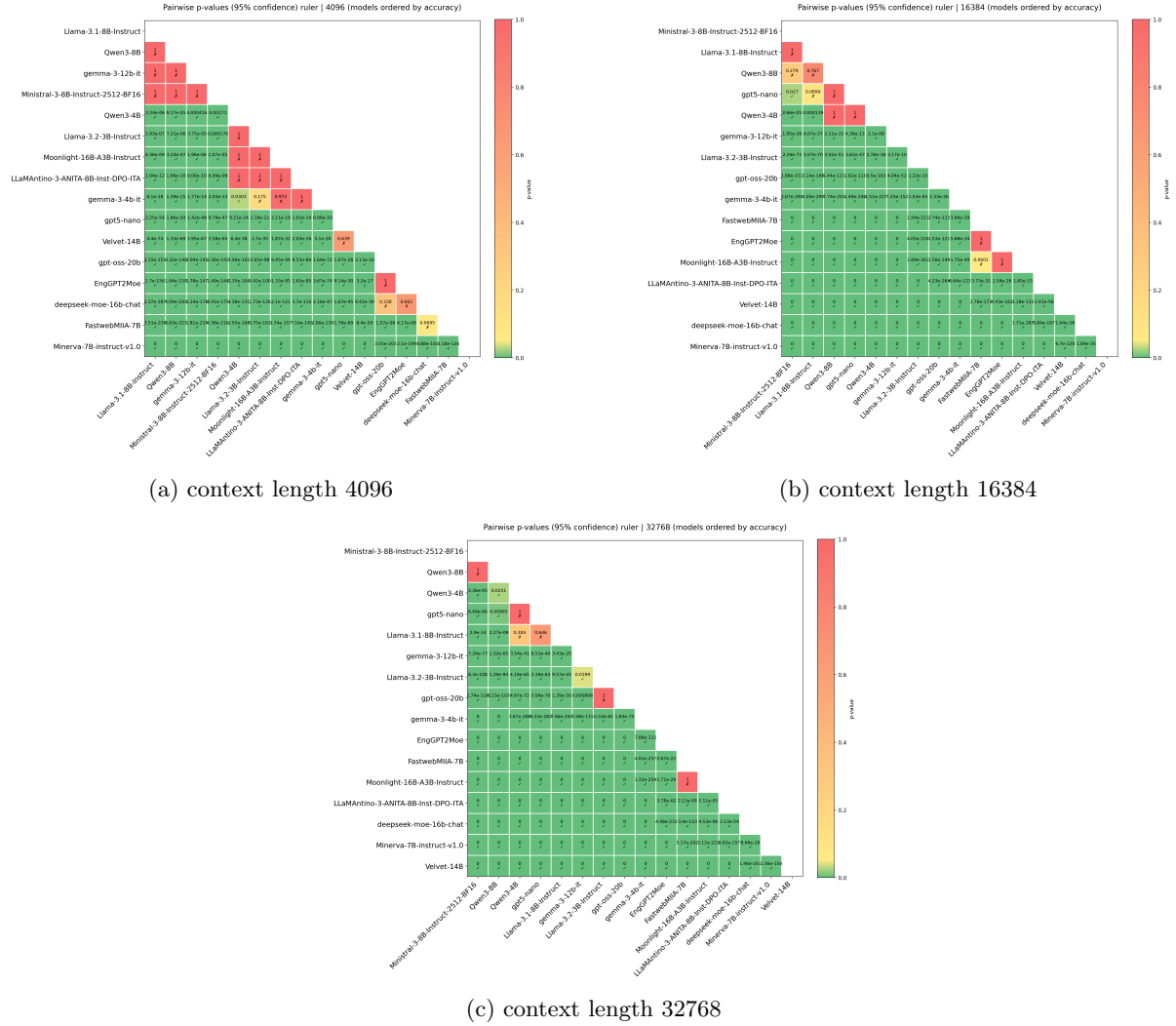


Figure 18: Bonferroni-adjusted pairwise significance heatmaps for RULER at context lengths 4096, 16384, and 32768.

## C BFCL dataset detailed structure

Table 26 contains the detailed description of each single task in the BFCL dataset, including number of samples, subdivision in task groups, and aggregation functions used to combine together scores obtained in each task towards group scores and then a single overall score. **Unweighted Average** means averaging each task score (accuracy) directly - meaning each tasks is considered equally, regardless of the number of samples it contains -, while **Weighted Average** means averaging each task score using the number of samples in the task as weights. All details were taken from the original BFCL implementation.

| Non-Live (Single-Turn)           |                                                    |                               |                                        |                             |                                         |  |
|----------------------------------|----------------------------------------------------|-------------------------------|----------------------------------------|-----------------------------|-----------------------------------------|--|
| Unweighted Average (1150)        |                                                    |                               |                                        |                             |                                         |  |
| Simple<br>Unweighted Average     |                                                    |                               | Python<br>Multiple<br>(200)            | Python<br>Parallel<br>(200) | Python<br>Parallel<br>Multiple<br>(200) |  |
| Python<br>(400)                  | Java<br>(100)                                      | JavaScript<br>(50)            |                                        |                             |                                         |  |
| Live (Single-Turn)               |                                                    |                               |                                        |                             |                                         |  |
| Weighted Average (1351)          |                                                    |                               |                                        |                             |                                         |  |
| Python<br>Simple<br>(258)        | Python<br>Multiple<br>(1053)                       | Python<br>Parallel<br>(16)    | Python<br>Parallel<br>Multiple<br>(24) |                             |                                         |  |
| Multi-Turn                       |                                                    |                               |                                        |                             |                                         |  |
| Unweighted Average<br>(800)      |                                                    |                               |                                        |                             |                                         |  |
| Multi Turn<br>Base<br>(200)      | Augmented Cases<br>Independent Categories<br>(600) |                               |                                        |                             |                                         |  |
|                                  | Missing<br>Function<br>(200)                       | Missing<br>Parameter<br>(200) | Long Context<br>(200)                  |                             |                                         |  |
| Hallucination Measurement        |                                                    |                               |                                        |                             |                                         |  |
| Unweighted Average<br>(1122)     |                                                    |                               |                                        |                             |                                         |  |
| Non-Live<br>Irrelevance<br>(240) |                                                    |                               | Live<br>Irrelevance<br>(882)           |                             |                                         |  |

Table 26: Detailed structure of the BFCL dataset used for the evaluations, in terms of task composition within each task group. The number in parentheses indicate the number of samples for each task and task group, while the averaging rules indicate how scores were grouped together for each task group.

## D Custom Task and Model Configurations for `lm_eval`

The following sections provide detailed descriptions of the modifications and extensions applied to the `lm_eval` harness tasks during our evaluation. We outline the base configuration template used to define the generation parameters passed to the evaluated models. Any tasks not explicitly discussed in these sections were executed using their default configurations as provided by the evaluation framework, without further modification.

```
8 doc_to_target: "{Answer}"
9 metric_list:
10 - metric: !function utils.pass_at_k
11   aggregation: mean
12   higher_is_better: true
13   k: [1,2,4,8]
14 - metric: !function utils.avg_at_k
15   aggregation: mean
16   higher_is_better: true
17   k: [1,2,4,8]
18 filter_list:
19 - name: "exact_match"
20   filter:
21     - function: "custom"
22       filter_fn: !function utils.extract_all_responses
23 generation_kwargs:
24   until:
25     - "Question:"
26     - "</s>"
27     - "<|im_end|>"
28     - "<|eot_id|>"
29 do_sample: false
30 temperature: 0.0
31 max_gen_toks: 32768
32 repeats: 8
```

Listing 1: Configuration for `aime24_pass_avg_8`

```
8 doc_to_target: "{answer}"
9 metric_list:
10 - metric: !function utils.pass_at_k
11   aggregation: mean
12   higher_is_better: true
13   k: [1,2,4,8]
14 - metric: !function utils.avg_at_k
15   aggregation: mean
16   higher_is_better: true
17   k: [1,2,4,8]
18 filter_list:
19 - name: "exact_match"
20   filter:
21     - function: "custom"
22       filter_fn: !function utils.extract_all_responses
23 generation_kwargs:
24   until:
25     - "Question:"
26     - "</s>"
27     - "<|im_end|>"
28     - "<|eot_id|>"
29 do_sample: false
30 temperature: 0.0
31 max_gen_toks: 32768
32 repeats: 8
```

Listing 2: Configuration for `aime25_pass_avg_8`

### D.1 AIME-2024 and AIME-2025

The AIME-2024 and AIME-2025 tasks were complemented with additional metrics, as shown in Listing 1 and 2 respectively, including `pass@k` and `avg@k` with `k` taking the values of 1, 2, 4, and 8. For evaluations performed through API calls (i.e., for OpenAI’s GPT-5-nano model), the prompt was changed (Listing 3 and 4) to clarify the output format as the final number had to be expressed with `\boxed{}` syntax.

```

8 fewshot_split: train
9 test_split: train
10 doc_to_text: "Given the following question, always answer using latex syntax for mathematics.
11 The final answer must be included in '\boxed{}'.\nQuestion: {{Problem}}\nAnswer:"
12 doc_to_target: "{{Answer}}"

```

Listing 3: Configuration for aime24\_openai

```

8 fewshot_split: test
9 test_split: test
10 doc_to_text: "Given the following question, always answer using latex syntax for mathematics.
11 The final answer must be included in '\boxed{}'.\nQuestion: {{problem}}\nAnswer:"
12 doc_to_target: "{{answer}}"

```

Listing 4: Configuration for aime25\_openai

## D.2 ARC-Challenge

The code in Listing 5 shows the modified regex rule to correctly parse the output in case of evaluations performed through API calls (i.e., for OpenAI’s GPT-5-nano model).

A similar change is shown in Listings 6 and 7. The latter was applied to the arc\_challenge task to improve the filtering functions.

```

26 ignore_punctuation: true
27 filter_list:
28   - name: strict_match
29     filter:
30       - function: "regex"
31         regex_pattern: The best answer is [^A-D]*([A-D])
32         group_select: -1
33       - function: take_first
34 metadata:
35   version: 1.0

```

Listing 5: Configuration for arc\_challenge\_chat\_openai

## D.3 MMLU

The code in Listing 8 shows the new regex rule added to compute accuracy also under flexible extraction.

## D.4 ITALIC

Since ITALIC is not natively supported by the `lm_evaluation_harness` library, we first performed the original chat-formatted evaluation, maintaining the two distinct prompting strategies available in the original implementation (a "fast" variant and a "chain-of-thought" (*CoT*) variant). While we adopted the original implementation, we introduced some enhancements to the `extract_answer` and `extract_answer_fast` for better answer extraction (both functions are located in the `ITALIC/run_eval.py` script). Specifically, we implemented more robust parsing rules to mitigate false negatives caused by functionally correct answers being misclassified (see Listing 9 and 10). Furthermore, we refined the few-shot formatting in the `configure_payload` function, explicitly formulating the example answers as 'Risposta: A' rather than a standalone 'A' (see Listing 11).

The ITALIC task was then made compatible with the `lm_evaluation_harness` library by creating custom configuration files. First, mirroring the original setup, both a standard non-*CoT* variant named `italic_fast` (Listing 12) and a *CoT* variant named `italic_cot` (Listing 13) were implemented. Second, we used more structured prompting by adapting templates from established benchmarks. This resulted in a non-*CoT* variant (`italic_fast_custom`) and a *CoT* variant (`italic_cot_custom`). To ensure methodological consistency, the prompt templates for these configurations were directly translated from

---

```

1 task: arc_challenge_cot_fixed
2 dataset_path: allenai/ai2_arc
3 dataset_name: ARC-Challenge
4 output_type: generate_until
5 training_split: train
6 validation_split: validation
7 test_split: test
8 fewshot_split: train
9 doc_to_text: "Given the following question and four candidate answers (A, B, C and D), choose the best ..."
10 fewshot_delimiter: "\n\n"
11 doc_to_target: "{ 'ABCD'[answerKey|int - 1] if answerKey|string in '1234' else answerKey }"
12 generation_kwargs:
13   until:
14     - "</s>"
15   max_gen_toks: 4096
16 metric_list:
17   - aggregation: mean
18     higher_is_better: true
19     ignore_case: true
20     ignore_punctuation: true
21     metric: exact_match
22 filter_list:
23   - name: flexible_extract
24     filter:
25       - function: regex
26         regex_pattern: "(?~|\\b)([A-D])(?:\\.|\\b)"
27         group_select: -1
28       - function: take_first
29   - name: strict_match
30     filter:
31       - function: "regex"
32         regex_pattern: The correct answer is [^A-D]*([A-D])
33         group_select: -1
34       - function: take_first
35 num_fewshot: 5
36 metadata:
37   version: 1.0

```

---

Listing 6: Configuration for arc\_challenge\_cot\_custom

the established `arc_challenge_chat` and `mmlu_cot_llama` tasks, respectively. We report only the code for the `italic_fast` and `italic_cot` tasks as the `italic_fast_custom` and `italic_cot_custom` tasks follow the same configurations except for the `doc_to_text` and `gen_prefix` parameters.

## D.5 Model Configurations

Listing 14 illustrates the base configuration template used to define the execution parameters for the evaluations. While this structure remained largely consistent across our experiments, specific arguments were dynamically adapted depending on the architecture and capabilities of the model under evaluation. The `model` parameter designates the inference backend used to serve the model. As previously discussed, `vllm` was our default choice for high-throughput evaluation. However, this parameter was modified to `hf-mistral3` for the Ministral-3-8B-Instruct-2512-BF16 evaluation and to `azure-openai-chat-completions` for the API-based evaluation of `gpt-5-nano`. Within the `model_args` definition, the reasoning-specific parameters—namely `enable_thinking` and `think_end_token`—were conditionally included. These were passed strictly to reasoning-capable models designed to output explicit intermediate thought traces (e.g., delimited by `<think>` tags) before providing a final answer. Additionally, the `include_path` directive was utilized exclusively when evaluating custom versions of the tasks, pointing to our local task implementations. Finally, to prevent excessively long generation and manage computational overhead, we optionally appended the `max_gen_toks` parameter within the generation keyword arguments (`gen_kwargs`) to strictly limit the maximum number of tokens a model could produce per sample.

```

1 task: arc_challenge_cot_it_fixed
2 dataset_path: alexandrinst/m_arc
3 dataset_name: it
4 output_type: generate_until
5 training_split: train
6 validation_split: val
7 test_split: test
8 fewshot_split: train
9 doc_to_text: "Data la seguente domanda e quattro possibili risposte (A, B, C e D), scegli la risposta ..."
10 fewshot_delimiter: "\n\n"
11 doc_to_target: "{answer}"
12 generation_kwargs:
13     until:
14         - "</s>"
15 metric_list:
16     - metric: exact_match
17       aggregation: mean
18       higher_is_better: true
19       ignore_case: true
20       ignore_punctuation: true
21 filter_list:
22     - name: flexible_extract
23       filter:
24         - function: regex
25           regex_pattern: "(?~|\\b)([A-D])(?:\\.|\\b)"
26           group_select: -1
27         - function: take_first
28     - name: strict_match
29       filter:
30         - function: "regex"
31           regex_pattern: "La risposta corretta è (.+?)(?:\\.)?$"
32           group_select: -1
33         - function: take_first
34 num_fewshot: 5
35 metadata:
36     version: 2.0

```

Listing 7: Configuration for arc\_challenge\_cot\_it\_custom

```

10 until: []
11 filter_list:
12     - name: strict_match
13       filter:
14         - function: "regex"
15           regex_pattern: "best answer is ([A-Z])"
16           group_select: -1
17         - function: take_first
18     - name: flexible_extract
19       filter:
20         - function: regex
21           regex_pattern: "(?~|\\b)([A-D])(?:\\.|\\b)"
22           group_select: -1
23         - function: take_first
24 metric_list:

```

Listing 8: Configuration for mmlu\_cot\_llama\_custom

```

1 def extract_answer_fast(output: str) -> str:
2     # truncate at the first USER: string to avoid hallucinations
3     output = re.split(r"\n?USER:", output)[0]
4
5     # First try: ASSISTANT: X
6     match = re.search(r"ASSISTANT:\s*([A-Z])", output)
7     if match:
8         return match.group(1)
9
10    LETTERS = "ABCDEFGHIJKLMNOPQRSTUVWXYZ"

```

Listing 9: Changes to the extract\_answer\_fast function

---

```

17 def extract_answer(output: str) -> str:
18     # truncate at the first USER: string to avoid hallucinations
19     output = re.split(r"\n?USER:", output)[0]
20
21     def _find(pattern: str, text: str, ignore_case: bool = True) -> str:
22
23         ...
24
25         return answer
26
27     # Last resort try: ASSISTANT: X (only if X is basically alone or at the very end/beginning)
28     match = re.search(r"ASSISTANT:\s*([A-Z])\b", output)
29     if match:
30         return match.group(1)
31
32

```

---

Listing 10: Changes to the `extract_answer` function

---

```

87     shot_options, shot_merged_letters = format_options(shot["options"])
88     messages.extend(
89         [
90             {
91                 "role": "user",
92                 "content": USER_QUERY_TEMPLATE.format(
93                     topic=shot["category"],
94                     question=shot["question"],
95                     options=shot_options,
96                     merged_letters=shot_merged_letters,
97                 ),
98             },
99             {
100                 "role": "assistant",
101                 "content": shot["answer"] if fast else f"Risposta: {shot['answer']}",
102             },
103         ]
104     )

```

---

Listing 11: Changes to the `configure_payload` function

---

```

1 task: italic_fast
2 tag: fast_original
3 dataset_path: Crisp-Unimib/ITALIC
4 output_type: generate_until
5 test_split: train
6 doc_to_text: !function utils.doc_to_text_original_fast
7 gen_prefix: 'Risposta:'
8 doc_to_target: !function utils.doc_to_target
9 generation_kwargs:
10     max_gen_toks: 100
11     until:
12         - "\n\n"
13         - "."
14 filter_list:
15     - name: extract_letter
16       filter:
17         - function: regex
18           regex_pattern: "([A-E])"
19         - function: take_first
20 metric_list:
21     - metric: exact_match
22       aggregation: mean
23       higher_is_better: true
24       ignore_case: true
25       ignore_punctuation: true
26 num_fewshot: 5
27 fewshot_split: train
28 fewshot_delimiter: "\n\n"
29 metadata:
30     version: 1.0

```

---

Listing 12: Configuration for `italic_fast`

---

```

1 task: italic_cot
2 tag: cot_original
3 dataset_path: Crisp-Unimib/ITALIC
4 output_type: generate_until
5 test_split: train
6 doc_to_text: !function utils.doc_to_text_original_cot_fewshot
7 doc_to_target: !function utils.doc_to_target
8 generation_kwargs:
9   until:
10     - "</s>"
11   max_gen_toks: 4096
12 filter_list:
13   - name: strict_match
14     filter:
15       - function: "regex"
16         regex_pattern: "Risposta: [^A-E]*([A-E])"
17         group_select: -1
18       - function: take_first
19   - name: flexible_extract
20     filter:
21       - function: regex
22         regex_pattern: "(?:~|\\b)([A-E])(?:\\.|\\b)"
23         group_select: -1
24       - function: take_first
25 metric_list:
26   - metric: exact_match
27     aggregation: mean
28     higher_is_better: true
29     ignore_case: true
30     ignore_punctuation: true
31 num_fewshot: 0
32 metadata:
33   version: 1.0

```

---

Listing 13: Configuration for italic\_cot

---

```

1 model: vllm
2 model_args:
3   pretrained: # Will be injected by run script
4   dtype: float16
5   trust_remote_code: true
6   enable_thinking: true
7   think_end_token: </think>
8
9 tasks:
10   - none # Will be injected by run script
11
12 include_path: ... # Path to custom tasks
13
14 batch_size: auto
15 device: cuda:0
16
17 apply_chat_template: true
18 confirm_run_unsafe_code: true
19
20 output_path: ... # Will be appended by run script
21 log_samples: true

```

---

Listing 14: Base configuration template utilized for the *lm\_eval* harness.

## E Evaluation Environment Technical Specifications

Tables 27 and 28 contains, respectively, the specifications for the hardware used to run all of the reported evaluations, and for the software used for each evaluation suite, for full reproducibility.

Table 27: Hardware specifications used for the evaluation.

| Component / Software  | Specification    |
|-----------------------|------------------|
| GPU Model             | NVIDIA A100-SXM4 |
| GPU Memory (VRAM)     | 40 GB            |
| NVIDIA Driver Version | 590.48.01        |
| CUDA Version          | 13.1             |
| Max Power Limit (TDP) | 400W             |

Table 28: Software libraries used for the evaluation and their respective versions, grouped by the different evaluation suites that were used.

| LM Evaluation Harness |         | BFCL      |          | HF Evaluation |         |
|-----------------------|---------|-----------|----------|---------------|---------|
| python                | 3.12.11 | python    | 3.12.11  | python        | 3.12.11 |
| lm_eval               | 0.4.12  | bfcl-eval | 2026.2.9 | lm_eval       | 0.4.12  |
| torch                 | 2.9.1   | evalscope | 1.4.2    | torch         | 2.10.0  |
| transformers          | 4.57.6  | requests  | 2.32.5   | transformers  | 5.3.0   |
| accelerate            | 1.12.0  | vllm      | 0.16.0   | accelerate    | 1.12.0  |
| openai                | 2.20.0  |           |          | openai        | 2.16.0  |
| PyJWT                 | 2.11.0  |           |          | PyJWT         | 2.10.1  |
| requests              | 2.32.5  |           |          | wonderwords   | 3.0.1   |
| aiohttp               | 3.13.3  |           |          | nltk          | 3.9.2   |
| tenacity              | 9.1.4   |           |          |               |         |
| tiktoken              | 0.12.0  |           |          |               |         |

## F Example responses

In this Section, we present a selection of illustrative samples drawn from most of the evaluated benchmarks. Each example includes the original input and target response from the dataset, actual responses from some of the models under evaluation, and an indication of correctness or score. These examples should be interpreted as qualitative evidence to help understand the different characteristics of both evaluation datasets and evaluated models, rather than as definitive indication on how a specific model responds to all samples and prompts.

### F.1 Example responses for AIME-2024

This subsection reports examples from the AIME24 benchmark, specifically for the `aime24_pass_avg_8` task. Table 29 presents a selected input instance (ID 2), while Table 30 summarizes the corresponding model outputs. In particular, we provide pass@1 responses for a meaningful subset of models, which is composed of the best performing models.

| Problem                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Solution                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Answer |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|
| Jen enters a lottery by picking 4 distinct numbers from $\{1, 2, 3, \dots, 9, 10\}$ . 4 numbers are randomly chosen from $S$ . She wins a prize if at least two of her numbers were 2 of the randomly chosen numbers, and wins the grand prize if all four of her numbers were the randomly chosen numbers. The probability of her winning the grand prize given that she won a prize is $\frac{m}{n}$ where $m$ and $n$ are relatively prime positive integers. Find $m + n$ . | <p>This is a conditional probability problem. Bayes' Theorem states that: <math>[P(A B) = \frac{P(B A) \cdot P(A)}{P(B)}]</math>. In other words, the probability of <math>A</math> given <math>B</math> is equal to the probability of <math>B</math> given <math>A</math> times the probability of <math>A</math> divided by the probability of <math>B</math>. In our case, <math>A</math> represents the probability of winning the grand prize, and <math>B</math> represents the probability of winning a prize. Clearly, <math>P(B A) = 1</math>, since by winning the grand prize you automatically win a prize.</p> <p>Thus, we want to find <math>\frac{P(A)}{P(B)}</math>. Let us calculate the probability of winning a prize. We do this through casework: how many of Jen's drawn numbers match the lottery's drawn numbers To win a prize, Jen must draw at least 2 numbers identical to the lottery. Thus, our cases are drawing 2, 3, or 4 numbers identical. Let us first calculate the number of ways to draw exactly 2 identical numbers to the lottery. Let Jen choose the numbers <math>a, b, c</math>, and <math>d</math>; we have <math>\binom{4}{2}</math> ways to choose which 2 of these 4 numbers are identical to the lottery. We have now determined 2 of the 4 numbers drawn in the lottery; since the other 2 numbers Jen chose can not be chosen by the lottery, the lottery now has <math>10 - 2 - 2 = 6</math> numbers to choose the last 2 numbers from. Thus, this case is <math>\binom{6}{2}</math>, so this case yields <math>\binom{4}{2} \binom{6}{2} = 6 \cdot 15 = 90</math> possibilities. Next, let us calculate the number of ways to draw exactly 3 identical numbers to the lottery. Again, let Jen choose <math>a, b, c</math>, and <math>d</math>. This time, we have <math>\binom{4}{3}</math> ways to choose the identical numbers and again 6 numbers left for the lottery to choose from; however, since 3 of the lottery's numbers have already been determined, the lottery only needs to choose 1 more number, so this is <math>\binom{6}{1}</math>. This case yields <math>\binom{4}{3} \binom{6}{1} = 4 \cdot 6 = 24</math>. Finally, let us calculate the number of ways to all 4 numbers matching. There is actually just one way for this to happen. In total, we have <math>90 + 24 + 1 = 115</math> ways to win a prize. The lottery has <math>\binom{10}{4} = 210</math> possible combinations to draw, so the probability of winning a prize is <math>\frac{115}{210}</math>.</p> <p>There is actually no need to simplify it or even evaluate <math>\binom{10}{4}</math> or actually even know that it has to be <math>\binom{10}{4}</math>; it suffices to call it <math>a</math> or some other variable, as it will cancel out later. However, let us just go through with this. The probability of winning a prize is <math>\frac{115}{210}</math>. Note that the probability of winning a grand prize is just matching all 4 numbers, which we already calculated to have 1 possibility and thus have probability <math>\frac{1}{210}</math>. Thus, our answer is <math>\frac{\frac{115}{210}}{\frac{1}{210}} = \frac{115}{1}</math>.</p> <p>Therefore, our answer is <math>1 + 115 = \boxed{116}</math>.</p> | 116    |

Table 29: Representative example (Sample ID 2) from the AIME24 benchmark. The Problem column reports the original question, the Solution column provides the reference step-by-step derivation, and the Answer column contains the final ground-truth value used for evaluation.

Table 30: Representative model responses on the AIME24 benchmark (sample ID 2). For each model, the Answer column reports the pass@1 response, while the Result column presents the corresponding pass@1, pass@8, and avg@8 metrics for the given data sample. NOTE: the response generated by FastwebMIIA-7B is truncated due to repetitive token generation. Minor formatting adjustments have been applied to model responses for clarity and consistency of presentation.

| Model           | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Result                                                |
|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------|
| EngGPT2-16B-A3B | <p>To solve the problem, we are given a scenario where:</p> <ul style="list-style-type: none"> <li>• <b>Jen</b> selects <b>4</b> distinct numbers from the set <math>S = 1, 2, 3, \dots, 10</math>.</li> <li>• <b>4</b> numbers are randomly selected from the same set <math>S</math>.</li> <li>• Jen <b>wins a prize</b> if <b>at least two</b> of her selected numbers are among the 4 randomly chosen.</li> <li>• She <b>wins the grand prize</b> if <b>all four</b> of her selected numbers are among the 4 randomly chosen.</li> </ul> <p>We are asked to compute the <b>conditional probability</b> that Jen <b>won the grand prize</b>, <b>given</b> that she won a prize. That is:</p> $P(\text{Grand Prize} \mid \text{Prize}) = \frac{P(\text{Grand Prize})}{P(\text{Prize})}$ <p>Step 1: Total number of ways to choose 4 numbers from 10<br/> The total number of possible 4-number combinations that can be drawn from the 10-number set is:</p> $\binom{10}{4} = 210$ <p>Step 2: Count the number of favorable outcomes for the <b>grand prize</b><br/> Jen's 4 numbers are a specific set of 4 distinct numbers. To win the <b>grand prize</b>, all 4 of her numbers must be among the 4 randomly selected numbers. There is <b>only one</b> such combination (the set of her 4 numbers), so:</p> $\text{Number of grand prize outcomes} = 1$ <p>Step 3: Count the number of favorable outcomes for the <b>prize</b><br/> Jen wins a <b>prize</b> if <b>at least two</b> of her 4 numbers are among the 4 selected. So, we need to count the number of 4-number combinations that include <b>at least 2</b> of her 4 numbers. To compute this, we use the <b>complement</b>:</p> <ul style="list-style-type: none"> <li>• <b>Number of combinations with 0 of her numbers</b> (i.e., all 4 numbers are from the other 6 numbers in <math>S</math>):</li> </ul> $\binom{6}{4} = 15$ <ul style="list-style-type: none"> <li>• <b>Number of combinations with 1 of her numbers</b> (i.e., 1 from her 4, and 3 from the other 6):</li> </ul> $\binom{4}{1} \cdot \binom{6}{3} = 4 \cdot 20 = 80$ <p>So, the number of combinations with <b>fewer than 2</b> of her numbers is:</p> $15 + 80 = 95$ <p>Therefore, the number of combinations with <b>at least 2</b> of her numbers is:</p> $210 - 95 = 115$ <p>Step 4: Compute the conditional probability<br/> We now compute the conditional probability that the <b>grand prize</b> was won, <b>given</b> that a prize was won: <math>P(\text{Grand Prize} \mid \text{Prize}) = \frac{1}{115}</math>. This is already in lowest terms, since 115 = 5 × 23, and 1 has no common factors with 115.<br/> Final Step: Find <math>m + n</math><br/> Given that the probability is <math>\frac{m}{n} = \frac{1}{115}</math>, and <math>m = 1</math>, <math>n = 115</math>, we have:</p> $m + n = 1 + 115 = \boxed{116}$ | <p>pass@1: 100%<br/> pass@8: 100%<br/> avg@8: 75%</p> |
| Velvet-14B      | <p>There are 10 choices for the first number, 9 choices for the second number, 8 choices for the third number, and 7 choices for the fourth number, for a total of <math>10 \cdot 9 \cdot 8 \cdot 7 = 5040</math> ways to choose the numbers. There are 10 ways to choose the numbers if all four are 2, and <math>9 \cdot 8 \cdot 7 = 504</math> ways to choose the numbers if at least two are 2. Thus, the probability of winning the grand prize given that she won a prize is <math>\frac{504}{5040} = \frac{1}{10}</math>. The answer is <math>1 + 10 = \boxed{11}</math></p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | <p>pass@1: 0%<br/> pass@8: 0%<br/> avg@8: 0%</p>      |

|                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                                                |
|----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------|
| FastwebMIIA-7B | <p>Let us outline the process to tackle this challenge:</p> <ol style="list-style-type: none"> <li>1. Start by understanding the problem: Jen has to pick 4 distinct numbers from a set of 10 numbers. She wins a prize if at least two of her numbers match the randomly chosen numbers, and the grand prize if all four of her numbers match the randomly chosen numbers.</li> <li>2. Calculate the total number of ways to choose 4 numbers from 10, which is <math>\binom{10}{4}</math>.</li> <li>3. Calculate the number of ways to choose 4 numbers such that none of them match the randomly chosen numbers, which is <math>\binom{6}{4}</math>.</li> <li>4. Calculate the number of ways to choose 4 numbers such that exactly two of them match the randomly chosen numbers, which is <math>\binom{4}{2} \cdot \binom{6}{2}</math>.</li> <li>5. Calculate the number of ways to choose 4 numbers such that exactly three of them match the randomly chosen numbers, which is <math>\binom{2}{3} \cdot \binom{6}{1}</math>.</li> <li>6. Calculate the number of ways to choose 4 numbers such that all four of them match the randomly chosen numbers, which is <math>\binom{1}{1} \cdot \binom{6}{4}</math>.</li> <li>7. Add up all these cases to get the total number of ways to win a prize, which is <math>\binom{10}{4} - \binom{6}{4} + \binom{4}{2} \cdot \binom{6}{2} - \binom{2}{3} \cdot \binom{6}{1} + \binom{1}{1} \cdot \binom{6}{4}</math>.</li> <li>8. Calculate the total number of ways to choose 4 numbers from 10, which is <math>\binom{10}{4}</math>.</li> <li>9. Calculate the number of ways to choose 4 numbers such that at least two of them match the randomly chosen numbers, which is <math>\binom{10}{4} - \binom{6}{4} + \binom{4}{2} \cdot \binom{6}{2} - \binom{2}{3} \cdot \binom{6}{1} + \binom{1}{1} \cdot \binom{6}{4}</math>.</li> <li>10. The probability of winning the grand prize given that she won a prize is the number of ways to choose 4 numbers such that all four of them match the randomly chosen numbers divided by the total number of ways to choose 4 numbers such that at least two of them match the randomly chosen numbers.</li> <li>11. Simplify this to <math>\frac{\binom{1}{1} \cdot \binom{6}{4}}{\binom{10}{4} - \binom{6}{4}}</math></li> </ol> <p>To conclude the reasoning, let's move on with the detailed explanation of the solution. There are <math>\binom{10}{4}</math> ways to choose 4 numbers from 10. There are <math>\binom{6}{4}</math> ways to choose 4 numbers such that none of them match the randomly chosen numbers. There are <math>\binom{4}{2} \cdot \binom{6}{2}</math> ways to choose 4 numbers such that exactly two of them match the randomly chosen numbers. There are <math>\binom{2}{3} \cdot \binom{6}{1}</math> ways to choose 4 numbers such that exactly three of them match the randomly chosen numbers. There are <math>\binom{1}{1} \cdot \binom{6}{4}</math> ways to choose 4 numbers such that all four of them match the randomly chosen numbers. So the total number of ways to win a prize is <math>\binom{10}{4} - \binom{6}{4} + \binom{4}{2} \cdot \binom{6}{2} - \binom{2}{3} \cdot \binom{6}{1} + \binom{1}{1} \cdot \binom{6}{4}</math>. The total number of ways to choose 4 numbers from 10 is <math>\binom{10}{4}</math>. The number of ways to choose 4 numbers such that at least two of them match the randomly chosen numbers is <math>\binom{10}{4} - \binom{6}{4} + \binom{4}{2} \cdot \binom{6}{2} - \binom{2}{3} \cdot \binom{6}{1} + \binom{1}{1} \cdot \binom{6}{4}</math>. The probability of winning the grand prize given that she won a prize is <math>\frac{\binom{1}{1} \cdot \binom{6}{4}}{\binom{10}{4} - \binom{6}{4} + \binom{4}{2} \cdot \binom{6}{2} - \binom{2}{3} \cdot \binom{6}{1} + \binom{1}{1} \cdot \binom{6}{4}}</math>. This simplifies to <math>\frac{\binom{6}{4}}{\binom{10}{4} - \binom{6}{4} + \binom{4}{2} \cdot \binom{6}{2} - \binom{2}{3} \cdot \binom{6}{1} + \binom{1}{1} \cdot \binom{6}{4}}</math>. This further simplifies to <math>\frac{\binom{6}{4}}{\binom{10}{4} - \binom{6}{4} + \binom{4}{2} \cdot \binom{6}{2} - \binom{2}{3} \cdot \binom{6}{1} + \binom{1}{1} \cdot \binom{6}{4}}</math>. This further simplifies to <math>\frac{\binom{6}{4}}{\binom{10}{4} - \binom{6}{4} + \binom{4}{2} \cdot \binom{6}{2} - \binom{2}{3} \cdot \binom{6}{1} + \binom{1}{1} \cdot \binom{6}{4}} \dots</math></p> | <p>pass@1: 0%<br/>pass@8: 0%<br/>avg@8: 0%</p> |
|----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------|

|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                             |
|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------|
| Qwen3-4B | <p>To solve the problem, we are given that Jen picks 4 distinct numbers from the set <math>S = 1, 2, 3, \dots, 10</math>, and then 4 numbers are randomly chosen from <math>S</math>. She wins a prize if <b>at least two</b> of her numbers match the 4 chosen numbers, and she wins the <b>grand prize</b> if <b>all four</b> of her numbers match the 4 chosen numbers.</p> <p>We are asked to find the probability that she <b>wins the grand prize</b> given that she <b>won a prize</b>, and express this probability as a reduced fraction <math>\frac{m}{n}</math>, then compute <math>m + n</math>.</p> <p>Step 1: Total Number of Possible Outcomes</p> <p>The total number of ways to choose 4 numbers from the set <math>S</math> of 10 elements is:</p> $\binom{10}{4} = 210$ <p>Step 2: Probability of Winning the Grand Prize</p> <p>This is the probability that all 4 of her numbers match the 4 chosen numbers. Since there is only <b>one</b> such favorable outcome (her exact set), the probability is:</p> $P(\text{Grand Prize}) = \frac{1}{210}$ <p>Step 3: Probability of Winning a Prize (At Least Two Matches)</p> <p>We compute the probability that <b>at least two</b> of her numbers match the 4 chosen numbers. This is the complement of the event that <b>fewer than two</b> numbers match, i.e., <b>0</b> or <b>1</b> match.</p> <p>Let's compute the number of favorable outcomes for each case:</p> <ul style="list-style-type: none"> <li><b>0 matches</b>: Choose 4 numbers from the 6 that are <b>not</b> in her set.</li> </ul> $\binom{6}{4} = 15$ <ul style="list-style-type: none"> <li><b>1 match</b>: Choose 1 number from her 4, and 3 from the 6 not in her set.</li> </ul> $\binom{4}{1} \cdot \binom{6}{3} = 4 \cdot 20 = 80$ <p>So, the number of outcomes with <b>at least two matches</b> is:</p> $210 - 15 - 80 = 115$ <p>Thus, the probability of winning a prize is:</p> $P(\text{At least 2 matches}) = \frac{115}{210} = \frac{23}{42}$ <p>Step 4: Conditional Probability</p> <p>We are asked to find the probability that she <b>won the grand prize</b> given that she <b>won a prize</b>. That is:</p> $P(\text{Grand Prize} \mid \text{At least 2 matches}) = \frac{P(\text{Grand Prize})}{P(\text{At least 2 matches})}$ <p>Substituting the values:</p> $P = \frac{1/210}{115/210} = \frac{1}{115}$ <p>Step 5: Final Answer</p> <p>The probability is <math>\frac{1}{115}</math>, so <math>m = 1</math>, <math>n = 115</math>, and:</p> $m + n = 1 + 115 = \boxed{116}$ | pass@1: 100%<br>pass@8: 100%<br>avg@8: 100% |
|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------|

|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |                                             |
|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------|
| Qwen3-8B | <p>To solve the problem, we are given a lottery scenario where:</p> <ul style="list-style-type: none"> <li>• Jen picks <b>4</b> distinct numbers from the set <math>S = 1, 2, 3, \dots, 10</math>.</li> <li>• The lottery randomly selects <b>4</b> numbers from the same set <math>S</math></li> <li>• Jen <b>wins a prize</b> if <b>at least two</b> of her numbers match the lottery numbers.</li> <li>• She <b>wins the grand prize</b> if <b>all four</b> of her numbers match the lottery numbers.</li> </ul> <p>We are asked to compute the <b>conditional probability</b> that Jen wins the <b>grand prize</b>, <b>given</b> that she has won a prize. That is:</p> $P(\text{Grand Prize} \mid \text{Prize}) = \frac{P(\text{Grand Prize} \cap \text{Prize})}{P(\text{Prize})}$ <p>Since winning the grand prize implies that she has also won a prize (because all four numbers match), the intersection is just the event of winning the grand prize. Therefore:</p> $P(\text{Grand Prize} \mid \text{Prize}) = \frac{P(\text{Grand Prize})}{P(\text{Prize})}$ <p>Step 1: Total Number of Possible Lottery Combinations<br/>The total number of ways to choose 4 numbers from 10 is:</p> $\binom{10}{4} = 210$ <p>Step 2: Probability of Winning the Grand Prize<br/>Jen wins the grand prize if the lottery draws <b>exactly</b> the 4 numbers she picked. Since there is only <b>one</b> such combination (her chosen numbers), the probability is:</p> $P(\text{Grand Prize}) = \frac{1}{210}$ <p>Step 3: Probability of Winning a Prize (At Least Two Matches)<br/>To compute this, we consider the number of lottery combinations that have <b>at least two</b> of Jen's numbers. This is the complement of the cases where <b>fewer than two</b> of her numbers match.<br/>Let's compute the number of combinations with:</p> <ul style="list-style-type: none"> <li>• <b>0 matches</b>: Choose 4 numbers from the 6 not in her set:</li> </ul> $\binom{6}{4} = 15$ <ul style="list-style-type: none"> <li>• <b>1 match</b>: Choose 1 number from her 4, and 3 from the 6 not in her set:</li> </ul> $\binom{4}{1} \cdot \binom{6}{3} = 4 \cdot 20 = 80$ <p>So, the number of combinations with <b>fewer than two matches</b> is:</p> $15 + 80 = 95$ <p>Therefore, the number of combinations with <b>at least two matches</b> is:</p> $210 - 95 = 115$ <p>Thus, the probability of winning a prize is:</p> $P(\text{Prize}) = \frac{115}{210} = \frac{23}{42}$ <p>Step 4: Compute the Conditional Probability<br/>Now, we compute the conditional probability:</p> $P(\text{Grand Prize} \mid \text{Prize}) = \frac{P(\text{Grand Prize})}{P(\text{Prize})} = \frac{\frac{1}{210}}{\frac{23}{42}} = \frac{1}{210} \cdot \frac{42}{23} = \frac{42}{210 \cdot 23} = \frac{1}{5 \cdot 23} = \frac{1}{115}$ <p>Step 5: Final Answer<br/>The probability is <math>\frac{1}{115}</math>, where <math>m = 1</math> and <math>n = 115</math>. Since they are relatively prime, the final answer is:</p> $m + n = 1 + 115 = \boxed{116}$ | pass@1: 100%<br>pass@8: 100%<br>avg@8: 100% |
|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------|

|                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                      |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------|
| deepseek-moe-16b-chat | <p>First, let's find the probability of Jen winning a prize. There are a total of <math>\binom{10}{4} = 210</math> possible ways to choose 4 numbers from the set <math>S</math>. If Jen wins a prize, she must have chosen 2 of the 4 randomly chosen numbers. So, the probability of winning a prize is <math>\frac{210}{210} = \frac{1}{1}</math>.</p> <p>Now, let's find the probability of Jen winning the grand prize given that she won a prize. There are <math>\binom{10}{4} = 210</math> ways to choose 4 numbers from the set <math>S</math> if Jen wins the grand prize. Since she won a prize, she must have chosen 2 of the 4 randomly chosen numbers. So, the probability of winning the grand prize given that she won a prize is <math>\frac{210}{210} = \frac{1}{1}</math>.</p> <p>Therefore, the probability of Jen winning the grand prize given that she won a prize is <math>\frac{1}{1} = \frac{1}{1}</math>.</p> <p>So, <math>m + n = 1 + 1 = \boxed{2}</math>.</p> <p>The answer is: 2</p> | <p>pass@1: 0%</p> <p>pass@8: 0%</p> <p>avg@8: 0%</p> |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------|

|             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                                             |
|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------|
| gpt-oss-20b | <p>&lt; channel &gt;analysis&lt; message &gt;We need probability that all four numbers match given that at least two match. So conditional probability. We have sample space: all possible 4-number sets chosen from 10 numbers. Jen picks a set of 4 distinct numbers. The lottery draws 4 numbers uniformly from 10. Equivalent to random draw of 4 numbers; we can treat Jen's set fixed. We want probability that intersection size <math>\geq 2</math>. Among those, probability that intersection size = 4. Compute: total number of possible draws: <math>C(10, 4) = 210</math>. Number of draws that match exactly <math>k</math> of Jen's numbers: choose <math>k</math> from her 4, and choose <math>4 - k</math> from remaining 6. So count = <math>C(4, k) * C(6, 4 - k)</math>. We need <math>k \geq 2</math>. So favorable for prize: <math>k = 2, 3, 4</math>. Grand prize: <math>k = 4</math>. Compute counts:</p> $k = 2 : C(4, 2) * C(6, 2) = 6 * 15 = 90.$ $k = 3 : C(4, 3) * C(6, 1) = 4 * 6 = 24.$ $k = 4 : C(4, 4) * C(6, 0) = 1 * 1 = 1.$ <p>Total prize draws: <math>90 + 24 + 1 = 115</math>.<br/> Grand prize draws: 1.<br/> Probability grand prize given prize = <math>1/115</math>.<br/> Simplify fraction: <math>\frac{1}{115}</math> already in lowest terms. <math>m = 1</math>, <math>n = 115</math>, <math>sum = 116</math>.<br/> But check: Are we missing that Jen picks 4 distinct numbers; but lottery picks 4 numbers from <math>S</math>. Yes. But maybe the lottery picks 4 numbers with replacement? No, 4 numbers are randomly chosen from <math>S</math> likely without replacement. So correct. Thus answer 116.<br/> But double-check: Are we assuming that Jen's numbers are distinct? Yes. So intersection counts as above. Thus answer 116.</p> <p>&lt; end &gt;&lt; start &gt;assistant&lt; channel &gt;final&lt; message &gt;For a fixed choice of Jen's four distinct numbers, the lottery draws a 4-element subset of the set</p> $[S = 1, 2, \dots, 10]$ <p>uniformly at random. The total number of possible draws is</p> $[\binom{10}{4} = 210].$ <p>Count the draws that give a prize<br/> Let <math>(k)</math> be the number of Jen's numbers that appear in the lottery draw. For a given <math>(k)</math> the number of draws is</p> $[\binom{4}{k} \binom{6}{4-k}],$ <p>because we choose <math>(k)</math> of Jen's 4 numbers and <math>(4 - k)</math> of the remaining 6 numbers.<br/> We need</p> $(k \geq 2).$ $k = 2 : \binom{4}{2} \binom{6}{2} = 6 \cdot 15 = 90,$ $k = 3 : \binom{4}{3} \binom{6}{1} = 4 \cdot 6 = 24,$ $k = 4 : \binom{4}{4} \binom{6}{0} = 1 \cdot 1 = 1.$ <p>Total draws that give at least a prize:</p> $[90 + 24 + 1 = 115].$ <p>Count the draws that give the grand prize<br/> The grand prize occurs only when all four of Jen's numbers are drawn:</p> $[k = 4 \implies 1 \text{ draw}].$ <p>Conditional probability</p> $P(\text{grand prize} \mid \text{prize}) = \frac{\# \text{ grand prize draws}}{\# \text{ prize draws}} = \frac{1}{115}.$ <p>The fraction <math>(\frac{1}{115})</math> is already in lowest terms, so</p> $m = 1, n = 115$ $m + n = 1 + 115 = 116]$ <div style="border: 1px solid black; padding: 2px; width: fit-content; margin: 0 auto;">116</div> | pass@1: 100%<br>pass@8: 100%<br>avg@8: 100% |
|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------|

|                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                             |
|----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------|
| Moonlight-16B-A3B-Instruct | <p>To solve the problem, we need to determine the probability that Jen wins the grand prize given that she wins a prize. We will break this down into several steps.</p> <p>Step 1: Calculate the total number of ways to choose 4 numbers from 10.<br/>The total number of ways to choose 4 numbers from a set of 10 is given by the binomial coefficient:</p> $\left[ \binom{10}{4} = \frac{10!}{4!(10-4)!} = \frac{10!}{4!6!} = 210 \right]$ <p>Step 2: Calculate the number of ways to choose 4 numbers such that none of them are among the 4 numbers Jen picked.<br/>If Jen does not win a prize, none of her 4 numbers are among the 4 numbers chosen. The number of ways to choose 4 numbers from the remaining 6 numbers is:</p> $\left[ \binom{6}{4} = \frac{6!}{4!(6-4)!} = \frac{6!}{4!2!} = 15 \right]$ <p>Step 3: Calculate the number of ways to choose 4 numbers such that at least two of them are among the 4 numbers Jen picked.<br/>To find this, we subtract the number of ways to choose 4 numbers such that none of them are among the 4 numbers Jen picked from the total number of ways to choose 4 numbers:</p> $\left[ \binom{10}{4} - \binom{6}{4} = 210 - 15 = 195 \right]$ <p>Step 4: Calculate the number of ways to choose 4 numbers such that all four of them are among the 4 numbers Jen picked.<br/>This is the number of ways to choose 4 numbers from the 4 numbers Jen picked:</p> $\left[ \binom{4}{4} = 1 \right]$ <p>Step 5: Calculate the probability that Jen wins the grand prize given that she wins a prize.<br/>The probability that Jen wins the grand prize given that she wins a prize is the ratio of the number of ways to choose 4 numbers such that all four are among the 4 numbers Jen picked to the number of ways to choose 4 numbers such that at least two are among the 4 numbers Jen picked:</p> $\left[ \frac{\binom{4}{4}}{\binom{10}{4} - \binom{6}{4}} = \frac{1}{195} \right]$ <p>Step 6: Simplify the fraction and find <math>(m + n)</math>.<br/>The fraction <math>(\frac{1}{195})</math> is already in simplest form, where <math>(m = 1)</math> and <math>(n = 195)</math>. Therefore, <math>(m + n = 1 + 195 = 196)</math>.<br/>Thus, the final answer is:</p> $\boxed{196}$ | pass@1: 0%<br>pass@8: 0%<br>avg@8: 0%       |
| gpt5-nano                  | <p>Let <math>X</math> = number of matches between Jen's 4 numbers and the lottery's 4 numbers. For fixed Jen set,</p> $(X \sim \text{Hypergeometric}(N = 10, K = 4, n = 4)),$ <p>so</p> $[P(X = k) = \frac{\binom{4}{k} \binom{6}{4-k}}{\binom{10}{4}}, \quad k = 0, 1, 2, 3, 4].$ <p>Thus</p> $[P(X = 0) = \frac{15}{210}, \quad P(X = 1) = \frac{80}{210},$ $P(X = 2) = \frac{90}{210}, \quad P(X = 3) = \frac{24}{210}, \quad P(X = 4) = \frac{1}{210}].$ <p>Win a prize</p> $(\iff X \geq 2.$ <p>Hence</p> $[P(\text{prize}) = P(X \geq 2) = \frac{90 + 24 + 1}{210} = \frac{115}{210} = \frac{23}{42}].$ <p>Grand prize</p> $(\iff X = 4).$ <p>Therefore</p> $[P(\text{grand prize} \mid \text{prize}) = \frac{P(X = 4)}{P(X \geq 2)} = \frac{1/210}{115/210} = \frac{1}{115}].$ <p>Thus</p> $(m = 1, n = 115)$ $(gcd((1, 115) = 1)),$ <p>and</p> $[m + n = 1 + 115 = \boxed{116}].$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | pass@1: 100%<br>pass@8: 100%<br>avg@8: 100% |

## F.2 Example responses for AIME-2025

This section reports representative examples from the AIME25 benchmark, specifically for the `aime25_pass_avg_8` task. Table 31 presents a selected input instance (ID 23), while Table 32 summarizes the corresponding model outputs. In particular, we provide `pass@_` responses for a meaningful subset of models.

| Problem                                                                                                                                                                                                                                                                                                                                   | Answer |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|
| <p>There are <math>n</math> values of <math>x</math> in the interval <math>0 &lt; x &lt; 2\pi</math> where</p> $f(x) = \sin(7\pi \cdot \sin(5x)) = 0.$ <p>For <math>t</math> of these <math>n</math> values of <math>x</math>, the graph of <math>y = f(x)</math> is tangent to the <math>x</math>-axis.<br/>Find <math>n + t</math>.</p> | 149    |

Table 31: Representative example (Sample ID 23) from the AIME25 benchmark. The Problem column reports the original question, while the Answer column contains the final ground-truth value used for evaluation.

Table 32: Representative model responses on the AIME25 benchmark (sample ID 23). For each model, the Answer column reports the `pass@1` response, while the Result column presents the corresponding `pass@1`, `pass@8`, and `avg@8` metrics for the given data sample. We report `pass@1` outputs for each model. NOTE: the responses generated by LLaMAntino-3-ANITA-8B and Ministral-3-8B-Instruct-2512-BF16 are truncated due to repetitive token generation. Minor formatting adjustments have been applied to model responses for clarity and consistency of presentation.

| Model           | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Result                                              |
|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------|
| EngGPT2-16B-A3B | <p>To solve the problem, we are given the function:</p> $f(x) = \sin(7\pi \cdot \sin(5x)), \quad \text{for } 0 < x < 2\pi$ <p>We are to find the number <math>n</math> of values of <math>x</math> in this interval where <math>f(x) = 0</math>, and the number <math>t</math> of those values where the graph of <math>f(x)</math> is <b>tangent</b> to the <math>x</math>-axis (i.e., where <math>f(x) = 0</math> and <math>f'(x) = 0</math>), and then compute <math>n + t</math>.<br/>Step 1: Count the number of zeros of <math>f(x)</math><br/>The function <math>f(x) = \sin(7\pi \cdot \sin(5x))</math> is zero when the argument of the sine function is an integer multiple of <math>\pi</math>:</p> $7\pi \cdot \sin(5x) = k\pi \quad \Rightarrow \quad \sin(5x) = \frac{k}{7}$ <p>So, for each integer <math>k</math>, we solve <math>\sin(5x) = \frac{k}{7}</math> in the interval <math>(0, 2\pi)</math>. The sine function <math>\sin(\theta) = c</math> has:</p> <ul style="list-style-type: none"> <li>• <b>Two solutions</b> in <math>(0, 2\pi)</math> if <math>c \in (-1, 1)</math></li> <li>• <b>One solution</b> if <math>c = \pm 1</math></li> <li>• <b>No solution</b> if <math> c  &gt; 1</math></li> </ul> <p>So, for each <math>k \in \mathbb{Z}</math>, we check if <math> \frac{k}{7}  \leq 1 \Rightarrow  k  \leq 7</math>. Thus, <math>k \in \{-7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7\}</math>, a total of <b>15 values</b>. Now, for each <math>k</math>, we count the number of <math>x \in (0, 2\pi)</math> such that <math>\sin(5x) = \frac{k}{7}</math>:</p> <ul style="list-style-type: none"> <li>• <b><math>k = \pm 7</math>:</b> <math>\sin(5x) = \pm 1</math>, which has <b>1 solution per period</b> of the sine function. Since <math>\sin(5x)</math> has period <math>\frac{2\pi}{5}</math>, and <math>x \in (0, 2\pi)</math> spans 5 full periods, we get <b>5 solutions</b> for each <math>k = 7</math> and <math>k = -7</math>.</li> <li>• <b><math>k = \pm 1</math>:</b> <math>\sin(5x) = \pm 1/7</math>, which is in <math>(-1, 1)</math>, so <b>2 solutions per period</b>. With 5 periods, that's <b>10 solutions</b> for each <math>k = 1</math> and <math>k = -1</math>.</li> <li>• <b>For <math> k  = 2, 3, 4, 5, 6</math>:</b> <math>\sin(5x) = \pm \frac{k}{7}</math>, still in <math>(-1, 1)</math>, so <b>10 solutions</b> for each of these 5 values of <math>k</math> (excluding <math>\pm 1</math>)</li> <li>• <b><math>k = 0</math>:</b> <math>\sin(5x) = 0</math>, which has <b>2 solutions per period</b>, 5 periods <b>9 solutions</b> in <math>(0, 2\pi)</math></li> </ul> | <p>pass@1: 0%<br/>pass@8: 100%<br/>avg@8: 62.5%</p> |



|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                        |
|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------|
| Qwen3-4B | <p>To solve the problem, we are given the function:</p> $f(x) = \sin(7\pi \cdot \sin(5x))$ <p>We are to find the number of values of <math>x</math> in the interval <math>0 &lt; x &lt; 2\pi</math> such that <math>f(x) = 0</math>, and then determine how many of those values correspond to points where the graph of <math>f(x)</math> is <b>tangent</b> to the x-axis. Finally, we are to compute <math>n + t</math>, where <math>n</math> is the total number of zeros of <math>f(x)</math>, and <math>t</math> is the number of those zeros where the graph is tangent to the x-axis.</p> <p>Step 1: Solve <math>f(x) = 0</math><br/> We start by solving the equation:</p> $\sin(7\pi \cdot \sin(5x)) = 0$ <p>This implies: <math>7\pi \cdot \sin(5x) = k\pi</math> for some integer <math>k</math><br/> Dividing both sides by <math>\pi</math>, we get: <math>\sin(5x) = \frac{k}{7}</math><br/> Now, since the sine function is bounded between <math>-1</math> and <math>1</math>, the values of <math>k</math> must satisfy:</p> $-1 \leq \frac{k}{7} \leq 1 \Rightarrow -7 \leq k \leq 7$ <p>So <math>k \in \{-7, -6, \dots, 6, 7\}</math>, a total of 15 values. For each such <math>k</math>, we solve:</p> $\sin(5x) = \frac{k}{7}$ <p>Let <math>\theta = 5x</math>, so <math>\theta \in (0, 10\pi)</math>, since <math>x \in (0, 2\pi)</math>. We now count the number of solutions to <math>\sin(\theta) = \frac{k}{7}</math> in <math>\theta \in (0, 10\pi)</math>.</p> <ul style="list-style-type: none"> <li>- For <math>k = \pm 7</math>, we have <math>\sin(\theta) = \pm 1</math>, which has <b>1</b> solution per <math>2\pi</math> interval. Since <math>\theta \in (0, 10\pi)</math>, which is 5 full periods of <math>2\pi</math>, there are <b>5</b> solutions for each of <math>k = \pm 7</math>.</li> <li>- For <math>k = 0</math>, we have <math>\sin(\theta) = 0</math>, which has <b>1</b> solution per <math>\pi</math> interval. Since <math>\theta \in (0, 10\pi)</math>, there are <b>10</b> intervals of <math>\pi</math>, but the endpoints <math>\theta = 0</math> and <math>\theta = 10\pi</math> are <b>excluded</b>, so we only count the <b>9</b> interior solutions.</li> <li>- For all other <math>k \in \{-6, \dots, 6\} \setminus \{0\}</math>, we have <math>\sin(\theta) = \frac{k}{7} \in (-1, 1) \setminus \{0, \pm 1\}</math>, which has <b>2</b> solutions per <math>2\pi</math> interval. Since <math>\theta \in (0, 10\pi)</math>, there are <b>5</b> intervals of <math>2\pi</math>, so <b>10</b> solutions per such <math>k</math>.</li> </ul> <p>Now, we count the total number of solutions:</p> <ul style="list-style-type: none"> <li>- <math>k = \pm 7</math>: 5 solutions each <math>\rightarrow 2 \times 5 = 10</math></li> <li>- <math>k = 0</math>: 9 solutions</li> <li>- <math>k = \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6</math>: 6 values of <math>k</math>, each with 10 solutions <math>\rightarrow 6 \times 10 = 60</math></li> </ul> <p>So total number of solutions:</p> $n = 10 + 9 + 60 = 139$ <p>Step 2: Determine the number of tangent points <math>t</math><br/> A point where the graph is <b>tangent</b> to the x-axis must satisfy both:</p> <ol style="list-style-type: none"> <li>1. <math>f(x) = 0</math></li> <li>2. <math>f'(x) = 0</math></li> </ol> <p>We compute the derivative:</p> $f(x) = \sin(7\pi \cdot \sin(5x)) \Rightarrow f'(x) = 35\pi \cos(7\pi \cdot \sin(5x)) \cdot \cos(5x)$ <p>So, <math>f'(x) = 0</math> when either:</p> <ul style="list-style-type: none"> <li>- <math>\cos(7\pi \cdot \sin(5x)) = 0</math>, or</li> <li>- <math>\cos(5x) = 0</math></li> </ul> <p>But from the condition <math>f(x) = 0</math>, we have:</p> $\sin(5x) = \frac{k}{7} \Rightarrow 7\pi \cdot \sin(5x) = k\pi \Rightarrow \cos(7\pi \cdot \sin(5x)) = \cos(k\pi) = (-1)^k$ <p>So <math>\cos(7\pi \cdot \sin(5x)) \neq 0</math>, unless <math>k = \pm 7</math>, in which case <math>\sin(5x) = \pm 1</math>, and <math>\cos(5x) = 0</math>. Thus, the only values of <math>k</math> for which <math>f(x) = 0</math> and <math>f'(x) = 0</math> are <math>k = \pm 7</math>, and for each of these, we have <b>5</b> solutions in <math>x \in (0, 2\pi)</math>.<br/> So: <math>t = 5 + 5 = 10</math><br/> Final Step: Compute <math>n + t</math></p> $n + t = 139 + 10 = \boxed{149}$ | <p>pass@1: 100%<br/> pass@8: 100%<br/> avg@8: 100%</p> |
|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------|

|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |                                                         |
|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------|
| Qwen3-8B | <p>To solve the problem, we are given the function:</p> $f(x) = \sin(7\pi \cdot \sin(5x))$ <p>We are to find the number of values of <math>x</math> in the interval <math>0 &lt; x &lt; 2\pi</math> for which <math>f(x) = 0</math>, and among these, how many are points where the graph of <math>f(x)</math> is tangent to the x-axis (i.e., where <math>f(x) = 0</math> and <math>f'(x) = 0</math>). Finally, we are to compute <math>n + t</math>, where <math>n</math> is the total number of zeros and <math>t</math> is the number of tangent points.</p> <p>Step 1: Find the number of solutions to <math>f(x) = 0</math><br/> We start by solving: <math>\sin(7\pi \cdot \sin(5x)) = 0</math>. This occurs when the argument of the sine function is an integer multiple of <math>\pi</math>: <math>7\pi \cdot \sin(5x) = k\pi</math> for some integer <math>k</math>. Dividing both sides by <math>\pi</math>: <math>\sin(5x) = \frac{k}{7}</math>. Since the sine function has a range of <math>[-1, 1]</math>, we require:</p> $-1 \leq \frac{k}{7} \leq 1 \Rightarrow -7 \leq k \leq 7$ <p>So <math>k \in -7, -6, \dots, 6, 7</math>, which gives 15 possible values of <math>k</math>. Now, for each <math>k</math>, we solve: <math>\sin(5x) = \frac{k}{7}</math>. Let <math>\theta = 5x</math>, so <math>\theta \in (0, 10\pi)</math>, since <math>x \in (0, 2\pi)</math>. The number of solutions to <math>\sin(\theta) = c</math> in <math>(0, 10\pi)</math> depends on the value of <math>c</math>:</p> <ul style="list-style-type: none"> <li>• If <math>c = \pm 1</math>, then <math>\sin(\theta) = \pm 1</math> has <b>5 solutions</b> in <math>(0, 10\pi)</math></li> <li>• If <math>c = 0</math>, then <math>\sin(\theta) = 0</math> has <b>9 solutions</b> in <math>(0, 10\pi)</math></li> <li>• If <math>c \in (-1, 1) \setminus 0</math>, then <math>\sin(\theta) = c</math> has <b>10 solutions</b> in <math>(0, 10\pi)</math></li> </ul> <p>Now, we count the number of solutions for each <math>k</math>:</p> <ul style="list-style-type: none"> <li>• <math>k = \pm 7</math>: <math>\sin(5x) = \pm 1</math> \u2192 5 solutions each \u2192 total: <math>2 \times 5 = 10</math></li> <li>• <math>k = 0</math>: <math>\sin(5x) = 0</math> \u2192 9 solutions</li> <li>• <math>-k = \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6</math>: 12 values, each with 10 solutions \u2192 <math>12 \times 10 = 120</math></li> </ul> <p>So the total number of solutions is:</p> $n = 10 + 9 + 120 = 139$ <p>Step 2: Find the number of tangent points <math>t</math>. A point is a <b>tangent point</b> to the x-axis if <math>f(x) = 0</math> and <math>f'(x) = 0</math>. We compute the derivative:</p> $f'(x) = \frac{d}{dx} [\sin(7\pi \cdot \sin(5x))] = \cos(7\pi \cdot \sin(5x)) \cdot 35\pi \cdot \cos(5x)$ <p>So: <math>f'(x) = 35\pi \cdot \cos(5x) \cdot \cos(7\pi \cdot \sin(5x))</math>. For <math>f'(x) = 0</math>, we need:</p> $\cos(5x) = 0 \quad \text{or} \quad \cos(7\pi \cdot \sin(5x)) = 0$ <p>But from earlier, we know that <math>f(x) = 0</math> implies <math>\sin(5x) = \frac{k}{7}</math>, so:</p> $7\pi \cdot \sin(5x) = k\pi \Rightarrow \cos(7\pi \cdot \sin(5x)) = \cos(k\pi) = (-1)^k$ <p>This is never zero, so the only way for <math>f'(x) = 0</math> is:</p> $\cos(5x) = 0$ <p>So we solve:</p> $\cos(5x) = 0 \Rightarrow 5x = \frac{\pi}{2} + n\pi \Rightarrow x = \frac{\pi}{10} + \frac{n\pi}{5}$ <p>We find all such <math>x \in (0, 2\pi)</math>:</p> $0 < \frac{\pi}{10} + \frac{n\pi}{5} < 2\pi \Rightarrow 0 < \frac{1+2n}{10} < 2 \Rightarrow -0.5 < n < 9.5$ <p>So <math>n = 0, 1, 2, \dots, 9</math>, giving <b>10</b> values of <math>x</math>. Now, we check whether these 10 values also satisfy <math>f(x) = 0</math>: From earlier, when <math>\cos(5x) = 0</math>, we have <math>\sin(5x) = \pm 1</math>, so:</p> $f(x) = \sin(7\pi \cdot \sin(5x)) = \sin(7\pi \cdot (\pm 1)) = \sin(\pm 7\pi) = 0$ <p>So all 10 values satisfy both <math>f(x) = 0</math> and <math>f'(x) = 0</math>, hence they are <b>tangent points</b>. Thus, <math>t = 10</math>.</p> <p>Final Step: Compute <math>n + t</math></p> $n + t = 139 + 10 = \boxed{149}$ | <p>pass@1: 100%<br/> pass@8: 100%<br/> avg@8: 87.5%</p> |
|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------|

|                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                                       |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------|
| Llama-3.1-8B-Instruct | <p>To find the values of <math>x</math> where <math>f(x) = \sin(7\pi \cdot \sin(5x)) = 0</math>, we need to find the values of <math>x</math> where <math>\sin(5x) = 0</math> or where <math>7\pi \cdot \sin(5x)</math> is an integer multiple of <math>\pi</math>. First, let's find the values of <math>x</math> where <math>\sin(5x) = 0</math>. This occurs when <math>5x = k\pi</math> for some integer <math>k</math>. Since <math>0 &lt; x &lt; 2\pi</math>, we have <math>0 &lt; 5x &lt; 10\pi</math>, so <math>0 &lt; k &lt; 20</math>. Therefore, there are 20 values of <math>x</math> where <math>\sin(5x) = 0</math>. Next, let's find the values of <math>x</math> where <math>7\pi \cdot \sin(5x)</math> is an integer multiple of <math>\pi</math>. This occurs when <math>\sin(5x) = \frac{m}{7}</math> for some integer <math>m</math> with <math>1 \leq m \leq 6</math>. Since <math>\sin(5x)</math> is between <math>-1</math> and <math>1</math>, we have <math>-\frac{7}{7} \leq \frac{m}{7} \leq \frac{7}{7}</math>, so <math>-1 \leq m \leq 1</math>. However, since <math>m</math> is an integer, we have <math>m = -1, 0, 1</math>. Therefore, there are 3 values of <math>x</math> where <math>7\pi \cdot \sin(5x)</math> is an integer multiple of <math>\pi</math>. However, we have counted the values of <math>x</math> where <math>\sin(5x) = 0</math> and <math>7\pi \cdot \sin(5x)</math> is an integer multiple of <math>\pi</math> twice, so we need to subtract these values once. There are 3 values of <math>x</math> where <math>\sin(5x) = 0</math> and <math>7\pi \cdot \sin(5x)</math> is an integer multiple of <math>\pi</math>. Therefore, there are <math>20 + 3 - 3 = 20</math> values of <math>x</math> where <math>f(x) = 0</math>. To find the values of <math>x</math> where the graph of <math>y = f(x)</math> is tangent to the <math>x</math>-axis, we need to find the values of <math>x</math> where <math>f'(x) = 0</math>. We have <math>f(x) = \sin(7\pi \cdot \sin(5x))</math>, so <math>f'(x) = 7\pi \cdot 5 \cos(7\pi \cdot \sin(5x)) \cos(5x)</math>. Therefore, <math>f'(x) = 0</math> when <math>\cos(7\pi \cdot \sin(5x)) = 0</math> or when <math>\cos(5x) = 0</math>. We have <math>\cos(7\pi \cdot \sin(5x)) = 0</math> when <math>7\pi \cdot \sin(5x) = \frac{(2k+1)\pi}{2}</math> for some integer <math>k</math>. This occurs when <math>\sin(5x) = \frac{2k+1}{14}</math> for some integer <math>k</math> with <math>0 \leq k \leq 13</math>. Since <math>\sin(5x)</math> is between <math>-1</math> and <math>1</math>, we have <math>-1 \leq \frac{2k+1}{14} \leq 1</math>, so <math>-7 \leq k \leq 6</math>. However, since <math>k</math> is an integer, we have <math>k = -7, -6, \dots, 6</math>. Therefore, there are 14 values of <math>x</math> where <math>\cos(7\pi \cdot \sin(5x)) = 0</math>. We have <math>\cos(5x) = 0</math> when <math>5x = \frac{(2k+1)\pi}{2}</math> for some integer <math>k</math>. Since <math>0 &lt; x &lt; 2\pi</math>, we have <math>0 &lt; 5x &lt; 10\pi</math>, so <math>0 &lt; k &lt; 20</math>. Therefore, there are 20 values of <math>x</math> where <math>\cos(5x) = 0</math>. However, we have counted the values of <math>x</math> where <math>\cos(7\pi \cdot \sin(5x)) = 0</math> and <math>\cos(5x) = 0</math> twice, so we need to subtract these values once. There are 4 values of <math>x</math> where <math>\cos(7\pi \cdot \sin(5x)) = 0</math> and <math>\cos(5x) = 0</math>. Therefore, there are <math>14 + 20 - 4 = 30</math> values of <math>x</math> where <math>f'(x) = 0</math>. Therefore, <math>n = 20</math> and <math>t = 30</math>, so <math>n + t = 20 + 30 = \boxed{50}</math>.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | pass@1: 0%<br>pass@8: 0%<br>avg@8: 0% |
| gemma-3-12b-it        | <p>Let <math>f(x) = \sin(7\pi \sin(5x))</math>. We want to find the number of values of <math>x</math> in the interval <math>0 &lt; x &lt; 2\pi</math> such that <math>f(x) = 0</math>. We have <math>\sin(7\pi \sin(5x)) = 0</math>. This means that <math>7\pi \sin(5x) = k\pi</math> for some integer <math>k</math>. Thus, <math>\sin(5x) = \frac{k}{7}</math> for some integer <math>k</math>. Since <math>-1 \leq \sin(5x) \leq 1</math>, we have <math>-1 \leq \frac{k}{7} \leq 1</math>, which means <math>-7 \leq k \leq 7</math>. So <math>k</math> can take values from <math>-7</math> to <math>7</math>, which is a total of <math>7 - (-7) + 1 = 15</math> values. For each value of <math>k</math>, we have <math>\sin(5x) = \frac{k}{7}</math>. Since <math>0 &lt; x &lt; 2\pi</math>, we have <math>0 &lt; 5x &lt; 10\pi</math>. For each value of <math>\frac{k}{7}</math>, there are two values of <math>5x</math> in the interval <math>(0, 2\pi)</math> except when <math>\frac{k}{7} = \pm 1</math>. If <math>\frac{k}{7} = 1</math>, then <math>\sin(5x) = 1</math>, so <math>5x = \frac{\pi}{2} + 2m\pi</math> for some integer <math>m</math>. Then <math>x = \frac{\pi}{10} + \frac{2m\pi}{5}</math>. Since <math>0 &lt; x &lt; 2\pi</math>, we have <math>0 &lt; \frac{\pi}{10} + \frac{2m\pi}{5} &lt; 2\pi</math>. <math>0 &lt; \frac{1}{10} + \frac{2m}{5} &lt; 2</math>, so <math>0 &lt; 1 + 4m &lt; 20</math>, which means <math>-1 &lt; 4m &lt; 19</math>, so <math>-\frac{1}{4} &lt; m &lt; \frac{19}{4} = 4.75</math>. Thus <math>m</math> can be <math>0, 1, 2, 3, 4</math>, so there are 5 values of <math>x</math>. If <math>\frac{k}{7} = -1</math>, then <math>\sin(5x) = -1</math>, so <math>5x = \frac{3\pi}{2} + 2m\pi</math> for some integer <math>m</math>. Then <math>x = \frac{3\pi}{10} + \frac{2m\pi}{5}</math>. Since <math>0 &lt; x &lt; 2\pi</math>, we have <math>0 &lt; \frac{3\pi}{10} + \frac{2m\pi}{5} &lt; 2\pi</math>. <math>0 &lt; \frac{3}{10} + \frac{2m}{5} &lt; 2</math>, so <math>0 &lt; 3 + 4m &lt; 20</math>, which means <math>-3 &lt; 4m &lt; 17</math>, so <math>-\frac{3}{4} &lt; m &lt; \frac{17}{4} = 4.25</math>. Thus <math>m</math> can be <math>0, 1, 2, 3, 4</math>, so there are 5 values of <math>x</math>. For each value of <math>k</math> between <math>-7</math> and <math>7</math>, we have two solutions for <math>5x</math> in the interval <math>(0, 10\pi)</math> except for <math>k = \pm 7</math>, where we have one solution. For <math>k = \pm 7</math>, we have <math>\sin(5x) = \pm 1</math>, so <math>5x = \frac{\pi}{2} + m\pi</math> or <math>5x = \frac{3\pi}{2} + m\pi</math>. For <math>k = 7</math>, <math>\sin(5x) = 1</math>, so <math>5x = \frac{\pi}{2} + 2m\pi</math>, <math>x = \frac{\pi}{10} + \frac{2m\pi}{5}</math>. For <math>k = -7</math>, <math>\sin(5x) = -1</math>, so <math>5x = \frac{3\pi}{2} + 2m\pi</math>, <math>x = \frac{3\pi}{10} + \frac{2m\pi}{5}</math>. For <math>k \in \{-6, -5, \dots, -1, 0, 1, \dots, 6\}</math>, we have two solutions for <math>5x</math> in <math>(0, 10\pi)</math>, so we have <math>2 \times 13 = 26</math> solutions. For <math>k = \pm 7</math>, we have one solution for <math>5x</math> in <math>(0, 10\pi)</math>, so we have <math>2 \times 1 = 2</math> solutions. Thus, the total number of solutions is <math>26 + 2 = 28</math>. However, we have <math>n = 28</math>. Now we want to find the number of values of <math>x</math> such that <math>f'(x) = 0</math>. <math>f'(x) = \cos(7\pi \sin(5x)) \cdot 7\pi \cdot \cos(5x) \cdot 5 = 35\pi \cos(7\pi \sin(5x)) \cos(5x)</math>. For <math>f'(x) = 0</math>, we have either <math>\cos(7\pi \sin(5x)) = 0</math> or <math>\cos(5x) = 0</math>. If <math>\cos(5x) = 0</math>, then <math>5x = \frac{\pi}{2} + m\pi</math>, so <math>x = \frac{\pi}{10} + \frac{m\pi}{5}</math>. Since <math>0 &lt; x &lt; 2\pi</math>, we have <math>0 &lt; \frac{\pi}{10} + \frac{m\pi}{5} &lt; 2\pi</math>, so <math>0 &lt; \frac{1}{10} + \frac{m}{5} &lt; 2</math>, so <math>0 &lt; 1 + 2m &lt; 20</math>, so <math>-1 &lt; 2m &lt; 19</math>, so <math>-\frac{1}{2} &lt; m &lt; \frac{19}{2} = 9.5</math>. Thus <math>m</math> can be <math>0, 1, 2, \dots, 9</math>, so there are 10 values of <math>x</math>. If <math>\cos(7\pi \sin(5x)) = 0</math>, then <math>7\pi \sin(5x) = \frac{\pi}{2} + m\pi</math>, so <math>\sin(5x) = \frac{1}{14} + \frac{m}{7}</math>. Since <math>-1 \leq \sin(5x) \leq 1</math>, we have <math>-1 \leq \frac{1}{14} + \frac{m}{7} \leq 1</math>, so <math>-1 \leq \frac{1}{14} + \frac{2m}{14} \leq 1</math>, so <math>-14 \leq 1 + 2m \leq 14</math>, so <math>-15 \leq 2m \leq 13</math>, so <math>-\frac{15}{2} \leq m \leq \frac{13}{2} = 6.5</math>. Thus <math>m</math> can be <math>-7, -6, \dots, 6</math>, so there are 14 values of <math>m</math>. For each value of <math>m</math>, we have two solutions for <math>5x</math> in <math>(0, 10\pi)</math>, so we have <math>14 \times 2 = 28</math> solutions. However, we want to find the number of values of <math>x</math> such that <math>f(x) = 0</math> and <math>f'(x) = 0</math>. We have <math>f(x) = 0</math> when <math>\sin(7\pi \sin(5x)) = 0</math>, so <math>7\pi \sin(5x) = k\pi</math>, so <math>\sin(5x) = \frac{k}{7}</math>. We have <math>f'(x) = 0</math> when <math>\cos(5x) = 0</math> or <math>\cos(7\pi \sin(5x)) = 0</math>. If <math>\cos(5x) = 0</math>, then <math>5x = \frac{\pi}{2} + m\pi</math>, so <math>x = \frac{\pi}{10} + \frac{m\pi}{5}</math>. We want <math>\sin(5x) = \frac{k}{7}</math>, so <math>\sin(\frac{\pi}{2} + m\pi) = \frac{k}{7}</math>. If <math>m</math> is even, then <math>\sin(\frac{\pi}{2}) = 1 = \frac{k}{7}</math>, so <math>k = 7</math>. If <math>m</math> is odd, then <math>\sin(\frac{3\pi}{2}) = -1 = \frac{k}{7}</math>, so <math>k = -7</math>. So we have <math>x = \frac{\pi}{10} + \frac{m\pi}{5}</math> where <math>m</math> is even or odd. If <math>m</math> is even, <math>m = 2j</math>, so <math>x = \frac{\pi}{10} + \frac{2j\pi}{5}</math>. If <math>m</math> is odd, <math>m = 2j + 1</math>, so <math>x = \frac{\pi}{10} + \frac{(2j+1)\pi}{5}</math>. We have <math>t = 10</math>. <math>n = 28</math>, <math>t = 10</math>. <math>n + t = 28 + 10 = 38</math>.<br/> Final Answer: The final answer is <math>\boxed{38}</math> </p> | pass@1: 0%<br>pass@8: 0%<br>avg@8: 0% |

gpt-oss-20b

<|channel|>analysis<|message|>We need to find number of solutions to  $\sin(7\pi \sin(5x)) = 0$  in  $(0, 2\pi)$ . That is  $\sin(7\pi \sin(5x)) = 0 \Rightarrow \text{argument } 7\pi \sin(5x) = k\pi$ , where  $k$  integer. So  $\sin(5x) = k/7$ . Since  $\sin(5x) \in [-1, 1]$ ,  $k$  must satisfy  $|k| \leq 7$ . Also  $k$  integer. So  $k = -7, -6, \dots, 6, 7$ . But  $\sin(5x) = k/7$ . For each  $k$ , we need solutions  $x$  in  $(0, 2\pi)$  to  $\sin(5x) = k/7$ . For each  $k$ , number of solutions?  $\sin(5x) = c$  has 10 solutions in  $(0, 2\pi)$  if  $c \in (-1, 1)$  because  $\sin(5x)$  has period  $2\pi/5$ . Wait  $\sin(5x)$  period =  $2\pi/5$ . Over interval  $(0, 2\pi)$ , number of periods = 5. In each period,  $\sin(\theta) = c$  has 2 solutions if  $c \in (-1, 1)$ . So total 10 solutions for each  $c$  in  $(-1, 1)$ . For  $c = \pm 1$ ,  $\sin(5x) = \pm 1$  has 5 solutions? Let's check:  $\sin(\theta) = 1$  occurs at  $\theta = \pi/2 + 2\pi n$ . Over  $0 < \theta < 10\pi$ ? Wait  $5x$  ranges from 0 to  $10\pi$ . So  $\sin(5x) = 1$  occurs at  $5x = \pi/2 + 2\pi n \Rightarrow x = \pi/10 + (2\pi/5)n$ . For  $n = 0, 1, 2, 3, 4$  gives 5 solutions. Similarly  $\sin(5x) = -1$  at  $5x = 3\pi/2 + 2\pi n \Rightarrow x = 3\pi/10 + (2\pi/5)n$ , also 5 solutions. So for  $k = \pm 7$ ,  $c = \pm 1$ , we have 5 solutions each. For  $k = \pm 6$ ,  $c = \pm 6/7$ , inside  $(-1, 1)$ , 10 solutions each. Similarly for  $k = \pm 5, \pm 4, \pm 3, \pm 2, \pm 1$ , 10 each. For  $k = 0$ ,  $c = 0$ ,  $\sin(5x) = 0$  has solutions at  $5x = n\pi \Rightarrow x = n\pi/5$ . For  $n = 1, 2, 3, 4, 5, 6, 7, 8, 9$ ? Wait  $0 < x < 2\pi \Rightarrow 0 < 5x < 10\pi \Rightarrow n = 1..9$  gives 9 solutions. But  $\sin(5x) = 0$  also at  $n = 0$ ?  $x = 0$  excluded.  $n = 10$  gives  $x = 2\pi$  excluded. So 9 solutions. Thus total  $n = \text{sum over } k \text{ of number of solutions}$ . Let's compute: For  $k = \pm 7$ :  $5 + 5 = 10$ . For  $k = \pm 6$ :  $10 + 10 = 20$ . For  $k = \pm 5$ : 20. For  $k = \pm 4$ : 20. For  $k = \pm 3$ : 20. For  $k = \pm 2$ : 20. For  $k = \pm 1$ : 20. For  $k = 0$ : 9. Sum:  $10 + 20 * 7 + 9 = 10 + 140 + 9 = 159$ . Wait  $20 * 7 = 140$ .  $10 + 140 + 9 = 159$ . So  $n = 159$ .

Now  $t$ : number of these  $n$  values where graph is tangent to x-axis. Tangent occurs when derivative  $f'(x) = 0$  at that root.  $f(x) = \sin(7\pi \sin(5x))$ .  $f'(x) = \cos(7\pi \sin(5x)) * 7\pi * \cos(5x) * 5$ ? Wait derivative:  $d/dx \sin(g(x)) = \cos(g(x)) * g'(x)$ .  $g(x) = 7\pi \sin(5x)$ .  $g'(x) = 7\pi * \cos(5x) * 5 = 35\pi \cos(5x)$ . So  $f'(x) = \cos(7\pi \sin(5x)) * 35\pi \cos(5x)$ . At root,  $\sin(7\pi \sin(5x)) = 0 \Rightarrow 7\pi \sin(5x) = k\pi \Rightarrow \sin(5x) = k/7$ . Then  $\cos(7\pi \sin(5x)) = \cos(k\pi) = (-1)^k$ . So  $f'(x) = 35\pi(-1)^k \cos(5x)$ . For tangent, we need  $f'(x) = 0$ . Since  $35\pi \neq 0$ , need  $\cos(5x) = 0$ . So tangent occurs when  $\sin(5x) = k/7$  and  $\cos(5x) = 0$  simultaneously. But  $\sin$  and  $\cos$  cannot both be zero. Wait  $\cos(5x) = 0$  implies  $\sin(5x) = \pm 1$ . So we need  $\sin(5x) = \pm 1$ . That corresponds to  $k = \pm 7$ . So only for  $k = \pm 7$ . For those, we also need  $\cos(5x) = 0$ . But  $\sin(5x) = \pm 1$  implies  $\cos(5x) = 0$  indeed. So all solutions with  $k = \pm 7$  are tangent? Let's check: For  $k = \pm 7$ ,  $\sin(5x) = \pm 1$ . At those points,  $\cos(5x) = 0$ . So  $f'(x) = 35\pi(-1)^k * 0 = 0$ . So yes tangent. So  $t$  equals number of solutions with  $k = \pm 7$ . We counted 5 each, total 10. So  $t = 10$ .

Thus  $n + t = 159 + 10 = 169$ .

Check if any other tangent? Could derivative zero also when  $\cos(7\pi \sin(5x)) = 0$ ? Wait  $f'(x) = \cos(7\pi \sin(5x)) * 35\pi \cos(5x)$ . For root,  $\sin(7\pi \sin(5x)) = 0$ . That implies  $7\pi \sin(5x) = k\pi$ . Then  $\cos(7\pi \sin(5x)) = \cos(k\pi) = (-1)^k$ . So not zero unless  $k$  is half-integer? But  $k$  integer, so  $\cos(k\pi) = \pm 1$ , not zero. So only zero from  $\cos(5x)$ . So indeed only  $k = \pm 7$ .

Thus answer 169. Let's double-check counts: For  $k = \pm 7$ , we had 5 solutions each. Let's verify:  $\sin(5x) = 1$  solutions:  $5x = \pi/2 + 2\pi n$ ,  $n = 0..4 \Rightarrow x = \pi/10 + (2\pi/5)n$ . That's 5 solutions.  $\sin(5x) = -1$ :  $5x = 3\pi/2 + 2\pi n$ ,  $n = 0..4 \Rightarrow x = 3\pi/10 + (2\pi/5)n$ . 5 solutions. So 10 total. Good.

Now check  $k = 0$ :  $\sin(5x) = 0$  solutions:  $5x = n\pi$ ,  $n = 1..9 \Rightarrow 9$  solutions. Good.

For  $k = \pm 1$ :  $\sin(5x) = \pm 1/7$ . For each, 10 solutions. So 20. Similarly for  $\pm 2.. \pm 6$ : each 20. So 7 pairs  $\Rightarrow 140$ . Add  $10 + 9 = 159$ . Good.

Thus answer 169.</end|><|start|>assistant<|channel|>final<|message|>

The equation  $[f(x) = \sin(7\pi \sin(5x)) = 0]$  is satisfied when

$$7\pi \sin(5x) = k\pi \quad (k \in \mathbb{Z})$$

i.e.

$$\sin(5x) = \frac{k}{7}, \quad |k| \leq 7.$$

For each integer  $k$  with  $|k| \leq 7$  we count the solutions of  $\sin(5x) = k/7$  in the interval  $0 < x < 2\pi$ .

### Counting the solutions

The function  $\sin(5x)$  has period  $\frac{2\pi}{5}$ .

Over  $0 < x < 2\pi$  it completes 5 periods.\* If  $|k| < 7$  (i.e.  $|k/7| < 1$ ), the equation  $\sin(5x) = k/7$  has

two solutions in each period, hence  $2 \times 5 = 10$  solutions.

\* If  $k = \pm 7$  (i.e.  $\sin(5x) = \pm 1$ ), the equation has one solution per period, hence 5 solutions for each sign.

\* For  $k = 0$  ( $\sin(5x) = 0$ ), the solutions are  $5x = n\pi$  with  $n = 1, 2, \dots, 9$  (excluding  $x = 0, 2\pi$ ), giving 9 solutions.

Summing:

$$k = \pm 7 : 5 + 5 = 10 \quad k = \pm 6, \pm 5, \pm 4, \pm 3, \pm 2, \pm 1 : 6 \times 20 = 120 \quad k = 0 : 9$$

$$n = 10 + 120 + 9 = 159.$$

### Tangent points

A root is a point of tangency to the  $x$ -axis when  $f'(x) = 0$ .

$$f'(x) = \cos(7\pi \sin(5x)) \cdot 35\pi \cos(5x).$$

pass@1: 0%  
pass@8: 100%  
avg@8: 62.5%

|                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                       |
|----------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------|
| gpt-oss-20b<br>(continued) | <p>At a root, <math>\sin(7\pi \sin(5x)) = 0 \Rightarrow 7\pi \sin(5x) = k\pi</math>, so<br/> <math>\cos(7\pi \sin(5x)) = \cos(k\pi) = (-1)^k \neq 0</math>.<br/> Thus <math>f'(x) = 0</math> only when <math>\cos(5x) = 0</math>.<br/> <math>\cos(5x) = 0</math> occurs exactly when <math>\sin(5x) = \pm 1</math>, i.e. for <math>k = \pm 7</math>.<br/> All solutions with <math>k = \pm 7</math> are therefore tangent points.<br/> Number of tangent points:</p> $t = 5 + 5 = 10.$ <p>—<br/> ### Result</p> $n + t = 159 + 10 = \boxed{169}.$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                       |
| Moonlight-16B-A3B-Instruct | <p>To solve the problem, we need to determine the number of values of <math>(x)</math> in the interval <math>(0 &lt; x &lt; 2\pi)</math> where <math>(f(x) = \sin(7\pi \cdot \sin(5x)) = 0)</math>, and then find how many of these values correspond to the graph being tangent to the <math>(x)</math>-axis.<br/> Step 1: Determine when <math>(f(x) = 0)</math><br/> The function <math>(f(x) = \sin(7\pi \cdot \sin(5x)))</math> is zero when its argument is an integer multiple of <math>(\pi)</math>:</p> $[7\pi \cdot \sin(5x) = k\pi \quad \text{for some integer } k.]$ <p>This simplifies to: <math>[\sin(5x) = \frac{k}{7}.]</math> Since <math>(\sin(5x))</math> must be in the interval <math>([-1, 1])</math>, <math>(k)</math> must be an integer such that: <math>[-7 \leq k \leq 7].</math> Thus, <math>(k)</math> can take any integer value from <math>(-7)</math> to <math>(7)</math>, inclusive. This gives us <math>(15)</math> possible values for <math>(k)</math>. Step 2: Determine when the graph is tangent to the <math>(x)</math>-axis<br/> The graph of <math>(y = f(x))</math> is tangent to the <math>(x)</math>-axis when <math>(f(x) = 0)</math> and <math>(f'(x) = 0)</math>. We need to find the derivative <math>(f'(x))</math> and solve for <math>(x)</math> where both <math>(f(x) = 0)</math> and <math>(f'(x) = 0)</math>.<br/> First, compute the derivative <math>(f'(x))</math>:</p> $[f(x) = \sin(7\pi \cdot \sin(5x)).]$ <p>Using the chain rule:</p> $[f'(x) = \cos(7\pi \cdot \sin(5x)) \cdot 7\pi \cdot \cos(5x) \cdot 5.]$ <p>Simplify:</p> $[f'(x) = 35\pi \cos(7\pi \cdot \sin(5x)) \cos(5x).]$ <p>For <math>(f(x) = 0)</math>, we have:</p> $[\sin(7\pi \cdot \sin(5x)) = 0 \implies 7\pi \cdot \sin(5x) = k\pi \implies \sin(5x) = \frac{k}{7}.]$ <p>For <math>(f'(x) = 0)</math>, we need:</p> $[\cos(7\pi \cdot \sin(5x)) = 0 \implies 7\pi \cdot \sin(5x) = \frac{\pi}{2} + n\pi \implies$ $\implies \sin(5x) = \frac{1}{14} + \frac{n}{7}.]$ <p>We need to find the values of <math>(x)</math> where both conditions are satisfied. This requires solving:</p> $[\sin(5x) = \frac{k}{7} \quad \text{and} \quad \sin(5x) = \frac{1}{14} + \frac{n}{7}.]$ <p>Step 3: Count the number of solutions<br/> For each <math>(k)</math> from <math>(-7)</math> to <math>(7)</math>, we need to check if there is a corresponding <math>(n)</math> such that:</p> $[\frac{k}{7} = \frac{1}{14} + \frac{n}{7}.]$ <p>Solving for <math>(n)</math>:</p> $[k = \frac{1}{2} + n.]$ <p>Since <math>(k)</math> must be an integer, <math>(\frac{1}{2} + n)</math> must also be an integer. This is only possible if <math>(n = -\frac{1}{2})</math>, which is not an integer. Therefore, there are no values of <math>(x)</math> where both conditions are satisfied simultaneously.<br/> Conclusion<br/> There are <math>(15)</math> values of <math>(x)</math> where <math>(f(x) = 0)</math>, and none of these values make the graph tangent to the <math>(x)</math>-axis. Therefore, <math>(n = 15)</math> and <math>(t = 0)</math>. Thus, the final answer is:</p> $[n + t = 15 + 0 = 15]$ | pass@1: 0%<br>pass@8: 0%<br>avg@8: 0% |



These examples highlight how model behavior varies under direct answer generation (Chat) versus structured reasoning (Chain-of-Thought) and how such differences impact prediction accuracy. Table 33 presents the input data sample with ID 12, while Table 34 reports the corresponding model responses for both `arc_challenge` settings.

| Problem                                                                                                                      | Choices                                  | Answer |
|------------------------------------------------------------------------------------------------------------------------------|------------------------------------------|--------|
| The morning temperature in a city is 41°F. If a sunny, mild day is forecast, which temperature is most likely for 2:00 p.m.? | A. 32°F<br>B. 41°F<br>C. 78°F<br>D. 98°F | C      |

Table 33: Sample 12 from the ARC Challenge benchmark. The Problem column reports the original question, the Choices column lists the candidate answer options, and the Answer column indicates the ground-truth label used for evaluation. Minor formatting adjustments have been applied to model responses for clarity and consistency of presentation.

Table 34: Representative model outputs on the `arc_challenge_chat` and `arc_challenge_cot_custom` evaluation settings. For each model, we report the raw generated response, the extracted (filtered) answer choice, and a correctness indicator. The Chat setting requires the model to produce a direct answer, typically a single option label, whereas the Chain-of-Thought (CoT) setting encourages intermediate reasoning steps prior to the final answer.

| Model                    | Setting | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Filtered | Correct |
|--------------------------|---------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|---------|
| EngGPT2-16B-A3B          | Chat    | D                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | D        | ✗       |
| EngGPT2-16B-A3B          | CoT     | <p>Step 1: Understand the temperature conversion<br/>The morning temperature is 41°F. Converting this to Celsius: <math>(41 - 32) \times \frac{5}{9} \approx 5^\circ\text{C}</math>.</p> <p>Step 2: Determine likely temperature range for a mild day<br/>A “mild day” implies a comfortable temperature range, not too hot or cold. Since 5°C is quite cold (<math>\approx 41^\circ\text{F}</math>), the afternoon temperature is more likely to be significantly warmer, avoiding freezing or extreme heat. Among the options, 78°F (25.5°C) is a typical mild outdoor temperature.</p> <p>The correct answer is C.</p> | C        | ✓       |
| FastwebMIIA-7B           | Chat    | D                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | D        | ✗       |
| FastwebMIIA-7B           | CoT     | D                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | D        | ✗       |
| Minerva-7B-instruct-v1.0 | Chat    | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | C        | ✓       |
| Minerva-7B-instruct-v1.0 | CoT     | D                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | D        | ✗       |
| Velvet-14B               | Chat    | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | C        | ✓       |
| Velvet-14B               | CoT     | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | C        | ✓       |
| LLaMAntino-3-ANITA-8B    | Chat    | B                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | B        | ✗       |
| LLaMAntino-3-ANITA-8B    | CoT     | <p>C “a more, not less, heat is expected so the temperature will likely increase” &amp; for a brief finish<br/>The correct answer is C.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | C        | ✓       |
| Qwen3-4B                 | Chat    | B                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | B        | ✗       |
| Qwen3-4B                 | CoT     | The correct answer is C.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | C        | ✓       |
| Qwen3-8B                 | Chat    | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | C        | ✓       |
| Qwen3-8B                 | CoT     | <p>The correct answer is C.<br/>The temperature typically rises during the day, especially with sunny conditions. A rise from 41°F in the morning to 78°F by 2:00 p.m. is reasonable. Other options suggest cooling or no change, which contradicts a sunny, mild forecast.<br/>The correct answer is C.</p>                                                                                                                                                                                                                                                                                                              | C        | ✓       |
| Llama-3.2-3B-Instruct    | Chat    | B                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | B        | ✗       |

Continued on next page

Continued on next page

Table 34 – continued from previous page

| Setting                           | Setting | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | Filtered | Correct |
|-----------------------------------|---------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|---------|
| Llama-3.2-3B-Instruct             | CoT     | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| Llama-3.1-8B-Instruct             | Chat    | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| Llama-3.1-8B-Instruct             | CoT     | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| gemma-3-4b-it                     | Chat    | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| gemma-3-4b-it                     | CoT     | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| gemma-3-12b-it                    | Chat    | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| gemma-3-12b-it                    | CoT     | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| deepseek-moe-16b-chat             | Chat    | B                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | B        | ✗       |
| deepseek-moe-16b-chat             | CoT     | B                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | B        | ✗       |
| gpt-oss-20b                       | Chat    | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| gpt-oss-20b                       | CoT     | < channel >analysis< message ><br>We have last problem: morning temp 41°F, sunny mild day forecast, which temp most likely for 2pm? Likely warmer, maybe around 78°F? The options: 32°F, 41°F, 78°F, 98°F. 98°F is extreme hot; 32°F same as morning; 78°F plausible increase. So best answer C. Provide minimal explanation: The temperature will rise as day progresses. Provide \"C\".<br>End with the required sentence.<br>< end >< start >assistant< channel >final< message >C | C        | ✓       |
| Moonlight-16B-A3B-Instruct        | Chat    | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| Moonlight-16B-A3B-Instruct        | CoT     | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| Ministral-3-8B-Instruct-2512-BF16 | Chat    | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| Ministral-3-8B-Instruct-2512-BF16 | CoT     | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| gpt5-nano                         | Chat    | Reasoning: A morning temperature of 41°F with a forecast for a sunny, mild day suggests warming by the afternoon, making 78°F the most plausible choice.<br>The best answer is C                                                                                                                                                                                                                                                                                                      | C        | ✓       |
| gpt5-nano                         | CoT     | C. 78°F<br>Reason: On a sunny mild day, temperatures typically rise from morning to afternoon, so 2:00 p.m. is likely around 78°F.<br>The correct answer is C                                                                                                                                                                                                                                                                                                                         | C        | ✓       |

## F.4 Example responses for GSM8K

In this section, we present representative examples from the GSM8K benchmark, specifically for the `gsm8k_llama` setting. Table 35 presents the input data sample with ID 2, while Table 36 reports the corresponding model responses.

| Problem                                                                                                                                                                              | Solution                                                                                                                                                                                                                                                                                       | Answer |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|
| Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did he make? | The cost of the house and repairs came out to $\$80,000 + \$50,000 = \$130,000$ . He increased the value of the house by $\$80,000 \times 1.5 = \$120,000$ . So the new value of the house is $\$120,000 + \$80,000 = \$200,000$ . So he made a profit of $\$200,000 - \$130,000 = \$70,000$ . | 70000  |

Table 35: Sample 2 from the `gsm8k_llama` benchmark. The Problem column contains the original question, the Solution column provides a step-by-step derivation of the result, and the Answer column indicates the final value used for evaluation against model outputs.

Table 36: Example responses on the `gsm8k_llama` benchmark under the `strict_match` evaluation setting. The Model column reports the evaluated models, the Answer column contains the full generated rationale and final response, while the Filtered Answer column extracts the filtered numerical answer used for evaluation. The Correct column indicates whether the extracted answer matches the ground-truth solution.

| Model                    | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Filtered | Correct |
|--------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|---------|
| EngGPT2-16B-A3B          | Josh bought the house for \$80,000 and spent \$50,000 on repairs, totaling a cost of \$130,000. The value increased by 150%, which is $1.5 \times \$80,000 = \$120,000$ . The new value of the house is $\$80,000 + \$120,000 = \$200,000$ . Profit is the selling price minus the total cost: $\$200,000 - \$130,000 = \$70,000$ . The final answer is 70000. The final answer is 70000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 70000.   | ✓       |
| FastwebMIAA-7B           | Josh bought the house for \$80,000 and then spent \$50,000 on repairs. So the total cost was $\$80,000 + \$50,000 = \$130,000$ . The house is now worth $\$130,000 \times 1.5 = \$195,000$ . So he made a profit of $\$195,000 - \$130,000 = \$65,000$ . The final answer is 65000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 65000.   | ✗       |
| Minerva-7B-instruct-v1.0 | Josh bought a house for \$80,000. He put in \$50,000 in repairs, which increased the value of the house by 150%. So the new value of the house is $80,000 + 50,000 = 130,000$ . The final answer is 130,000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 130,000. | ✗       |
| Velvet-14B               | The house was originally worth 80,000. A 150% increase means the house is now worth $80,000 + (80,000 \times 150/100) = 240,000$ dollars. He spent 50,000 dollars on repairs, so his profit is $240,000 - 80,000 - 50,000 = 110,000$ dollars. The final answer is 110,000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 110,000. | ✗       |
| LLaMAntino-3-ANITA-8B    | The house was worth 80000 dollars initially. After the repairs, the value of the house increased by 150%, so the new value is $80000 \times (3/2) = 120000$ . The cost of the house and the repairs is $80000 + 50000 = 130000$ . The profit is the new value minus the cost, so $120000 - 130000 = -10000$ , but the problem is asking for the positive profit, which is the opposite of the loss, so the answer is 10000 of loss in profit (made of debt, since it's a loss), and sold in the form of the word of 'negative of that which is sold in the form of the word of the term of the amount of the loss of the profit' as 'of the amount of the profit of the kind of that which is of the negative of the kind', that's 'of the positive of it', that's 10000, as 'of the answer of the kind of it', that's 'the answer of it of the kind', that's the form, that's 'the final answer is 10000 of the kind of the positive of the of the of the of it of the answer' but the form is not that, the form is 'The final answer is [answer]' where [answer] is the response to the problem, so the form is 'The final answer is 10000'. | 10000    | ✗       |
| Qwen3-4B                 | Josh's total cost is $\$80,000(\text{purchase}) + \$50,000(\text{repairs}) = \$130,000$ . The house's value increased by 150% of its original value ( $\$80,000$ ), so the increase is $\$80,000 \times 1.5 = \$120,000$ . The new value of the house is $\$80,000 + \$120,000 = \$200,000$ . Profit is $\$200,000 - \$130,000 = \$70,000$ . The final answer is 70000. The final answer is 70000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | 70000.   | ✓       |
| Qwen3-8B                 | Josh initially spent \$80,000 to buy the house and an additional \$50,000 on repairs, totaling \$130,000 in costs. The house's value increased by 150% of its original purchase price (\$80,000). A 150% increase means the new value is $\$80,000 + (1.5 \times \$80,000) = \$200,000$ . Profit is calculated as the final value minus total costs: $\$200,000 - \$130,000 = \$70,000$ . The final answer is 70000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 70000.   | ✓       |
| Llama-3.2-3B-Instruct    | Josh bought the house for \$80,000. He put in \$50,000 in repairs. So the total cost is $\$80,000 + \$50,000 = \$130,000$ . The value of the house increased by 150%. To find the new value, we need to calculate 150% of the original value. 150% of \$80,000 is $1.5 \times \$80,000 = \$120,000$ . The new value of the house is $\$80,000 + \$120,000 = \$200,000$ . To find the profit, we subtract the cost from the new value. $\$200,000 - \$130,000 = \$70,000$ . The final answer is 70000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | 70000.   | ✓       |
| Llama-3.1-8B-Instruct    | Josh bought the house for \$80,000 and put in \$50,000 in repairs. So the total cost is $80,000 + 50,000 = 130,000$ dollars. The repairs increased the value of the house by 150%. This means the new value of the house is 150% of the original value plus the original value. The original value is 80,000 dollars. 150% of 80,000 is $0.15 \times 80,000 = 12,000$ . So the new value is $80,000 + 12,000 = 92,000$ dollars. The profit is the new value minus the total cost. $92,000 - 130,000 = -38,000$ dollars. However, since Josh made a loss, we can say he made a negative profit of 38,000 dollars. The final answer is -38000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | -38000.  | ✗       |
| gemma-3-4b-it            | The cost of the house was \$80,000. The repairs cost \$50,000. So the total investment is $\$80,000 + \$50,000 = \$130,000$ . The value increased by 150%, which means the increase is 150% of the original value. Increase = $1.50 \times \$80,000 = \$120,000$ . The new value of the house is $\$80,000 + \$120,000 = \$200,000$ . Profit = New value - Total investment = $\$200,000 - \$130,000 = \$70,000$ . The final answer is 70000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 70000.   | ✓       |
| gemma-3-12b-it           | First, calculate the total cost: $\$80,000(\text{purchaseprice}) + \$50,000(\text{repairs}) = \$130,000$ . Next, calculate the increase in value: $\$50,000 \times 1.50 = \$75,000$ . Then, calculate the new value of the house: $\$80,000 + \$75,000 = \$155,000$ . Finally, calculate the profit: $\$155,000(\text{sellingprice}) - \$130,000(\text{totalcost}) = \$25,000$ . The final answer is 25000"                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | 25000    | ✗       |
| deepseek-moe-16b-chat    | Josh bought a house for \$80,000 and then put in \$50,000 in repairs. The total cost of the house is now $\$80,000 + \$50,000 = \$130,000$ . The value of the house increased by 150%, so the new value is $\$130,000 \times 150\% = \$195,000$ . The profit is $\$195,000 - \$130,000 = \$65,000$ . The final answer is 65,000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 65,000.  | ✗       |

Continued on next page

Table 36 – continued from previous page

| Model                             | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Filtered  | Correct |
|-----------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|---------|
| gpt-oss-20b                       | < channel >analysis< message >We need to parse the problem: Josh buys a house for \$80,000, then spends \$50,000 on repairs. This increased the value of the house by 150%. We need to find profit. Let's interpret: He buys for 80k, repairs cost 50k, total investment 130k. After repairs, the value increased by 150%. Does that mean the new value is 150% of the original purchase price? Or 150% increase relative to original? Usually "increased the value by 150%" means the value increased by 150% of the original value. So $newvalue = originalvalue + 150\%oforiginal = 1 + 1.5 = 2.5 \times original$ . So $newvalue = 2.5 \times 80k = 200k$ . Alternatively, could be 150% of the original purchase price? That would be $150\%of80k = 120k$ . But "increased the value by 150%" implies increase of 150% of original, so new value 200k. Let's compute profit: sale price minus total cost. $Profit = 200k - 130k = 70k$ . So answer 70,000. Let's double-check: If new value 200k, profit 70k. If new value 120k, profit -10k (loss). That seems unlikely. So answer 70k. Provide final answer. < end >< start >assistant< channel >final< message >Josh's total investment was the purchase price plus repairs: $(80,000 + 50,000 = 130,000)$ . The repairs increased the house's value by 150% of the original price. A 150% increase on the original $(80,000)$ is $(1.5 \times 80,000 = 120,000)$ . Adding that to the original price gives the new value: $(80,000 + 120,000 = 200,000)$ . Profit is the selling price minus the total cost: $(200,000 - 130,000 = 70,000)$ . The final answer is 70,000. | 70,000.   | ✓       |
| Moonlight-16B-A3B-Instruct        | The house was bought for \$80,000 and \$50,000 was put into repairs, so the total investment was \$80,000 + \$50,000 = \$130,000. The value of the house increased by 150%, so the new value is $130,000 \times 150\% = 130,000 \times 1.5 = 195,000$ . The profit is the new value minus the total investment, so $195,000 - 130,000 = 65,000$ . The final answer is \$65,000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | [invalid] | ✗       |
| Ministral-3-8B-Instruct-2512-BF16 | Let's break this down step-by-step: 1. <b>Initial Costs:</b> - Purchase price of the house: <b>**\$80,000**</b> - Repair costs: <b>**\$50,000**</b> - <b>**Total investment:**</b> $\$80,000 + \$50,000 = \$130,000$ 2. <b>Increase in Value:</b> - The house's value increased by <b>**150%**</b> of its <b>**total investment**</b> (\$130,000). - <b>**150%of\$130,000 = <math>1.5 \times \\$130,000 = \\$195,000</math>**</b> 3. <b>Final Value of the House:</b> - Original investment + increase = <b>**\$130,000 + \$195,000 = \$325,000**</b> 4. <b>Profit Calculation:</b> - <b>**Profit = Final value - Total investment**</b> - <b>**Profit = <math>\\$325,000 - \\$130,000 = \\$195,000</math>**</b> The final answer is <b>**195000**</b> .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | [invalid] | ✗       |
| GPT-5 nano                        | The house was bought for \$80,000 and \$50,000 was spent on repairs, for a total cost of \$130,000. The repairs increased the value by 150% of the original \$80,000, which is \$120,000, so the final value is \$200,000. $Profit = finalvalue(\$200,000) - totalcost(\$130,000) = \$70,000$ . The final answer is 70000.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 70000.    | ✓       |

## F.5 Example responses for HumanEval

In this section, we present representative examples from the HumanEval benchmark, focusing on the `humaneval_8_instruct_cot` setting. Table 37 shows the input data sample with ID 8, while Table 38 reports the corresponding model outputs. For brevity, Table 38 includes results only for the highest-performing models.

| Problem                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Solution                                                                                                                            | Test                                                                                                                                                                                                                                                                                                                        |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <pre>from typing import List, Tuple  def sum_product(numbers: List[int]) -&gt; Tuple[int, int]:     """ For a given list of integers,         return a tuple consisting of a         sum and a product of all the         integers in a list. Empty sum         should be equal to 0 and empty         product should be equal to 1.     """     &gt;&gt;&gt; sum_product([])     (0, 1)     &gt;&gt;&gt; sum_product([1, 2, 3, 4])     (10, 24)     """</pre> | <pre>sum_value = 0 prod_value = 1  for n in numbers:     sum_value += n     prod_value *= n  return sum_value,     prod_value</pre> | <pre>METADATA = {     'author': 'jt',     'dataset': 'test' }  def check(candidate):     assert candidate([]) == (0, 1)     assert candidate([1, 1, 1]) == (3, 1)     assert candidate([100, 0]) == (100, 0)     assert candidate([3, 5, 7]) == (3 + 5 +         7, 3 * 5 * 7)     assert candidate([10]) == (10, 10)</pre> |

Table 37: Sample 8 from the HumanEval benchmark. The Problem column reports the original function specification, including its docstring and example usages. The Solution column shows a reference implementation that satisfies the specification, while the Test column provides the corresponding unit tests used to automatically verify the correctness of model-generated solutions. Minor formatting adjustments have been applied to model responses for clarity and consistency of presentation.

Table 38: Example responses on the humaneval\_8\_instruct\_cot benchmark. The Model column identifies the evaluated models, the Filtered Answer column reports the generated code solutions after regex filtering, and the Result column provides the corresponding evaluation metrics. The reported filtered answers correspond to first-pass outputs (pass@1) for space constraints and to improve readability.

| Model           | Filtered Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Result                                                               |
|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------|
| EngGPT2-16B-A3B | <pre> from typing import List, Tuple  def sum_product(numbers: List[int]) -&gt; Tuple[int, int]:     """ For a given list of integers, return a tuple consisting of a         sum and a product of all the integers in a list. Empty sum         should be equal to 0 and empty product should be equal to 1.     """     &gt;&gt;&gt; sum_product([])     (0, 1)     &gt;&gt;&gt; sum_product([1, 2, 3, 4])     (10, 24)     """     total_sum = sum(numbers)     product = 1     for num in numbers:         product *= num     return (total_sum, product) </pre>                                                            | <pre> pass@1: 62.50% pass@2: 89.29% pass@4: 100% pass@8: 100% </pre> |
| Qwen3-4B        | <pre> from typing import List, Tuple  def sum_product(numbers: List[int]) -&gt; Tuple[int, int]:     """ For a given list of integers, return a tuple consisting of a         sum and a product of all the integers in a list. Empty sum         should be equal to 0 and empty product should be equal to 1.     """     &gt;&gt;&gt; sum_product([])     (0, 1)     &gt;&gt;&gt; sum_product([1, 2, 3, 4])     (10, 24)     """     if not numbers:         return (0, 1)     total_sum = sum(numbers)    total_product = 1     for num in numbers:         total_product *= num     return (total_sum, total_product) </pre> | <pre> pass@1: 100%, pass@2: 100% pass@4: 100% pass@8: 100% </pre>    |
| Qwen3-8B        | <pre> from typing import List, Tuple  def sum_product(numbers: List[int]) -&gt; Tuple[int, int]:     """ For a given list of integers, return a tuple consisting of a         sum and a product of all the integers in a list. Empty sum         should be equal to 0 and empty product should be equal to 1.     """     &gt;&gt;&gt; sum_product([])     (0, 1)     &gt;&gt;&gt; sum_product([1, 2, 3, 4])     (10, 24)     """     if not numbers:         return (0, 1)     total_sum = sum(numbers)     product = 1     for num in numbers:         product *= num     return (total_sum, product) </pre>                  | <pre> pass@1: 75%, pass@2: 96.43% pass@4: 100% pass@8: 100% </pre>   |
| gemma-3-4b-it   | <pre> from typing import List, Tuple  def sum_product(numbers: List[int]) -&gt; Tuple[int, int]:     """ For a given list of integers, return a tuple consisting of a         sum and a product of all the integers in a list. Empty sum         should be equal to 0 and empty product should be equal to 1.     """     sum_product([])     (0, 1)     sum_product([1, 2, 3, 4])     (10, 24)     """     sum_numbers = 0     product_numbers = 1      for number in numbers:         sum_numbers += number         product_numbers *= number     return (sum_numbers, product_numbers) </pre>                                | <pre> pass@1: 75%, pass@2: 96.43% pass@4: 100% pass@8: 100% </pre>   |

Continued on next page

Table 38 – continued from previous page

| Model                      | Filtered Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Result                                                           |
|----------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------|
| gemma-3-12b-it             | <pre> from typing import List, Tuple  def sum_product(numbers: List[int]) -&gt; Tuple[int, int]:     """ For a given list of integers, return a tuple consisting of a         sum and a product of all the integers in a list. Empty sum         should be equal to 0 and empty product should be equal to 1.     """     sum_product([])     (0, 1)     sum_product([1, 2, 3, 4])     (10, 24)     """     sum_numbers = 0     product_numbers = 1      for number in numbers:         sum_numbers += number         product_numbers *= number      return (sum_numbers, product_numbers) </pre>                                                                                                                                                                                                                                                                                                                                                                                                  | pass@1: 100%,<br>pass@2: 100%,<br>pass@4: 100%,<br>pass@8: 100%  |
| gpt-oss-20b                | <pre> from typing import List, Tuple  def sum_product(numbers: List[int]) -&gt; Tuple[int, int]:     """ For a given list of integers, return a tuple consisting of a         sum and a product of all the integers in a list. Empty sum         should be equal to 0 and empty product should be equal to 1.     """     &gt;&gt;&gt; sum_product([])     (0, 1)     &gt;&gt;&gt; sum_product([1, 2, 3, 4])     (10, 24)     """     # The sum of an empty list is 0 by definition of Python's sum()     total_sum = sum(numbers)      # Compute product manually, starting with 1 (empty product     # convention)     total_product = 1     for number in numbers:         total_product *= number     return total_sum, total_product </pre>                                                                                                                                                                                                                                                   | pass@1: 100%,<br>pass@2: 100%,<br>pass@4: 100%,<br>pass@8: 100%  |
| Moonlight-16B-A3B-Instruct | <pre> from typing import List, Tuple  def sum_product(numbers: List[int]) -&gt; Tuple[int, int]:     """ For a given list of integers, return a tuple consisting of a         sum and a product of all the integers in a list. Empty sum         should be equal to 0 and empty product should be equal to 1.     """     &gt;&gt;&gt; sum_product([])     (0, 1)     &gt;&gt;&gt; sum_product([1, 2, 3, 4])     (10, 24)     """     # Initialize sum and product variables to handle edge cases of     # empty list     total_sum = 0     total_product = 1     # Iterate through the list of numbers to calculate sum and product     for number in numbers:         total_sum += number         total_product *= number     # Return the computed sum and product as a tuple     return total_sum, total_product  # Example usage and test cases if __name__ == "__main__":     print(sum_product([]))     # Output: (0, 1)     print(sum_product([1, 2, 3, 4]))     # Output: (10, 24) </pre> | pass@1: 87.5%,<br>pass@2: 100%,<br>pass@4: 100%,<br>pass@8: 100% |

Continued on next page

Table 38 – continued from previous page

| Model                             | Filtered Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Result                                                              |
|-----------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------|
| Ministral-3-8B-Instruct-2512-BF16 | <pre> from typing import List, Tuple  def sum_product(numbers: List[int]) -&gt; Tuple[int, int]:     """ For a given list of integers, return a tuple consisting of a         sum and a product of all the integers in a list. Empty sum         should be equal to 0 and empty product should be equal to 1.     """     sum_product([])     (0, 1)     sum_product([1, 2, 3, 4])     (10, 24)     """     total_sum = 0     total_product = 1      for num in numbers:         total_sum += num         total_product *= num      return (total_sum, total_product) </pre>                                | pass@1: 62.50%,<br>pass@2: 89.29%,<br>pass@4: 100%,<br>pass@8: 100% |
| gpt5-nano                         | <pre> from typing import List, Tuple  def sum_product(numbers: List[int]) -&gt; Tuple[int, int]:     """ For a given list of integers, return a tuple consisting of a         sum and a product of all the integers in a list. Empty sum         should be equal to 0 and empty product should be equal to 1.     """     &gt;&gt;&gt; sum_product([])     (0, 1)     &gt;&gt;&gt; sum_product([1, 2, 3, 4])     (10, 24)     """     total_sum = 0     total_product = 1     for number in numbers:         total_sum += number         total_product *= number     return total_sum, total_product </pre> | pass@1: 100%,<br>pass@2: 100%,<br>pass@4: 100%,<br>pass@8: 100%     |

## F.6 Example responses for MMLU

In this section, we present representative examples from the MMLU benchmark, specifically for the `mmlu_cot_llama` setting. Table 39 presents the input data sample with ID 76, while Table 40 reports the corresponding model responses.

| Problem                                                                                                                                                                                                                                                                                                                                                                                                      | Choices                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | Answer |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|
| <p>In the procedure <code>Mystery</code> below, the parameter <code>number</code> is a positive integer.</p> <pre> PROCEDURE Mystery (number) {   REPEAT UNTIL (number &lt;= 0) {     number \u2190 number - 2   }   IF (number = 0) {     RETURN (true)   }   ELSE {     RETURN (false)   } } </pre> <p>Which of the following best describes the result of running the procedure <code>Mystery</code>?</p> | <p>A. The procedure returns true when the initial value of <code>number</code> is 2, and it otherwise returns false.</p> <p>B. The procedure returns true when the initial value of <code>number</code> is greater than 2, and it otherwise returns false.</p> <p>C. The procedure returns true when the initial value of <code>number</code> is even, and it otherwise returns false.</p> <p>D. The procedure returns true when the initial value of <code>number</code> is odd, and it otherwise returns false.</p> | C      |

Table 39: Sample 76 from the MMLU benchmark. The Problem column contains the original question. The Choices column lists the possible answer options provided to the model, and the Answer column reports the correct option used for evaluation against model outputs.

Table 40: Example model responses for sample 76 from the `mmlu_cot_llama` dataset (flexible\_extract setting). The table compares the raw outputs, the corresponding filtered answers, and their correctness. NOTE: the response generated by LLaMAntino-3-ANITA-8B-Inst-DPO-ITA is truncated due to repetitive token generation, preventing it from producing a valid final answer and resulting in an incorrect prediction. Minor formatting adjustments have been applied to model responses for clarity and consistency of presentation.

| Model                              | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | Filtered    | Correct |
|------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|---------|
| EngGPT2-16B-A3B                    | <p>## Step 1: Analyze the loop behavior<br/>The procedure repeatedly subtracts 2 from the number until it is less than or equal to 0. This continues until the number is non-positive.</p> <p>## Step 2: Determine the final value of the number<br/>- For even numbers: The number will eventually reach 0 after repeated subtraction of 2 (e.g., 2 \u2192 0, 4 \u2192 0).<br/>- For odd numbers: The number will eventually become negative (e.g., 3 \u2192 -1, 5 \u2192 -1), which does not equal 0.</p> <p>The procedure returns true only if the final value is 0 (i.e., the initial number was even) and false otherwise.<br/>The best answer is C.</p>                                                                                                                                                                                                                                                                                                       | C           | ✓       |
| FastwebMIA-7B                      | <p>## Step 1: The procedure Mystery is a loop that repeatedly subtracts 2 from the input number until it becomes less than or equal to 0.</p> <p>## Step 2: If the number becomes 0, the procedure returns true. Otherwise, it returns false.<br/>The best answer is A.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | A           | ✗       |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | <p>Let's think step by step for this problem.</p> <p>## Step 1: Understand the procedure<br/>The procedure keeps subtracting 2 from the input number until the number is less than 1.<br/>Then it checks the final value of the number.</p> <p>## Step 2: Analyze the cases of the input<br/>If the initial number is even and greater than 2, the procedure will return false (not true after a even number less than of 2 in a sub of two each time of 'not of less of 1') of a of of of of of of of of of of ...</p>                                                                                                                                                                                                                                                                                                                                                                                                                                             | "[invalid]" | ✗       |
| Qwen3-4B                           | The best answer is C.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | C           | ✓       |
| Qwen3-8B                           | <p>## Step 1: Analyze the loop behavior<br/>The loop subtracts 2 from the number repeatedly until it becomes less than or equal to 0. This effectively reduces the number by 2 until it reaches 0 or a negative value.</p> <p>## Step 2: Determine the condition for returning true<br/>After the loop, the procedure checks if the number is 0. If it is, it returns true. This happens only when the initial number is even, as subtracting 2 repeatedly from an even number will eventually reach 0. For odd numbers, the loop stops when the number becomes 1 (which is not \u2264 0), and the procedure returns false.<br/>The best answer is C.</p>                                                                                                                                                                                                                                                                                                           | C           | ✓       |
| Llama-3.2-3B-Instruct              | <p>## Step 1: Understand the procedure Mystery<br/>The procedure Mystery takes a positive integer as input and repeatedly subtracts 2 from it until it reaches 0 or less.</p> <p>## Step 2: Analyze the condition for returning true<br/>The procedure returns true when the input number becomes 0, which means the initial value of the number must be even, as subtracting 2 from an even number will always result in an even number.</p> <p>## Step 3: Analyze the condition for returning false<br/>The procedure returns false when the input number does not become 0, which means the initial value of the number must be odd, as subtracting 2 from an odd number will always result in an odd number.</p> <p>## Step 4: Determine the correct answer<br/>Based on the analysis, the procedure returns true when the initial value of the number is even, and it returns false when the initial value of the number is odd.<br/>The best answer is C.</p> | C           | ✓       |

Continued on next page

Table 40 – continued from previous page

| Model                 | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Filtered | Correct |
|-----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|---------|
| Llama-3.1-8B-Instruct | <p>## Step 1: Analyze the procedure Mystery<br/>The procedure Mystery takes a positive integer as input and decrements it by 2 in a loop until it reaches 0 or less.</p> <p>## Step 2: Determine the condition for the procedure to return true<br/>The procedure returns true when the input number is a multiple of 2 (i.e., even), because in this case, the loop will decrement the number to 0, and the procedure will return true.</p> <p>## Step 3: Determine the condition for the procedure to return false<br/>The procedure returns false when the input number is not a multiple of 2 (i.e., odd), because in this case, the loop will decrement the number to an odd number, and the procedure will return false.</p> <p>## Step 4: Consider the edge case where the initial value of number is 2<br/>When the initial value of number is 2, the procedure will decrement it to 0, and return true.</p> <p>## Step 5: Consider the edge case where the initial value of number is greater than 2<br/>When the initial value of number is greater than 2, the procedure will decrement it to an odd number, and return false.</p> <p>## Step 6: Consider the edge case where the initial value of number is even and greater than 2<br/>When the initial value of number is even and greater than 2, the procedure will decrement it to an even number, and then to 0, and return true.</p> <p>## Step 7: Consider the edge case where the initial value of number is odd<br/>When the initial value of number is odd, the procedure will decrement it to an even number, and then to 0, and return true.</p> <p>## Step 8: Analyze the results from the previous steps<br/>From the previous steps, we can see that the procedure returns true when the initial value of number is even, and it otherwise returns false.</p> <p>The best answer is C.</p> | C        | ✓       |
| gemma-3-4b-it         | <p>## Step 1: Analyze the procedure's logic.<br/>The procedure repeatedly subtracts 2 from the input 'number' until 'number' becomes less than or equal to 0.</p> <p>## Step 2: Trace the procedure with different initial values.<br/>- If 'number' is 2: <math>2 - 2 = 0</math>. The loop terminates.<br/>- If 'number' is 3: <math>3 - 2 = 1</math>. The loop terminates because <math>1 \leq 0</math>.<br/>- If 'number' is 4: <math>4 - 2 = 2</math>. The loop terminates because <math>2 \leq 0</math>.<br/>The procedure returns 'true' only when the initial value of 'number' results in 'number' becoming 0 after the repeated subtraction.<br/>This happens when the initial value is even.<br/>The best answer is C.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | C        | ✓       |
| gemma-3-12b-it        | <p>## Step 1: Trace the execution with different inputs<br/>Let's trace the execution of the procedure with different initial values of 'number'.<br/>If number = 2: REPEAT UNTIL (2 &lt;= 0) is false. number = 2 - 2 = 0. IF (number = 0) is true. RETURN true.<br/>If number = 3: REPEAT UNTIL (3 &lt;= 0) is false. number = 3 - 2 = 1. REPEAT UNTIL (1 &lt;= 0) is false. number = 1 - 2 = -1. REPEAT UNTIL (-1 &lt;= 0) is true. IF (number = 0) is false. RETURN false.<br/>If number = 4: REPEAT UNTIL (4 &lt;= 0) is false. number = 4 - 2 = 2. REPEAT UNTIL (2 &lt;= 0) is false. number = 2 - 2 = 0. IF (number = 0) is true. RETURN true.<br/>If number = 5: REPEAT UNTIL (5 &lt;= 0) is false. number = 5 - 2 = 3. REPEAT UNTIL (3 &lt;= 0) is false. number = 3 - 2 = 1. REPEAT UNTIL (1 &lt;= 0) is false. number = 1 - 2 = -1. REPEAT UNTIL (-1 &lt;= 0) is true. IF (number = 0) is false. RETURN false.</p> <p>## Step 2: Identify the pattern<br/>The procedure returns true if, after repeatedly subtracting 2 from the initial 'number' until it becomes less than or equal to 0, the final value of 'number' is 0. This happens when the initial 'number' is even. If the initial 'number' is odd, the loop terminates with 'number' being negative, and the procedure returns false.</p> <p>The best answer is C.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | C        | ✓       |

Continued on next page

Table 40 – continued from previous page

| Model                      | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | Filtered | Correct |
|----------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|---------|
| deepseek-moe-16b-chat      | <p>## Step 1: The procedure Mystery is defined to repeat until the number is less than or equal to 0.</p> <p>## Step 2: In each iteration of the REPEAT UNTIL loop, the number is decreased by 2.</p> <p>## Step 3: If the number becomes 0, the procedure returns true.</p> <p>## Step 4: If the number is not 0, the procedure returns false.</p> <p>Now, let's analyze the given options:</p> <p>A. The procedure returns true when the initial value of number is 2, and it otherwise returns false.</p> <p>B. The procedure returns true when the initial value of number is greater than 2, and it otherwise returns false.</p> <p>C. The procedure returns true when the initial value of number is even, and it otherwise returns false.</p> <p>D. The procedure returns true when the initial value of number is odd, and it otherwise returns false.</p> <p>None of these options directly match the behavior of the procedure. However, we can see that the procedure returns true when the number is 0, and false otherwise. This means that the procedure returns true when the initial value of number is even, and false otherwise.</p> <p>Therefore, the best answer is:</p> <p>The best answer is C.</p>                                                                                                                                                                                                                                | C        | ✓       |
| gpt-oss-20b                | <p>&lt; channel &gt;analysis&lt; message &gt;We need to analyze the procedure. The procedure takes a positive integer number. It repeats until number &lt;= 0, subtracting 2 each time. So it subtracts 2 repeatedly until number &lt;= 0. After loop, if number == 0 return true else false. So essentially, if the initial number is even, subtracting 2 repeatedly will eventually reach 0. If odd, subtracting 2 repeatedly will eventually reach -1 (since odd minus 2 repeatedly will go to 1, -1). So returns true for even numbers, false for odd. So answer C. Let's produce final.&lt; end &gt;&lt; start &gt;assistant&lt; channel &gt;final&lt; message &gt;</p> <p>The loop subtracts 2 until the value is no longer positive.</p> <ul style="list-style-type: none"> <li>- If the starting number is even, the sequence ends exactly at 0, so the function returns <b>true</b>.</li> <li>- If the starting number is odd, the sequence ends at -1, so the function returns <b>false</b>.</li> </ul> <p>Thus the procedure returns true precisely for even initial values.</p> <p>The best answer is <b>C</b>.</p>                                                                                                                                                                                                                                                                                                                          | C        | ✓       |
| Moonlight-16B-A3B-Instruct | <p>## Step 1: Understand the Procedure</p> <p>The procedure 'Mystery' takes a positive integer 'number' as input. It repeatedly subtracts 2 from 'number' until 'number' is less than or equal to 0. If 'number' is exactly 0 after this process, it returns 'true'; otherwise, it returns 'false'.</p> <p>## Step 2: Analyze the Condition</p> <p>The key part of the procedure is the 'REPEAT UNTIL' loop:</p> <ul style="list-style-type: none"> <li>- The loop continues as long as 'number' is greater than 0.</li> <li>- Inside the loop, 'number' is decreased by 2 each iteration.</li> </ul> <p>## Step 3: Determine the Final Value of 'number'</p> <ul style="list-style-type: none"> <li>- If the initial value of 'number' is even, it will be reduced to 0 after a certain number of iterations.</li> <li>- If the initial value of 'number' is odd, it will become negative after the first iteration and thus will not satisfy the condition 'number &lt;= 0'.</li> </ul> <p>## Step 4: Evaluate the Return Statements</p> <ul style="list-style-type: none"> <li>- If 'number' is 0 after the loop, the procedure returns 'true'.</li> <li>- If 'number' is not 0 after the loop, the procedure returns 'false'.</li> </ul> <p>## Step 5: Conclusion</p> <p>The procedure returns 'true' when the initial value of 'number' is even, and it returns 'false' when the initial value of 'number' is odd.</p> <p>The best answer is C.</p> | C        | ✓       |

Continued on next page

Table 40 – continued from previous page

| Model                             | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Filtered | Correct |
|-----------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|---------|
| Ministral-3-8B-Instruct-2512-BF16 | <pre> ## Step 1: Understand the loop behavior     The loop 'REPEAT UNTIL (number &lt;= 0)' subtracts 2 from 'number'     repeatedly until 'number' becomes less than or equal to 0.     This means the loop continues as long as 'number &gt; 0'. ## Step 2: Determine the final value of 'number'     Since 2 is subtracted repeatedly, the final value of 'number' after     the loop depends on the initial value modulo 2:     - If the initial 'number' is even, subtracting 2 repeatedly       will eventually reach exactly 0 (e.g., 4 \u2192 2 \u2192       0).     - If the initial 'number' is odd, subtracting 2 repeatedly will       reach 1 (e.g., 3 \u2192 1 \u2192 loop stops because 1 &gt;       0). ## Step 3: Check the return condition     The procedure returns 'true' only if 'number == 0' after the loop.     This happens <b>exactly</b> when the initial 'number' is even.     The best answer is <b>C</b>. </pre> | C        | ✓       |
| GPT-5 nano                        | <pre> It returns true iff the starting number is even (0 after the loop for even numbers; negative for odd numbers). So the correct choice is C. The best answer is C. </pre>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | C        | ✓       |

## F.7 Example responses for ITALIC

This section presents a comparison between Fast and Chain-of-Thought (CoT) prompting strategies, both in the original setting and in the custom *italic\_fast* and *italic\_cot\_custom* settings. The Fast setting constrains the model to produce a direct answer, while the CoT setting encourages intermediate reasoning before producing the final response. Table 43 reports the corresponding model outputs for all Italian models (both native and fine-tuned) as well as the best-performing international models, highlighting the impact of the prompting strategy on response quality. Table 41 and 42 present the input data sample with ID 2 for both Fast and CoT original and custom settings, respectively.

| Setting | Question                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fast    | <pre> "messages": [   {     "role": "system",     "content": "Sei un assistente utile."   },   ...   {     "role": "user",     "content": "Rispondi alla seguente domanda a scelta multipla sull'argomento '     literature'. La tua risposta deve essere nel seguente formato: 'LETTERA'     (senza virgolette) dove LETTERA é una tra ABCD. Come si chiamava il     periodico, pubblicato dal 1764 al 1766, principale strumento di     diffusione del pensiero illuminista in Italia?     A) Athenaeum     B) L'Asino     C) Il Caffé     D) L'Arcadia     Risposta:"   } ] </pre>                                                          |
| CoT     | <pre> "messages": [   {     "role": "system",     "content": "Sei un assistente utile."   },   ...   {     "role": "user",     "content": "Rispondi alla seguente domanda a scelta multipla sull'argomento '     literature'. L'ultima riga della tua risposta deve essere nel seguente     formato: 'Risposta: LETTERA' (senza virgolette) dove LETTERA é una tra     ABCD. Ragiona brevemente prima di rispondere. Come si chiamava il     periodico, pubblicato dal 1764 al 1766, principale strumento di     diffusione del pensiero illuminista in Italia?     A) Athenaeum     B) L'Asino     C) Il Caffé     D) L'Arcadia"   } ] </pre> |

Table 41: Sample id 2 from the ITALIC benchmark: comparison of Fast vs CoT settings in the original version of the task. The Setting column indicates the prompting strategy used (e.g., fast prompting or Chain-of-Thought prompting), while the Question column contains the actual prompt provided to the model. NOTE: the content shown in the question column includes only the last part of the full prompt, removing any few-shot examples, so as to report only the final question asked to the models.

| Setting | Question                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fast    | <p>Rispondi alla seguente domanda a scelta multipla sull'argomento 'synonyms_and_antonyms'. La tua risposta deve essere nel seguente formato: 'LETTERA' (senza virgolette) dove LETTERA \u00e8 una tra A, B, C o D. Scrivi solo la lettera corrispondente alla tua risposta senza spiegazioni.</p> <p>...</p> <p>Come si chiamava il periodico, pubblicato dal 1764 al 1766, principale strumento di diffusione del pensiero illuminista in Italia?</p> <p>A. Athenaeum<br/> B. L'Asino<br/> C. Il Caff\u00e8<br/> D. L'Arcadia</p> <p>Risposta:</p>                                                                                                                                                                                                                                                                                                                         |
| CoT     | <p>Data la seguente domanda e 4 possibili risposte (A, B, C e D), scegli la risposta migliore.</p> <p>- Per problemi semplici:<br/> Fornisci direttamente la risposta con una spiegazione minima.</p> <p>- Per problemi complessi:\nUsa questo formato passo dopo passo:<br/> ## Passo 1: [Descrizione concisa]<br/> [Breve spiegazione]<br/> ## Passo 2: [Descrizione concisa]<br/> [Breve spiegazione]<br/> Indipendentemente dall'approccio, concludi sempre con:<br/> La risposta corretta \u00e8 [lettera_della_risposta].<br/> dove la [lettera_della_risposta] \u00e8 una tra A, B, C o D.<br/> Pensa passo dopo passo.</p> <p>...</p> <p>Domanda: Come si chiamava il periodico, pubblicato dal 1764 al 1766, principale strumento di diffusione del pensiero illuminista in Italia?</p> <p>A. Athenaeum<br/> B. L'Asino<br/> C. Il Caff\u00e8<br/> D. L'Arcadia</p> |

Table 42: Sample id 2 from the ITALIC benchmark: comparison of Fast (taken from *italic\_fast*) vs CoT (taken from *italic\_cot\_custom*) custom settings. The Setting column indicates the prompting strategy used (e.g., fast prompting or Chain-of-Thought prompting), while the Question column contains the actual prompt provided to the model. NOTE: the content shown in the question column includes only the last part of the full prompt, removing any few-shot examples, so as to report only the final question asked to the models.

Table 43: Example responses on an ITALIC benchmark sample comparing FAST and CoT settings.

| Model                    | Setting       | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Filtered    | Correct |
|--------------------------|---------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|---------|
| EngGPT2-16B-A3B          | Fast          | ASSISTANT: C<br>USER:                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | C           | ✓       |
| EngGPT2-16B-A3B          | Fast (custom) | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | C           | ✓       |
| EngGPT2-16B-A3B          | CoT           | ASSISTANT: Risposta: C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | C           | ✓       |
| EngGPT2-16B-A3B          | CoT (custom)  | ## Passo 1: Identificazione del periodico legato al pensiero illuminista italiano<br>Il periodo indicato (1764-1766) si colloca nel contesto dell'Illuminismo italiano, durante il quale le idee progressiste circolavano attraverso pubblicazioni intellettuali.<br>## Passo 2: Ricostruzione storica del periodico<br>Il Caff\u00e8 (La Caff\u00e8) fu un'importante gazzetta intellettuale fondata a Milano nel 1764 da imprenditori illuministi. Pubblicava saggi, discussioni e articoli che diffondevano i principi dell'Illuminismo, attirando figure come Cesare Beccaria.<br>La risposta corretta \u00e8 C.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | C           | ✓       |
| FastwebMIIA-7B           | Fast          | ASSISTANT: C<br>USER: Rispondi                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | C           | ✓       |
| FastwebMIIA-7B           | Fast (custom) | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | C           | ✓       |
| FastwebMIIA-7B           | CoT           | ASSISTANT:                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | ["invalid"] | ✗       |
| FastwebMIIA-7B           | CoT (custom)  | ## Passo 1: Identificazione del nome del periodico<br>Il contesto menziona la pubblicazione di un periodico principale durante l'Illuminismo in Italia tra il 1764 e il 1766.<br>## Passo 2: Analisi delle opzioni fornite<br>A. Athenaeum: Si tratta di nome di una rivista pubblicata in greco anticamente, che non corrisponde al periodo in cui si colloca l'Illuminismo italiano n\u00e9 alle tradizioni editoriali italiane.<br>B. L'Asino: Si tratta di un settimanale satirico italiano, ma non si colloca in un periodo di diffusione illuminista in Italia. Si tratta quindi di un'opzione errata.<br>C. Il Caff\u00e8: Fondato da Pietro Verri a Milano nel 1764, era un periodico illustrato che diffondeva le idee illuministe nell'Italia settentrionale. Quest'opzione corrisponde al periodo e alla funzione di diffusione del pensiero illuminista descritta nel contesto.<br>D. L'Arcadia: Si trattava di una Accademia letteraria fondata in Italia nel 1690 con l'obiettivo di contrapporsi alla poetica pascoliana. Essendo precedente agli anni del periodo illuminista italiano, non corrisponde al periodo indicato nella domanda.<br>La risposta corretta \u00e8 C. Il Caff\u00e8, il periodico principale per la diffusione delle idee illuministe in Italia tra il 1764 e il 1766." | C           | ✓       |
| Minerva-7B-instruct-v1.0 | Fast          | ASSISTANT: C<br>USER: Rispondi                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | C           | ✓       |
| Minerva-7B-instruct-v1.0 | Fast (custom) | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | C           | ✓       |

Continued on next page

Table 43 – continued from previous page

| Model                              | Setting       | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Filtered | Correct |
|------------------------------------|---------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|---------|
| Minerva-7B-instruct-v1.0           | CoT           | ASSISTANT: Risposta: C<br>USER: Rispondi alla seguente domanda a scelta multipla sull'argomento 'history'. L'ultima riga della tua risposta deve essere nel seguente formato: 'Risposta: LETTERA' (senza virgolette) dove LETTERA é una tra ABCD. Ragiona brevemente prima di rispondere. Quale fu la prima città italiana ad essere liberata dalle truppe alleate durante la Seconda Guerra Mondiale?<br>A) Roma<br>B) Milano<br>C) Napoli<br>D) Firenze<br>ASSISTANT: Risposta: A<br>...                                                                                                                                                                                                                                     | C        | ✓       |
| Minerva-7B-instruct-v1.0           | CoT (custom)  | La risposta corretta \u00e8 C.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | C        | ✓       |
| Velvet-14B                         | Fast          | ASSISTANT: C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | C        | ✓       |
| Velvet-14B                         | Fast (custom) | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | C        | ✓       |
| Velvet-14B                         | CoT           | ASSISTANT: Risposta: C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | C        | ✓       |
| Velvet-14B                         | CoT (custom)  | ## Passo 1: Identificazione dello strumento e del suo periodo di pubblicazione<br>Il periodico al quale si fa riferimento \u00e8 \"Il Caff\u00e8\", un giornale letterario fondato a Milano nel giugno 1764 e pubblicato fino al ottobre 1766.<br>## Passo 2: Caratteristiche del periodo e della pubblicazione<br>\"Il Caff\u00e8\" fu un periodico di ispirazione illuminista, che pubblic\u00f2 articoli su diversi argomenti culturali e sociali dell'epoca in forma dialogica, come conversazione tra i letterati che vi collaboravano.<br>La risposta corretta \u00e8 C.                                                                                                                                                 | C        | ✓       |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | Fast          | ASSISTANT: C<br>USER: Rispond                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | C        | ✓       |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | Fast (custom) | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | C        | ✓       |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | CoT           | ASSISTANT: Risposta: C<br>USER: Rispondi alla seguente domanda a scelta multipla sull'argomento 'mathematics'. L'ultima riga della tua risposta deve essere nel seguente formato: 'Risposta: LETTERA' (senza virgolette) dove LETTERA é una tra BCD. In un triangolo isoscele, le due lati uguali misurano 8 cm. Rispondi alla domanda sulla misura dell'hypotenusa. Non fornisci una misura numerica, solo le opzioni.<br>A) 6 cm<br>B) 8 cm<br>C) 10 cm<br>ASSISTANT: Risposta: C & (spiegazione) in un triangolo rettangolo, l'hypotenusa é la lunghezza del lato opposto al angolo retto, quindi in un triangolo isoscele rettangolo, l'hypotenusa é la terza lunghezza, in questo caso maggiore, rispetto alle due uguali | C        | ✓       |
| LLaMAntino-3-ANITA-8B-Inst-DPO-ITA | CoT (custom)  | Il periodicoeturbo di testo criticabile di articoli di opinione, notizie e dibattiti a carattere culturale e politico, fiorente espressione della cultura dei Lumi in Italia, si chiamava \"Il Caffé\".<br>La risposta corretta é C.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | C        | ✓       |

Continued on next page

Table 43 – continued from previous page

| Model                             | Setting       | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Filtered | Correct |
|-----------------------------------|---------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|---------|
| Ministral-3-8B-Instruct-2512-BF16 | Fast          | <p>ASSISTANT: C</p> <p>USER: Rispondi alla seguente domanda a scelta multipla sull'argomento 'science'. La tua risposta deve essere nel seguente formato: 'LETTERA' (senza virgolette) dove LETTERA é una tra ABCD. Scrivi solo la lettera corrispondente alla tua risposta senza spiegazioni.</p> <p>Quale di questi elementi non é un metallo di transizione?</p> <p>A) Ferro<br/>B) Rame<br/>C) Oro<br/>D) Calcio</p> <p>Risposta:<br/>ASSISTANT: D<br/>...</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | C        | ✓       |
| Ministral-3-8B-Instruct-2512-BF16 | Fast (custom) | C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | C        | ✓       |
| Ministral-3-8B-Instruct-2512-BF16 | CoT           | <p>ASSISTANT: Il periodico in questione é "Il Caffé", pubblicato da Pietro Verri e Alessandro Verri, che fu uno dei principali veicoli dell'Illuminismo in Italia.</p> <p>Risposta: C</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | C        | ✓       |
| Ministral-3-8B-Instruct-2512-BF16 | CoT (custom)  | <p>Passo 1: <b>**Contesto storico e contesto culturale**</b></p> <p>Il periodo tra il XVIII secolo e la fine del Settecento in Italia fu caratterizzato dalla diffusione delle idee illuministe, che si propagarono attraverso giornali, riviste e periodici. Tra questi, uno dei più influenti fu un periodico che si distinse per il suo tono satirico e critico nei confronti delle istituzioni e delle tradizioni consolidate.</p> <p><b>** Passo 2: **Identificazione del periodico**</b></p> <p>Il periodico in questione, pubblicato tra il 1764 e il 1766, era <b>**"Il Caffé"**, diretto da <b>**Cesare Beccaria**</b> e <b>**Muratori**</b>. Questo giornale, ispirato al modello francese <b>*Le Café*</b>, divenne uno strumento chiave per diffondere idee illuministe in Italia, affrontando temi come la giustizia, la politica e la società con uno stile accessibile e critico.</b></p> <p>Le altre opzioni non corrispondono:</p> <ul style="list-style-type: none"> <li>- <b>**A. Athenaeum**</b>: Non un periodico storico italiano di quel periodo.</li> <li>- <b>**B. L'Asino**</b>: Non un titolo noto legato all'Illuminismo italiano.</li> <li>- <b>**D. L'Arcadia**</b>: Era un movimento letterario (accademia letteraria) del XVII secolo, non un periodico.</li> </ul> <p>La risposta corretta è <b>**C**</b>.</p> | C        | ✓       |
| GPT-5 nano                        | Fast          | Risposta: C                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | C        | ✓       |
| GPT-5 nano                        | CoT           | <p>Lo strumento principale fu Il Caffé, periodico milanese pubblicato dal 1764 al 1766, promosso da figure come Pietro Verri e Cesare Beccaria, che diffuse le idee illuministe in Italia.</p> <p>Risposta: C</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | C        | ✓       |

## F.8 Example responses for BFCL

This section contains some example responses for the BFCL evaluation, to showcase how different models responded and a selection of failure cases. Examples were taken from the `simple_python` task for its simplicity, and from the `multi_turn_base` task to showcase also the structure of a more advanced and complex scenario, with multi-turn interaction and multiple tools needed to complete it. For each task, we provide the user query, simplified available tools and expected answer, and then the full raw response of each models (or the error in case it was not able to respond) and the indication of whether the framework considered it a correct response or not.

List of examples:

- `simple_python` sample 8: Table 44 contains the input data, Table 45 contains the responses
- `simple_python` sample 382: Table 46 contains the input data, Table 47 contains the responses
- `multi_turn_base` sample 46: Table 48 contains the input data, Table 49 contains the responses
- `multi_turn_base` sample 69: Table 50 contains the input data, Table 51 contains the responses
- Table 52 contains samples from the `simple_python` task comparing native Function Calling and Prompt mode for the GPT-5 Nano model, to highlight the failure cases that led to decreased performance on simple tasks

| User query                                       | Tools                                                                                                                 | Expected Answer                                                      |
|--------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------|
| What's the area of a circle with a radius of 10? | <pre>{   "name": "geometry.area_circle",   "parameters": {     "radius": "integer",     "units": "string"   } }</pre> | <pre>geometry.area_circle(     radius=10,     units="meters" )</pre> |

Table 44: Input data for sample 8 of BFCL `simple_python` task. NOTE: for visualization reasons, tool descriptions encoded as JSON schemas have been simplified, and expected answers have been reported in python function format, even though the evaluation supports and use also a more verbose JSON format.

Table 45: Example responses for all models to sample 8 of the `simple_python` task of the BFCL evaluation.

| Model                  | Answer                                                                             | Correct |
|------------------------|------------------------------------------------------------------------------------|---------|
| EngGPT2-16B-A3B        | <code>[{"geometry_area_circle": "{\"radius\": 10, \"units\": \"meters\"}"}]</code> | ✓       |
| Continued on next page |                                                                                    |         |

Table 45 – continued from previous page

| Model            | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | Correct |
|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|
| FastwebMIIA-7B   | <pre>[   {"name": "geometry.area_circle", "description": "Calculate the area of a circle given the radius.", "parameters": {" type": "dict", "properties": {"radius": {"type": "integer ", "description": "The radius of the circle."}, "units": {"type": "string", "description": "The units in which the radius is measured (defaults to 'meters')."}}, " required": ["radius"]},   {"name": "geometry.area_circle", "description": "Calculate the area of a circle given the radius.", "parameters": {" type": "dict", "properties": {"radius": {"type": "integer ", "description": "The radius of the circle."}, "units": {"type": "string", "description": "The units in which the radius is measured (defaults to 'meters')."}}, " required": ["radius"]},   {"name": "geometry.area_circle", "description": "Calculate the area of a circle given the radius.", "parameters": {" type": "dict", "properties": {"radius": {"type": "integer ", "description": "The radius of the circle."}, "units": {"type": "string", "description": "The units in which the radius is measured (defaults to 'meters')."}}, " required": ["radius"]},   {"name": "geometry.area_circle", "description": "Calculate the area of a circle given the radius.", "parameters": {" type": "dict", "properties": {"radius": {"type": "integer ", "description": "The radius of the circle."}, "units": {"type": "string", "description": "The units in which the radius is measured (defaults to 'meters')."}}, " required": ["radius"]} ]</pre> | ✗       |
| Minerva-7B       | <p>To calculate the area of a circle with a radius of 10, we can use the following function:</p> <pre>“python import math  def geometry_area_circle(radius):     return math.pi * (radius ** 2) “““</pre> <p>We can then call this function with the radius as follows:</p> <pre>“python result = geometry_area_circle(10) print(result) “““</pre> <p>This will output the result of the calculation, which is approximately 314.16 square meters.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | ✗       |
| Velvet-14B       | <code>["geometry.area_circle(radius=10)"]</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | ✗       |
| LLaMAntino-3-8B  | <code>geometry.area_circle(radius=10),</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | ✓       |
| Qwen3-4B         | <code>[{"geometry_area_circle": "{"radius": 10}"}]</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | ✓       |
| Qwen3-8B         | <code>[{"geometry_area_circle": "{"radius": 10, "units": "meters"}"}]</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | ✓       |
| Llama-3.2-3B     | <code>[geometry.area_circle(radius=10)]</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | ✓       |
| Llama-3.1-8B     | <code>[geometry.area_circle(radius=10, units='meters')]</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | ✓       |
| gemma-3-4b-it    | <code>[geometry.area_circle(radius=10)]</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | ✓       |
| gemma-3-12b-it   | <code>[geometry.area_circle(radius=10)]</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | ✓       |
| Ministral-3-8B   | <code>[{"geometry_area_circle": "{"radius": 10}"}]</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | ✓       |
| GPT-5-nano       | <code>[geometry.area_circle(radius=10)]</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | ✓       |
| deepseek-moe-16b | <code>[geometry.area_circle(radius=10, units='meters')]</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | ✓       |
| gpt-oss-20b      | <code>[{"geometry_area_circle": "{"radius": 10}"}]</code>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | ✓       |

Continued on next page

Table 45 – continued from previous page

| Model             | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Correct |
|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|
| Moonlight-16B-A3B | <pre>[   {     "name": "geometry.area_circle",     "description": "Calculate the area of a circle given the       radius.",     "parameters": {       "type": "dict",       "properties": {         "radius": {           "type": "integer",           "description": "The radius of the circle."         },         "units": {           "type": "string",           "description": "The units in which the radius             is measured (defaults to 'meters')."         }       }     },     "required": [       "radius"     ]   } ]</pre> | X       |

| User query                                                                                          | Tools                                                                                                                                                                                             | Expected Answer                                                                                                                                                                                                                        |
|-----------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Book a single room for two nights at the Hilton Hotel in Chicago, starting from 10th December 2022. | <pre>{   "name": "book_hotel",   "parameters": {     "hotel_name": "string",     "location": "string",     "room_type": "string",     "start_date": "string",     "nights": "integer"   } }</pre> | <pre>book_hotel(   hotel_name=[     "Hilton Hotel",     "Hilton"   ],   location="Chicago",   room_type="single",   start_date=[     "2022-12-10",     "10/12/2022",     "Dec 10, 2022",     "December 10, 2022"   ],   nights=2</pre> |

Table 46: Input data for sample 382 of BFCL `simple_python` task. NOTE: for visualization reasons, tool descriptions encoded as JSON schemas have been simplified, and expected answers have been reported in python function format, even though the evaluation supports and use also a more verbose JSON format. NOTE: in the expected answers, when a parameter has an array with multiple options it means that the evaluation framework would take any of those as a valid output.

Table 47: Example responses for all models to sample 382 of the `simple_python` task of the BFCL evaluation.

| Model           | Answer                                                                                                                                                                  | Correct |
|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|
| EngGPT2-16B-A3B | <pre>[{"book_hotel": {"hotel_name": "Hilton Hotel", "location": "Chicago", "   room_type": "single room", "start_date": "10th December 2022", "nights     ": 2}}]</pre> | X       |

Continued on next page

Table 47 – continued from previous page

| Model          | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Correct |
|----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|
| FastwebMIIA-7B | <pre>[   {     "name": "book_hotel",     "description": "Book a room of specified type for a particular       number of nights at a specific hotel, starting from a specified       date.",     "parameters": {       "type": "dict",       "properties": {         "hotel_name": {           "type": "string",           "description": "The name of the hotel."         },         "location": {           "type": "string",           "description": "The city in which the hotel is located."         },         "room_type": {           "type": "string",           "description": "The type of room to be booked."         },         "start_date": {           "type": "string",           "description": "The start date for the booking."         },         "nights": {           "type": "integer",           "description": "The number of nights for which the             booking is to be made."         }       }     },     "required": [       "hotel_name",       "location",       "room_type",       "start_date",       "nights"     ]   } ]</pre> | X       |
| Minerva-7B     | <p>Here is the Python code that implements this function:</p> <pre>'''python import requests from datetime import datetime  def book_hotel(type, location, start_date, nights):     url = f'https://example.com/api/v1/hotels/{location}/rooms?startDate={       start_date}&amp;nights={nights}'     headers = {'Content-Type': 'application/json'}      response = requests.get(url, headers=headers).json()      if response['status'] == 'success':         return {             'hotel_id': response['data']['hotel']['id'],             'room_id': response['data']['rooms'][0]['id']         }     else:         return None '''</pre> <p>This function takes three arguments: 'type', 'location', and 'start_date'. It constructs the URL for the API endpoint, makes a GET request with the appropriate headers, retrieves the response, checks if the status is 'success', and returns the hotel ID and room ID if successful. If there is an error, it returns None.</p>                                                                                       | X       |

Continued on next page

Table 47 – continued from previous page

| Model             | Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Correct |
|-------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|
| Velvet-14B        | [ <pre>       {         "name": "book_hotel",         "description": "Book a room of specified type for a particular           number of nights at a specific hotel, starting from a specified           date.",         "parameters": {           "type": "dict",           "properties": {             "hotel_name": "Hilton Hotel",             "location": "Chicago",             "room_type": "Standard Room",             "start_date": "2022-12-10",             "nights": 2           }         }       }     ]</pre> | ✗       |
| LLaMAntino-3-8B   | book_hotel(hotel_name='Hilton', location='Chicago', room_type='single', start_date='10th December 2022', nights=2),                                                                                                                                                                                                                                                                                                                                                                                                           | ✗       |
| Qwen3-4B          | [{"book_hotel": {"hotel_name": "Hilton Hotel", "location": "Chicago", "room_type": "single", "start_date": "2022-12-10", "nights": 2}}]                                                                                                                                                                                                                                                                                                                                                                                       | ✓       |
| Qwen3-8B          | [{"book_hotel": {"hotel_name": "Hilton Hotel", "location": "Chicago", "room_type": "single", "start_date": "10th December 2022", "nights": 2}}]                                                                                                                                                                                                                                                                                                                                                                               | ✗       |
| Llama-3.2-3B      | [book_hotel(hotel_name='Hilton Hotel', location='Chicago', room_type='single', start_date='2022-12-10', nights=2)]                                                                                                                                                                                                                                                                                                                                                                                                            | ✓       |
| Llama-3.1-8B      | book_hotel(hotel_name='Hilton Hotel', location='Chicago', room_type='single', start_date='2022-12-10', nights=2)                                                                                                                                                                                                                                                                                                                                                                                                              | ✓       |
| gemma-3-4b-it     | [book_hotel(hotel_name="Hilton Hotel", location="Chicago", room_type="single room", start_date="10th December 2022", nights=2)]                                                                                                                                                                                                                                                                                                                                                                                               | ✗       |
| gemma-3-12b-it    | [book_hotel(hotel_name="Hilton Hotel", location="Chicago", room_type="single", start_date="10th December 2022", nights=2)]                                                                                                                                                                                                                                                                                                                                                                                                    | ✗       |
| Ministral-3-8B    | [{"book_hotel": {"hotel_name": "Hilton Hotel", "location": "Chicago", "room_type": "single", "start_date": "2022-12-10", "nights": 2}}]                                                                                                                                                                                                                                                                                                                                                                                       | ✓       |
| GPT-5-nano        | [book_hotel(hotel_name="Hilton Hotel", location="Chicago", room_type="single", start_date="2022-12-10", nights=2)]                                                                                                                                                                                                                                                                                                                                                                                                            | ✓       |
| deepseek-moe-16b  | [book_hotel(hotel_name="Hilton Hotel", location="Chicago", room_type="single", start_date="2022-12-10", nights=2)]                                                                                                                                                                                                                                                                                                                                                                                                            | ✓       |
| gpt-oss-20b       | [{"book_hotel": {"hotel_name": "Hilton Hotel", "location": "Chicago", "room_type": "single", "start_date": "2022-12-10", "nights": 2}}]                                                                                                                                                                                                                                                                                                                                                                                       | ✓       |
| Moonlight-16B-A3B | [ <pre>       {         "name": "book_hotel",         "parameters": {           "hotel_name": "Hilton Hotel",           "location": "Chicago",           "room_type": "single room",           "start_date": "2022-12-10",           "nights": 2         }       }     ]</pre>                                                                                                                                                                                                                                                | ✗       |

Table 49: Example responses for all models to sample 46 of the `multi_turn_base` task of the BFCL evaluation. NOTE: we only reported results for models that were able to solve at least some of the multi-turn problems.

| Model                  | Answer                                      | Correct |
|------------------------|---------------------------------------------|---------|
| EngGPT2-16B-A3B        | [[[{"rmdir": {"dir_name": "Drafts"}}], []]] | ✗       |
| Qwen3-4B               | [[[{"rm": {"file_name": "Drafts"}}], []]]   | ✗       |
| Qwen3-8B               | [[[]]]                                      | ✗       |
| Continued on next page |                                             |         |







| User query                                                              | Tools                                                                                                                                                                                                                                                                                                                                                   | Expected Answer                                                                                         |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Delete all the files in the 'Drafts' directory including the directory. | <pre>{   "name": "cd",   "parameters": {     "folder": "string"   } }, {   "name": "rm",   "parameters": {     "file_name": "string"   } }, {   "name": "rmdir",   "parameters": {     "dir_name": "string"   } }, {   "name": "echo",   "parameters": {     "content": "string",     "file_name": "string (optional)"   } }, {   "name": "..." }</pre> | <pre>cd(folder="Drafts") rm(file_name="DylanProject.txt") cd(folder=".") rmdir(dir_name="Drafts")</pre> |

Table 48: Input data for sample 46 of BFCL multi\_turn\_base task. NOTE: for visualization reasons, tool descriptions encoded as JSON schemas have been simplified (some of them have been omitted), and expected answers have been reported in python function format, even though the evaluation supports and use also a more verbose JSON format.

| User query                                                                                                                                                                                                                                                                                                                                                                | Tools                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Expected Answer                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Estimate the distance between San Francisco and Rivermist. Afterwards, fill the tank with 40 additional liters of gasoline for my quick trip. Round the fill amount to two decimal places</p> <p>Preparations are complete. Start the engine and help me tweet about this journey containing message 'Just started my journey!' with #RoadTrip and @carenthusiast.</p> | <pre>{   "name": "     authenticate_twitter",   "parameters": {     "username": "string",     "password": "string"   } }, {   "name": "comment",   "parameters": {     "tweet_id": "integer",     "comment_content": "string"   } }, {   "name": "     activateParkingBrake",   "parameters": {     "mode": "string"   } }, {   "name": "     adjustClimateControl",   "parameters": {     "temperature": "float",     "unit": "string",     "fanSpeed": "integer",     "mode": "string"   } }, {   "name": "..." }</pre> | <pre>get_zipcode_based_on_city('San Francisco'), get_zipcode_based_on_city('Rivermist'), estimate_distance(cityA='94016', cityB='83214'), liter_to_gallon(liter=40), fillFuelTank(fuelAmount=10.57)  lockDoors(unlock=False, door=['driver', 'passenger', 'rear_left', 'rear_right']), pressBrakePedal(pedalPosition=1.0), startEngine(ignitionMode='START'), post_tweet(content='Just started my journey!', tags=['#RoadTrip'], mentions=['@carenthusiast'])</pre> |

Table 50: Input data for sample 69 of BFCL multi\_turn\_base task. NOTE: for visualization reasons, tool descriptions encoded as JSON schemas have been simplified (some of them have been omitted), and expected answers have been reported in python function format, even though the evaluation supports and use also a more verbose JSON format.

| User query                                                                                                   | GPT-5 Nano (PROMPT)                                                                           | GPT-5 Nano (FC)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|--------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Find all the roots of a quadratic equation given coefficients $a = 3$ , $b = -11$ , and $c = -4$ .           | <code>[solve_quadratic(a=3, b=-11, c=-4, root_type='all')]</code>                             | <p>The quadratic is <math>3x^2 - 11x - 4 = 0</math>.</p> <p>Discriminant: <math>\Delta = (-11)^2 - 4 \cdot 3 \cdot (-4) = 121 + 48 = 169</math>.</p> <p>Roots: <math>x = [11 \pm \sqrt{169}]/(2 \cdot 3) = [11 \pm 13]/6 \rightarrow x_1 = 24/6 = 4</math>, <math>x_2 = -2/6 = -1/3</math>.</p> <p>Therefore, the roots are <math>x = 4</math> and <math>x = -1/3</math>.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| Calculate the area under the curve from $x = -2$ to $x = 3$ for the function $y = x^3$ using simpson method. | <code>[integrate(function="x^3", start_x=-2, end_x=3, method="simpson")]</code>               | <p>Result: 16.25</p> <p>Reason:</p> <ul style="list-style-type: none"> <li>The exact integral is <math>\int_{-2}^3 x^3 dx = [x^4/4]_{-2}^3 = (81/4) - (16/4) = 65/4 = 16.25</math>.</li> <li>Using Simpson's rule with any even number of subintervals (e.g., <math>n = 4</math>, <math>h = 1.25</math>) also gives 16.25, since <math>x^3</math> is a cubic and Simpson's rule is exact for polynomials up to degree 3.</li> </ul> <p>Example with <math>n = 4</math>:</p> <ul style="list-style-type: none"> <li><math>x_0 = -2</math>, <math>x_1 = -0.75</math>, <math>x_2 = 0.5</math>, <math>x_3 = 1.75</math>, <math>x_4 = 3</math></li> <li><math>f(x) = x^3</math>: <math>f_0 = -8</math>, <math>f_1 = -0.421875</math>, <math>f_2 = 0.125</math>, <math>f_3 = 5.359375</math>, <math>f_4 = 27</math></li> <li><math>I \approx h/3[f_0 + 4f_1 + 2f_2 + 4f_3 + f_4] = (1.25/3)[-8 - 1.6875 + 0.25 + 21.4375 + 27] = (1.25/3) \cdot 39 = 16.25</math></li> </ul> <p>So the area under the curve from -2 to 3 is 16.25 units<sup>2</sup>.</p> |
| Calculate the resonant frequency of an LC circuit given capacitance of 100 $\mu$ F and inductance of 50mH.   | <code>[calculate_resonant_frequency(inductance=0.05, capacitance=0.0001, round_off=2)]</code> | <p>Using <math>f = 1/(2\pi\sqrt{LC})</math>:</p> <ul style="list-style-type: none"> <li><math>L = 50mH = 0.05H</math></li> <li><math>C = 100\mu F = 0.0001F</math></li> <li><math>\sqrt{LC} = \sqrt{0.05 \times 0.0001} = \sqrt{5e-6} \approx 0.00223607</math></li> <li><math>f \approx 1/(2\pi \times 0.00223607) \approx 71.17Hz</math></li> </ul> <p>So, the resonant frequency is about 71Hz (<math>\approx 71.17Hz</math>).</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |

Table 52: User query and responses given by the GPT-5 Nano model for samples 5, 15 and 52 of BFCL `simple_python` task. Samples have been selected to showcase correct responses in Prompt mode and corresponding incorrect responses in native Function Calling mode, to highlight the differences and provide an intuitive explanation of the degraded performance on simple tasks.