# TAILORED PRIVATE LARGE LANGUAGE MODELS FOR CYBERSECURITY COPILOTS

Powered By Penfield.AI

## Abstract

A method to boost privately hosted Llama 2's reasoning to outperform out-of-the-box LLMs for tailored cybersecurity tasks

Penfield.AI Inc.
info@penfield.ai

# Tailored Private LLMs for Cybersecurity Copilots

Powered by Penfield.AI

## Table of Contents

# Tailored Private LLMs for Cybersecurity Copilots

Powered by Penfield.AI

## Author(s):

- Tahseen Shabab, CEO of Penfield.AI Inc. (tahseen@penfield.ai)
- Sohail Habib, Head of AI at Penfield.AI Inc. (sohail@penfield.ai)
- Long Tan, Head of Engineering at Penfield.AI Inc. (long@penfield.ai)
- Manan Singh, Head of IT at Penfield.AI Inc. (manan@penfield.ai)

Penfield.AI Website: https://penfield.ai/

## 1. Abstract

In recent years, the ability of Large Language Models (LLMs) to Reason, Adopt, and Act, has ushered in a new era of Cybersecurity Copilots. LLMs aspire to augment cybersecurity efforts by understanding threats and interacting with cyber tools, potentially bridging the cyber talent gap. However, realizing their full potential requires overcoming several hurdles.

**First,** out-of-the-box LLMs lack knowledge of organization-specific cybersecurity processes, policies, tools, and the IT infrastructure. This hinders its ability to make contextual decisions, leading to unreliable automation [1].

**Second,** there's reluctance from enterprises and governments to adopt publicly hosted LLMs, driven by privacy, regulatory, and control concerns. A recent survey concluded that 75% of Enterprises don’t plan to use commercial LLMs in Production due to this reason [2].

**Third,** bilingual support specifically French for Canadian organizations.

**Fourth**, the need for cost-effective, versatile hosting solutions to support multiple cybersecurity team roles like Red Team, Blue Team, Threat Hunting, etc.

While privately hosted open-source models like Llama2 ensure ownership and privacy, they initially lack the sophisticated reasoning abilities of models like GPT 3.5/4. Llama2 (70B) reasoning was shown to be 81% worse than GPT4 in a recent benchmark [3].

The paper proposes a method to boost Llama2's reasoning to outperform out-of-the-box LLMs for tailored cybersecurity tasks, using Penfield.AI. It details continuous fine-tuning techniques with Penfield’s curated data from Sr. Analysts (Section 3.3), applies Penfield's curated data from prior tasks for enhancing model calls (Section 3.4), and uses Penfield’s Process Prompt for clear process instructions (Section 3.6). It also covers a hosting architecture enabling customized features like bilingual support (Section 4).

## 2. Limitations of generic LLMs in client-specific Cybersecurity applications

Key limitations of generic LLMs to drive client-specific Security copilots have been stated below.

### 2.1 Making Organization-specific Contextual Decisions

Out-of-the-box LLMs like GPT-4, trained on extensive data including documents, images, and videos, excel in text prediction but often fall short in tasks specific to organizations. They may generate generic, inaccurate, or even harmful content without tailored data or instructions relevant to a specific domain [5].

Without training on domain and client-specific data, models risk generating inaccurate or irrelevant text, known as hallucination [6]. This is exemplified in the decoding process of language models, where they convert input to output, often using beam search to estimate the most likely word sequences. This process, as shown in Figure 1, selects the most suitable sequences based on training data, highlighting the importance of domain-relevant training.

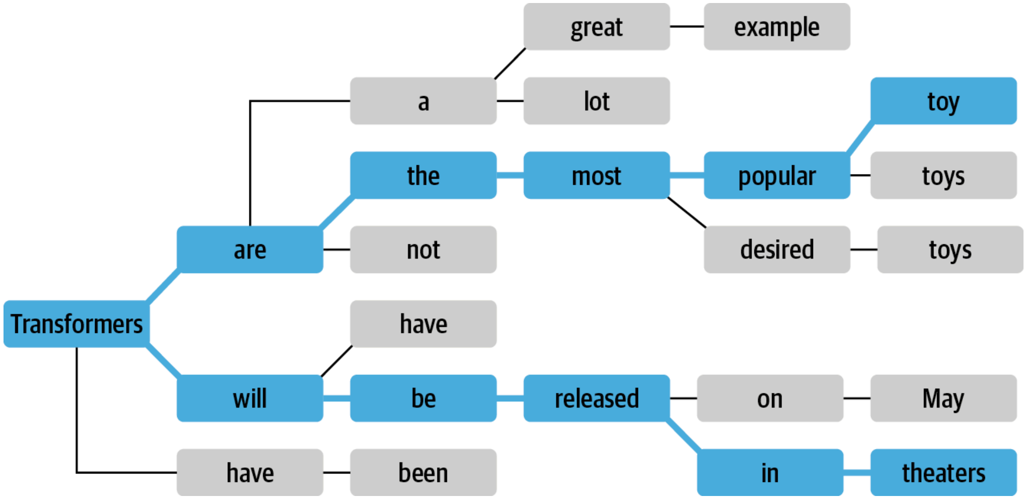
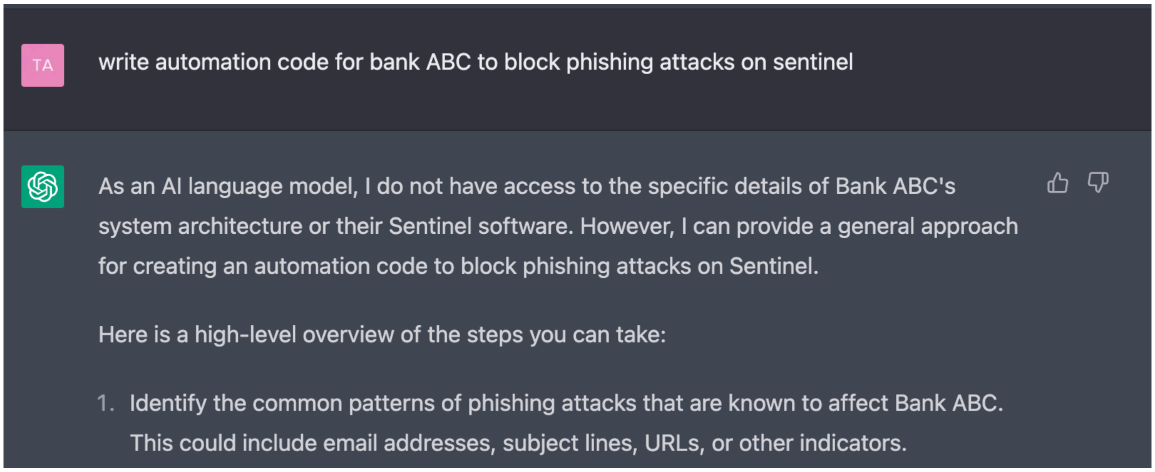


Figure 1. Beam Search [6].

For tasks such as developing anti-phishing code for a particular bank, Language Models (LLMs) might be ineffective without deep knowledge of the bank's specific operations and systems, as shown in Figure 2.



TA write automation code for bank ABC to block phishing attacks on sentinel

As an AI language model, I do not have access to the specific details of Bank ABC's system architecture or their Sentinel software. However, I can provide a general approach for creating an automation code to block phishing attacks on Sentinel.

Here is a high-level overview of the steps you can take:

1. Identify the common patterns of phishing attacks that are known to affect Bank ABC. This could include email addresses, subject lines, URLs, or other indicators.

Figure 2. ChatGPT attempting to author automation code for a client-specific task.

The challenge is further aggravated in cybersecurity due to context-sensitive processes. Process gaps have already hindered the initial goal of SOAR (Security Orchestration, Automation, and Response) solutions to automate everything using off-the-shelf playbooks. For instance, a bank might react differently to the same cyber-attack on different servers, like online banking versus rewards, due to differences in technology, policy, and business importance [1].

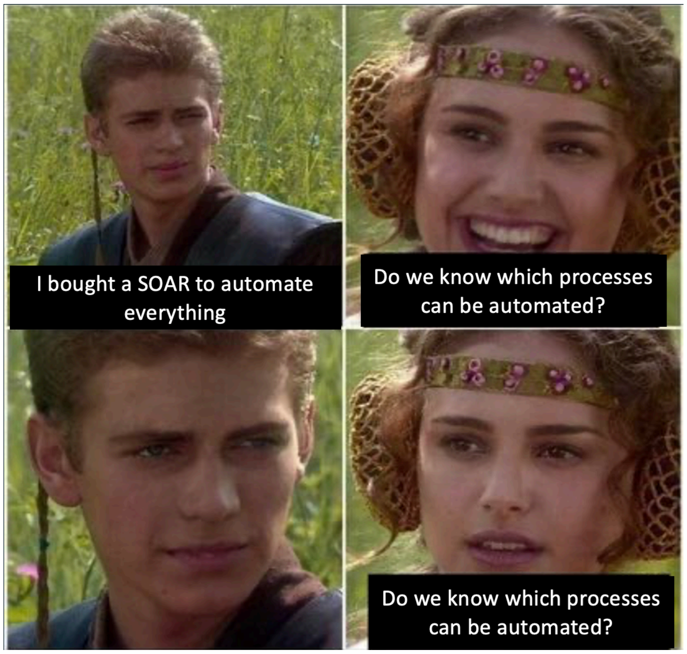

I bought a SOAR to automate everything
Do we know which processes can be automated?
Do we know which processes can be automated?

Figure 3. The dependency of SOAR tools on defined processes [1]

This general challenge is further echoed by recent research that highlights while LLMs exhibit human reasoning, they falter with complex tasks. Karthik et. al. notes that current LLM benchmarks prioritize simple reasoning, neglecting more intricate problems [7].

This paper explores strategies to enhance LLMs with domain-specific data and instructions, enabling them to perform human-like tasks and detailed reasoning [8]. It also discusses how integrating Penfield.AI's AI-generated documentation and process knowledge from senior analysts can facilitate complex reasoning in LLMs.

### 2.2 Privacy and Control

Recent data shows that over three-quarters of enterprises are hesitant to implement commercial Large Language Models (LLMs) like GPT-4 in product due to data privacy concerns [9]. This reluctance is mainly due to the need to share sensitive information, like IP addresses and security vulnerabilities, via internet-based APIs, conflicting with many firms' privacy needs. This paper explores how privately hosted, open-source LLMs could mitigate these risks.

### 2.3 Bilingual Support

Bilingual support is crucial for organizations dealing with multi-language data. Open-source models like Llama2-chat are often English-centric and lack inherent multilingual abilities 10]. Multilingual transformers fill this gap by training on texts in over a hundred languages, enabling understanding of multiple languages without extra fine-tuning, known as zero-shot cross-lingual transfer [6]. Despite their primary language focus, models like Llama2 can be fine-tuned for multilingual support, which we will examine in this paper.

### 2.4 Cost-effective and Modular Deployment

High costs hinder the broad adoption of Large Language Models (LLMs) by enterprises. Developing and training such models demands substantial GPU investment, with examples like OpenAI's GPT-3 needing over $5 million in GPUs. Operational costs, including cloud services and API usage, add to this financial strain [11]. Additionally, fine-tuning open-source models can be costly due to the high compute, storage, and hosting expenses, especially when full retraining is required for various applications and teams within an organization [8], as depicted in Figure 4.

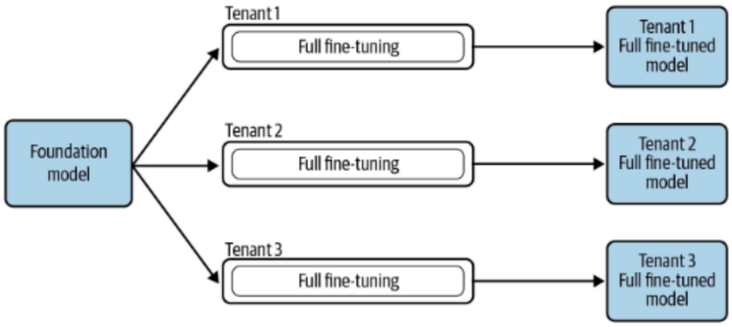

Foundation model
Tenant 1
Full fine-tuning
Tenant 1 Full fine-tuned model
Tenant 2
Full fine-tuning
Tenant 2 Full fine-tuned model
Tenant 3
Full fine-tuning
Tenant 3 Full fine-tuned model

Figure 4. Full Finetuning of a Foundation Model across different tenants [8]

The paper will address how Perimeter-efficient Fine-tuning (PEFT) offers techniques to fine-tune models with fewer resources, focusing on using human context data from Penfield.AI. This approach is also relevant when managing multilingual capabilities [6], as maintaining multiple monolingual models substantially raises costs and complexity for engineering teams.

## 3. Building Tailored and Performant Security Copilots

This section outlines the approach for creating customized security copilots.

### 3.1 Model Selection

Privately hosted open-source models are chosen due to ownership and privacy needs, despite their initial inferior reasoning capabilities compared to models like GPT 3.5/4—for instance, Llama2's reasoning is 81% less effective than GPT-4 [3]. This performance gap is illustrated in The paper intends to detail techniques to enhance the reasoning of these models to surpass the performance of standard LLMs for cybersecurity tasks tailored to specific organizations.

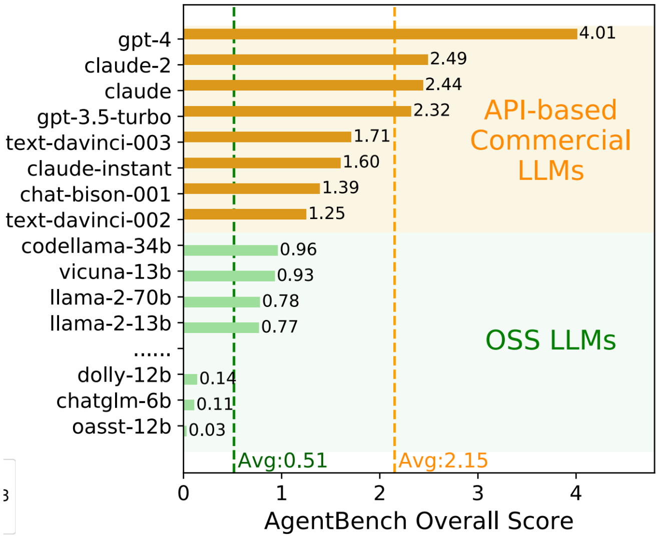
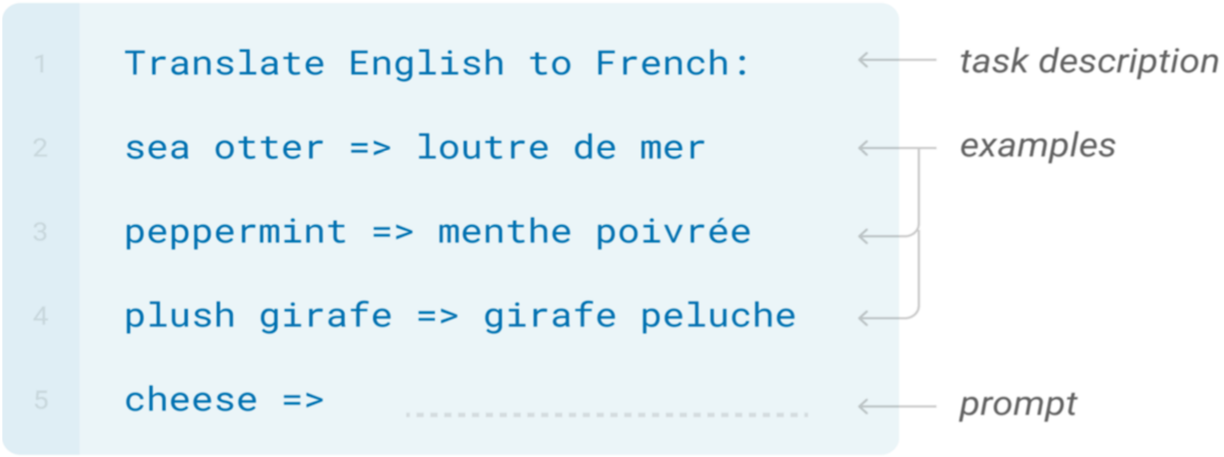


Figure 5. Reasoning Benchmark of different Generic LLMs [3]

### 3.2 Fine-Tuning Techniques and Hosting

Training Large Language Models (LLMs) fully is a resource-intensive task, demanding significant GPU memory and increasing both the computing budget and costs [8]. To mitigate this, a method known as Parameter Efficient Fine Tuning (PEFT) is employed.

In-context learning has become a norm for training LLMs, where a model is fed relevant examples to learn from within its context window. This technique enables the LLM to become context-aware. An illustration of this is when an LLM is given examples of translations, word corrections, or arithmetic operations, which it then applies to new sentences, words, or problems, a process demonstrated in Figure 6 [13].



```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

Figure 6. Example of In-Context Learning

However, the small context size of Transformers, usually less than 100 examples, limits their efficiency, making the search for an effective fine-tuning mechanism essential [14].

Parameter Efficient Fine Tuning (PEFT) addresses this by freezing the foundational model and fine-tuning only a small set of parameters, sometimes as little as 1-2% of the original LLM weights. This process often requires just one GPU [8]. PEFT fine-tunes minimal weights for distinct uses, significantly reducing the model's size.

Additionally, a pre-trained model can serve as the basis for several small PEFT modules tailored to different tasks. By freezing the core model and switching out only the fine-tuned matrices or layers, storage needs and task-switching complexity are greatly reduced. This approach also maintains inference speed, avoiding the latency typically introduced with full fine-tuning [15]. This setup is detailed in Figure 7.

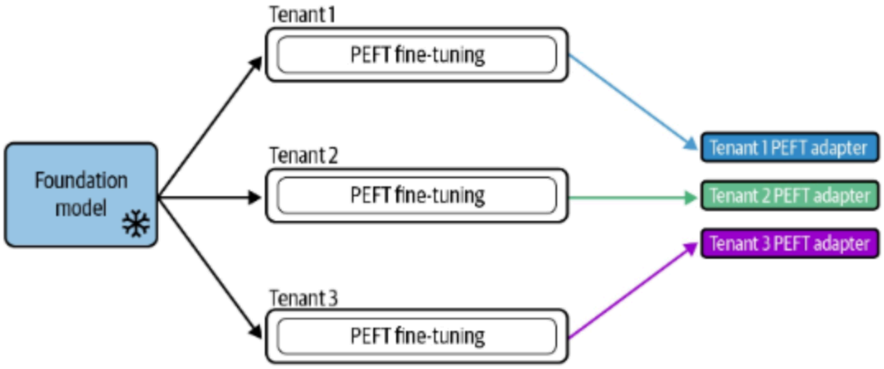


Figure 7. Merging original LLM with task-specific PEFT model weights at inference [8]

The performance of models fine-tuned with this method can be impressive. The LoRA Paper [15] reports that such models achieve or exceed the results of fully tuned counterparts like GPT-3 while utilizing fewer trainable parameters and maintaining a quicker training pace without adding any extra time to the inference process.

### 3.3 Curating Data for Fine-Tuning with Penfield.AI

Earlier, we explored methods for fine-tuning open-source LLMs economically. The question now is, how do we source client-specific cybersecurity data to enhance these copilots?

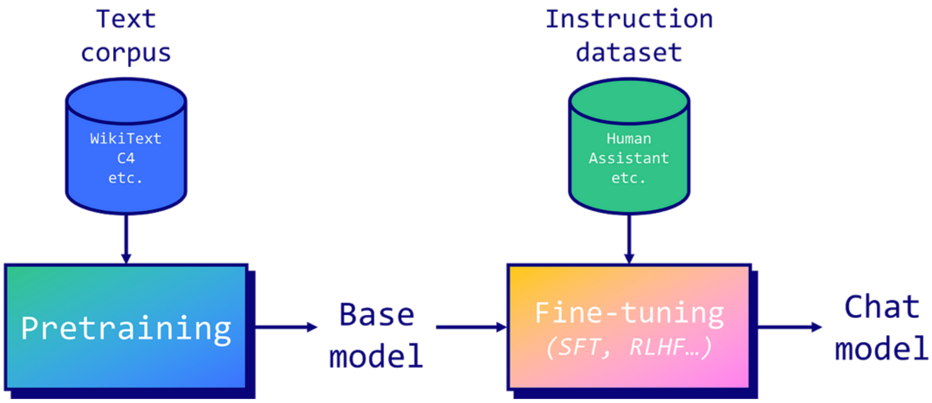
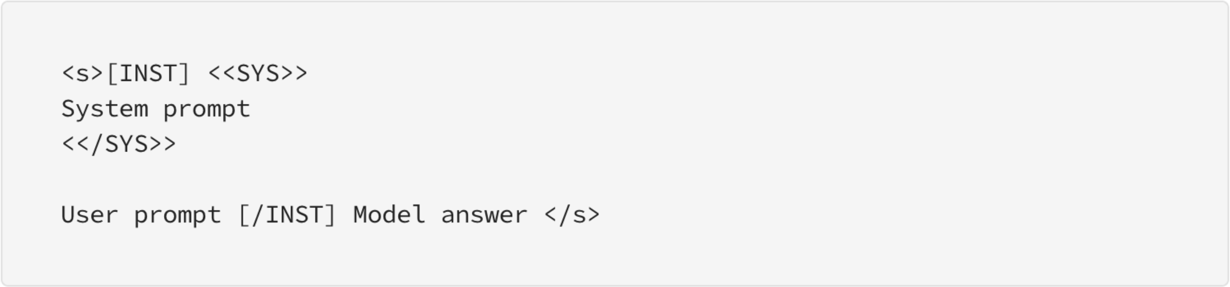


Figure 8. Data for training LLMs [16]

Foundation models are pre-trained on vast datasets, encompassing text, images, and audio. These models are then fine-tuned for specific customer and domain tasks. Research has shown that fine-tuning LLMs with a limited set of high-quality data samples can significantly enhance their performance. According to the LIMA paper [17], a fine-tuned LLaMA (v1) model with just 1000 high-quality samples outperformed GPT-3. LIMA has demonstrated impressive abilities, learning to handle complex queries and follow specific formats with minimal examples. It also generalizes well to new tasks do not present in its training data. In a human study, LIMA's responses were found to be equivalent to or preferred over GPT-4 in 43% of cases.

For fine-tuning, data must be formatted according to a 'Prompt Template.' Figure 9 illustrates this template for Llama2 in chat mode [18]. Generally, a dataset of User Questions and Sample Answers is required for such fine-tuning processes.



```
<s>[INST] <<SYS>>
System prompt
<</SYS>>

User prompt [/INST] Model answer </s>
```

Figure 9. Llama2 Prompt Template [18]

The Penfield.AI platform utilizes interaction data from Senior Analysts to generate high-quality data, enabling the continuous fine-tuning of LLMs. The Penfield Pathfinder Browser Extension captures this interaction data from web-based applications, including proprietary in-house solutions, and transforms it into procedural documentation.

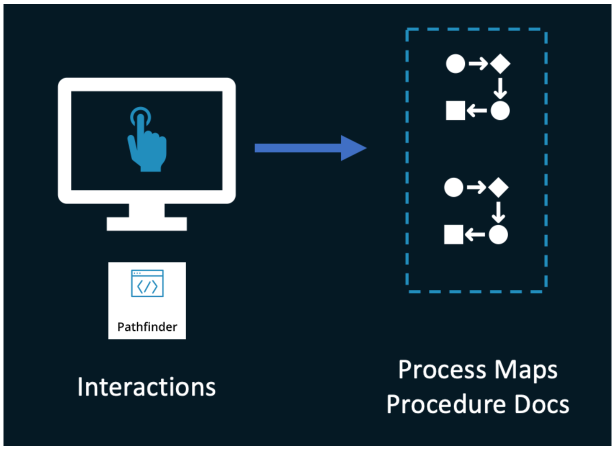


Figure 10. Penfield Pathfinder converts analyst interactions into Process Maps and Documents

Penfield transforms interactions into Process Maps, organizing them as graph data structures for advanced downstream analysis through Process Mining. This graph data is then enriched into detailed text documents, incorporating both the client's existing organizational data and open-source repositories. This enrichment is guided by analysts' interactions and tool usage, all managed within the Penfield product suite.

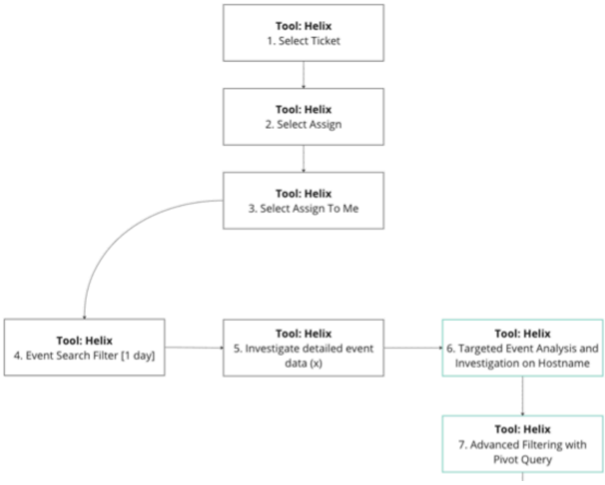

Tool: Helix
1. Select Ticket
Tool: Helix
2. Select Assign
Tool: Helix
3. Select Assign To Me
Tool: Helix
4. Event Search Filter [1 day]
Tool: Helix
5. Investigate detailed event data (x)
Tool: Helix
6. Targeted Event Analysis and Investigation on Hostname
Tool: Helix
7. Advanced Filtering with Pivot Query

Figure 11. A Penfield.AI Process Map

Pathfinder captures not just basic user interactions like clicks and copy-paste actions but also complex queries and decisions, distilling the nuanced knowledge from human experts to enhance the fine-tuning of LLMs.

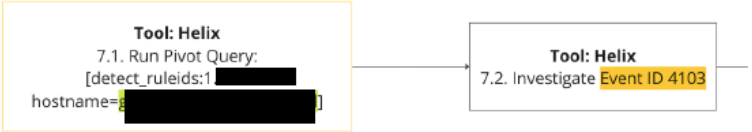


Figure 12. Penfield Pathfinder Capturing an Advanced Filter with Pivot Query

Analysts carry on with their usual tasks as data is seamlessly collected and used to enhance the models.

## 3.4 Few-Shot Learning with Penfield.AI

Figure 6 demonstrates Few-Shot learning, where the model is exposed to a handful of task examples to enhance its response quality on unfamiliar tasks. Penfield streamlines this process in Cybersecurity by guiding a copilot with Penfield.AI's Generated Documentation, which is crafted from senior analysts' problem-solving interactions and procedures. This strategy boosts copilot performance on tasks tailored to specific cybersecurity needs.

## 3.5 Tailored Prompting

Another method to enhance copilot action quality is customizing prompts used by the copilot to align with client requirements. This involves examining a key component of Copilot Architecture: the interaction between the AI Agent and the Tools available to it.

The fundamental principle of agents is to empower an AI with the capacity for reasoning and selecting a series of actions [19]. Conversely, Tools serve as interfaces, like APIs to software products, that allow the Agent to engage with its environment. Outputs from these Tool interactions, such as API responses, are then integrated into the LLM that informs the Agent, thereby refining its reasoning and guiding subsequent actions. This concept is visualized in Figure 20.

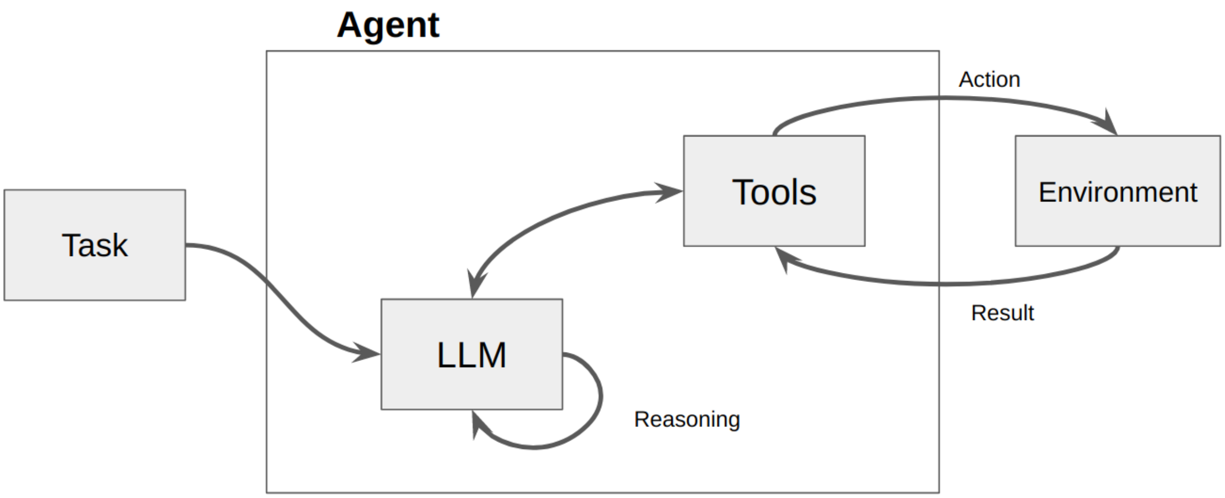


Figure 13. Agent Architecture [20]

Figure 14 provides a more intricate illustration of the process. Here, a user submits a query to the Agent, which is processed by the LLM at the core of the agent. The LLM deliberates and formulates a sequence of steps that prompts interactions with external data sources and the execution of tasks using external tools, such as Firewalls and Scanners. Following task completion, the Agent delivers a response to the user.

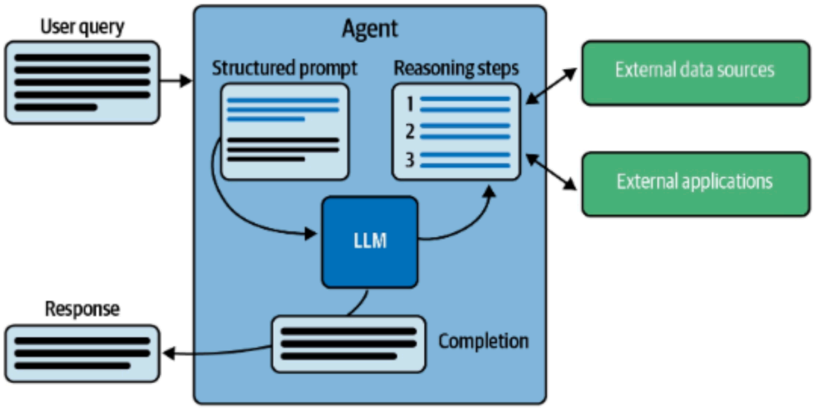
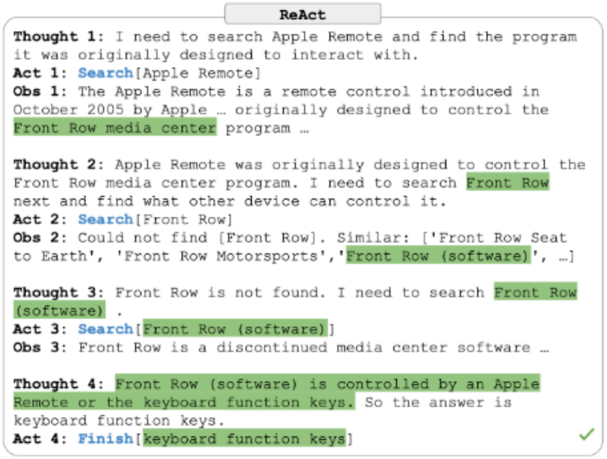


Figure 14. Agent Workflow [8]

This section examines two types of prompts to enhance copilot efficiency: **ReAct Prompting** and **Tool Prompting**.

**ReAct Prompting** incorporates a set of task resolution examples, showcasing few-shot learning with traces of human reasoning, actions taken, and observed environmental reactions, as shown in Figure 15.

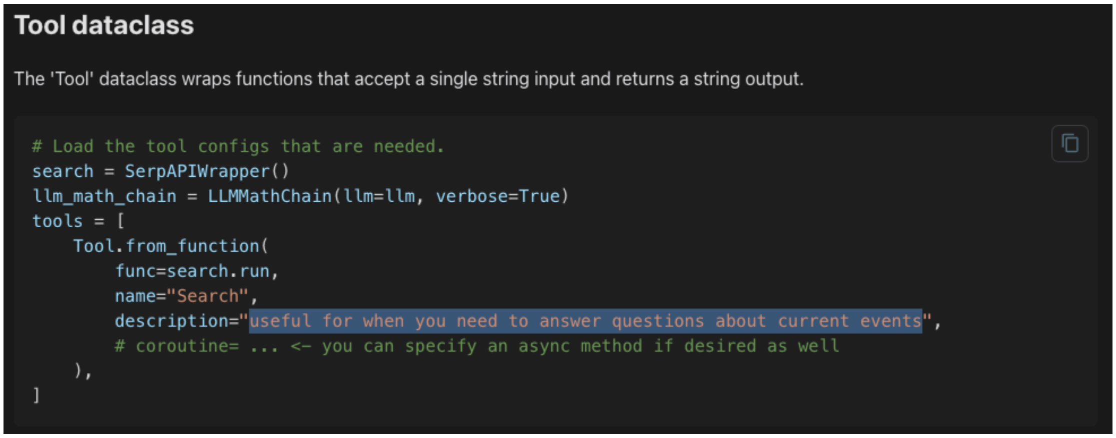
This method is particularly user-friendly and adaptable, yielding top-tier few-shot learning outcomes in various tasks, from question-answering to online shopping. Customizing ReAct prompts is thus a potent way to bolster copilot performance.



```
ReAct
Thought 1: I need to search Apple Remote and find the program
it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in
October 2005 by Apple … originally designed to control the
Front Row media center program …

Thought 2: Apple Remote was originally designed to control the
Front Row media center program. I need to search Front Row
next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat
to Earth', 'Front Row Motorsports','Front Row (software)', …]

Thought 3: Front Row is not found. I need to search Front Row
(software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Thought 4: Front Row (software) is controlled by an Apple
Remote or the keyboard function keys. So the answer is
keyboard function keys.
Act 4: Finish[keyboard function keys]                        ✓
```

Figure 15. Example of ReAct prompting [21]

A **Tool Prompt** encapsulates a textual description that serves as an interface for Tools, such as API calls to various products, guiding the Agent on the timing and method of using specific tools for problem-solving, as illustrated in Figure 16. Fine-tuning Tool Prompts is especially advantageous for environments with an abundance of tools, preventing the Agent from being overwhelmed by choice and maintaining the copilot's effectiveness.



**Tool dataclass**

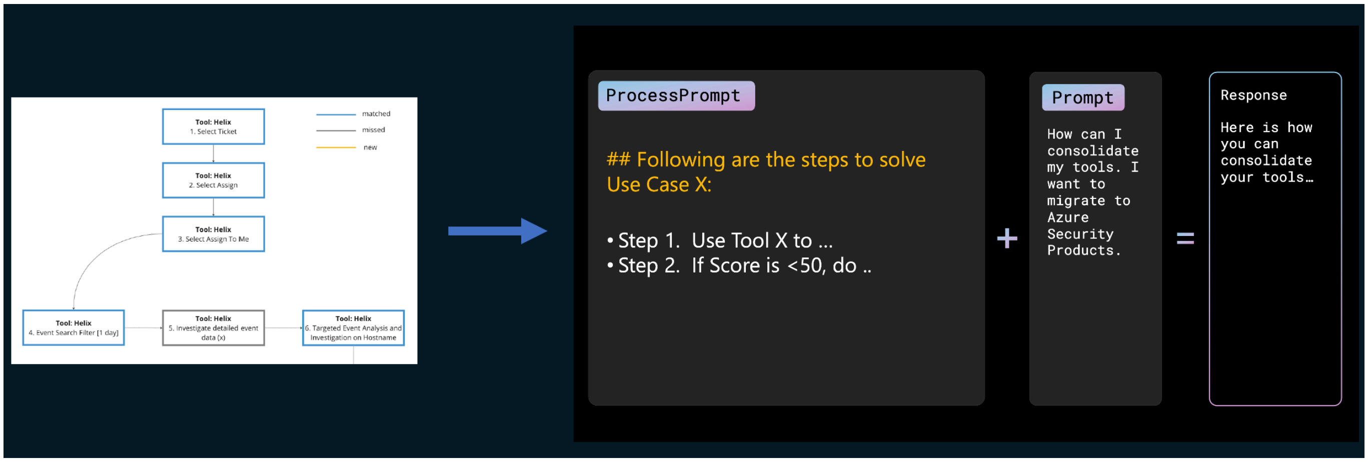
The 'Tool' dataclass wraps functions that accept a single string input and returns a string output.

```
# Load the tool configs that are needed.
search = SerpAPIWrapper()
llm_math_chain = LLMMathChain(llm=llm, verbose=True)
tools = [
    Tool.from_function(
        func=search.run,
        name="Search",
        description="useful for when you need to answer questions about current events",
        # coroutine= ... <- you can specify an async method if desired as well
    ),
]
```

Figure 16. Tool Data class on LangChain

## 3.6 Infusing Process Knowledge with Penfield.AI ProcessPrompt

Penfield's ProcessPrompt, designed to synchronize Security Copilots with organizational practices, continuously updates with procedures, tools, and operations by learning from senior analyst interactions. This constantly refreshed knowledge guides Agents to adhere to organizational protocols and tool usage, enhancing their reasoning focus and alignment without compromising their decision-making capabilities. By knowing which tools analysts use, when, and how, the Agents' performance is substantially improved.



## Following are the steps to solve Use Case X:

- Step 1. Use Tool X to ...
- Step 2. If Score is <50, do ..

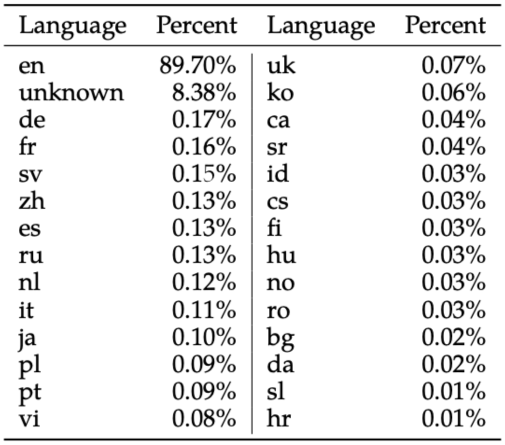
How can I consolidate my tools. I want to migrate to Azure Security Products.

Here is how you can consolidate your tools…

Figure 17. Making LLMs process aware with Penfield ProcessPrompt

## 3.7 Bilingual Support

Penfield.AI can enable and enhance bilingual support of open-sourced models like Llama 2 through fine-tuning. Llama 2, straight out of the box, faces challenges with non-English languages, attributed to its predominantly English training dataset. Although Llama 2-Chat shows a degree of multilingual ability from our experimental observations, its capabilities are restricted. This is largely because non-English pretraining data is less abundant, a limitation reflected in Figure 18 [10].



| Language | Percent | Language | Percent |
|---|---|---|---|
| en | 89.70% | uk | 0.07% |
| unknown | 8.38% | ko | 0.06% |
| de | 0.17% | ca | 0.04% |
| fr | 0.16% | sr | 0.04% |
| sv | 0.15% | id | 0.03% |
| zh | 0.13% | cs | 0.03% |
| es | 0.13% | fi | 0.03% |
| ru | 0.13% | hu | 0.03% |
| nl | 0.12% | no | 0.03% |
| it | 0.11% | ro | 0.03% |
| ja | 0.10% | bg | 0.02% |
| pl | 0.09% | da | 0.02% |
| pt | 0.09% | sl | 0.01% |
| vi | 0.08% | hr | 0.01% |

Figure 18. A large segment of Llama  2's training data was in English

However, Llama 2 can be efficiently fine-tuned for bilingual support with the PEFT method. Research has shown successful fine-tuning of Llama 2 for Portuguese [22] and German [23], achieving notable results. For these adaptations, the baseline LLaMA-7b model underwent fine-tuning via PEFT on a single A100 GPU over a brief period. Penfield-driven fine-tuning can leverage this method to provide improved bilingual support and continuously funnel bilingual data and documentation into the model to achieve even better results.

## 4. Summary

In summary, this paper outlines a strategy to elevate the Llama 2 model above standard LLMs for cybersecurity, utilizing Penfield.AI's. It focuses on enhancing reasoning through continuous Penfield-driven fine-tuning (Section 3.3), leveraging Penfield's historical data for more informed model interactions (Section 3.4), and providing precise Penfield Process Prompts for procedural tasks (Section 3.6). The paper also details a hosting architecture (Section 4) that enables tailored, advanced copilot functionalities. This methodology is poised to refine the application of LLMs in cybersecurity with Penfield.AI's integration at each step.

## 5. About Penfield.AI

Penfield.AI, established at the University of Waterloo in 2017, specializes in Human Machine Intelligence in cybersecurity. The company excels in transforming the interactions of analysts with various tools into comprehensive knowledge, which is then leveraged to train both AI and humans. For more information, please visit our website at https://penfield.ai/, or reach out to the authors for further details.

## References

1. Shabab, T. (2023). "Continuously Improving the Capability of Human Defenders with AI". Penfield.AI. Available at: https://youtu.be/sRm4uWS7kkg [Feb 16, 2023].
2. Business Wire. (2023). "Survey: More than 75% of Enterprises Don't Plan to Use Commercial LLMs in Production Citing Data Privacy as Primary Concern". Available at: https://www.businesswire.com/news/home/20230823249705/en/Survey-More-than-75-of-Enterprises-Don%E2%80%99t-Plan-to-Use-Commercial-LLMs-in-Production-Citing-Data-Privacy-as-Primary-Concern [Aug 23, 2023].
3. Xiao Liu et. al. (2023). "AgentBench: Evaluating LLMs as Agents". Available at: https://arxiv.org/pdf/2308.03688.pdf [Oct 25, 2023].
4. Jie Huang et. al. (2023). "Towards Reasoning in Large Language Models: A Survey". [May 26, 2023]
5. OpenAI. (2022). "Aligning LLMs to follow instructions". Available at: https://openai.com/research/instruction-following [Jan 27, 2022].
6. Lewis et. al. (2023). "Natural Language Processing with Transformers". Oreilly.
7. Karthik et. al. (2023). "PlanBench Paper". Journal/Conference. Available at: https://arxiv.org/pdf/2206.10498.pdf
8. Chris Fregly et. al. (2023). "Generative AI on AWS". Oreilly.
9. Business Wire. (2023). "Survey: More than 75% of Enterprises Don't Plan to Use Commercial LLMs in Production Citing Data Privacy as Primary Concern". Available at: https://www.businesswire.com/news/home/20230823249705/en/Survey-More-than-75-of-Enterprises-Don%E2%80%99t-Plan-to-Use-Commercial-LLMs-in-Production-Citing-Data-Privacy-as-Primary-Concern.
10. Meta Research. (2023). "Llama 2 Paper". Available at: https://arxiv.org/pdf/2307.09288.pdf
11. Smith, C. (2023). "What Large Models Cost You – There Is No Free AI Lunch". Forbes. Available at: https://www.forbes.com/sites/craigsmith/2023/09/08/what-large-models-cost-you--there-is-no-free-ai-lunch/?sh=6b26181c4af7
12. Xiao Liu et. al. (2023). "AgentBench: Evaluating LLMs as Agents". Available at: https://arxiv.org/pdf/2308.03688.pdf [Oct 25, 2023].
13. The Gradient. "In-context Learning in Context". Available at: https://thegradient.pub/in-context-learning-in-context/.
14. Alexa AI. (2023). "Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning". Available at: https://arxiv.org/pdf/2303.15647.pdf
15. Microsoft Research. (2021). "LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS". Available at: https://arxiv.org/pdf/2106.09685.pdf
16. Towards Data Science. (2023). "Fine-Tune Your Own Llama 2 Model in a Colab Notebook". Available at: https://towardsdatascience.com/fine-tune-your-own-llama-2-model-in-a-colab-notebook-df9823a04a32.

17. Meta AI. (2023). "LIMA: Less Is More for Alignment". Available at: https://arxiv.org/abs/2305.11206.
18. Towards Data Science. (2023). "Fine-Tune Your Own Llama 2 Model in a Colab Notebook". Available at: https://towardsdatascience.com/fine-tune-your-own-llama-2-model-in-a-colab-notebook-df9823a04a32.
19. Langchain Documentation. (2023). Available at: https://python.langchain.com/docs/modules/agents/
20. Better Programming. (Year). "Make Langchain Agent Actually Works with Local LLMs". Available at: https://betterprogramming.pub/make-langchain-agent-actually-works-with-local-llms-vicuna-wizardlm-etc-da42b6b1a97?gi=ed1dcdb4b6b9.
21. Shunyu Yao et. Al. (2023). "ReAct: Synergizing Reasoning and Acting in Language Models". Available at: https://react-lm.github.io/
22. GitHub. (2023). "Cabrita: A Portuguese finetuned instruction Llama". Available at: https://github.com/22-hours/cabrita.
23. GitHub. (2023). "Zicklein: A German finetuned instruction Llama ". Available at: https://github.com/avocardio/Zicklein.