What is Bella GPT Guard?

Bella GPT Guard is a cutting-edge classification model designed to predict if a given document was produced by a large language model. It analyzes content at varying granularities, including the sentence, paragraph, and entire document. Bella GPT Guard's training incorporated a diverse array of human-written and AI-generated texts, with a special emphasis on English prose.

How can I utilize Bella GPT Guard effectively?

Bella GPT Guard has proven useful in a wide range of sectors, including but not limited to education, hiring, social writing platforms, and even tackling disinformation. Its main function is to shed light on potential instances of AI-generated text within prose.

To get the most out of Bella GPT Guard, we suggest focusing on the document-level score, named completely_generated_prob. This score reflects the likelihood of a document being entirely AI-generated and should guide you in deciding whether AI was significantly involved in the text's generation.

The sentence-level classification feature should be used when Bella GPT Guard identifies a mix of human-written and AI-generated content in a document. A highlighted sentence only indicates the potential AI involvement in that particular segment, not the entire essay. If a large portion of a document is flagged as AI-generated, the highlighted sentences will pinpoint where we believe this occurred.

We intend for Bella GPT Guard to flag potential AI-generated content and foster meaningful conversations, such as those between educators and students, to raise awareness about the implications of using AI in writing.

Is Bella GPT Guard limited to detecting outputs from ChatGPT?

Not at all. Bella GPT Guard is capable of identifying content produced by a variety of AI language models, including but not limited to ChatGPT, GPT-3, GPT-2, LLaMA, and AI services built on these models.

What limitations does Bella GPT Guard have?

As AI-generated content evolves rapidly, our model's results should not be used as a sole determinant in any form of punitive action, such as penalizing students. We continue to refine Bella GPT Guard to increase its robustness. Meanwhile, we encourage educators to utilize these results as part of a comprehensive assessment of student work.

Bella GPT Guard's accuracy improves as the text input length increases, meaning its document-level classification is generally more accurate than its paragraph-level or sentence-level classifications. Furthermore, Bella GPT Guard is most accurate with text that resembles our training dataset, primarily adult-written English prose.

Our model does not currently detect heavily modified AI-generated text, and it may misclassify certain machine-generated or highly procedural text as AI-generated. Therefore, it's recommended to use Bella GPT Guard primarily on more descriptive sections of text.

How can I respond if Bella GPT Guard flags AI-generated text in my students' work?

We recognize that no AI detection tool is flawless, and there may be false positives and negatives. That said, if Bella GPT Guard detects AI-generated content, we recommend the following steps:

Engage students in a controlled environment where they can demonstrate their understanding. This can be through an in-person assessment or using an editor that monitors their edit history.

Ask students to provide evidence of their writing process, such as drafts, revision histories, or brainstorming notes.

Investigate if there's a recurring pattern of AI use in a student's work. It's more beneficial to identify long-term AI usage rather than a one-time instance.

How can I, as an educator, mitigate the risk of AI misuse?

Proactive collaboration between educators and students to understand and address AI misuse can lead to better educational outcomes. As an educator, you can:

Educate students about the risks and implications of using AI in their work.

Create assessments that cannot be solved by AI, such as tasks requiring personal experiences or citing primary sources.

Set clear expectations that student work will be examined with an AI detector like Bella GPT Guard, to discourage AI misuse.

Can you share details about the data Bella GPT Guard was trained on?

Bella GPT Guard was trained on a dataset consisting of human-written and AI-generated texts. The human-written portion spans various categories, including student articles, news pieces, and Q&A datasets from multiple disciplines. For every human-written text, corresponding AI-generated articles were produced to eliminate topic-level bias. Our training model maintains an equal balance of human and AI-written texts.

How is the performance of Bella GPT Guard validated?

We test Bella GPT Guard on a unique set of human and AI articles, including some that challenge its training distribution. Our classifier correctly categorizes 99% of the human-written and 85% of the AI-generated articles.

How can I interpret the probabilities generated by Bella GPT Guard's API?

For determining whether a document is entirely AI-generated, we recommend utilizing the completely_generated_prob. This score represents the likelihood of a document being completely AI-generated. We suggest a threshold of 0.65 or higher to minimize false positives, as erroneously categorizing human writing as AI-generated is presently more harmful.

Does Bella GPT Guard store data from API calls?

No, we don't store or collect documents sent via our API calls, maintaining our commitment to user data privacy. However, we do store inputs from dashboard calls to help improve Bella GPT Guard. For more details, please refer to our privacy policy.

Why should I choose Bella GPT Guard over other detection models?

In contrast to many competitors, Bella GPT Guard is specifically fine-tuned for student writing and academic prose, resulting in significantly improved accuracies.

Our single, dedicated mission is to equip every individual with the tools needed to detect and safely adopt AI technologies, earning us the trust of numerous users, especially educators.