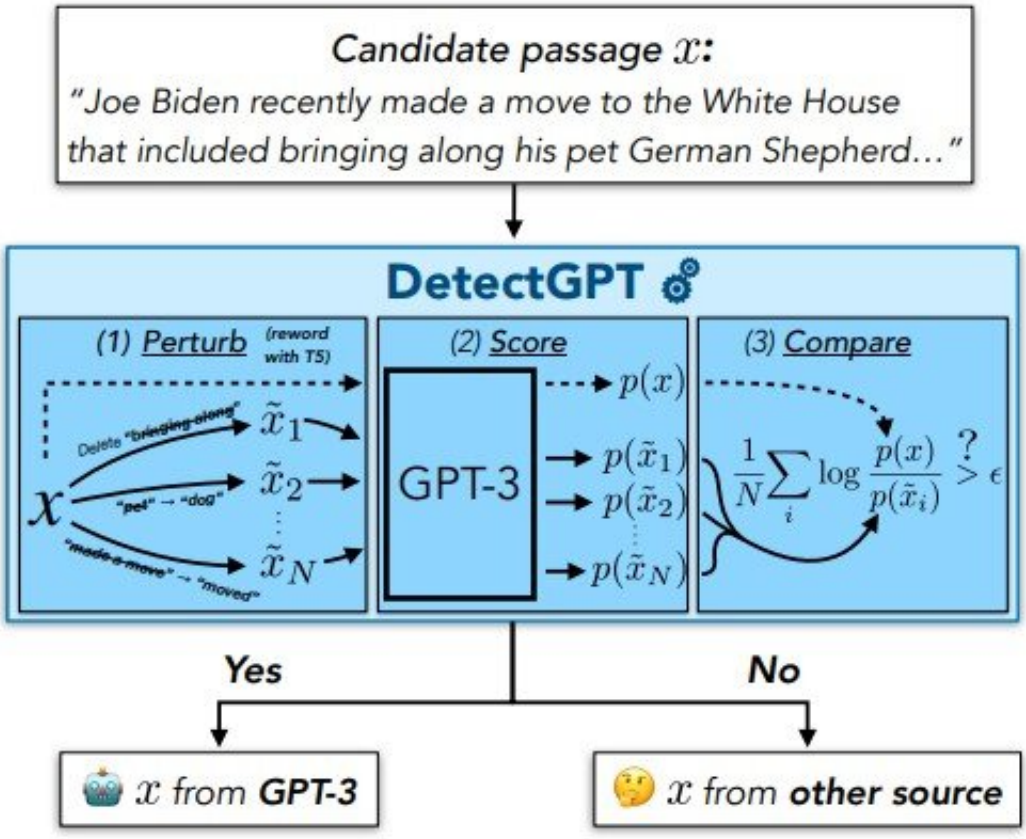
Human writer or AI? Scholars build a detection tool

February 15 2023, by Katharine Miller



We aim to determine whether a piece of text was generated by a particular LLM p, such as GPT-3. To classify a candidate passage x, DetectGPT first generates minor perturbations of the passage x˜i using a generic pre-trained model such as T5. Then DetectGPT compares the log probability under p of the original sample x with each perturbed sample x˜i. If the average log ratio is high, the sample is likely from the source model. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2301.11305

The launch of OpenAI's ChatGPT, with its remarkably coherent responses to questions or prompts, catapulted large language models (LLMs) and their capabilities into the public consciousness. Headlines captured both excitement and cause for concern: Can it write a cover letter? Allow people to communicate in a new language? Help students cheat on a test? Influence voters across social media? Put writers out of a job?

Now with similar models coming out of Google, Meta, and more, researchers are calling for more oversight.

"We need a new level of infrastructure and tools to provide guardrails around these models," says Eric Anthony Mitchell, a fourth-year computer science graduate student at Stanford University whose Ph.D. research is focused on developing such an infrastructure.

One key guardrail would provide teachers, journalists, and citizens a way to know when they are reading text generated by an LLM rather than a human. To that end, Mitchell and his colleagues have developed DetectGPT, released as a demo and a paper last week, which distinguishes between human- and LLM-generated text. In initial experiments, the tool accurately identifies authorship 95% of the time across five popular open-source LLMs.

While the tool is in its early stages, Mitchell hopes to improve it to the point that it can benefit society.

"The research and deployment of these language models is moving quickly," says Chelsea Finn, assistant professor of computer science and of electrical engineering at Stanford University and one of Mitchell's advisors. "The general public needs more tools for knowing when we are reading model-generated text."

An intuition

Barely two months ago, fellow graduate student and co-author Alexander Khazatsky texted Mitchell to ask: Do you think there's a way to classify whether an essay was written by ChatGPT? It set Mitchell thinking.

Researchers had already tried several general approaches to mixed effect. One—an approach used by OpenAI itself—involves training a model with both human- and LLM-generated text and then asking it to classify whether another text was written by a human or an LLM. But, Mitchell thought, to be successful across multiple subject areas and languages, this approach would require a huge amount of data for training.

A second existing approach avoids training a new model and simply uses the LLM that likely generated the text to detect its own outputs. In essence, this approach asks an LLM how much it "likes" a text sample, Mitchell says. And by "like," he doesn't mean this is a sentient model that has preferences. Rather, a model's "liking" of a piece of text is a shorthand way to say "scores highly," and it involves a single number: the probability of that specific sequence of words appearing together, according to the model. "If it likes it a lot, it's probably from the model. If it doesn't, it's not from the model." And this approach works reasonably well, Mitchell says. "It does much better than random guessing."

But as Mitchell pondered Khazatsky's question, he had the initial intuition that because even powerful LLMs have subtle, arbitrary biases for using one phrasing of an idea over another, the LLM will tend to "like" any slight rephrasing of its own outputs less than the original. By contrast, even when an LLM "likes" a piece of human-generated text, meaning it gives it a high probability rating, the model's evaluation of slightly modified versions of that text would be much more varied. "If

we perturb a human-generated text, it's roughly equally likely that the model will like it more or less than the original."

Mitchell also realized that his intuition could be tested using popular open-source models including those available through OpenAI's API. "Calculating how much a model likes a particular piece of text is basically how these models are trained," Mitchell says. "They give us this number automatically, which turns out to be really useful."

Testing the intuition

To test Mitchell's idea, he and his colleagues ran experiments in which they evaluated how much various publicly available LLMs liked human-generated text as well as their own LLM-generated text, including fake news articles, creative writing, and academic essays. They also evaluated how much the LLMs, on average, liked 100 perturbations of each LLM- and human-generated text. When the team plotted the difference between these two numbers for LLM- compared to human-generated texts, they saw two bell curves that barely overlapped. "We can discriminate between the source of the texts pretty well using that single number," Mitchell says. "We're getting a much more robust result compared with methods that simply measure how much the model likes the original text."

In the team's initial experiments, DetectGPT successfully classified human- vs. LLM-generated text 95% of the time when using GPT3-NeoX, a powerful open-source variant of OpenAI's GPT models. DetectGPT was also capable of detecting human- vs. LLM-generated text using LLMs other than the original source model, but with slightly less confidence. (As of this time, ChatGPT is not publicly available to test directly.)

More interest in detection

Other organizations are also looking at ways to identify AI-written text. In fact, OpenAI released its new text classifier last week and reports that it correctly identifies AI-written text 26% of the time and incorrectly classifies human-written text as AI-written 9% of the time.

Mitchell is reluctant to directly compare the OpenAI results with those of DetectGPT because there is no standardized dataset for evaluation. But his team did run some experiments using OpenAI's previous generation pre-trained AI detector and found that it worked well on English news articles, performed poorly on PubMed articles, and failed completely on German language news articles. These kinds of mixed results are common for models that depend on pre-training, he says. By contrast, DetectGPT worked out of the box for all three of these domains.

Evading detection

Although the DetectGPT demo has been publicly available for only about a week, the feedback has already been helpful in identifying some vulnerabilities, Mitchell says. For example, a person can strategically design a ChatGPT prompt to evade detection, such as by asking the LLM to speak idiosyncratically or in ways that seem more human. The team has some ideas for how to mitigate this problem, but hasn't tested them yet.

Another concern is that students using LLMs like ChatGPT to cheat on assignments will simply edit the AI-generated text to evade detection. Mitchell and his team explored this possibility in their work, finding that although there is a decline in the quality of detection for edited essays, the system still did a pretty good job of spotting machine-generated text

when fewer than 10%–15% of the words had been modified.

In the long run, Mitchell says, the goal is to provide the public with a reliable, actionable prediction as to whether a text—or even a portion of a text—was machine generated. "Even if a model doesn't think an entire essay or news article was written by a machine, you'd want a tool that can highlight a paragraph or sentence that looks particularly machine-crafted," he says.

To be clear, Mitchell believes there are plenty of legitimate use cases for LLMs in education, journalism, and elsewhere. However, he says, "giving teachers, newsreaders, and society in general the tools to verify the source of the information they're consuming has always been useful, and remains so even in the AI era."

Building guardrails for LLMs

DetectGPT is only one of several guardrails that Mitchell is building for LLMs. In the past year he also published several approaches for editing LLMs, as well as a strategy called "self-destructing models" that disables an LLM when someone tries to use it for nefarious purposes.

Before completing his Ph.D., Mitchell hopes to refine each of these strategies at least one more time. But right now, Mitchell is grateful for the intuition he had in December. "In science, it's rare that your first idea works as well as DetectGPT seems to. I'm happy to admit that we got a bit lucky."

The study is published on the *arXiv* preprint server.

More information: Eric Mitchell et al, DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, *arXiv* (2023). DOI: 10.48550/arxiv.2301.11305

Provided by Stanford University