Intuition Against the Machine: Detecting AI Generated Text Like a Detective

Erik Amundsen 1/4/2024

The GPT Large Language Models (LLMs) like ChatGPT, Bing Copilot, and Google Bard have been widely available for a little more than a year at time of writing. By now, you've likely read some AI generated text. For professionals whose credibility centers on knowledge, familiarity with these tools and the ability to spot their outputs has become a key skill.

There are several reasons why the ability to detect AI generated content has become so important, but there are two that stand out above the others. These are:

1. AI generated content is often factually incorrect.
2. There is a lot of it.

These alone create a dynamic in which it becomes very difficult to trust information presented as factual and authoritative, and difficult to wade through the volume of sources now available thanks to the popularity of LLM applications.

Disseminating factually incorrect information is an efficient way to lose hard-won credibility and authority as a best case, and it can have greater consequences. The foraging community has been coping with a surge in popularity for foraging wild foods following the pandemic that prompted a series of at least partly AI generated field guides with misinformation that could sicken unwary hobbyists.

The volume of AI generated content has caused its own problems, as when the speculative fiction magazine Clarkesworld was forced to shut down their submissions for 3 weeks in early 2023 as their editors and slush readers were overwhelmed by AI generated submissions.

Alone, these two issues arising from AI generated content warrant a savvy approach to all content, and knowledge of the skills and tools to detect content generated by LLMs.

The Intuitive Approach

LLMs work by recognizing patterns in their dataset, and using those patterns to generate what should come next. They are getting more sophisticated at this process all the time.

Humans also recognize patterns, we're good at it, and we have millions of years' worth of head-start. Machine generated writing can feel uncanny to an intuitive reader, and sometimes that is enough. We know it when we see it.

You might be tempted to stop there, trust your instincts, cringe at a few uncanny samples of AI generated content, and feel confident in your ability to tell human from machine. The tools and those who use them for good or ill are refining their techniques, however, and a gut check may not always suffice.

Looking for Tells

AI generated text has some "tells," clues to the provenance of the content that are helpful to look for when reading.

1. Coherence – Does the text build on what has come before in a logical way that flows through the whole document? Do paragraphs build on one another toward a conclusion or merely list information? Poor human writing can lack coherence, but this can also be a tell.
2. Voice – The word you choose and the word you do not choose are equally important in a text. A key component to the writer's craft is cultivating consistency in those choices without leaning on repetition. An LLM will generate text that often contains both repetition (especially if the model is geared toward SEO) and word choices that feel inconsistent with what has come before.
3. Point of View – AI generated text does not do a good job developing arguments, venturing opinions, or taking sides. If a piece of writing equivocates or tries to persuade with a list of pieces of evidence presented in an order that doesn't flow from one item to the next, this can be a tell.
4. Shallowness – AI generated text will stick very close to the parameters of its prompts. It's unlikely to extrapolate, add details from outside the scope of the prompt that contribute to the point the piece is making, or go in-depth for important details.

Like a detective questioning a person of interest, focusing on specific tells to look for when detecting AI generated content is a good way to augment your intuition.

Tools of the Trade

There are several tools designed to help detect AI generated text. Three popular ones are Copyleaks, Originality.ai, and GPTZero. These tools use predictive text algorithms like the LLMs to look for repetition and passages that match their model of what an AI would write.

These tools are the bloodhound and black light of the toolkit. Remember for each bot-sniffing tool created, another tool conceals the provenance of the AI generated text. The tools are powerful but can't do your detective work for you. You are still the one who must detect.

True Detective Work

A good detective, like a good journalist or academic, checks their sources. Machine generated content prompts us all to be detectives, lest we squander our credibility or mislead our readers. If there is information presented in a document and your gut, its tells, or the tools you employ tell you to doubt its source, then check the facts it presents. If you employ an LLM at any point in your process, check the facts it presents.

When you read a piece of writing, consider the author and what they want to gain from creating the piece. Consider the incentives they might have to rely on AI tools to compose the text. Consider the risks they might take if they are discovered to be relying upon the writing of an LLM.

The importance of media literacy cannot be overstated, and that importance is likely to grow. Answering these questions about a piece of writing will help you spot misleading or inauthentic writing of many kinds, human and machine.

The challenge of being a human author is finding something to care about within a topic and then communicating why you care and why the reader should as well. A lot of the weaknesses of AI generated writing and the means of detecting it stem from the fact that an LLM cannot care, and for now, that remains hard to simulate.

@mycomutant r/behindthebastards (8/2023). This is worse than the dinosaur colouring books - AI generated mushroom foraging/cooking books. https://www.reddit.com/r/behindthebastards/comments/15xclq9/this_is_worse_than_the_dinosaur_colouring_books/?rdt=55676

@newyorkmyc (8/27/2023) "@Amazon and other retail outlets have been inundated with AI foraging and identification books. Please only buy books of known authors and foragers, it can literally mean life or death." https://twitter.com/newyorkmyc/status/1695689778224594959

Attard, Simon (11/20/2023). *The Evolution of LLMs Over the Last 12 Months*. Medium. https://medium.com/@simon_attard/the-evolution-of-llms-over-the-last-12-months-188a04edb3ac

Bluehost (11/11/2023). *How to Make AI Content Undetectable in 2024*. https://www.bluehost.com/blog/how-to-make-ai-content-undetectable/

Delouya, Samantha (1/17/2023). *A tech news site issued a string of corrections after an article quietly written by AI got key facts wrong*. Business Insider. https://www.businessinsider.com/tech-site-issued-corrections-after-ai-writing-got-facts-wrong-2023-1

Garces, Karen (7/18/2023). *How Does an AI Checker Work: 5 Reliable AI Detectors*. Penji. https://penji.co/ai-checker/

Geib, Claudia (10/10/2023). *AI Is Writing Books About Foraging. What Could Go Wrong?* Civil Eats. https://civileats.com/2023/10/10/ai-is-writing-books-about-foraging-what-could-go-wrong/

Kulkarni, Ninad (12/3/2023). *Building basic intuition for Large Language Models (LLMs)*. Medium. https://medium.com/@thefrankfire/building-basic-intuition-for-large-language-models-llms-91f7ca92dfe7

McLean, Deanna (8/7/2023). *How to Detect AI Writing in 2023*. Elegant Themes. https://www.elegantthemes.com/blog/business/how-to-detect-ai-writing

Montclair State University, Office for Faculty Excellence. *AI Writing Detection: Red Flags*. https://www.montclair.edu/faculty-excellence/teaching-resources/clear-course-design/practical-responses-to-chat-gpt/red-flags-detecting-ai-writing/

Thorbecke, Catherine (8/29/2023) *AI tools make things up a lot, and that's a huge problem.* CNN. https://www.cnn.com/2023/08/29/tech/ai-chatbot-hallucinations/index.html