OpenAI won’t watermark ChatGPT text because its users could get caught

OpenAI has had a system for watermarking ChatGPT-created text and a tool to detect the watermark ready for about a year, reports *The Wall Street Journal*. But the company is divided internally over whether to release it. On one hand, it seems like the responsible thing to do; on the other, it could hurt its bottom line.

OpenAI’s watermarking is described as adjusting how the model predicts the most likely words and phrases that will follow previous ones, creating a detectable pattern. (That’s a simplification, but you can check out Google’s more in-depth explanation for Gemini’s text watermarking for more).

Offering any way to detect AI-written material is a potential boon for teachers trying to deter students from turning over writing assignments to AI. The *Journal* reports that the company found watermarking didn’t affect the quality of its chatbot’s text output. In a survey the company commissioned, “people worldwide supported the idea of an AI detection tool by a margin of four to one,” the *Journal* writes.

After the *Journal* published its story, OpenAI confirmed it’s worked on watermarking text in a blog post update today that was spotted by *TechCrunch.* In it, the company says its method is very accurate (“99.9% effective,” according to documents the *Journal* saw) and resistant to “tampering, such as paraphrasing.” But it says techniques like rewording with another model make it “trivial to circumvention by bad actors.” The company also says it’s concerned about the

stigmatization AI tools’ usefulness for non-native speakers.

But it seems OpenAI is also worried that using watermarking could turn off surveyed ChatGPT users, almost 30 percent of whom evidently told the company that they’d use the software less if watermarking was implemented.

Despite that, some employees still reportedly feel that watermarking is effective. In light of nagging user sentiments, though, the *Journal* says some suggested trying methods that are “potentially less controversial among users but unproven.” In its blog post update today, the company said it’s “in the early stages” of exploring embedding metadata. It says it’s still “too early” to know how well it will work, but that because it’s cryptographically signed, there would be no false positives.

https://www.theverge.com/2024/8/4/24213268/openai-chatgpt-text-watermark-cheat-detection-tool