Evelyn Drake

Dr. Demeter

ENGL 301

1 May 2025

LLM-Assisted Plagiarism and Strategies to Detect It

Large language models, or LLMs for short, are advanced statistical language models capable of generating and interpreting written language. OpenAI's ChatGPT and Google Gemini are widely-adopted examples of large language models that can quickly generate coherent, intelligible language due to their vast knowledge base of human-written training data (Wichuta Chompurach et al. 93). While these tools have enhanced productivity in various fields, their increasing use has sparked ethical concerns, particularly in education. Unlike traditional plagiarism, which often involves direct copying, AI-generated plagiarism is more difficult to detect because LLMs typically paraphrase existing information rather than reproducing it verbatim. Furthermore, the stochastic nature of these models causes their output to be inherently random and unpredictable, making conventional plagiarism detection systems, which rely on comparing input text to known sources, largely ineffective (Weber-Wulff et al. 6). This issue has become especially pressing in higher education, where students frequently utilize LLMs for writing assistance (Wichuta Chompurach et al. 93). Some institutions have responded by restricting access to AI tools on their networks, but such measures fail to address the underlying challenge: detecting AI-assisted plagiarism reliably and ethically (Weber-Wulff et al. 2).

Although computational linguists and software developers have made progress with AI detection software like GPTZero and enhanced versions of Turnitin, their reliability remains uncertain. As this essay argues, addressing LLM-assisted plagiarism requires a multifaceted

approach that combines automated detection technologies, human judgment, and student education to keep pace with evolving AI capabilities and preserve academic integrity.

In response to the growing concerns of educators, computer scientists have been training a variety of different language processing models to detect AI-generated text, which fall into three categories: simple classifiers, zero-shot classifiers, and fine-tuned classifiers (Ibrahim 7). Simple classifiers are machine-learning algorithms that attempt to find subtle differences between machine-generated and human-written texts, although texts generated by larger LLMs become more difficult to detect (Ibrahim 7). Zero-shot classifiers make use of an existing generative LLM (e.g. GPT-2) in an attempt to classify text generated by itself or similar language models, but these are generally less effective than simple classifiers (Ibrahim 8). Finally, fine-tuned classifiers are similar to zero-shot classifiers in that they both take advantage of pre-trained generative models, but these classifiers, such as RoBERTa and GROVER, receive additional training to detect machine-generated text (Ibrahim 8). These are easily the most effective and promising resources for detecting LLM-assisted plagiarism, and are generally used to power the majority of commercial AI detection software (Ibrahim 9).

While fine-tuned classifiers seem promising, the efficacy of these products is controversial, especially when they are used to support accusations of academic misconduct. One study attempted to discern the efficacy of two RoBERTa-based classifier products—GPT Output Detector Demo and Crossplag AI Content Detector. They tested these products with 120 human-written essays and 120 ChatGPT-generated essays, finding that these detectors were able to discern between the two with an average of 89% accuracy, also noticing that Crossplag was better at detecting machine-generated texts while GPT Output Detector was more sensitive to human-written texts (Ibrahim 22). Though impressive, this margin of error is significant when

academic consequences are at stake. The researchers concluded that while these detectors were capable of accurately identifying AI-generated texts, educators should not rely solely on their output, instead incorporating other approaches such as interviewing students about their work or comparing their out-of-class work with samples of their in-class writing (Ibrahim 25). This reinforces the argument that human involvement is critical to ensure fair assessments.

Another study focused on the fact that students who submit AI-generated writing are more likely to attempt to obfuscate its origin—something known as an adversarial attack (Fishchuk and Braun 862). They identified three common types of adversarial attacks: prompt engineering (e.g. asking an LLM to generate more humanlike output), hyperparameter tweaking (changing different technical parameters of the model, such as temperature and presence penalty, which control the level of randomness in the output), and post processing (e.g. replacing and inserting unusual characters, translating text into a different language and then back into English, and paraphrasing) (Fishchuk and Braun 864). They then tested these attack strategies against a variety of AI detection programs, including GPT-2 Output Detector, the OpenAI Text Classifier, and Turnitin, finding that while hyperparameter tweaking was an effective strategy, prompt engineering and post processing were slightly less effective (Fishchuk and Braun 867). Thus, they concluded that while no detection software is completely immune to adversarial attacks, the latest generation of neural text detectors including Copyleaks and GPTZero are becoming increasingly robust (Fishchuk and Braun 871). Even so, researchers emphasize the need for cautious implementation, warning against overreliance and pointing out the serious implications of false accusations, especially for non-native English speakers (Fishchuk and Braun 872). Thus, detection must be paired with careful, context-sensitive evaluation.

Similarly, a different study also evaluated the performance of publicly available commercial AI detection software, although this time the researchers found that these tools were neither accurate nor reliable, especially when faced with the adversarial attacks mentioned in the previous paper. They tested 14 AI detection solutions, including GPTZero, GPT-2 Output Detector, ZeroGPT, and Turnitin, finding that they may not be as reliable as they claim—all of them scored below 80% accuracy and only 5 scored over 70%, and approximately 50% of AI-generated texts that employ adversarial attacks would likely be misattributed to humans (Weber-Wulff et al. 25–26). Again, the authors of this study also emphasized the grave danger of making false accusations of plagiarism, especially because of the propensity of these tools to give false positives. Researchers found that over half of the tools had a non-zero risk of false positives, with one tool (GPTZero) having a false positive risk of 50%, therefore arguing that these tools are unsuitable for the academic environment (Weber-Wulff et al. 20). Because these tools do not provide clear explanations or evidence for their results, the researchers advocate for a prevention-based model, emphasizing education and awareness over punitive action (Weber-Wulff et al. 26–27). This supports the broader argument that these automated solutions are insufficient alone.

While AI detection tools have made significant progress, several challenges hinder their effectiveness. The adaptability of large language models, ethical concerns surrounding their use in academic settings, and computational limitations present obstacles to reliably identifying AI-generated plagiarism. One of the most significant of these challenges is the fact that AI itself is constantly improving. As models become more advanced, they generate text that is increasingly coherent, contextually aware, and stylistically diverse, making it harder to differentiate from human writing (Wu et al. 281). Unlike earlier AI-generated texts, which often contained

repetitive phrasing or unnatural sentence structures, modern LLMs produce content that mimics the nuances of human authors. Beyond these improvements in AI, students seeking to evade detection can employ a variety of adversarial techniques to further obscure the origin of AI-generated content. As Fishchuk and Braun identify, three common strategies include prompt engineering, hyperparameter tweaking, and post-processing techniques (Fishchuk and Braun 864). Prompt engineering involves crafting detailed and specific instructions that guide an LLM toward producing more human-like writing. Hyperparameter tweaking, such as adjusting the "temperature" of an LLM's output, can modify the level of randomness, resulting in responses that are less robotic. Finally, post-processing techniques, including paraphrasing with synonym replacement tools or using machine translation, can further distort detectable AI markers. The success of these methods highlights the metaphorical "arms race" between AI-generated plagiarism and efforts to detect it. To meet this challenge, institutions must incorporate both technological and human strategies, continuously adapting to new developments in the field of AI.

Furthermore, the widespread use of AI detection tools in academic settings raises ethical and pedagogical concerns. One major issue is the risk of false positives, where AI detection software incorrectly labels human-written work as machine-generated. One previously-mentioned study found that many detection tools exhibited false positive rates as high as 50%, disproportionately affecting students whose writing style deviates from traditional norms (Weber-Wulff et al. 20). This is particularly concerning for non-native English speakers, whose writing may be mistakenly flagged due to structural differences from conventional English (Fishchuk and Braun 872). False accusations of AI plagiarism carry severe consequences, including academic penalties and damage to students' reputations. Another concern is the

overreliance on automated detection tools in academic integrity policies. Some educators and institutions treat AI detection results as definitive proof of misconduct, neglecting the possibility of errors or ambiguous cases. Blind trust in these tools risks unjustly penalizing students while arguably fostering a climate of mistrust between educators and their students. Researchers argue that AI-assisted plagiarism detection should be supplemented with human judgement, such as reviewing students' writing processing, conducting oral interviews, or comparing in-class and out-of-class writing samples, rather than serving as the sole basis for accusations (Ibraham 25). Ultimately, the stakes involved in academic misconduct allegations demand an approach that blends both automated detection and human dialogue.

Overall, while AI-assisted plagiarism detection has made notable strides in the past few years, it remains an evolving challenge due to the rapid advancement of large language models and the increasing sophistication of adversarial techniques. Despite the promise shown by fine-tuned classifiers and commercial AI detection programs, a variety of studies have highlighted their limitations, particularly regarding the risk for false positives and their vulnerability to manipulation. Ethical concerns, including the risk of wrongful accusations and biases against non-native English speakers, further complicate their use in academic settings. Given these challenges, educators and institutions must adopt a balanced approach that integrates AI detection tools with human judgement and prioritizes education on responsible AI usage. Rather than relying solely on detection-based strategies, fostering critical thinking, ethical awareness, and transparent discussions about generative AI's role in academia will be essential in addressing the complexities of LLM-assisted plagiarism.

Works Cited

Fishchuk, Vitalii, and Daniel Braun. “Robustness of Generative AI Detection: Adversarial Attacks on Black-Box Neural Text Detectors.” *International Journal of Speech Technology*, vol. 27, no. 4, Dec. 2024, pp. 861–74. *EBSCOhost*, https://doi.org/10.1007/s10772-024-10144-2.

Ibrahim, Karim. “Using AI-Based Detectors to Control AI-Assisted Plagiarism in ESL Writing: ‘The Terminator versus the Machines.’” *Language Testing in Asia*, vol. 13, Jan. 2023. *EBSCOhost*, https://doi.org/10.1186/s40468-023-00260-2.

Weber-Wulff, Debora, et al. “Testing of Detection Tools for AI-Generated Text.” *International Journal for Educational Integrity*, vol. 19, no. 1, 1, Dec. 2023, pp. 1–39. *edintegrity.biomedcentral.com*, https://doi.org/10.1007/s40979-023-00146-z.

Wichuta Chompurach, et al. “OpenAI ChatGPT vs Google Gemini: A Study of AI Chatbots’ Writing Quality Evaluation and Plagiarism Checking.” *English Language Teaching Educational Journal*, vol. 7, no. 2, Jan. 2024, pp. 90–108.

Wu, Junchao, et al. “A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions.” *Computational Linguistics*, vol. 51, no. 1, Mar. 2025, pp. 275–338. *EBSCOhost*, https://doi.org/10.1162/coli_a_00549.