Large Language Models As Zero-Shot AI-Generated Text Detectors for Academic Writing

DANIEL ALLAN, Vassar College, USA
WILLIAM MILLER, Vassar College, USA
NORA PHELAN, Vassar College, USA

Abstract. In this study, we determine the validity of several popular LLMs as AI detection tools for undergraduate-level student writing. As previous work has focused on the development of new AI text classifiers, we extend the field by testing LLM capability. We generate adversarial examples with LLMs, and using predetermined prompts, we measure the accuracy with which different LLMs identify senior undergraduate-level human academic writing. We also compare these models against a baseline popular online AI text detector.

1 Introduction

Take a peek inside most modern classrooms, and you'll bear witness to an ever-evolving arms race: large language model (LLM) text generation against AI text detection. Since the explosion in popularity of ChatGPT, students have been relying on LLMs to produce classwork that appears human-written. In response, many teachers and professors have turned to online AI text detection tools, sometimes turning to the very same LLMs to classify student submissions.

Previous literature has largely focused on developing better techniques for AI text classification. These models tend to be small and efficient, and vary wildly in efficacy. This is not necessarily useful for teachers who need to test only a classes worth of essays. A few authors have tackled LLMs as AI-generated text classifiers. However, most of these experiments were small in scale, focused largely on the capability of ChatGPT, and have become dated.

In this study, we tackle the validity of several popular LLMs for zero-shot classification of academic writing as human-written or AI-generated. We find a dataset of senior undergraduate-level papers (n=100). We also generate a dataset of essays with ChatGPT (n=100), matching each one to a paper from our original dataset. We test each of these papers on a set of popular LLMs to determine accuracy on human-written versus AI-generated essays. We also compare against the baseline of a popular online model.

We find that validity of the tested LLMs was extremely variable, but the most accurate (Claude Sonnet 4.5) was extremely effective. We also find that the online baseline model was similarly accurate, which may reduce the utility of LLMs in this area. Importantly, we find that all tested LLMs were unlikely to misclassify human-written essays as AI-generated.

2 Related Works

Prior work on AI-generated text detection has often focused on two approaches: feature-based classifiers and probabilistic methods. In 2023, Nguyen et al. [4] created a text detection model that uses feature based machine learning. They extracted about 30 features to train simple classifiers. Even simple classifiers trained on these features can achieve high accuracy in categorizing text. The authors think that simpler models are preferable for explaining classifications. However, feature-based models are weak against paraphrasing and adversarial writing. More work needs to be done to create robust yet explainable models.

Daniel Allan, dallan@vassar.edu; William Miller, wmiller@vassar.edu; Nora Phelan, nphelan@vassar.edu.

In the same year, Mitchell et al. also demonstrated impressive results in their "DetectGPT" paper[3]. Using just log probabilities computed from models, they were able to predict the model that originated passages of text with astonishing accuracy. This model is both simple and effective, but there remain fears that improved models or better prompting could evade detection by such means.

Other projects have researched the efficacy of existing publicly available models. Chaka Chaka [2] performed a review of papers testing online AI detection tools. Most articles tested text generated by ChatGPT, and many articles reported the inconsistencies of AI detectors. This paper recommended that AI detectors be combined with human reviewers to reduce the margin of error.

Some work has been done to study ChatGPT's ability to identify AI-generated text. In early 2024, Bhattacharjee and Liu [1] tested whether different models of ChatGPT could successfully determine if a provided text was produced by a human or an LLM. The study found that GPT-3.5 never classified GPT-3.5 generated articles as AI, and GPT-4 only correctly classified the writing as AI about 40% of the time. GPT-4 correctly classified over 95% of articles written by older language models but misclassified almost all human articles. GPT-3.5 classified human articles correctly about 90% of the time, and had mixed results against articles generated by older language models. Overall, GPT-3.5 was surprisingly more accurate in distinguishing between human-written and AI-generated text. GPT-4 was apparently very sensitive to noise and dataset artifacts, possibly leading to its poor performance. The researchers suggested that new, large models could be effective for identifying text generated by older models.

The subject of our work is closely related to these past papers. Many have tested the efficacy of ChatGPT, private models, and online AI detectors. However, their works have not heavily tested the classification abilities of publicly available LLMs outside of ChatGPT. Inspired by Bhattacharjee and Liu's article [1], our work tests these new, large models, keeping in mind how the cleanliness of data may affect the results.

3 Dataset and Resources

Each model is tested on a dataset of 100 human-written and 100 AI-generated essays. Each human-written essay was selected randomly from the University of Michigan's *Michigan Corpus of Upper-Level Student Papers*, after filtering to only include senior-level undergraduate student's who are native English speakers, and removing any papers with problematic non-textual features such as figures, charts, or graphs. These essays span a diverse set of fields including Biology, English, Philosophy, and Political Science. Each essay ranges from 1000 to 6000 words and all include cited references. For each student-written paper, we employ GPT-4o to generate an equivalent essay, as detailed in our Methods section. Because our dataset is on the smaller side, we are able to more easily look over the essays to catch any artifacts that could hamper the performances of the models.

In this study, we test several popular LLMs: OpenAI's GPT-4o, Google's Gemini 2.5 Flash, Anthropic's Claude Sonnet 4.5, and High-Flyer's DeepSeek-R1. For each model, we ensure that no previous prompts would be saved. We also test our dataset on the online AI-text detector GPTinf to act as a baseline for our research.

4 Methods

We follow this same method for generation and classification of each essay by the tested LLMs.

4.1 Setup

For each model, we employ a specific method to ensure that the correct model is used and no prompts are saved between steps and iterations of our process.

GPT-4o: We use the ChatGPT interface in logged-in, temporary chat mode. After each prompt and response, we reload the page.
Gemini 2.5 Flash: We use the Gemini interface in logged-out, Fast chat mode. After each prompt and response, we reload the page.
DeepSeek-R1: We use the DeepSeek interface and turn on DeepThink and Search tools. After every prompt and response, we delete all chats (Settings → Data → Delete all chats).
Claude Sonnet 4.5: We use the Claude interface in logged-in, Sonnet 4.5, incognito chat mode. After every prompt and response, we reload the page.
GPTinf: We use the GPTinf detector interface. After every prompt and response, we reload the page.

4.2 Essay Generation

We generate each essay to correspond with a human-written essay from our dataset. For each human essay, we extract key features: word count, discipline, paper type (e.g. argumentative essay, report, etc.), and title. We then input this prompt into GPT-4o to obtain an AI-generated essay:

> Generate a roughly [word count] word [discipline] [paper type] titled [title]. Include in-text citations and a references section. If sections have a header, use only text and no numbers. Do not use bullet points.

4.3 Essay Classification

To classify an essay with an LLM, we copy and paste a prompt followed by the full text of the essay into the interface of the LLM:

> Based on the contents of this essay, classify this essay as (0) human-written or (1) AI-generated. Include a confidence score from 0 to 1, with 0 representing complete confidence that this essay was written by a human and 1 representing complete confidence that this essay was generated by AI. [essay]

If the confidence score is flipped from what is expected (e.g. model states that it is very confident in its prediction of human, but confidence score is 0.95), then we subtract it from 1. This handles the case that the LLM misinterprets the prompt and thinks that a higher confidence score necessarily means that it is more confident in its prediction.

5 Results

The accuracy over 100 human essays and 100 AI essays of each model's classification is recorded, along with the confidence score provided by each classifier. The accuracy and confidence on human essays are calculated separately from those on AI essays.

We observe that GPT-4o classified human essays with perfect accuracy, but struggled slightly to classify AI essays. It reported a similar confidence score for both human-written and AI-generated essays.

Gemini was more confident and less accurate overall and had relatively low accuracy on human-written essays, given how well other models performed in that field.

Table 1. Accuracy and confidence of each model when predicting essay classification. Categories are split by essay type: human-written essays (H-E) and AI-generated essays (A-E). All results are rounded to the third decimal.

Classification Model	Accuracy on H-E	Correct Confidence on H-E	Incorrect Confidence on H-E	Accuracy on A-E	Correct Confidence on A-E	Incorrect Confidence on A-E	Overall Average Prediction	Overall Average Confidence
GPT-4o	1.000	0.769	#N/A	0.870	0.753	0.743	0.435	0.460
Gemini 2.5 Flash	0.850	0.842	0.833	0.860	0.929	0.871	0.505	0.538
DeepSeek-R1	0.940	0.845	0.775	0.180	0.814	0.837	0.120	0.236
Claude Sonnet 4.5	1.000	0.881	#N/A	0.980	0.858	0.850	0.490	0.481
GPTinf	1.000	0.998	#N/A	0.980	0.785	0.965	0.490	0.386

DeepSeek mostly classified human essays correctly, but strongly tended toward also classifying AI-generated text as human-written. This bias is visible in its average prediction and in its average confidence.

Claude was the best performing LLM, not only perfectly classifying all human-written essays, but also achieving 98% accuracy for AI-generated text.

GPTinf consistently classified human essays correctly and identified AI-generated text roughly 98% of the time, making the same number of mistakes as Claude.

6 Conclusions and Limitations

We draw several key conclusions from these results. First, the high accuracies of GPT-4o and Claude Sonnet 4.5 indicate that they could be useful for AI-generated text classification. They were effective at identifying AI-generated papers (.87 and .98 respectively), and never falsely classified a human-written essay. This is particularly important because a falsely identified student paper has more potential consequences than a falsely identified AI text.

The average prediction and average confidence of almost all models except Gemini-2.5 Flash falling under 0.5 indicates a possible bias towards labeling texts as human-written to avoid harmful outcomes. This is especially evident in the results from DeepSeek-R1, as almost all AI-generated essays were classified as human-written.

The efficacy of GPTinf demonstrates that specialized models may still be better suited for AI-generated text classification problems. However, this model was actually slightly slower than the tested LLMs, and did not outperform Claude Sonnet 4.5.

We glean from these results that there may be a place for LLMs in academic text classification. However, we recommend that work be done to improve the interpretability of classification responses before LLMs are relied upon in any situation with potential repercussions. Any current LLM results should be subject to human review before action is taken.

In this study, the labor-intensive nature of our essay generation and classification prevented us from developing a larger dataset. In the future, similar studies should be repeated with a wider dataset including generated texts from from other LLMs. This will help to address any biases in the classification and clarify results by increasing sample size. We also suggest that a wider variety of academic text be tested, as LLMs may perform differently on high school, undergraduate, and professional writing.

Furthermore, our study failed to compare against many popular AI detectors that required a purchase to test past a maximum word count. We recommend that studies compare LLMs against these AI detectors, as well as other

state-of-the-art models such as Mitchell et al.'s DetectGPT [3]. This could provide a more comprehensive view of the large-scale validity of LLM academic text classification.

Finally, we suggest that future research investigate the effect of various prompts (both for generation and classification) on the accuracy of LLM classification.

References

[1] Amrita Bhattacharjee and Huan Liu. 2024. Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text? *ACM SIGKDD Explorations Newsletter*, 25, 2, (Mar. 26, 2024), 14–21. doi:10.1145/3655103.3655106.

[2] Chaka Chaka. 2024. Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning and Teaching*, 7, 1, (Feb. 6, 2024), 115–126. doi:10.37074/jalt.2024.7.1.14.

[3] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: zero-shot machine-generated text detection using probability curvature. (July 23, 2023). arXiv: 2301.11305[cs]. doi:10.48550/arXiv.2301.11305.

[4] Trung T. Nguyen, Amartya Hatua, and Andrew H. Sung. 2023. How to Detect AI-Generated Texts? In *2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. 2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). (Oct. 2023), 0464–0471. doi:10.1109/UEMCON59035.2023.10316132.