Who is speaking? Discriminating Artificial and Human-Generated Text with A Natural Language Processing Approach

Mingsheng Sun
University of Minnesota
CSCI5541 NLPer
sun00544@umn.edu

Yutong Sun
University of Minnesota
CSCI5541 NLPer
sun00545@umn.edu

Moyan Zhou
University of Minnesota
CSCI5541 NLPer
zhou0972@umn.edu

Abstract

The rapid advancement of AI technology, particularly ChatGPT and GPT-4, has raised concerns about the ethical implications and potential misuse of AI-generated texts. This study examined the accuracy of detection models, such as GPT-2 Output Detector Demo, DetectGPT, and GPTzero, and human judgements in discerning between human-written and AI-generated content. Our findings revealed that the effectiveness of detection methods varies based on the dataset and language model, and text processing techniques can significantly impact accuracy. Human judgements, on the other hand, were found to be inconsistent and inaccurate. The results underscore the need for more robust, adaptive, and comprehensive detection methods to maintain the integrity of academic, professional, and informational environments in the era of AI-generated content.

1 Introduction

The rapid advancement of artificial intelligence (AI) technology, particularly with the emergence of ChatGPT and GPT-4, has revolutionized the field of natural language processing. Significantly reducing the costs associated with generating human-like text, making it accessible to a broader audience, AI models have demonstrated its potential to cater to a wide range of user requirements, from daily communication to professional documentation (Goyal and Saxena, 2021).

While the benefits of AI-generated texts are numerous, concerns have arisen regarding the ethical implications of their use. The pervasive nature of these technologies raises questions about potential misuse and the blurring of lines between human and machine-generated content. A prominent example of this is the decision by some universities to ban students from utilizing ChatGPT for completing assignments, citing concerns over academic integrity (Mearian, 2023). This highlights the need to address the moral and ethical considerations associated with AI-generated text.

In order to mitigate these concerns and ensure the responsible application of AI-generated content, it is essential to develop robust methods for accurately and reliably distinguishing between AI-generated and human-generated text. This not only helps maintain the integrity of academic and professional environments, but also safeguards against the spread of disinformation and malicious use of AI-generated content (Loh, 2023).

Enhancing the capacity to detect AI-generated text has broader implications beyond the realm of academia. With the proliferation of AI-generated content in various spheres, such as journalism, social media, and marketing, the ability to differentiate between human and machine-generated content becomes increasingly important. The implications of undetected AI-generated content can lead to the erosion of trust in information sources and, ultimately, have significant consequences on public opinion and decision-making.

These lead to the development of detection models such as GPTZero (GPTZero, 2023) and DetectGPT (Mitchell et al., 2023). However, the accuracy of these tools remains uncertain (Krishna et al., 2023). In this paper, we investigate the accuracy of these detection models across various data domains and compare their performance with human judgements. Our study reveals that both detection tools and human judgements exhibit inconsistency and inaccuracies.

Therefore, many detection models have been implemented, either in black-box detection or white-box detection, such as GPTZero and DetectGPT (Tang et al., 2023). However, accuracy of these tools remain ambiguous. In this paper, we examine the accuracy of detection models across different data domains. We also integrate human detection capabilities into this discussion, adding another critical dimension to our understanding of

AI-generated text. Our study aims to address the following research questions:

RQ1: What is the accuracy of GPT detection tools when distinguishing human-written and AI-generated text?

RQ2: Do human agree with each other when they make decisions about text source?

RQ3: What is the accuracy for human judgement when distinguishing human-written and AI-generated text?

The rest of the paper is organized in the following structure: We first provide an overview of related work, and then we describe our methodology. We also present our results and discuss the implications of our findings. We conclude by outlining potential directions for future research and next steps.

2 Literature survey

One study (Clark et al., 2021) investigates non-experts' ability to distinguish between human and machine-generated text (GPT-2 and GPT-3) in three domains: stories, news articles, and recipes. Untrained evaluators were found to perform at random chance levels when identifying GPT-3-authored text. The researchers tested three approaches for quickly training evaluators to better identify GPT-3-authored text, including detailed instructions, annotated examples, and paired examples. Although evaluators' accuracy improved up to 55%, it did not significantly improve across all domains. Our research team plans to draw inspiration from the referenced study. We aim to divide our research scope into several areas, such as academia, daily life, poetry, and prose, to investigate the differences between AI-generated and human-generated texts within these domains. Texts will be classified into five levels: human-generated, possibly human-generated, co-generated by AI and humans, possibly AI-generated, and AI-generated.

Another study (Ma et al., 2023) explores the gap between AI-generated and human-written scientific text. The researchers collected scientific text from the OpenAI API, designed fine-grained prompts, and conducted human evaluations to analyze the ability to distinguish AI-generated content. They constructed a feature description framework to analyze differences in syntax, semantics, and pragmatics, and used a logistic regression model for analysis. They also fine-tuned the RoBERTa large OpenAI detector for detection purposes. Our research group intends to adopt the methodology employed in the referenced study. We plan to analyze the differences between AI-generated texts and human-generated texts by examining factors such as keyword usage, grammatical errors, sentence fluency, and logical coherence.

The study (Köbis and Mossink, 2021) investigates people's ability to distinguish between human-written and algorithm-generated poetry and their preferences for either. Two experiments were conducted using GPT-2 generated poems alongside human-written poems. Participants failed to reliably detect algorithmically generated poems in the Human-in-the-loop treatment, but succeeded in the Human-out-of-the-loop treatment. People revealed a slight aversion to algorithm-generated poetry, regardless of whether they knew about its origin. This study offers us a novel perspective. The capabilities of NLP algorithms to mimic human-like creative texts are steadily increasing. We can differentiate between AI-generated and human-generated texts by examining the sentiment conveyed in the language and the presence of sentences that demonstrate creativity.

3 Broader impact

The broad impact of the technology to distinguish between human-generated text and AI-generated text extends to various fields and sectors, which greatly contributes to the development of a secure, accountable and transparent digital environment. Developing this technology will empower users. Providing users, and groups of people, with the ability to identify AI-generated content enables them to make informed decisions about the credibility and authenticity of the information they receive, promoting critical thinking and discernment in the digital age. Secondly, it also protects intellectual property and prevents plagiarism, ensuring that the public maintains a positive view of AI technologies. This likewise contributes to the development of AI technology, which encourages AI developers to focus on responsible and ethical AI applications and reduces the potential for harmful consequences or misuse of the technology. In conclusion, the broader implications of distinguishing human-generated text from AI-generated text go beyond recognition. This technology also has far-reaching implications in terms of promoting the security and development of AI technologies, and the development of a transparent digital ecosystem.

4 Method

4.1 Dataset

The dataset used in this study combines two parts to provide a comprehensive evaluation of the performance for detection models and human judgement.

GPT Wiki Intro: This dataset contains Wikipedia introductions and GPT (Curie) generated introductions for 150k topics. The inclusion of this dataset is essential for two reasons. First, considering that the training datasets of GPT-2 Output Detector Demo and DetectGPT are from GPT-2, we need to ensure that the accuracy of these detection tools remains reasonably high without any processing when confronted with text generated by GPT (Curie). Second, it allows for a thorough evaluation of the detection methods across a large volume of text from various domains, ensuring their generalizability and robustness.

ChatGPT (GPT-4) Generated Passages: The second part of the dataset consists of 20 paragraphs from different domains generated by author using ChatGPT (GPT-4). The prompt used in this case is "Please provide 20 passages from different domains with a length of about 100 words." Domains included are Technology, Climate Change and Sustainability, Literature and Writing, Psychology and Mental Health, Economics and Finance, Education and Learning, etc. The choice to include text generated by the GPT-4 model allows us to assess the detection tools' capability to identify artificially generated text from the latest, state-of-the-art language model.

4.2 RQ1: Detection model's performance

In this section, we describe detection methods, and treatments used in our study to discriminate between artificial and human-generated text. The ultimate goal is to determine the effectiveness of these detection methods and their susceptibility to manipulation using rewriting tools and GPTZZZs (Declipsonator, 2023). Three different detection methods were employed in our study:

GPT-2 Output Detector Demo: This model was obtained by fine-tuning a RoBERTa model with the outputs of the 1.5B-parameter GPT-2 model. **DetectGPT**: DetectGPT is a general-purpose method for using a language model to detect its own generations. It defines a new curvature-based criterion for judging if a passage is generated from a given LLM.

GPTzero: GPTzero is another detection method that checks for perplexity and burstiness in text to determine if it is artificially generated or produced by a human, and they claim on their website that GPTZero is the world's first artificial intelligence detector with over 1 million users.

These three detection tools were chosen because the GPT-2 Output Detector Demo originated from OpenAI, lending credibility to its effectiveness and robustness in text detection tasks. DetectGPT, on the other hand, was chosen for its general applicability and its innovative curvature-based criteria. GPTzero was finally chosen because of its focus on perplexity and burstiness to determine the origin of texts and its wide user base. By incorporating these three different detection tools, the study benefits from a comprehensive and diverse evaluation of its capabilities, allowing a comprehensive analysis of its effectiveness in distinguishing between human and human-generated texts. Initially, the accuracy of these detection tools were measured without applying any treatment to the dataset.

To evaluate the resilience of the detection methods, we performed two different treatments on the text segments to attempt to fool the detection tools:

Paraphrasing: The first processing method is to use a text Paraphrasing tool to modify the original paragraphs generated by GPT-4. It is the most common processing method normally used to fool the detection tools while maintaining the meaning and structure of the content.

GPTZZZs: GPTZzzs work by downloading a dictionary of synonyms and replacing some words with their counterparts. While this works in most cases, sometimes the synonyms can be so strange that they need to be corrected by a human.

After performing these treatments on the dataset, we separately measured the accuracy of the detection tools again to assess their effectiveness in the face of manipulation. We compare the performance of these three detection methods by getting the accuracy of the two datasets before and after processing. This comparison allows us to determine the robustness of each method and to what extent they can be deceived by different treatments.

To determine the likelihood that the input text was generated by an AI model, we developed an automated web scraping program that interacts with three different online AI text detection tools . We used Selenium, a popular web testing library, to programmatically control a web browser and simulate user interaction. The program uses the

ChromeDriver executable file to control a Google Chrome browser instance. It leverages the Selenium WebDriver API to locate and interact with web elements by ID, name, or XPath attribute. To ensure that the required elements exist before attempting to interact with them, the program uses the WebDriverWait class to manage timeout issues. By running this program, we are able to obtain detection results from all three tools simultaneously after a single input. By adopting this automated approach, we are able to collect and analyze AI text detection results from multiple sources, greatly improving detection efficiency.

4.3 RQ2& RQ3: Human judgement's Accuracy

In order to evaluate the performance and decision making process of humans distinguish text that is generated from AI and human, we sent surveys to students at the University of Minnesota. The survey consists of 20 text paragraphs asking them to decide whether it is human-written or AI-generated. These 20 randomly selected 20 examples (including 10 human-written and 10 AI-generated texts) from **GPT Wiki Intro** dataset from Huggingface. Additionally, and 1 question about confidence level and 1 open-ended question about how they made the judgement. With the responses we received, we calculated both Fleiss' Kappa and the accuracy. The results are reported in the next section.

5 Result

5.1 Detection model's performance

In this section, we present the accuracy results for the different detection tools across the two datasets, Wikipedia GPT(Curie) and GPT-4 generated passages.

Without any processing, GPT-2 Output Detector Demo was the most accurate in detecting GPT (Curie) generated introductions, followed by DetectGPT and GPTzero. It is worth noting that GPTzero had the lowest accuracy for this dataset. For GPT-4 generated text, GPTzero outperformed the other two detection methods significantly, achieving the highest accuracy in detecting GPT-4 generated text. Both GPT-2 Output Detector Demo and DetectGPT showed relatively lower accuracy for this dataset. When comparing the detection tools' performance across the two datasets (as illustrated in Figure 1), we can observe that the

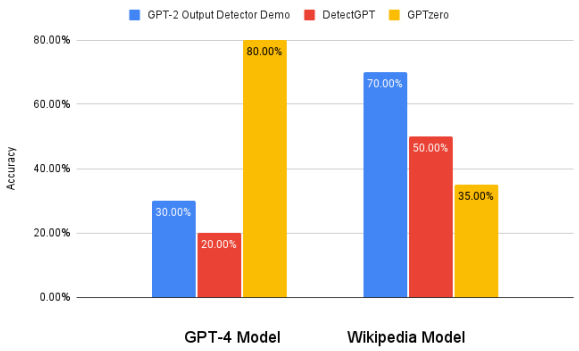

Figure 1: Accuracy of the two datasets for different detection tools

GPT-2 Output Detector Demo performs well on the Wikipedia GPT dataset but struggles with the more advanced GPT-4 generated passages. DetectGPT demonstrates modest accuracy for both datasets, with a slightly better performance on the Wikipedia GPT dataset. In contrast, GPTzero shows a remarkable improvement in accuracy when detecting GPT-4 generated text, despite its lower accuracy on the Wikipedia GPT dataset.

Subsequently, we present the accuracy results for the different detection tools across the two datasets, Wikipedia GPT (Curie) and GPT-4 generated passages, both before and after applying two different processing methods: Paraphraser and GPTZZZs. The findings are illustrated in Figure 2 (Wikipedia GPT dataset) and Figure 3 (GPT-4 generated passages dataset), which provide a visual comparison of the detection methods' performance on both datasets and after applying the processing techniques.

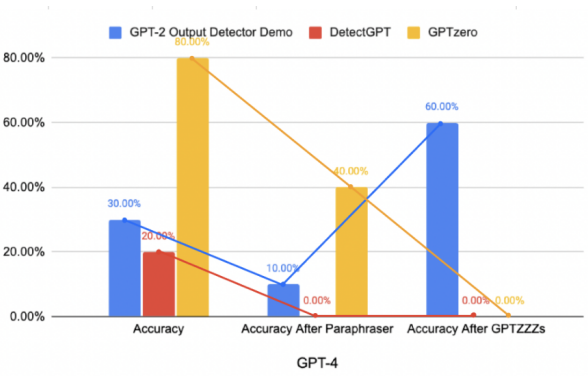

Figure 2: Accuracy of GPT-4 model after applying different processing methods using three detection tools

As seen in Figure 2, for Wikipedia GPT, the accuracy of all three detection tools decreases after processing the text with both Paraphraser and

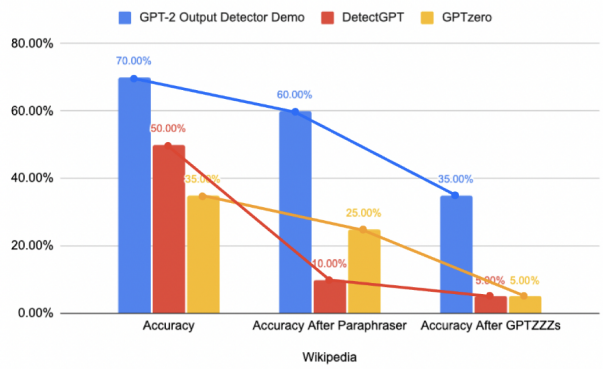

Figure 3: Accuracy of Wikipedia model after applying different processing methods using three detection tools

GPTZZZs, with GPTZZZs having a more significant impact on the detection accuracy. The reason is that GPTZZZs operate by replacing some words with their counterparts, which can lead to a significant increase in the perplexity of the sentence, resulting in a sharp decrease in the accuracy of GPTzero, which uses perplexity as a basis for judgment. Futhermore, as depicted in Figure 3, for GPT-4 the accuracies of DetectGPT and GPTzero drop dramatically after applying both processing methods. However, GPT-2 Output Detector Demo's accuracy increases after applying GPTZZZs, indicating that the processing method might introduce patterns or features that this detector can identify. The results indicate that both Paraphraser and GPTZZZs have a considerable impact on the detection accuracy, making it more challenging for the tools to discriminate between artificial and human-generated text. it becomes apparent that the effectiveness of these detection methods varies depending on the dataset, language model, and text processing techniques.

5.2 Survey result

We received 25 responses in total from the survey. However, we exclude 1 response from our analysis due to incomplete survey. The Fleiss' Kappa is 0.04 with p-value of 0.016, which indicates significant of 0.04.

Among 24 participants, on a scale from 1 to 10 to indicate confidence for their answer to be correct, 4 of them answered with a number that is larger than 6, thus we categorized them as having "high confidence level." 15 participants answered a number between 4 and 6, and we labeled them as having "medium confidence level." For participants who answered a number less than 4, we considered them as having "low confidence level." Figure 4 shows a pie chart of the distribution of high, medium and low confidence level.

We calculated the accuracy to each question, and

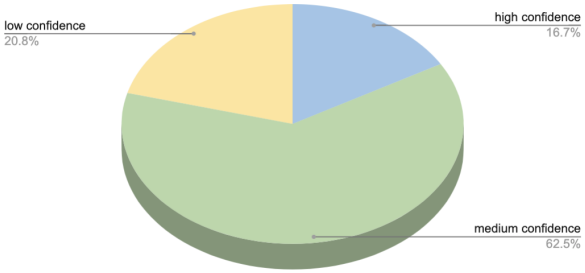

Figure 4: Confidence level distribution

the result is shown in Figure 5. We observed that the overall accuracy is relatively inconsistent for the 20 questions. While the accuracy between medium and low confidence level shows no obvious difference, we note that high confidence level exhibit a large variation in the trend as the accuracy shows extremes: either extremely high accuracy (where all 4 participants answered correctly) or extremely low accuracy (where all 4 participants answered incorrectly).

Figure 6 compares accuracy between human judgements and detection models. As shown in the graph, though neither gptzero, detectGPT or GPT-2 output detector is perfectly accurate, all of them outperform human judgement, whose overall accuracy is less than 0.5. We also performed a qualitative analysis to the open-ended question which asks the participants how they make decision in terms of distinguishing human-written and AI-generated texts. Common themes include sentence structure,

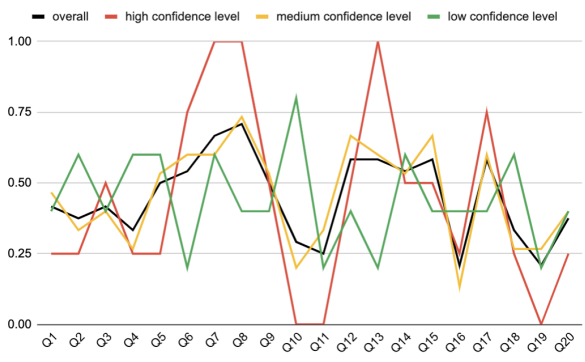

Figure 5: Accuracy for each question given high, medium and low confidence levels

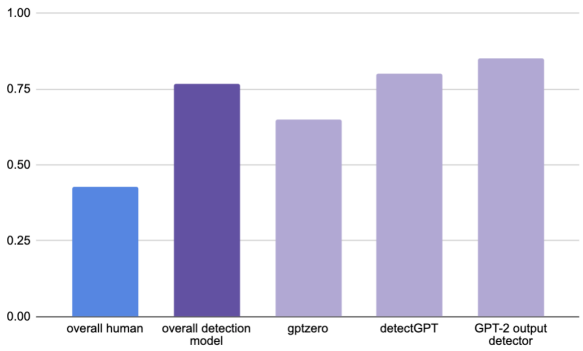

Figure 6: Comparison between accuracy of human and detection models

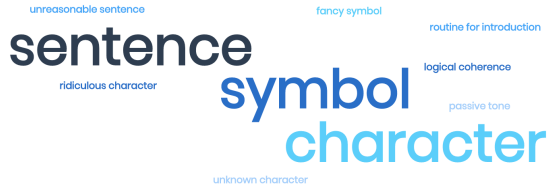

Figure 7: Word cloud generated from open-ended question

uncommon symbols or characters. In addition, they also mentioned logical coherence, passive tone, routine for introduction.

6 Discussion

The results of this study illuminate the capabilities and limitations of three detection methods—GPT-2 Output Detector Demo, DetectGPT, and GPTzero—in discriminating between artificial and human-generated text. Our findings indicate that the detection methods' effectiveness varies depending on the dataset, language model.

The results also demonstrate that text processing techniques, such as Paraphraser and GPTZZZs, can have a substantial impact on detection accuracy. This poses a challenge for detection methods, as obfuscation techniques might be used to evade detection and potentially spread misinformation or other harmful content generated by AI models. Consequently, it is crucial to develop robust and adaptive detection methods that can account for such obfuscation techniques.

Given the performance of these detection tools, we also investigated whether humans are able to distinguish between artificial and human-generated text. To answer **RQ2**, the result shows that they do not agree with each other when they make decisions about text source. We also found that human judgement about distinguishing human-written and AI-generated text is inconsistent and inaccurate, regardless of their perception on confidence level.

To summarize, our results imply that there is no accurate way of distinguishing human-written and AI-generated text from GPT-4 (or GPT-3) output. Neither human judgement nor Detection models can tell the differences in a coherent manner.

6.1 Replicability and Datasets

For the replicability, the results of this study can be reproduced by others, provided they have access to the necessary resources, such as the GPT-wiki-intro dataset and the GPT-4 generated passages. Additionally, the detection methods—GPT-2 Output Detector Demo, DetectGPT, and GPTzero—are publicly available, which allows other researchers to test and validate our findings.

Our choice of datasets, GPT-wiki-intro and GPT-4 generated passages, reflects the need to understand the performance of detection methods across different language models and AI-generated content. These datasets may affect other researchers' choices of projects, as they demonstrate the importance of evaluating detection tools in diverse contexts.

6.2 Limitation

Dataset Regarding the AI detection tools portion of our study, it is important to note several limitations. Firstly, the dataset used in this study was relatively small, indicating the need for further evaluation using a larger, more diverse, and representative dataset. Secondly, due to the dataset's size, our study was not able to conduct a comprehensive analysis of the impact of text domains on accuracy, which is a crucial factor affecting the performance of detection tools.

Participants We surveyed University students, mainly because of the convenience and ensure that we will get responses. However, they level of expertise may explain some part of why they are not accurate in terms of distinguishing Wikipedia text.

7 Conclusion

This study has examined the accuracy of detection models, such as GPT-2 Output Detector Demo, DetectGPT, and GPTzero, in distinguishing between human-written and AI-generated text across different data domains. Our findings suggest that the effectiveness of these detection methods varies

depending on the dataset and language model. Additionally, we observed that text processing techniques, such as Paraphraser and GPTZZZs, can significantly impact detection accuracy, highlighting the need for robust and adaptive detection methods that can account for obfuscation techniques.

Furthermore, our analysis of human judgements demonstrates that individuals do not consistently agree with one another when determining the source of a text, and their ability to accurately distinguish between human-written and AI-generated text is limited. This finding underscores the challenges in maintaining the integrity of academic, professional, and informational environments in an era of pervasive AI-generated content.

Despite the limitations of our study, such as the relatively small dataset and the narrow participant pool, our findings contribute valuable insights into the current state of AI-generated text detection. Our study demonstrates the importance of evaluating detection tools in diverse contexts and highlights the need for further research using larger, more representative datasets and a broader range of participants.

In conclusion, our results emphasize that accurately distinguishing between human-written and AI-generated text remains a challenge for both detection models and human judgements. To address this issue, it is crucial to continue developing more robust, adaptive, and comprehensive detection methods that can effectively differentiate between human-generated and AI-generated content across various data domains. This will help ensure the responsible use of AI-generated content, preserve the integrity of academic and professional environments, and maintain trust in information sources in the digital age.

Future Work For next step, we want to send the survey to Wikipedia experts, namely, participants who have made edits on Wikipedia, and test whether their expertise knowledge improves human judgement accuracy.

The text processing techniques employed in this study (Paraphrase tool and GPTZZZs) are only two examples of potential confusion methods; future research could explore the impact of other techniques on detection accuracy and identify the most effective methods for spoofing detection tools.

References

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.

Declipsonator. 2023. https://github.com/Declipsonator/GPTZzzs. [link].

Somya Goyal and Arti Saxena. 2021. Creditworthiness assessment using natural language processing. *Advances in Computational Intelligence and Robotics*.

GPTZero. 2023. https://gptzero.me/team. [link].

Nils Köbis and Luca D Mossink. 2021. Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. *Computers in human behavior*, 114:106553.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense.

Erwin Loh. 2023. Chatgpt and generative ai chatbots: challenges and opportunities for science, medicine and medical leaders. *BMJ Leader*.

Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. Ai vs. human – differentiation analysis of scientific content generation.

Lucas Mearian. 2023. Schools look to ban chatgpt, students use it anyway.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts.