What Are We Detecting, Really? LLM-Generated Text Detection Remains an Unsolved Problem

Anonymous Author(s)
Affiliation
Address
email

Abstract

This position paper argues that, in most practical cases, it is not possible to accurately detect LLM-generated text. We consider that "LLM-generated text" refers to the content produced by LLMs through normal prompts. As implied by the names "LLM-generated text" and "human-written text", the difference lies in how they are produced, but in practice, we can only evaluate them based on the final output—the text itself—where there is often significant overlap between human- and machine-generated content.. The numerical results of LLM-generated text detection are often misunderstood and their significance is diminishing. The detectors can serve a purpose under specific conditions, whose results should only be used as a reference with greater caution rather than the decisive indicator.

1 Introduction

The rapid development of large language models (LLMs) has led to a rise of LLM-generated text, which has been observed across various fields, such as academia [Liang et al., 2024, Geng and Trotta, 2024], Wikipedia [Brooks et al., 2024, Huang et al., 2025], and numerous online texts [Sun et al., 2024, Liang et al., 2025]. The detection of LLM-generated text has attracted the attention of researchers, and many detectors have been proposed and studied [Yang et al., 2023, Wu et al., 2025]. Before starting our discussion, we want to clarify the following definition:

What exactly is "LLM-generated text"?

In fact, the term "LLM-generated text" is fairly new. Researchers also used expressions like "machine-generated text" or "AI-generated", as seen in Table 1. For simplicity, we use "LLM-generated text" to refer to the subject of study in this paper, as it is more precise than the other expressions. Apart from slight differences in terminology, the definition of "LLM-generated text" in most papers is quite broad, meaning the text can be produced in many ways using LLMs, like paraphrasing, translation, or generating long text from simple prompts. Therefore, we can consider that **"LLM-generated text" refers to the content produced by LLMs through normal prompts**.

The reliability of the detectors has also been widely discussed [Sadasivan et al., 2023, Chakraborty et al., 2024]. The indistinguishability between LLM-generated and human-written text is one of big challenges for LLMs [Kaddour et al., 2023]. Similar to the central question in Chakraborty et al. [2024]'s work, we explore the following key one:

Is it possible to detect the LLM-generated text in practice?

Literally speaking, LLM-generated text detectors need to cover all these different scenarios, but few detectors have tried to distinguish them [Cheng et al., 2025]. If we take into account the diversity of LLMs and prompts, as well as human-in-the-loop, the situation becomes even more intricate. Hence, we argue that, **in most practical cases, it is not possible to accurately detect LLM-generated text.**

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Table 1: Definition of LLM-generated text in different papers

Paper	Definition
Crothers et al. [2023]	*"Machine-generated text" is natural language text that is produced, modified, or extended by a machine.*
Kumarage et al. [2024]	*In this survey, we define AI-generated text as output produced by a natural language generation pipeline employing a neural probabilistic language model.*
Wu et al. [2025]	*LLM-generated Text is defined as cohesive, grammatically sound, and pertinent content generated by LLMs.*

While limitations of these detection methods have caused concern among researchers [Sadasivan et al., 2023, Liang et al., 2023, Doughman et al., 2024, Nicks et al., 2023, Saha and Feizi, 2025], they could be applied in diverse contexts. The emergence of the GPTZero platform is a good example, although it is currently unclear how frequently people use it. So the next question arises naturally:

Should we use these detectors?

We think that **the detectors can serve a purpose under specific conditions, whose results should only be used as a reference with greater caution rather than the decisive indicator.**

We will discuss them in detail in the following sections.

2 Detectors

Probably most people became aware of LLMs after the release of ChatGPT, but the research on detecting text generated by language models had started before that. For example, Gehrmann et al. [2019] proposed the GLTR tool to detect whether text was generated by models, with experiments involving GPT-2 [Radford et al., 2019] and BERT [Devlin et al., 2019]. Zellers et al. [2019] developed the Grover model to detect AI-generated fake news. Even GPT-3 [Brown et al., 2020] continued to face skepticism regarding its text-generation capabilities [Bender et al., 2021], making the detectors' performance unsurprising.

Another pioneering work by [Ippolito et al., 2019] demonstrated that humans have already encountered some difficulties in identifying texts generated by GPT-2. Later, Clark et al. [2021] found that untrained people at the time were not very good at recognizing text produced by GPT-3, and Wahle et al. [2022] noticed the similar situation for machine-paraphrased plagiarism.

In the past two or three years, the rapid development and spread of LLMs has drawn significant attention from researchers to the detection of LLM-generated text, and diverse methods have been proposed [Wu et al., 2025]: DetectGPT [Mitchell et al., 2023], Fast-DetectGPT [Bao et al., 2023], DetectLLM [Su et al., 2023], LLMDet [Wu et al., 2023], DeID-GPT [Liu et al., 2023] and some others [Dugan et al., 2023] in 2023; Binoculars [Hans et al., 2024], TOCSIN [Ma and Wang, 2024], Dpic [Yu et al., 2024b], Text Fluoroscopy [Yu et al., 2024a] in 2024. The examples listed above are illustrative, and the actual number of detectors is much larger.

In the meantime, specialized detectors have been developed, for instance, targeting journalistic news articles [Bhattacharjee et al., 2023] and tweets [Kumarage et al., 2023, Gambini et al., 2022]. Additionally, the detection of LLM-generated text is not limited to English [Wang et al., 2025]. we have also seen detectors for other languages, such as French [Antoun et al., 2023a], Japanese [Zaitsu and Jin, 2023], Chinese [Wang et al., 2024a].

While these techniques of detection performed well earlier on certain datasets, the ongoing progress of LLMs also makes detection harder [Wu et al., 2025]. A wide range of methods are utilized by these detectors, but the absence of universal benchmarks and different application scenarios limit meaningful comparison. We will address this issue in greater detail in Section 5.1.

These detection methods can be classified into many categories according to different criteria. For instance, Abdali et al. [2024] classifies them as supervised methods, zero-shot methods, retrieval-based methods, watermarking methods, discriminating features. Wu et al. [2025] mainly examines

them through the lens of watermarking techniques, statistics-based detectors, neural-based detectors, and human-assisted methods. It is difficult to provide a comprehensive summary of LLM-generated text detectors, but to our knowledge, no detector has been conclusively established as the best, particularly in practical deployment contexts.

There are other ways to categorize the detectors. For example, most studies only think about binary classification, and detectors with multi-category cases have rarely been explored, which will be further examined in Section 5.4.

3 Related Work

Benchmark One of the major challenges in establishing benchmarks for detecting LLM-generated text is that LLMs are continuously evolving, and their characteristics do not remain the same. For example, [Liyanage et al., 2022] created their benchmark with GPT-2, which should be quite differently from the current advanced LLMs. More LLMs were employed in subsequent benchmark construction [Wang et al., 2024b, He et al., 2024, Cornelius et al., 2024], but the number of prompts and scenarios used was limited. Some recent benchmarks [Tao et al., 2024, Wu et al., 2024] have incorporated a broader range of scenarios, and their impact and effectiveness remain to be seen. Similar challenge for the dataset [Gritsai et al., 2024].

Watermarking and Attack As mentioned earlier, watermarking LLMs is considered a category of detection methods, and it has shown good effectiveness in simulation [Kirchenbauer et al., 2023], which may also be an ethical necessity [Grinbaum and Adomaitis, 2022]. Researchers have proposed diverse watermarking techniques [Jovanović et al., 2024, Dathathri et al., 2024, Christ et al., 2024], but given the variety of attack techniques, this game is far from over [Krishna et al., 2023, Cai and Cui, 2023, Sadasivan et al., 2023, Lau and Zubiaga, 2024, Tufts et al., 2024, Macko et al., 2024].

Evaluation and Test These different methods have also been tested by various researchers [Weber-Wulff et al., 2023, Walters, 2023], in the beginning, some issues have been pointed out, such as biases [Liang et al., 2023] and false positives [Rashidi et al., 2023, Giray, 2024]. Although different detectors have been proposed over time, many challenges still exist [Kar et al., 2024, Tufts et al., 2024]. Given the diversity of usage scenarios and the variety of LLMs, such results are to be expected. We will provide a detailed analysis later.

4 Case Study

For a simple illustration of LLM-generated text and detection, we take the introduction from the website of NeurIPS[1] for simulations. This paragraph is processed with various prompts across different LLMs (DeepSeek-V3 [Liu et al., 2024], DeepSeek-R1 [Guo et al., 2025], GPT-3.5, GPT-4o-mini, GPT4o [Hurst et al., 2024]).

We present the original text and two generated texts with different prompts, and more examples are provided in the Appendix A.

Original Text *The conference was founded in 1987 and is now a multi-track interdisciplinary annual meeting that includes invited talks, demonstrations, symposia, and oral and poster presentations of refereed papers. Along with the conference is a professional exposition focusing on machine learning in practice, a series of tutorials, and topical workshops that provide a less formal setting for the exchange of ideas.*

DeepSeek-V3 ("Polish the following passage: ...") *Founded in 1987, the conference has grown into a premier multi-track interdisciplinary annual event. It features invited talks, demonstrations, symposia, and peer-reviewed oral and poster presentations. Alongside the conference, attendees can explore a professional exposition highlighting practical applications of machine learning, a series of tutorials, and topical workshops that foster informal idea exchange.*

[1] `https://neurips.cc/`

DeepSeek-R1 ("Polish the following passage: ...") *Established in 1987, the conference has grown into a premier interdisciplinary event held annually, featuring multiple tracks of programming. Attendees engage with invited talks, demonstrations, symposia, and peer-reviewed research presented through both oral and poster sessions. Complementing the core program are a professional exposition showcasing real-world applications of machine learning, hands-on tutorials, and topical workshops designed to foster informal dialogue and collaborative idea exchange.*

GPT-4o ("Polish the following passage: ...") *Established in 1987, the conference has evolved into a comprehensive, multi-track interdisciplinary annual event. It features invited talks, demonstrations, symposia, and both oral and poster presentations of peer-reviewed papers. In addition to the main conference, there is a professional exposition dedicated to practical applications of machine learning, as well as a series of tutorials and topical workshops. These workshops offer a more informal setting for the exchange of ideas.*

GPT-4o ("Rewrite the following passage without using complex expressions: ...") *The conference started in 1987 and now happens every year with multiple sessions. It features invited talks, demonstrations, discussions, and presentations of selected papers. There is also a professional expo about using machine learning, a series of tutorials, and workshops that offer a more relaxed space for sharing ideas.*

Although they modify the original text (for example, the underlined words above), the added words and expressions are not the same. Table 2 presents the results of detecting these texts using Fast-DetectGPT [Bao et al., 2023]. Even though these texts are all generated by LLMs, their detection outcomes vary widely. As we can easily find, in this case, the detector struggles to clearly identify text generated by DeepSeek-V3 and DeepSeek-R1, as the probability of texts by them being identified as machine-generated is even lower than that of the original text from NeurIPS website. The results of GPT-3.5, GPT-4o-mini, and GPT-4o show that prompts can easily affect the ouputs and the detection results.

Table 2: Detection results using Fast-DetectGPT. The last two columns correspond to the predictions of the machine-generated results when the Sampling/scoring model is gpt-neo-2.7b and falcon-7b, respectively.

Prompts	**Model**	**p1**	**p2**
(original text)	-	44%	23%
Polish the following passage:	DeepSeek-V3	34%	12%
	DeepSeek-R1	**21%**	**10%**
Polish the following passage:	GPT-3.5	50%	22%
	GPT-4o-mini	35%	20%
	GPT-4o	**84%**	**75%**
Rewrite the following passage without using complex expressions:	GPT-3.5	30 %	16 %
	GPT-4o-mini	51%	55%
	GPT-4o	38%	15%

These are merely a few basic examples of the issues and limitations faced by LLM-generated text detectors. A more detailed discussion will follow in the next section.

5 Issues and Limitations

As we briefly introduced before, the detection of text generated by LLMs has emerged as a widely discussed and actively pursued task in natural language processing. Such detection tools are often promoted for their potential utility in identifying instances of plagiarism [Pudasaini et al., 2024], academic dishonesty (e.g., cheating during examinations) [Wang and Li, 2025], the automatic generation of unethical peer reviews [Kumar et al., 2025], and other forms of content manipulation.

Meanwhile, many issues and challenges have been discussed [Tang et al., 2024, Wu et al., 2025, Fraser et al., 2024, Abdali et al., 2024]. Despite their apparent usefulness, there are some fundamental

limitations associated with these tools that raise serious ethical and methodological concerns. We will address these issues and limitations from various perspectives in this section.

5.1 Lack of Precise Definition and Gold-Standard Benchmark

Unlike most question-answering or classification tasks, "human-written text" itself lacks a clear and well-defined boundary compared to "LLM-generated text". **As implied by their names, the difference lies in how they are produced, but in practice, we can only assess them based on their final output i.e., the text, where in which a lot of overlap between them.**

Researchers often say that the text generated by LLMs is different from that written by humans [Muñoz-Ortiz et al., 2024, Reinhart et al., 2025]. Just as different people can write in different styles, LLMs can also generate varied outputs. We think that what is commonly referred to as "LLM-generated text" is only a subset of the text that LLMs can potentially produce, and it's likely the kind that corresponds to the most common and direct prompts. For instance, many detectors are trained on text generated by LLMs, which cannot represent all possibilities. Consequently, their detection capabilities are constrained. While different parameters can be set for various types of cases [Hans et al., 2024], such configurations can hardly cover all possible scenarios.

As we mentioned in related work, although some researchers have proposed benchmarks for detecting LLM-generated text [Liyanage et al., 2022, Wang et al., 2024b, He et al., 2024, Tao et al., 2024, Wu et al., 2024], their adoption has not yet become as widespread as other well-known LLM benchmarks, such as GLUE [Wang et al., 2018] and MMLU [Hendrycks et al., 2020]. Although these benchmarks have also faced some criticism [Hadi et al., 2023], there is still no highly universal benchmark for detecting LLM-generated texts.

Besides, due to the diversity of usage scenarios and the continuous updates of LLMs, a gold-standard benchmark is hard to realize, may even remain permanently absent.

5.2 Inherent Imperfection of Detection Tools

No existing LLM-detection system is infallible. In real-world conditions, a detection accuracy of 85% is typically considered outstanding. Yet, this figure necessarily implies a 15% error rate, which may include both false positives and false negatives.

False positives—in which human-written content is incorrectly flagged as machine-generated—are particularly problematic in high-stakes contexts such as academic integrity investigations. And the problem of false positives has already been observed and discussed. For example, Rashidi et al. [2023] found that the AI text detector erroneously identified up to 8% of the known real abstracts as AI-generated text, and Giray [2024] states that false positives disproportionately affect non-native English speakers and scholars with distinctive writing styles. In addition, Tufts et al. [2024] think that adversarial attacks can easily bypass these detectors, and balancing high sensitivity with a reasonable true positive rate remains challenging.

Accusing someone of misconduct based on an imperfect tool can lead to unjust outcomes, reputational damage, and institutional distrust. Therefore, even detectors with relatively high accuracy present significant risks when used for evaluative or disciplinary purposes.

Experiments also show that certain detectors may exhibit bias against non-native English writers [Liang et al., 2023] or against certain demographic groups [Kadoma et al., 2025]. The analyses from Li and Wan [2025] revealed that all the detectors they tested are highly sensitive to CEFR level and language environment. With LLMs being so widely used in academia [Eger et al., 2025, Russell et al., 2025], detecting AI-generated text must be handled with extreme care.

These detectors also face numerous other challenges, including difficulties in short-text detection [Gameiro et al., 2024, Shi et al., 2024] and the issues of modification and classification that we will discuss later. As such, current detectors are far from perfect and may never achieve perfection in the future either.

5.3 Poor Robustness to Textual Modifications

There have always been many doubts about the effectiveness of these detectors [Sadasivan et al., 2023, Weber-Wulff et al., 2023]. Another issue pertains to the brittleness of these tools in realistic scenarios. An early study has shown that while humans can reliably detect poetry produced by GPT-2, but they struggle to accurately recognize it after human selection [Köbis and Mossink, 2021]. If post-generation modifications are taken into account, the detection process should become more challenging.

Most current detectors are trained to recognize text that has been directly generated by an LLM without post-editing. While some recent systems claim to maintain performance when the LLM-generated text is lightly modified, empirical evidence shows that detection accuracy tends to decline as the extent of human revision increases.

In practice, LLM-generated content is often edited, paraphrased, or interwoven with human-written material, especially in academic contexts. Consequently, the tools' applicability to real-world use cases remains limited. This limitation exacerbates the concerns raised in the first point, as reliance on imperfect systems in nuanced or ambiguous situations increases the likelihood of erroneous judgments.

5.4 A Wide Variety of Use Cases and the Limits of Binary Classification

A fourth, and perhaps more fundamental, concern lies in the heterogeneity of LLM-generated content.

The ethical implications of LLM use depend heavily on the context and intent of usage. For instance, a researcher who uses LLMs to generate entire manuscripts with minimal intellectual input contributes to the proliferation of unoriginal work, thereby burdening peer-review systems and undermining the credibility of scholarly communication.

Such practices are clearly unethical. In contrast, a non-native speaker might use an LLM to translate, rephrase, or refine a manuscript that is otherwise the product of original research. In this case, the LLM acts as a language aid rather than a generator of substantive content. Yet most detection systems treat these qualitatively different scenarios in the same manner, reducing the complex spectrum of authorship to a binary classification of "human-written" versus "machine-generated". Similar problems have also been noted in very recent studies [Lepp and Smith, 2025].

Generally, most studies focus on the binary classification problem of determining whether a given text is generated by LLMs. While some detection methods could achieve good results on given datasets, the scenario becomes more much complicated in real-world settings. For example, people could edit LLM-generated text or mix it with human written text, which has also attracted considerable attention [Zhang et al., 2024a, Abassy et al., 2024, Kumar et al., 2025]. Only a small number of researchers have tried to identify specific roles of LLM in content generation [Cheng et al., 2025], and no universally accepted approaches have been established.

5.5 Diversity in LLMs

Even without considering the usage scenarios noted before, different LLMs generate text in different styles [Rosenfeld and Lazebnik, 2024, Sun et al., 2025]. Empirical studies have consistently demonstrated that different LLMs exhibit distinct stylistic patterns fingerprints, which could even be used for classifiction [McGovern et al., 2024, Sun et al., 2025, Bitton et al., 2025].

Studies indicate that the detectability of texts depends on the LLM used for text generation [Antoun et al., 2023b]. For example, Wu et al. [2024] pointed out that the Binoculars [Hans et al., 2024] only achieved a 55.15% AUROC in detecting texts generated by Claude, while for texts generated by several other models, it reached at least 88%. A comparable point is reflected in Table 2.

Detectors may more easily flag text from older and smaller models [Elkhatat et al., 2023, Saha and Feizi, 2025]. As we all know, the development of LLMs has not stopped, so the timeliness of detectors is also another challenge. Obviously, the same LLM can produce different texts in response to different prompts for the same task, as we have shown before. **Although these detectors may still be applicable in certain scenarios, their use requires greater caution.**

5.6 Others

Those familiar with LLMs and detectors are aware of the potential issues, but the public tends to be easily drawn to these numbers and brief conclusion. The lack of detector interpretability represents another concern [Ji et al., 2024], severely limiting the ability to provide transparent explanations to the public.

The appropriate use of LLMs has now been widely accepted, such as in the NeurIPS submission process [2]. In addition to the examples given earlier, the traces of LLM-generated text have now been found in various fields, such as student essays' answers [Leppänen et al., 2025] and words used in speaking [Yakura et al., 2024, Geng et al., 2024].

The abuse and misuse of these detectors can create ethical risks. Meanwhile, the numerical effectiveness of LLM-generated text detectors is declining. On the one hand, human may be influenced by LLMs and may create text resembling LLM-generated text. One the other hand, people may also adapt their language to bypass LLM detection tools [Geng and Trotta, 2025].

Therefore, **when interpreting the detection results of LLM-generated text, it is necessary to explicitly specify which kind of subset serves as the reference to establish the detector**.

6 Positive Impact of LLM Usage

The social impact of of LLMs has already been considered [Solaiman et al., 2019].

While LLM-generated text is frequently the subject of criticism—particularly due to concerns around academic dishonesty, plagiarism, and fraud, which have led to the development of various detection tools—it is equally important to emphasize the legitimate and ethical uses of large language models. As discussed earlier, LLMs can play a valuable role in numerous contexts. For instance, they help bridge linguistic divides by enabling non-native speakers to produce coherent and idiomatic texts in English or other target languages, thereby supporting greater inclusivity in academic and professional communication. They also facilitate high-quality machine translation, making content in multiple languages more accessible, and allow for the efficient synthesis of large textual corpora, which can aid research and knowledge production.

Rather than focusing solely on the detection and policing of LLM-generated text, it may be more productive to advocate for transparency regarding their use. In academic publishing, for example, it is increasingly common to disclose how LLMs have assisted in drafting, editing, or rephrasing portions of a manuscript.

Such uses are generally limited to improving expression or exploring alternative formulations; the substantive intellectual work remains the responsibility of human authors. Importantly, LLMs should not be considered co-authors nor used to autonomously generate scientific content in its entirety. Clear guidelines and disclosures can thus help normalize the ethical integration of LLMs into scholarly workflows without undermining academic integrity.

People began discussing ChatGPT's positive impact shortly after its emergence [Kasneci et al., 2023]. Non-native English speakers have to put in more effort as scientists, and there has been discrimination in the past [Amano et al., 2023, Lepp and Smith, 2025]. Automatic editing methods have shown promise in improving alignment between LLM-generated and human-written text [Chakrabarty et al., 2024]. If LLMs are applied properly and people assess detection tools reasonably, their positive influence can be greatly amplified.

7 Alternative Views

Researchers have not yet reached full agreement on the detectability of LLM-generated text.

Chakraborty et al. [2024] claim in their position paper: "Despite ongoing debate about the feasibility of such differentiation, we present evidence supporting its consistent achievability, except when human and machine text distributions are indistinguishable across their entire support. Drawing from

[2] `https://neurips.cc/Conferences/2025/LLM`

information theory, we argue that as machine-generated text approximates human-like quality, the sample size needed for detection increases."

But we believe that LLMs are fully capable of generating text that is nearly indistinguishable from human-written content. Furthermore, practical observations have shown that humans possess the capacity to identify LLM-generated text with reasonable accuracy [Russell et al., 2025], and such coevolution may already be occurring [Geng and Trotta, 2025]. These challenges in these real-world data cannot be resolved by increasing the sample size.

The key disagreement among researchers may not be technical in nature, but rather stems from differing perspectives on human intervention. Take watermarking studies as an example, if people edit the generated text (which is simple to do), the watermark's reliability may be greatly weakened [Dathathri et al., 2024].

There are also researchers who share similar views with us. For example, Zhang et al. [2024b] argue that "We believe that the issue of AI-generated text detection remains an unresolved challenge. As LLMs become increasingly powerful and humans become more proficient in using them, it is even less likely to detect AI text in the future." And Nicks et al. [2023] "advise against continued reliance on LLM-generated text detectors".

8 Future Perspectives and Predictions

LLMs were compared to stochastic parrot [Bender et al., 2021] a couple of years ago, but their capabilities are gradually being recognized [Srivastava et al., 2022], and now their competencies have reached or even surpassed those of human experts in various fields. This leads us to speculate that future LLM-generated texts could surpass current versions in human resemblance. The persona effect could be a good example [Hu and Collier, 2024].

The gap between LLM-generated and human-written text is expected to narrow further. As noted by several researchers, humans can learn to detect AI-generated texts [Milička et al., 2025] and can become accurate and robust detector of LLM-generated text [Russell et al., 2025].

And therefore, we can find more cases of coevolution [Pedreschi et al., 2024, Geng and Trotta, 2025] between human and Human in the future. Consequently, in the future, detecting LLM-generated text may become less important, particularly in terms of numerical interpretation.

Accounting for model collapse [Shumailov et al., 2024, Guo et al., 2023] and knowledge collapse [Peterson, 2025], the detection results may become even more intriguing. Detection efforts should target substantive content (e.g., fact-checking) rather than linguistic characteristics [Schuster et al., 2020].

9 Conclusions

Given the current state-of-the-art, existing tools are not equipped to make such fine-grained distinctions. They are structurally unable to assess the proportion, function, or ethical significance of LLM contributions in a given text. As a result, the development and deployment of LLM-generated text detectors raise serious concerns, not only due to technical limitations but also because they risk enforcing overly simplistic and potentially unjust frameworks for evaluating authorship and intellectual responsibility.

And some of these difficulties are simply unavoidable, not merely temporary technical challenges. Texts generated by LLMs and those generated by humans often overlap greatly, with no obvious separation. Moreover, as LLMs become more widely used and people may learn from their outputs, the difference between them may get further smaller. Since text is different from images, it is difficult to balance both the watermark and the original textual information.

While acknowledging that detectors for LLM-generated text can serve a purpose in certain scenarios, we recommend using them with greater caution. And the detection results should only be used as a reference rather than the decisive indicator.

References

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, et al. Llm-detectaive: a tool for fine-grained machine-generated text detection. *arXiv preprint arXiv:2408.04284*, 2024.

Sara Abdali, Richard Anarfi, CJ Barberan, and Jia He. Decoding the ai pen: Techniques and challenges in detecting ai-generated text. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6428–6436, 2024.

Tatsuya Amano, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, Flavia Montaño-Centellas, Kumar Paudel, Rachel Louise White, et al. The manifold costs of being a non-native english speaker in science. *PLoS Biology*, 21(7):e3002184, 2023.

Wissam Antoun, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah. Towards a robust detection of language model generated text: is chatgpt that easy to detect? *arXiv preprint arXiv:2306.05871*, 2023a.

Wissam Antoun, Benoît Sagot, and Djamé Seddah. From text to source: Results in detecting large language model-generated content. *arXiv preprint arXiv:2309.13322*, 2023b.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*, 2023.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. Conda: Contrastive domain adaptation for ai-generated text detection. *arXiv preprint arXiv:2309.03992*, 2023.

Yehonatan Bitton, Elad Bitton, and Shai Nisan. Detecting stylistic fingerprints of large language models. *arXiv preprint arXiv:2503.01659*, 2025.

Creston Brooks, Samuel Eggert, and Denis Peskoff. The rise of ai-generated content in wikipedia. *arXiv preprint arXiv:2410.08044*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Shuyang Cai and Wanyun Cui. Evade chatgpt detectors via a single space. *arXiv preprint arXiv:2307.02599*, 2023.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits. *arXiv preprint arXiv:2409.14509*, 2024.

Souradip Chakraborty, Amrit Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. Position: On the possibilities of ai-generated text detection. In *Forty-first International Conference on Machine Learning*, 2024.

Zihao Cheng, Li Zhou, Feng Jiang, Benyou Wang, and Haizhou Li. Beyond binary: Towards fine-grained llm-generated text detection via role recognition and involvement measurement. In *Proceedings of the ACM on Web Conference 2025*, pages 2677–2688, 2025.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1125–1139. PMLR, 2024.

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

Joseph Cornelius, Oscar Lithgow-Serrano, Sandra Mitrović, Ljiljana Dolamic, and Fabio Rinaldi. Bust: Benchmark for the evaluation of detectors of llm-generated text. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8029–8057, 2024.

Evan N Crothers, Nathalie Japkowicz, and Herna L Viktor. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002, 2023.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. Exploring the limitations of detecting machine-generated text. *arXiv preprint arXiv:2406.11073*, 2024.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12763–12771, 2023.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, et al. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *arXiv preprint arXiv:2502.05151*, 2025.

Ahmed M Elkhatat, Khaled Elsaid, and Saeed Almeer. Evaluating the efficacy of ai content detection tools in differentiating between human and ai-generated text. *International Journal for Educational Integrity*, 19(1):17, 2023.

Kathleen C Fraser, Hillary Dawkins, and Svetlana Kiritchenko. Detecting ai-generated text: Factors influencing detectability with current methods. *arXiv preprint arXiv:2406.15583*, 2024.

Margherita Gambini, Tiziano Fagni, Fabrizio Falchi, and Maurizio Tesconi. On pushing deepfake tweet detection capabilities to the limits. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 154–163, 2022.

Henrique Da Silva Gameiro, Andrei Kucharavy, and Ljiljana Dolamic. Llm detectors still fall short of real world: Case of llm-generated short news-like posts. *arXiv preprint arXiv:2409.03291*, 2024.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

Mingmeng Geng and Roberto Trotta. Is chatgpt transforming academics' writing style? *arXiv preprint arXiv:2404.08627*, 2024.

Mingmeng Geng and Roberto Trotta. Human-llm coevolution: Evidence from academic writing. *arXiv preprint arXiv:2502.09606*, 2025.

Mingmeng Geng, Caixi Chen, Yanru Wu, Dongping Chen, Yao Wan, and Pan Zhou. The impact of large language models in academia: from writing to speaking. *arXiv preprint arXiv:2409.13686*, 2024.

Louie Giray. The problem with false positives: Ai detection unfairly accuses scholars of ai plagiarism. *The Serials Librarian*, 85(5-6):181–189, 2024.

Alexei Grinbaum and Laurynas Adomaitis. The ethical need for watermarks in machine-generated language. *arXiv preprint arXiv:2209.03118*, 2022.

German Gritsai, Anastasia Voznyuk, Andrey Grabovoy, and Yury Chekhovich. Are ai detectors good enough? a survey on quality of datasets with machine-generated texts. *arXiv preprint arXiv:2410.14677*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. The curious decline of linguistic diversity: Training language models on synthetic text. *arXiv preprint arXiv:2311.09807*, 2023.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 2023.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*, 2024.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. Mgtbench: Benchmarking machine-generated text detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 2251–2265, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*, 2024.

Siming Huang, Yuliang Xu, Mingmeng Geng, Yao Wan, and Dongping Chen. Wikipedia in the era of llms: Evolution and risks. *arXiv preprint arXiv:2503.02879*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.

Jiazhou Ji, Ruizhe Li, Shujun Li, Jie Guo, Weidong Qiu, Zheng Huang, Chiyu Chen, Xiaoyu Jiang, and Xinru Lu. Detecting machine-generated texts: Not just" ai vs humans" and explainability is complicated. *arXiv preprint arXiv:2406.18259*, 2024.

Nikola Jovanović, Robin Staab, and Martin Vechev. Watermark stealing in large language models. *arXiv preprint arXiv:2402.19361*, 2024.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

Kowe Kadoma, Danaë Metaxa, and Mor Naaman. Generative ai and perceptual harms: Who's suspected of using llms? In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2025.

Sujita Kumar Kar, Teena Bansal, Sumit Modi, and Amit Singh. How sensitive are the free ai-detector tools in detecting ai-generated texts? a comparison of popular ai-detector tools. *Indian Journal of Psychological Medicine*, page 02537176241247934, 2024.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.

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

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36:27469–27500, 2023.

Sandeep Kumar, Samarth Garg, Sagnik Sengupta, Tirthankar Ghosal, and Asif Ekbal. Mixrevdetect: Towards detecting ai-generated content in hybrid peer reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 944–953, 2025.

Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*, 2023.

Tharindu Kumarage, Garima Agrawal, Paras Sheth, Raha Moraffah, Aman Chadha, Joshua Garland, and Huan Liu. A survey of ai-generated text forensic systems: Detection, attribution, and characterization. *arXiv preprint arXiv:2403.01152*, 2024.

Hiu Ting Lau and Arkaitz Zubiaga. Understanding the effects of human-written paraphrases in llm-generated text detection. *arXiv preprint arXiv:2411.03806*, 2024.

Haley Lepp and Daniel Scott Smith. " you cannot sound like gpt": Signs of language discrimination and resistance in computer science publishing. *arXiv preprint arXiv:2505.08127*, 2025.

Leo Leppänen, Lili Aunimo, Arto Hellas, Jukka K Nurminen, and Linda Mannila. How large language models are changing mooc essay answers: A comparison of pre-and post-llm responses. *arXiv preprint arXiv:2504.13038*, 2025.

Jiatao Li and Xiaojun Wan. Who writes what: Unveiling the impact of author roles on ai-generated text detection. *arXiv preprint arXiv:2502.12611*, 2025.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7), 2023.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*, 2024.

Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. The widespread adoption of large language model-assisted writing across society. *arXiv preprint arXiv:2502.09747*, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023.

Vijini Liyanage, Davide Buscaldi, and Adeline Nazarenko. A benchmark corpus for the detection of automatically generated text in academic publications. *arXiv preprint arXiv:2202.02013*, 2022.

Shixuan Ma and Quan Wang. Zero-shot detection of llm-generated text using token cohesiveness. *arXiv preprint arXiv:2409.16914*, 2024.

Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason Samuel Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. Authorship obfuscation in multilingual machine-generated text detection. *arXiv preprint arXiv:2401.07867*, 2024.

Hope McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. Your large language models are leaving fingerprints. *arXiv preprint arXiv:2405.14057*, 2024.

Jiří Milička, Anna Marklová, Ondřej Drobil, and Eva Pospíšilová. Humans can learn to detect ai-generated texts, or at least learn when they can't. *arXiv preprint arXiv:2505.01877*, 2025.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR, 2023.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265, 2024.

Charlotte Nicks, Eric Mitchell, Rafael Rafailov, Archit Sharma, Christopher D Manning, Chelsea Finn, and Stefano Ermon. Language model detectors are easily optimized against. In *The twelfth international conference on learning representations*, 2023.

Dino Pedreschi, Luca Pappalardo, Emanuele Ferragina, Ricardo Baeza-Yates, Albert-László Barabási, Frank Dignum, Virginia Dignum, Tina Eliassi-Rad, Fosca Giannotti, János Kertész, et al. Human-ai coevolution. *Artificial Intelligence*, page 104244, 2024.

Andrew J Peterson. Ai and the problem of knowledge collapse. *AI & SOCIETY*, pages 1–21, 2025.

Shushanta Pudasaini, Luis Miralles-Pechuán, David Lillis, and Marisa Llorens Salvador. Survey on plagiarism detection in large language models: The impact of chatgpt and gemini on academic integrity. *arXiv preprint arXiv:2407.13105*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Hooman H Rashidi, Brandon D Fennell, Samer Albahra, Bo Hu, and Tom Gorbett. The chatgpt conundrum: Human-generated scientific manuscripts misidentified as ai creations by ai text detection tool. *Journal of Pathology Informatics*, 14:100342, 2023.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122, 2025.

Ariel Rosenfeld and Teddy Lazebnik. Whose llm is it anyway? linguistic comparison and llm attribution for gpt-3.5, gpt-4 and bard. *arXiv preprint arXiv:2402.14533*, 2024.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. People who frequently use chatgpt for writing tasks are accurate and robust detectors of ai-generated text. *arXiv preprint arXiv:2501.15654*, 2025.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Shoumik Saha and Soheil Feizi. Almost ai, almost human: The challenge of detecting ai-polished writing. *arXiv preprint arXiv:2502.15666*, 2025.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510, 2020.

Yuhui Shi, Qiang Sheng, Juan Cao, Hao Mi, Beizhe Hu, and Danding Wang. Ten words only still help: Improving black-box ai-generated text detection via proxy-guided efficient re-sampling. *arXiv preprint arXiv:2402.09199*, 2024.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*, 2023.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J Zico Kolter, and Zhuang Liu. Idiosyncrasies in large language models. *arXiv preprint arXiv:2502.12150*, 2025.

Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. Are we in the ai-generated text world already? quantifying and monitoring aigt on social media. *arXiv preprint arXiv:2412.18148*, 2024.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated text. *Communications of the ACM*, 67(4):50–59, 2024.

Zhen Tao, Zhiyu Li, Dinghao Xi, and Wei Xu. Cudrt: Benchmarking the detection of human vs. large language models generated texts. *arXiv preprint arXiv:2406.09056*, 2024.

Brian Tufts, Xuandong Zhao, and Lei Li. A practical examination of ai-generated text detectors for large language models. *arXiv preprint arXiv:2412.05139*, 2024.

Jan Philip Wahle, Terry Ruas, Frederic Kirstein, and Bela Gipp. How large language models are transforming machine-paraphrased plagiarism. *arXiv preprint arXiv:2210.03568*, 2022.

William H Walters. The effectiveness of software designed to detect ai-generated writing: A comparison of 16 ai text detectors. *Open Information Science*, 7(1):20220158, 2023.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Quan Wang and Haoran Li. On continually tracing origins of llm-generated text and its application in detecting cheating in student coursework. *Big Data and Cognitive Computing*, 9(3):50, 2025.

Rongsheng Wang, Haoming Chen, Ruizhe Zhou, Han Ma, Yaofei Duan, Yanlan Kang, Songhua Yang, Baoyu Fan, and Tao Tan. Llm-detector: Improving ai-generated chinese text detection with open-source llm instruction tuning. *arXiv preprint arXiv:2402.01158*, 2024a.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, et al. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection. *arXiv preprint arXiv:2402.11175*, 2024b.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, et al. Genai content detection task 1: English and multilingual machine-generated text detection: Ai vs. human. *arXiv preprint arXiv:2501.11012*, 2025.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltỳnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, 19(1):1–39, 2023.

Junchao Wu, Runzhe Zhan, Derek Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia Chao. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *Advances in Neural Information Processing Systems*, 37:100369–100401, 2024.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pages 1–66, 2025.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. Llmdet: A third party large language models generated text detection tool. *arXiv preprint arXiv:2305.15004*, 2023.

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, and Iyad Rahwan. Empirical evidence of large language model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*, 2024.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. A survey on detection of llms-generated content. *arXiv preprint arXiv:2310.15654*, 2023.

Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, and Nenghai Yu. Text fluoroscopy: Detecting llm-generated text through intrinsic features. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15838–15846, 2024a.

Xiao Yu, Yuang Qi, Kejiang Chen, Guoqiang Chen, Xi Yang, Pengyuan Zhu, Xiuwei Shang, Weiming Zhang, and Nenghai Yu. Dpic: Decoupling prompt and intrinsic characteristics for llm generated text detection. *Advances in Neural Information Processing Systems*, 37:16194–16212, 2024b.

Wataru Zaitsu and Mingzhe Jin. Distinguishing chatgpt (-3.5,-4)-generated and human-written papers through japanese stylometric analysis. *PLoS One*, 18(8):e0288453, 2023.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, et al. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? *arXiv preprint arXiv:2401.05952*, 2024a.

Yuehan Zhang, Yongqiang Ma, Jiawei Liu, Xiaozhong Liu, Xiaofeng Wang, and Wei Lu. Detection vs. anti-detection: Is text generated by ai detectable? In *International Conference on Information*, pages 209–222. Springer, 2024b.

A Simulation Results

- **DeepSeek-V3 (Polish the following passage:).** Founded in 1987, the conference has grown into a premier multi-track interdisciplinary annual event. It features invited talks, demonstrations, symposia, and peer-reviewed oral and poster presentations. Alongside the conference, attendees can explore a professional exposition highlighting practical applications of machine learning, a series of tutorials, and topical workshops that foster informal idea exchange.
- **DeepSeek-R1 (Polish the following passage:).** Established in 1987, the conference has grown into a premier interdisciplinary event held annually, featuring multiple tracks of programming. Attendees engage with invited talks, demonstrations, symposia, and peer-reviewed research presented through both oral and poster sessions. Complementing the core program are a professional exposition showcasing real-world applications of machine learning, hands-on tutorials, and topical workshops designed to foster informal dialogue and collaborative idea exchange.
- **GPT-3.5 (Polish the following passage:).** Established in 1987, the conference has evolved into a multi-track interdisciplinary annual event featuring invited talks, demonstrations, symposia, and oral and poster presentations of refereed papers. In addition, there is a professional exposition dedicated to machine learning in practice, a variety of tutorials, and topical workshops that offer a more relaxed environment for the sharing of ideas.
- **GPT-4o-mini (Polish the following passage:).** Founded in 1987, the conference has evolved into a dynamic multi-track interdisciplinary annual event featuring invited talks, demonstrations, symposia, and both oral and poster presentations of refereed papers. Complementing the conference is a professional exposition dedicated to practical applications of machine learning, as well as a series of tutorials and topical workshops that create a more informal atmosphere for exchanging ideas.
- **GPT-4o (Polish the following passage:).** Established in 1987, the conference has evolved into a comprehensive, multi-track interdisciplinary annual event. It features invited talks, demonstrations, symposia, and both oral and poster presentations of peer-reviewed papers. In addition to the main conference, there is a professional exposition dedicated to practical applications of machine learning, as well as a series of tutorials and topical workshops. These workshops offer a more informal setting for the exchange of ideas.
- **GPT-3.5 (Rewrite the following passage without using complex expressions:).** The conference started in 1987 and has grown into a yearly event that covers different subjects and includes talks, demonstrations, symposiums, and presentations of reviewed papers. There is also an exhibition on machine learning, tutorials, and workshops for sharing ideas in a casual setting.
- **GPT-4o-mini (Rewrite the following passage without using complex expressions:).** The conference started in 1987 and is now held every year. It features various sessions that include invited talks, demonstrations, and both oral and poster presentations of reviewed papers. There is also a professional exhibit that highlights real-world applications of machine learning, along with a series of tutorials and workshops that offer a more relaxed atmosphere for sharing ideas.
- **GPT-4o (Rewrite the following passage without using complex expressions:).** The conference started in 1987 and now happens every year with multiple sessions. It features invited talks, demonstrations, discussions, and presentations of selected papers. There is also a professional expo about using machine learning, a series of tutorials, and workshops that offer a more relaxed space for sharing ideas.