Evaluating the Accuracy of AI Content Detectors*

Fatih Emre YILDIZ

- Muğla Sıtkı Koçman University, Engineering Faculty, Computer Engineering, Muğla, Turkiye.
- https://orcid.org/0009-0000-3871-8569
- fatihemreyildiz@posta.mu.edu.tr

Enis KARAARSLAN

- Muğla Sıtkı Koçman University, Engineering Faculty, Computer Engineering, Muğla, Turkiye.
- https://orcid.org/0000-0002-3595-8783
- enis.karaarslan@mu.edu.tr

Ömer AYDIN

- Manisa Celal Bayar University, Faculty of Engineering and Natural Sciences, Electrical-Electronics Engineering, Manisa, Türkiye.
- https://orcid.org/0000-0002-7137-4881
- omer.aydin@cbu.edu.tr
- Corresponding Author

Yıldız, F.E., Karaarslan, E., & Aydin, O. (2025). Evaluating the Accuracy of AI Content Detectors. In O. Aydin & E. Karaarslan (Eds.), *The Age of Generative Artificial Intelligence* (pp. 143-155). Izmir Academy Association. **Doi:** 10.5281/zenodo.16009433

1. Introduction

Large Language Models (LLMs), such as OpenAI's GPT [1], Google's Gemini [2], or Anthropic's Claude [3], and many others have become very popular in recent years. These use advanced natural language processing (NLP) techniques to generate human-like text. People now use LLMs in many areas, such as writing, journalism, essay writing, academic research, and even coding, because they can create content effectively and quickly. This study reveals that after the first release of ChatGPT, the number of scientific publications has increased [4].

LLMs are trained on huge amounts of data [5], which allows them to produce new data based on the data it is used to train them. While bringing many benefits, they also raise very important problems.

One of the most important problems with LLMs is the potential misuse. Especially in academic and professional work. The major problems are plagiarism and hidden AI authorship [6]: an academic work done with the help of LLM and publishing it without declaring it can lead to academic fraud and fake scientific publications. Another very serious problem is hallucination. Hallucination is when an LLM model produces fabricated or false information[7]. If this false information is trusted by the researchers without proper checks, it will make the publications wrong. This will spread incorrect knowledge.

They can also be used in generating false news. Thus, false information and propaganda are circulated for political campaigns or advertisements [8]. This sort of usage of LLMs can ruin people's confidence in online websites and in trusting the news.

Another concern is spamming and harmful content [9]. Such tools may be misused to spread large amounts of realistic comments or posts with the intent of abuse or harassment online. This type of misuse is particularly harmful to mental health and will discourage people from contributing to online communities.

LLMs can also create problems by repeating biased or stereotypical ideas from the data they are trained on [10]. This can result in unfair treatment of certain groups and make LLMs less effective for people from underrepresented communities [11]. These issues reduce people’s trust in information created by AI.

Criminal activities using LLMs are becoming a bigger problem [12]. People can use them for phishing attacks to steal personal information, for fraud, or to create deep fake [13, 14] videos and images to impersonate or harm others. These types of misuse can put people’s privacy and safety at risk and make it harder to trust digital systems.

Finally, AI-generated content can be used to train other models, which will lead the data on the internet to be more biased; therefore, the new models can suffer from this data [15].

Many AI content detectors have been developed in this regard to solve these problems [16]. These tools try to decide whether a given text was written by a human or generated by an AI system. However, these tools are not always reliable. They misclassify human-written text as AI-generated, or vice versa.

These misclassifications may not be just a technical issue; there could be serious problems, as identified in this research [17]. Misidentified cheating can occur with this technology when a content detector mistakenly decides that students or professionals are cheating. Such errors highlight the risks of depending too much on these tools.

In this study, the purpose is to evaluate the accuracy and reliability of widely used AI content detectors by testing them on human-written scientific papers published before the emergence of LLMs. The importance of this work lies in revealing how often these tools can misclassify legitimate academic texts, which can lead to unfair accusations or wrong decisions in critical fields such as education, publishing, or research. As its contribution, this study looks at real human-written texts

and tests how AI detection tools respond to them. This helps us better understand what these tools can and cannot do. The study also gives useful information for researchers, educators, and policymakers who use such tools. We also include real examples where AI detectors caused problems, showing why it is important to be careful when using them in serious situations.

In the following sections, we will look at the studies about AI content detectors and related topics, focusing on how they work, how effective they are, and their weaknesses. We will also discuss challenges related to AI-generated text detection and what can go wrong when people rely too heavily on these tools. Finally, we will report the performance of popular AI-detection tools with a reasonable dataset, together with their accuracy. This research is meant to make users understand how the AI detection tools can and cannot do, reminding them to use these tools carefully, at least in important situations.

2. Literature Review

There are various methods developed for AI content detection [18, 19]. These techniques can be watermarking, zero-shot detection, stylistic analysis, machine learning, and self-detection, which uses LLMs to detect, each with its own approach to identifying AI-generated content.

Watermarking is the hidden trace or message left behind in the output of LLMs. It has been widely used for video or image generation [8], which makes it easy to detect text. Watermarking involves making subtle patterns in the choice of words or structure, which turn out to be helpful in the identification of AI-generated content [20]. These watermarks can be groups of specific words, syntactic structures, uncommon phrases or words, also semantic anomalies that cannot be detected by humans placed in the output of a model [21].

This study [22] highlights that machine paraphrasing and even human paraphrasing cannot remove the watermarks, which means that watermarking is a strong detection tool for AI-generated content. Watermarking is not only an AI Detection Algorithm, but it can also be used to protect the big LLMs. Zhao et al. [23] introduce a novel technique that can be used to protect LLMs from model extraction attacks, without re-training the source model. Gu et al. [24] propose a watermarking model to protect the intellectual property of LLMs by combining rare words and common words as triggers with high accuracy, which is also robust to fine-tuning modifications and detection.

Zero-Shot Methods rely on pre-trained models, which identify contents that have not been seen before by the AI. They do not require specific datasets; hence, they are flexible. However, they tend to have low accuracy for complex content. This study [25] highlights the importance of the prompts for the AI content detection. There are a few more studies about the development of zero-shot methods. Hans et al. [26] present Binoculars, a novel zero-shot detector approach, which contrasts perplexity scores between two language models to identify AI-generated content. Mitchell et al. [27] introduce DetectGPT, same method that uses probability curvature, derived from log probabilities of text samples. Bao et al. [28] improve DetectGPT and introduce Fast-DetectGPT, which reduces the computational costs by replacing perturbations with the conditional probability metrics. It becomes 340x faster and outperforms DetectGPT.

There are a few more techniques to detect AI-generated content. Statistical and Stylistic Analysis checks for the content patterns, structures, and stylistic elements, including repeated phrases or even sentence length, common in AI outputs. Gehrmann et al. [29] worked on a tool before the first commercial LLMs were introduced. This tool combines statistical metrics with a visual interface to detect AI-generated texts. It improved the detection accuracy from 54% to 72%. On the other hand, it struggles to generalize different models that use specific sampling schemes. Ippolito et al. [30] benchmark the detectors by top-k and nucleus sampling. This study [31] identifies how LLMs follow natural language rules, such as sentence appearance, word patterns, and their relationship. They found out that different models match human language

in different ways. Tulchinskii et al. [32] show a new way to detect AI content by checking the shape of the text's patterns called "intrinsic dimensionality". According to their study, human-written texts have higher values than the LLM-written content. This method is correctly working in many kinds of text and is also robust against the paraphrasing method.

Machine Learning (ML) uses ML, DL, and NLP techniques to detect AI-generated content. It is based on labeled datasets that are used in training models to differentiate between human-written and AI-written content. Abdalla et al. [33] introduced a dataset for this purpose, using various machine learning classifiers, such as Random Forest and Logistic Regression. Prajapati et al. [34] studied the SVM, decision trees, and neural networks. Their results show high detection performance.

Self-detection is using LLMs to detect themselves. Such as, any LLM model can be used to examine if the text is produced by the model itself. This approach usually depends on internal knowledge of the model and does not require any additional training. In this study [35], ChatGPT's performance was evaluated as a content detector. ChatGPT was good at identifying human-written text; on the other hand, the model gave a significant amount of false negatives, especially with the good quality of AI-generated content. This shows the potential usage of LLMs for some part of the detection tasks, but also identifies the limitations for high-quality AI-generated outputs.

One of the challenges with AI content detection is that it can easily be bypassed by using common techniques like paraphrasing [36] and rewording techniques. There are also other techniques, such as translation [37, 38] or emoji insertion [39], to trick the content detectors. Furthermore, optimized prompts minimize the detectability of AI-generated text. With those techniques, the content is altered to sound more human-like.

The accuracy and performance of AI content generators and AI detectors vary widely.

Thorat et al. [40] identified that ChatGPT is better at generating more human-like content; this study has shown better performance on AI Detector tool tests compared to other popular models. Also, Hayawi et al. [41] found that ChatGPT is significantly better than BARD, now known as Gemini; they also developed a dataset for this study.

Recent research compares how well popular AI-text detectors work. Kar et al. [42] checked ten free tools, such as ZeroGPT, QuillBot, and Undetectable AI. They found that results were very different: some tools caught all simple AI-written text, but many failed when the text was paraphrased. Singh [43] also reviewed detectors like GLTR, GPTZero, and Copyleaks, showing where each one is useful and where it is weak. Elkhatat et al. [44] tested five big tools and saw that they recognized GPT-3.5 text better than GPT-4, and sometimes wrongly marked human writing as AI. Together, these studies show that AI detectors can help, but their accuracy is uneven, so they should not be trusted alone for important school or work decisions.

Finally, as AI models improve, their outputs become better [45] and more human-like, which makes detection more challenging. Continuous development and innovation are essential to enhance the reliability and effectiveness of AI content detection tools.

3. Methodology

We evaluate the accuracy and reliability of popular AI content detectors by following a systematic approach with a dataset of scientific articles published before the release of large-scale LLMs like ChatGPT. In this way, the dataset only includes human-written content, which can be taken as a reliable benchmark to test the performance of these tools. Our methodology is explained below:

We selected articles published before 2022 from top-quality journals and conferences. The articles were in many disciplines: computer science, biology, social sciences, engineering, and more. In only considering the publication of that period, we knew that all publications were created by a human author and did not run the risk of including AI-generated content.

This subset was then used as a ground truth for the detection tools. Selected articles were picked because of their high citation numbers. That means that they were representative enough of the formal academic writing styles.

Figure 1 shows the step-by-step process we used to choose and test the dataset.

- First, we collected scientific papers that were published before large language models (LLMs) became popular.
- Second, we focused on well-known journals and conferences in different fields. Then, we checked the publication years to make sure all papers were written before AI tools were commonly used.
- Third, we picked papers with many citations, because these are usually written in a formal and academic style.
- Finally, we tested these texts using different AI content detectors and recorded the results.

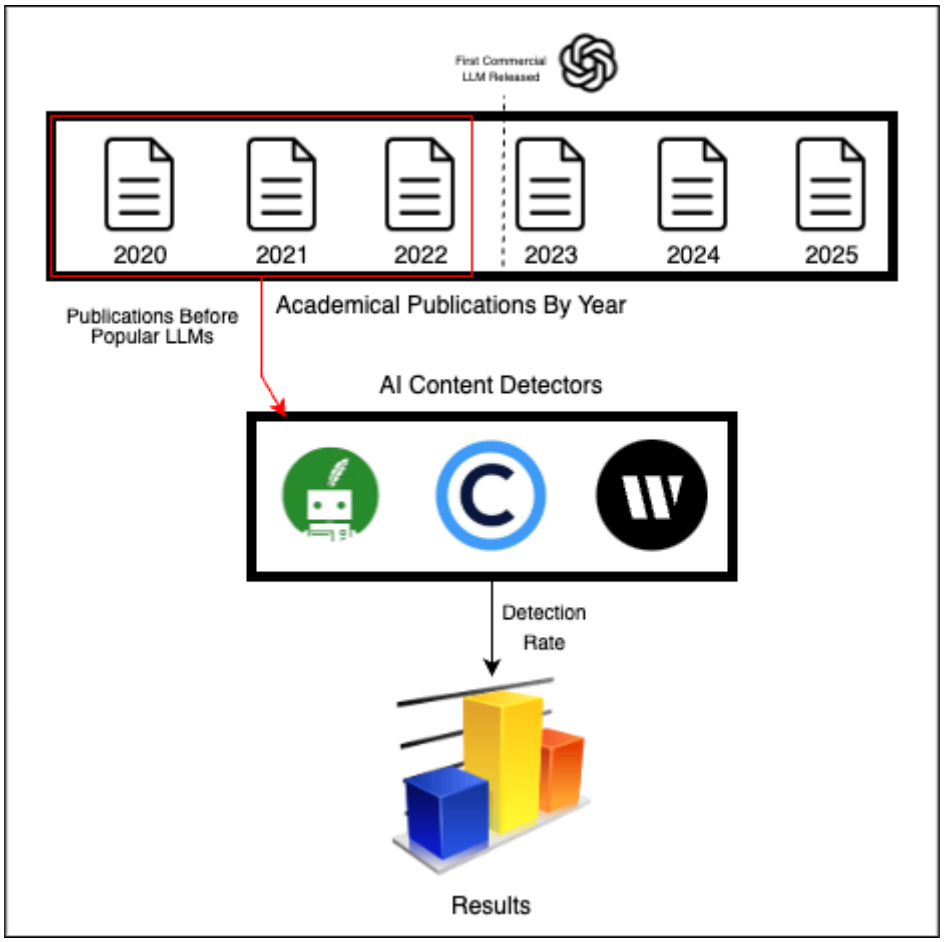

Figure 1. Method: Publications Before Popular LLMs and AI Content Detectors.

For this study, the following AI content detectors were tested:

- QuillBot [2]
- ZeroGPT[3]
- GPTZero[4]
- UndetectableAI[5]
- Writer[6]
- CopyLeaks[7]
- JustDone[8]

The tested papers are listed in Table 1.

Table 1. Tested papers

Article Title	Authors	Source	Year
Non-local Neural Networks [46]	Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He	Arxiv.org	2018
Attention Is All You Need [47]	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin	Arxiv.org	2023
Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning [48]	Sebastian Raschka	Arxiv.org	2020
Sketching for Principal Component Regression [49]	Liron Mor-Yosef, Haim Avron	Arxiv.org	2018
Water Bridging Dynamics of Polymerase Chain Reaction in the Gauge Theory Paradigm of Quantum Fields [50]	Luc Montagnier, Jamal Aïssa, Antonio Capolupo, Travis J. A. Craddock, Philip Kurian, Claude Lavallee, Albino Polcari, Paola Romano, Alberto Tedeschi, Giuseppe Vitiello	MDPI Water	2017
Squeeze-and-Excitation Networks [51]	Jie Hu, Li Shen, Samuel Albanie, Gang Sun, Enhua Wu	Arxiv.org	2019
Polymerase Chain Reaction [52]	Lilit Garibyan, Nidhi Avashia	Journal of Investigative Dermatology	2013

[2] https://quillbot.com/ai-content-detector

[3] https://www.zerogpt.com/

[4] https://gptzero.me/

[5] https://undetectable.ai/

[6] https://writer.com/ai-content-detector/

[7] https://copyleaks.com/

[8] https://justdone.com/

The tools selected are based on their popularity and adoption across a wide user base in academic, professional, and general-purpose applications. All detectors were run without changes to their normal operating configuration to closely simulate real-world usage.

4. Results

The results from our methodology are presented in Table 2.

Table 2. Results of the AI content detectors tested on the given papers

Publication	QuillBot	ZeroGPT	GPTZero	Writer	JustDone	Undetectable AI	CopyLeaks
[46]	0	0	1	0	73	AI Written	0
[47]	0	15	1	3	89	AI Written	0
[48]	0	0	2	4	88	AI Written	0
[49]	0	0	4	0	91	AI Written	0
[50]	0	0	1	0	79	AI Written	0
[51]	0	0	1	0	87	Not AI W.	0
[52]	44	78	100	16	71	Not AI W.	0

Note: Some of the detectors in this table provide results as percentages, while others return binary labels. ZeroGPT, GPTZero, Writer, JustDone, and QuillBot display percentage scores, which indicate how likely the text is AI-generated—the higher the value, the stronger the AI likelihood. On the other hand, UndetectableAI returns a binary label such as "AI Written" or "Not AI Written," while CopyLeaks uses a numerical binary output: 0 means human-written, and 1 means AI-generated. These two types of output must be interpreted differently when comparing overall detector performance across human-written texts.

It shows how each tool performed on the tested scientific papers. Some of the AI content detection tools, like Quillbot, Writer and CopyLeaks scored very accurate results, while others, like JustDone or UndetectableAI, show low accuracy results.

The results show that AI content detectors are not always accurate, differ in performance and some are not reliable at all. Some of the tools do better than others, but none were perfect. And, when it came to scientific papers written by actual humans, some detectors had a very hard time figuring them out. Sometimes, they would even misclassify these as being written by AI. After all, these tools work based on methods that we have discussed so far, which is not always the best way to decide whether something is human or AI-generated.

Another issue is that some tools gave random or strange results. For example, some tools misclassify texts on purpose to make users think they need to pay for a subscription to get "better" results. This creates an ethical problem, as people might pay for tools that are not reliable.

In short, such tools may be used to cross-check whether something was written by a machine, though they cannot be trusted as the sole criterion when making important decisions. They will just have to be handled with care, especially whenever a lot is at stake.

5. Discussion and Conclusion

In this study, we tested the popular AI content detectors. We provided human-written scientific articles published before the first commercial LLM, ChatGPT, was introduced. The results indicate that while capable, such tools are never reliable.

These findings are consistent with earlier studies discussed in the Literature Review section. Earlier studies showed that AI text detectors do not all work the same and often give wrong answers for both human-written and AI-written text. Our study found the same thing: the detectors we used were not always accurate. So, trusting them alone for school or work decisions is risky.

We also discovered that the detectors regularly misclassify papers written by human writers, at least in formal or technical styles. This indeed is a great problem in the areas of education or work, where such faults can ruin one's career or reputation.

AI detectors can assist in scanning the content, but cannot be used for any big decisions. They can be easily tricked or misclassified the given text, and. Always, a double check is needed for the results, when combined with human review or other checks, in cases like an academic investigation.

We lastly found that some tools may show incorrect results purposefully so that people will pay for a premium version of that tool. The situation has many ethical problems, and providers are supposed to be more honest about how their tools work.

As we discussed before, since the LLM increases in terms of performance, we need more robust detection tools. Also, there will have to be better and more accurate AI detection tools in the future, of course. Nevertheless, it will always be necessary for a human to make the final decision.

6. Limitations of the Study

This study has some limitations that readers should keep in mind. First, we used only a small group of research papers that were written before large language models became common. These papers were chosen because they are well-known, but the set is still small, and all of them are formal academic texts. We did not test blog posts, news stories, or creative writing, so the results might not work the same for those kinds of text.

All the papers were in English. We do not know how well today's AI-text detectors work with other languages. Some of the AI-text detectors don't even work with some of the most spoken languages. Such as Turkish, Italian, Chinese, etc. Also, we tested the papers with only free versions of the detectors. Paid versions might give different results.

Also, many detectors don't explain how they decide if the text is written by a person or AI. Earlier studies have already warned that these tools can be biased or unreliable. Finally, AI detectors change often. Our results show how they behaved at one moment in time, but future updates could improve or reduce their accuracy.

7. Future Studies

In the future, many different types of text can be used, such as social media posts, student essays, short stories, news, etc. Also, different languages can be tested, which can show whether the tools handle non-English content.

It would help to compare free and paid versions of the same detector to see if paying improves accuracy. Researchers could also look closely at the mistakes these tools make to learn why they call some human texts "AI-written."

An open website that stores human and AI examples and checks detectors regularly could make results clearer for everyone. Repeating the same tests from time to time and trying new detectors as they appear will show if the tools are getting better and help people use them wisely.

8. Declarations

Acknowledgement

We are grateful to Muğla Sıtkı Koçman University, Faculty of Engineering, Department of Computer Engineering, for the opportunities and resources that made this chapter possible.

Funding

This study did not receive any outside funding or support.

Authors` Contributions

All authors have participated in drafting the chapter. All authors read and approved the final version of the chapter.

Conflict of Interest

The authors certify that there is no conflict of interest with any financial organization regarding the material discussed in the chapter.

Data Availability

This review is entirely dependent on previously published literature and does not incorporate any novel datasets. All data that support the findings of this study are available within the cited articles and publicly accessible databases.

Ethics Approval

Authors followed the principles of scientific research and publication ethics. This study did not involve human or animal subjects and did not require additional ethics committee approval.

Declaration of AI Usage

AI-assisted tools were employed in this study for minor tasks such as grammar correction, language refinement, and proofreading. These tools were used transparently and in a manner that does not compromise the authors' intellectual contribution. The authors affirm that all substantive content reflects original thought and upholds academic integrity.

9. References

[1] OpenAI, *ChatGPT*. (2024). [Online]. Available: https://chat.openai.com/chat

[2] Google, *Gemini*. (2024). [Online]. Available: https://ai.google

[3] *Claude*. (2024). [Online]. Available: https://www.anthropic.com/claude

[4] Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., ... & Zou, J. Y. (2024). Mapping the increasing use of LLMs in scientific papers. *arXiv preprint arXiv:2404.01268.* https://doi.org/10.48550/arXiv.2404.01268

[5] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language

models: A survey. *arXiv preprint arXiv:2402.06196*. https://doi.org/10.48550/arXiv.2402.06196

[6] Huang, B., Chen, C., & Shu, K. (2025). Authorship attribution in the era of llms: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter*, *26*(2), 21-43. https://doi.org/10.1145/3715073.3715076

[7] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1-55. https://doi.org/10.1145/3703155

[8] Fraser, K. C., Dawkins, H., & Kiritchenko, S. (2025). Detecting ai-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research*, *82*, 2233-2278. https://doi.org/10.1613/jair.1.16665

[9] Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023, December). Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1236-1270).

[10] Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and machines*, *30*(4), 681-694. https://doi.org/10.1007/s11023-020-09548-1

[11] Jiao, J., Afroogh, S., Xu, Y., & Phillips, C. (2024). Navigating llm ethics: Advancements, challenges, and future directions. *arXiv preprint arXiv:2406.18841*. https://doi.org/10.48550/arXiv.2406.18841

[12] Mozes, M., He, X., Kleinberg, B., & Griffin, L. D. (2023). Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. *arXiv preprint arXiv:2308.12833*. https://doi.org/10.48550/arXiv.2308.12833

[13] Wagner, T. L., & Blewer, A. (2019). "The word real is no longer real": Deepfakes, gender, and the challenges of ai-altered video. *Open Information Science*, *3*(1), 32-46. https://doi.org/10.1515/opis-2019-0003

[14] Vahdati, D. S., Nguyen, T. D., Azizpour, A., & Stamm, M. C. (2024). Beyond deepfake images: Detecting ai-generated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4397-4408). http://dx.doi.org/10.1109/CVPRW63382.2024.00443

[15] Briesch, M., Sobania, D., & Rothlauf, F. (2023). Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop. *arXiv preprint arXiv:2311.16822*. https://doi.org/10.48550/arXiv.2311.16822

[16] Akram, A. (2023). An Empirical Study of AI-Generated Text Detection Tools. *Advances in Machine Learning & Artificial Intelligence, 4*(2), 44-55. https://doi.org/10.33140/AMLAI

[17] Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, *51*(1), 275-338. https://doi.org/10.1162/coli_a_00549

[18] Tang, R., Chuang, Y. N., & Hu, X. (2024). The science of detecting LLM-generated text. *Communications of the ACM*, *67*(4), 50-59. https://doi.org/10.1145/3624725

[19] Aydin, O., Karaarslan, E., Erenay, F. S., & Bacanin, N. (2025). Generative AI in Academic Writing: A Comparison of DeepSeek, Qwen, ChatGPT, Gemini, Llama, Mistral, and Gemma. arXiv preprint arXiv:2503.04765. https://doi.org/10.48550/arXiv.2503.04765

[20] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023, July). A watermark for large language models. In *International Conference on Machine Learning* (pp. 17061-17084). PMLR. https://proceedings.mlr.press/v202/kirchenbauer23a/kirchenbauer23a.pdf

[21] Christ, M., Gunn, S., & Zamir, O. (2024, June). Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory* (pp. 1125-1139). PMLR. https://proceedings.mlr.press/v247/christ24a.html

[22] Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., ... & Goldstein, T. (2023). On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*. https://doi.org/10.48550/arXiv.2306.04634

[23] Zhao, X., Li, L., & Wang, Y. X. (2022). Distillation-resistant watermarking for model protection in nlp. *arXiv preprint arXiv:2210.03312*. https://doi.org/10.48550/arXiv.2210.03312

[24] Gu, C., Huang, C., Zheng, X., Chang, K. W., & Hsieh, C. J. (2022). Watermarking pre-trained language models with backdooring. *arXiv preprint arXiv:2210.07543*. https://doi.org/10.48550/arXiv.2210.07543

[25] Taguchi, K., Gu, Y., & Sakurai, K. (2024). The impact of prompts on zero-shot detection of ai-generated text. *arXiv preprint arXiv:2403.20127*. https://doi.org/10.48550/arXiv.2403.20127

[26] Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., ... & Goldstein, T. (2024). Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*. https://doi.org/10.48550/arXiv.2401.12070

[27] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023, July). Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning* (pp. 24950-24962). PMLR. https://proceedings.mlr.press/v202/mitchell23a.html

[28] Bao, G., Zhao, Y., Teng, Z., Yang, L., & Zhang, Y. (2023). Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*. https://doi.org/10.48550/arXiv.2310.05130

[29] Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*. https://doi.org/10.48550/arXiv.1906.04043

[30] Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2019). Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*. https://doi.org/10.48550/arXiv.1911.00650

[31] Meister, C., & Cotterell, R. (2021). Language model evaluation beyond perplexity. *arXiv preprint arXiv:2106.00085*. https://doi.org/10.48550/arXiv.2106.00085

[32] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Nikolenko, S., Burnaev, E., ... & Piontkovskaya, I. (2023). Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, *36*, 39257-39276. https://doi.org/10.5555/3666122.3667828

[33] Abdalla, M. H. I., Malberg, S., Dementieva, D., Mosca, E., & Groh, G. (2023). A benchmark dataset to distinguish human-written and machine-generated scientific papers. *Information*, *14*(10), 522. https://doi.org/10.3390/info14100522

[34] Prajapati, M., Baliarsingh, S. K., Dora, C., Bhoi, A., Hota, J., & Mohanty, J. P. (2024, February). Detection of AI-generated text using large language model. In *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)* (pp. 735-740). IEEE. https://doi.org/10.1109/ESIC60604.2024.10481602

[35] Bhattacharjee, A., & Liu, H. (2024). Fighting fire with fire: can ChatGPT detect AI-generated text?. *ACM SIGKDD Explorations Newsletter*, *25*(2), 14-21. https://doi.org/10.1145/3655103.3655106

[36] Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of ai-generated

text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, *36*, 27469-27500. https://doi.org/10.5555/3666122.3667317

[37] Yong, Z. X., Menghini, C., & Bach, S. H. (2023). Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*. https://doi.org/10.48550/arXiv.2310.02446

[38] Ayoobi, N., Knab, L., Cheng, W., Pantoja, D., Alikhani, H., Flamant, S., ... & Mukherjee, A. (2024). Esperanto: Evaluating synthesized phrases to enhance robustness in ai detection for text origination. *arXiv preprint arXiv:2409.14285*. https://doi.org/10.48550/arXiv.2409.14285

[39] Wei, Z., Liu, Y., & Erichson, N. B. (2024). Emoji attack: Enhancing jailbreak attacks against judge llm detection. *arXiv preprint arXiv:2411.01077*. https://doi.org/10.48550/arXiv.2411.01077

[40] Thorat, S., & Yang, T. (2024). Which LLMs are Difficult to Detect? A Detailed Analysis of Potential Factors Contributing to Difficulties in LLM Text Detection. *arXiv preprint arXiv:2410.14875*. https://doi.org/10.48550/arXiv.2410.14875

[41] Hayawi, K., Shahriar, S., & Mathew, S. S. (2024). The imitation game: Detecting human and AI-generated texts in the era of ChatGPT and BARD. *Journal of Information Science*, 01655515241227531. https://doi.org/10.1177/01655515241227531

[42] Kar, S. K., Bansal, T., Modi, S., & Singh, A. (2025). 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*, *47*(3), 275-278. https://doi.org/10.1177/02537176241247934

[43] Singh, A. (2023, March). A comparison study on AI language detector. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0489-0493). IEEE. https://doi.org/10.1109/CCWC57344.2023.10099219

[44] Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, *19*(1), 1-16. https://doi.org/10.1007/s40979-023-00140-5

[45] Koralus, P., & Wang-Maścianica, V. (2023). Humans in humans out: On gpt converging toward common sense in both success and failure. *arXiv preprint arXiv:2303.17276*. https://doi.org/10.48550/arXiv.2303.17276

[46] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794-7803). https://doi.org/10.1109/CVPR.2018.00813

[47] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[48] Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*. https://doi.org/10.48550/arXiv.1811.12808

[49] Mor-Yosef, L., & Avron, H. (2019). Sketching for principal component regression. *SIAM Journal on Matrix Analysis and Applications*, *40*(2), 454-485. https://doi.org/10.1137/18M1188860

[50] Montagnier, L., Aïssa, J., Capolupo, A., Craddock, T. J., Kurian, P., Lavallee, C., ... & Vitiello, G. (2017). Water bridging dynamics of polymerase chain reaction in the gauge theory paradigm of quantum fields. *Water*, *9*(5), 339. https://doi.org/10.3390/w9050339

[51] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141). https://doi.org/10.1109/CVPR.2018.00745

[52] Garibyan, L., & Avashia, N. (2013). Polymerase chain reaction. *Journal of investigative dermatology*, *133*(3), 1-4. https://doi.org/10.1038/jid.2013.1