Survey for Detecting AI-generated Content

Yu Wang[1,a], Ziyan Wang[2,b]

[1] School of Foreign Languages, Dalian Jiaotong University, Dalian, China

[2]School of Civil Engineering, Dalian Jiaotong University, Dalian, China

[a] wangyu9337@163.com, [b] dljt227@163.com

Abstract. In large language models (LLMs) field, the rapid advancements have significantly improved text generation, which has blured the distinction between AI-generated and human-written texts. These developments have sparked concerns about potential risks, such as disseminating fake information or engaging in academic cheating. As the responsible use of LLMs becomes imperative, the detection of AI-generated content has become a crucial task. Most existing surveys on AI-generated text (AIGT) Detection have analysed the detection approaches from a computational perspective, with less attention to linguistic aspects. This survey seeks to provide a fresh perspective to drive progress in the area of LLM-generated text detection. Futhermore, in order to make the assessment more explainable, we emphasize the great importence of leveraging specific parameters or metrics to linguistically evaluate the candidate text.

Keywords: Detect AIGT, Linguistic Analysis, Explainable AI.

1. Introduction

1.1 Challenges and Evolution of AI-generated Text Detection

If the unchecked proliferation of AIGC within the field of computing continues, our beloved artificial intelligence will trigger another 'paperdemic'. [1] investigates the COVID-19 articles published in 2020, underscoring their detrimental effects on research quality and review efficiency. This serves as a warning sign of potential oversaturation in AIGC, which could impede the healthy development of the entire academic and information ecosystem.

The focus of most existing surveys on AIGT detection tends to center around computational models, often overlooking the importance of linguistic aspects. [2] provides a detailed look at how LLMs use data collection, feature selection, and classification models to boost detection accuracy. [3] explores various machine learning and deep learning algorithms, and uses features like TF-IDF to assess model performance on human and ChatGPT-generated texts. Methods such as model signatures and neural-based detectors are explored in [3] to distinguish AI-generated texts. Despite these advances, novel evasion techniques continue to emerge. [5] discusses 'spaceinfo,' which is a tactic that involves adding spaces before commas to reduce detectability, and [6] notes how paraphrasing can help bypass detectors. Advanced techniques like SICO, discussed in [7]. Various adversarial attacks, highlighted in [8] and [9] , are ongoing challenges and the critical need for enhanced AIGT detection methods.

Although many existing models are effective, even well-performing multitasking models(like ChatGPT) struggle with tasks like detecting AIGC. Recently, more models in vertical field have emerged, such as K2[10] for geosciences, GEOGALACTICA[11] for scientific accuracy, and EduChat[12] for education. The model described above customizes large language models (LLMs) for specific fields, potentially improving the generation of texts relevant to those domains. We collected several texts from Wikipedia, GPT-4, and Llama across different domains. Our simple Turing test revealed that distinguishing between human and AI-generated texts remains a challenge for both human and models. Table 1 illustrates the ongoing struggle in accurately identifying AI-generated content.

Table 1

	Detection Method	human	GPT-4

	Text Origin	human-written	GPT-generated	human-written	GPT-generated
"Geoscience" from	Wikipedia	√			√
	GPT-4	√		√	
	Llama		√		√
"Education" from	Wikipedia	√		√	
	GPT-4		√		√
	Llama	√		√	
"Cultural heritage" from	Wikipedia		√		√
	GPT-4	√		√	
	Llama		√		√

Today, many works treat AIGT detection as a classification and regression problem, which inherently introduces certain biases from a linguistic perspective. Therefore, by conducting some linguistic analysis and induction, we hope to motivate existing research to analyze detection task from linguistic perspectives and back to the essence of machine-generated content.

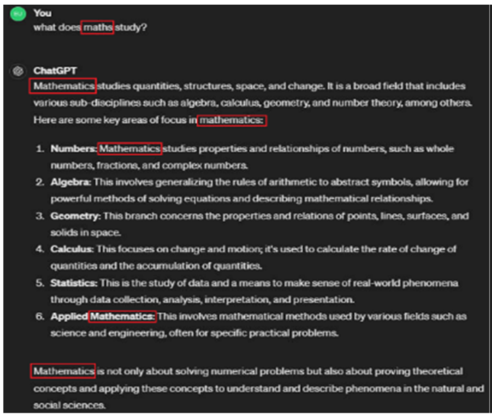
1.2 Technological Advances in AI-generated Text Detection

From a linguistic perspective, large language model text generation still needs improvement, analyzed from six angles:

1. In highly predictable contexts, ChatGPT's responses are usually formal, while human responses are more colloquial, using abbreviations, colloquialisms, metaphors, etc. In Figure 1, ChatGPT uses "mathematics" while humans tend to use "maths" [13].

2. Syntactic ambiguity is a major challenge in natural language processing, especially in contexts involving polysemous words and ambiguous structures. [14] tested the performance of language models in Chinese grammar acceptance using the CoLAC dataset, and the results showed that the models often failed to correctly resolve syntactic ambiguity. This result highlights that ChatGPT may overlook key contextual clues when dealing with sentences such as "The hunter killed the poacher with a gun" [13] that may trigger multiple interpretations.

Consider Figure 2: A travel story mentions a traveler and a guide. The sentence, "The traveler followed the guide with binoculars," can be interpreted differently based on context: If only one guide is mentioned, it might mean "The traveler used binoculars to follow the guide." If two guides are mentioned, it could mean "The traveler followed the guide who had binoculars."



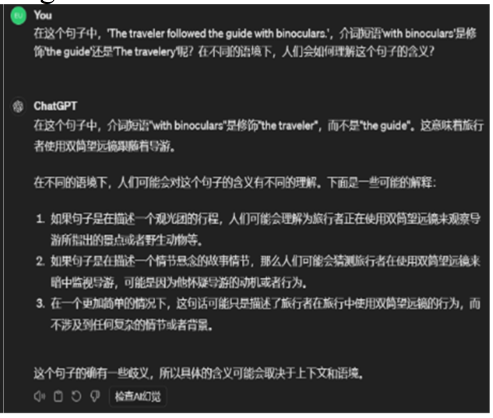
You
what does maths study?

ChatGPT
Mathematics studies quantities, structures, space, and change. It is a broad field that includes various sub-disciplines such as algebra, calculus, geometry, and number theory, among others. Here are some key areas of focus in mathematics:

1. **Numbers:** Mathematics studies properties and relationships of numbers, such as whole numbers, fractions, and complex numbers.
2. **Algebra:** This involves generalizing the rules of arithmetic to abstract symbols, allowing for powerful methods of solving equations and describing mathematical relationships.
3. **Geometry:** This branch concerns the properties and relations of points, lines, surfaces, and solids in space.
4. **Calculus:** This focuses on change and motion; it's used to calculate the rate of change of quantities and the accumulation of quantities.
5. **Statistics:** This is the study of data and a means to make sense of real-world phenomena through data collection, analysis, interpretation, and presentation.
6. **Applied Mathematics:** This involves mathematical methods used by various fields such as science and engineering, often for specific practical problems.

Mathematics is not only about solving numerical problems but also about proving theoretical concepts and applying these concepts to understand and describe phenomena in the natural and social sciences.

Figure 1



You
在这个句子中，'The traveler followed the guide with binoculars.'，介词短语'with binoculars'是修饰'the guide'还是'The travelery'呢？在不同的语境下，人们会如何理解这个句子的含义？

ChatGPT
在这个句子中，介词短语"with binoculars"是修饰"the traveler"，而不是"the guide"。这意味着旅行者使用双筒望远镜跟随着导游。

在不同的语境下，人们可能会对这个句子的含义有不同的理解。下面是一些可能的解释：

1. 如果句子是在描述一个观光团的行程，人们可能会理解为旅行者正在使用双筒望远镜来观察导游所指出的景点或者野生动物等。
2. 如果句子是在描述一个情节悬念的故事情节，那么人们可能会猜测旅行者在使用双筒望远镜来暗中监视导游，可能是因为他怀疑导游的动机或者行为。
3. 在一个更加简单的情况下，这句话可能只是描述了旅行者在旅行中使用双筒望远镜的行为，而不涉及到任何复杂的情节或者背景。

这个句子的确有一些歧义，所以具体的含义可能会取决于上下文和语境。

Figure 2

3. Vocabulary features: [15] compared the linguistic features of human-written and machine-generated articles, examining vocabulary richness, including vocabulary density, precision, and variability. Their study found that although machine-generated texts might be syntactically complex, they are usually less rich in vocabulary than human authors, especially for advanced English learners[15]. Human responses, though shorter, tend to have a higher density of used vocabulary[16].

This is because ChatGPT focuses on the question itself, while human thoughts are more divergent. According to [16] formula D = 100 × V /(L × N).

- Vocabulary size (V): 228 unique words.
- Average length (L): 156 words.
- N (number of answers): 5.
- Density (D): Approximately 29.23（ChatGPT）

- Vocabulary size (V): 134 unique words
- Average length (L): 201 words
- N (number of answers): 1.
- Density (D): Approximately 66.67（Wikipedia）

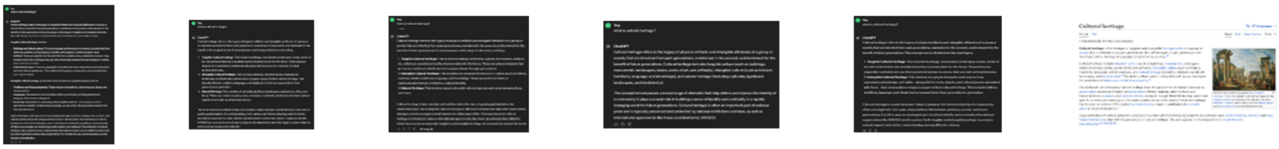

(a) (b) (c) (d) (e) (f)

Figure 3

4. When language users fail to use language appropriately to in specific contexts, pragmatic failures will occur, which affect the acceptability of language. Usually, humans can judge whether a sentence is grammatically correct and whether it is pragmatically appropriate. [14] tested the performance of LLM on CoLAC, the results on CoLAC was only at a random level. This indicates that current large-scale models still have significant limitations in accurately assessing the acceptability of Chinese language. While machine-generated text might be grammatically correct, it often lacks naturalness or contextual appropriateness at the pragmatic level.

5. Part-of-Speech Analysis: LLMs use more NOUN, VERB, DET, ADJ, AUX, CCONJ, and PART words, and fewer ADV and PUNCT words. LLMs tend to provide objective answers, while humans prefer subjective expressions. A high proportion of nouns usually indicates informativeness and objectivity, and frequent conjunctions indicate clear structural and logical relationships in the text[17].

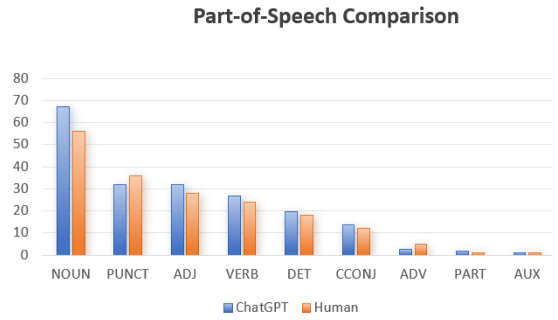


Figure 4

6. Dependency Analysis: ChatGPT has notably shorter conjunction (CONJ) relations. As mentioned, compared to humans, ChatGPT typically uses more conjunctions to make content more logical. Additionally, ChatGPT shows longer distances for punctuation and dependency relations. Because humans tend to use more punctuation and a variety of grammatical structures to express emotions, while ChatGPT tends to use more conjunctions and adverbs to express clear structures and logical relationships[16].

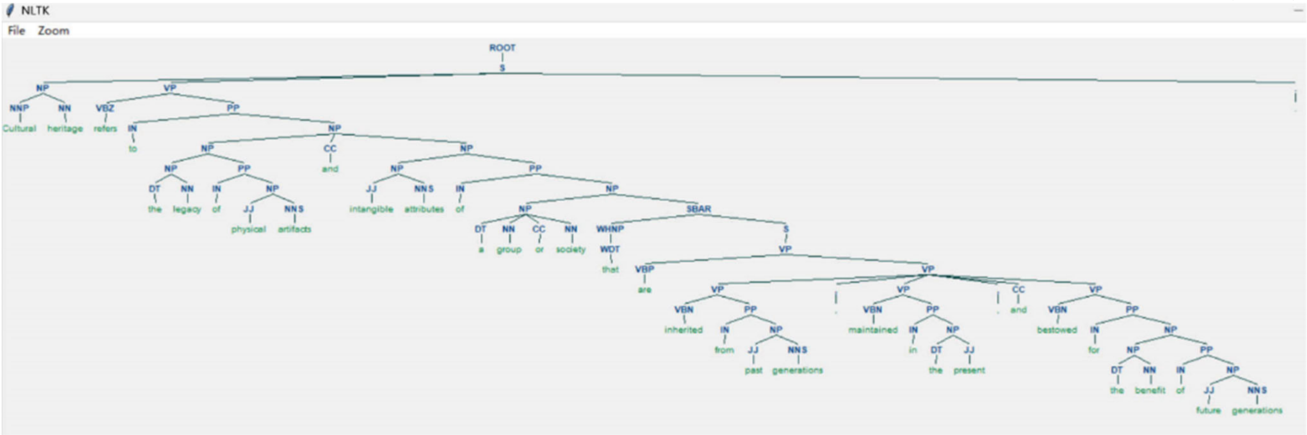


Figure 5

In this digital age, ChatGPT learns and answers relying on human knowledge. Excessive reliance may lead to serious problems, despite providing lots of convenience.

1. Knowledge Volume: In some key fields such as biology, education, and medicine, the risk of knowledge stagnation looms if experts and students neglect originality and innovation. With no fresh knowledge inputs, the reservoir of learning materials will drain, and human knowledge base will no longer expand even converge.

2. Language Evolution: Language serves as a medium for thought and cultural development. With dynamic inherent, Language evolves through diverse inputs, which promotes creativity and adaptability. Excessive dependence on ChatGPT may diminish language evolution and cultural diversity.

3. Cultural Diversity: Language contains subtle cultural differences. When ChatGPT dominates language input, homogenization occurs, marginalizing minority languages, weakening individual expression and cross-cultural communication. This impacts cultural inheritance and diminishes cultural diversity.

In summary, it is important to detect whether content is generated by LLM or not. Research indicates that current studies primarily focus on black-box model detectors, yet they lack of transparency and neglect the methods used by models to generate text, which is not convincing. Therefore, this paper proposes that the future advancements should employe specific parameters and metrics to evaluate AIGC, instead of model-based detection.

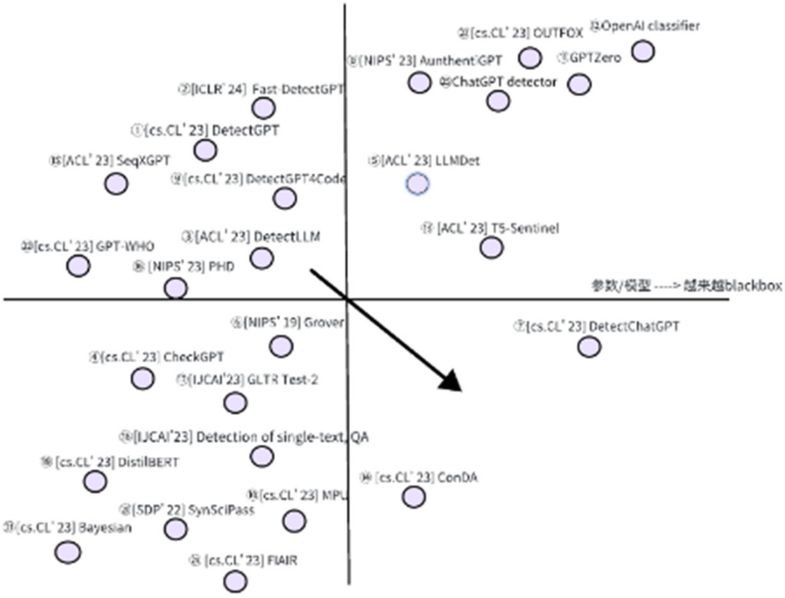


Figure 6

2. Existing Methods

With text generation technology thriving, identifying the authorship of texts generated by human or machine are increasingly sophisticated. Technology is undergoing a transformation. These changes not only fosters LLM development but also necessitates enhanced models for distinguishing their outputs.

Traditionally, text classification has been seen as the binary task. However, technological advances and growing needs demand more nuanced classification methods.[18], [19] propose multiclass classification for text attribution, which not only differentiates sources but also identifies them. For example, SynSciPass [18] labels texts by technological processes (e.g. translation), MGTBench [19] benchmarks machine-generated text detection methods, and GPT-who [20] leverages statistical UID to discern differences between LLM-generated and human texts, accurately determining authorship. Moreover, [21] set a boundary detection task where texts start as human-written and then transition to machine continuations, aiming to identify the point of transition. Additionally, [21] notes that using LLMs directly as detectors, like ChatGPT-turbo, is less effective. Proxy models[22], [23], and [24], however, show promise as detectors using smaller or partially trained LLMs.

This study introduces a novel approach to AIGT detection by reverting to basic linguistic analysis. We explore how linguistic indicators affect detection outcomes and underscore the importance of interpretability in classification tasks. The following models are categorized into open-source, closed-source, and metric-based models.

Table 2

Model name	basemodel	quadrant	citation	Open-source models	Close-source models	metric
DetectGPT	T5-3B, T5-11B	二	①	√		√
Fast-DetectGPT	GPT-2, OPT-2.7b, GPT-NEO-2.7B, GPT-J-6B, GPT-NEOX-20b	二	②	√		√
DetectLLM(LRR,NPR)	GPT-2, GPT-NEO, GPT-J, OPT, Llama	二	③	√		√
CheckGPT	RoBERTa	三	④	√		√
LLMDet	GPT-2, OPT, Llama	一	⑤	√		√
Grover	GPT-2	三	⑥	√		√
DetectChatGPT	GPT-2	四	⑦	√		√
AuthentiGPT	GPT-3.5-turbo	一	⑧		√	√
DetectGPT4Code	Incoder-6B	二	⑨	√		√
MPU	RoBERTa	三	⑩	√		√
GPTZero		一	⑪		√	√
OpenAI	GPT-3.5, GPT-4	一	⑫		√	√
T5-Sentinel	T5	二	⑬	√		√
ConDa	RoBERTa	四	⑭	√		√
SeqXGPT	GPT-2-XL, GPT-Neo, GPT-J, Llama	二	⑮	√		√
PHD	XLM-RoBERTa-base	二	⑯	√		√
GLTR Test-2		三	⑰	√		√
A deep classifier for single-text and QA detection.	RoBERTa	三	⑱	√		√

DistilBERT	BERT	三	⑲	√		√
SynSciPass	BERT-base	三	⑳	√		√
Bayesian Surrogate model	Gaussian Process model	三	㉑	√		√
GPT-who	GPT-2-XL	二	㉒	√		√
OUTFOX	ChatGPT (gpt-3.5-turbo-0613)	一	㉓		√	√
FLAIR	GPT-3、ChatGPT、Llama、Alpaca、Vicuna	三	㉔		√	√
ChatGPT detector	ChatGPT-3.5	一	㉕	√		√

2.1 recommend commercial models

OpenAI's text classifier [25] leverages fine-tuned versions of large language models, such as GPT-3 or possibly GPT-4, enhancing its ability to differentiate AI from human texts over zero-shot models that lack specific task tuning. The OUTFOX framework [26] involves a Detector and an Attacker, which works in mutual learning method. Focusing on input/output behaviors rather than the models' internal mechanics. The detector improves its recognition technology, while the attacker enhances its text generation capabilities. Based on text perplexity and patterns, GPT Zero [27], another AI content detector, determines whether content from sources like Google Bard or GPT4 generated by AI. Based on deviation from expected LLM outputs, AuthentiGPT [28] identifies human texts by employing a black-box model, which works in the method of denoising texts and measure semantic similarity comparing with artificially noised versions. Lastly, the ChatGPT detector [29] targets academic publications, it provides a high-accuracy AI-generated text detector specific to scholarly texts, although its use is restricted to this specific field.

2.2 open-source model

2.2.1 Fine-tuning

Fine tuning refers to pre training a model on a specific dataset that has already learned the general features of a large and diverse dataset, then changing the weight of the model parameters to obtain the desired output results.

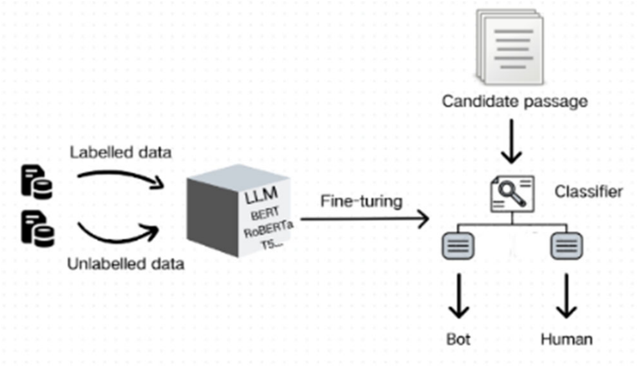


Figure 7

Grover[30] is designed for the detection of fake news. It utilizes a language model trained to generate and detect fake news. DistilBERT [31], a streamlined BERT model, is utilized to detect restaurant reviews generated by ChatGPT. It uses SHAP to enhance decision-making transparency by highlighting influential features. ConDa [32], a contrastive domain adaptation framework, is trained with labeled data to detect AI-generated texts based on a pre-trained RoBERTa model, especially in news articles.[16] developed three detection models using RoBERTa: GLTR Test-2, which has employed logistic regression; a deep classifier for single-text detection and a QA-specific deep classifier. T5-Sentinel [33] uses the T5 model's ability to predict the next token, which works in the method of transforming a multi-class classification issue into several binary decisions. Each question judges whether a text is produced by a particular LLM or not. MPU [34] reconfigures BERT and RoBERTa for text detection by fine-tuning on Positive Unlabeled problems. It has created an open-source detector that treats human-generated short texts as positive and machine-generated ones as 'unlabeled', which enhancing the detection accuracy for short texts.

2.2.2 zero-shot

On the contrary, zero sample detection method can be applied without the need for additional training data. It can analyze the average log probability of each word or the average ranking among possible options to determine the author.

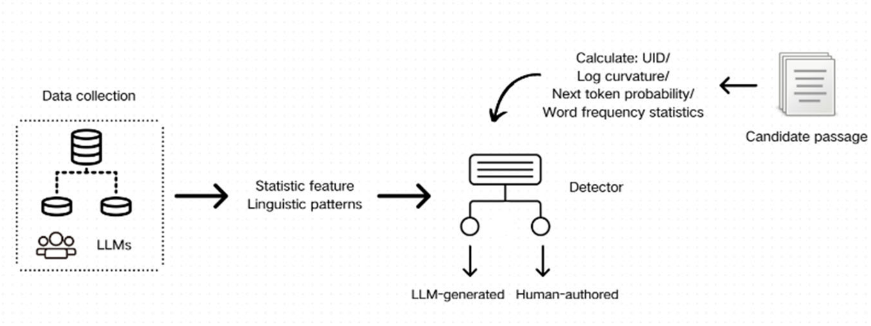


Figure 8

DetectGPT [35] utilizes the fundamental properties of pretrained language models for zero-shot detection. It analyzes the change in log probabilities caused by slight text perturbations to determine the likelihood of a text generated by machine. DetectChatGPT [36] working as a black-box, zero-shot model, queries internal probability distributions from models like GPT-2 or GPT-Neo. This method synthesizes outputs from various models to estimate text characteristics, making it suitable for situations lacking direct access to model internals. DetectGPT4Code [37] adapts the DetectGPT method to identify code source. It uses smaller models like PolyCoder-160M to analyze code-specific probability curves. Fast-DetectGPT [38] relies on existing large language models like GPT-3 or GPT-4 to compute log probabilities. It creates text samples with minor lexical differences to compare their conditional probabilities, identifying word choices indicative of machine generation.

SeqXGPT [39], a sentence-level AI-generated text detector, employs open-source language models to analyze and detect text using the log probabilities of words or tokens. CheckGPT [40] uses RoBERTa for deep textual analysis, incorporating an attention-BiLSTM classification module, and operates independently of the internal details of the generating models (model-agnostic). LLMDet [41] calculates proxy perplexity for LLMs by recording the probability of the next token in significant n-grams. This method, designed for closed source models, still faces challenges in constructing a "dictionary" for these models. The Bayesian Surrogate model [22] utilizes Gaussian processes to emulate the local structures and probabilities of texts from language models, requiring a thorough understanding of LLM outputs for white-box applications and adapting to different data scales without traditional parameter constraints. GPT-who [20] depends on understanding text generation, particularly through information density, to resolve text attribution issues and accurately determine authorship. DetectLLM (LRR, NPR) [42] enhances zero-shot detection methods based on log-rank statistical information.

In reviewing these models, we find that the evaluation metrics used are computer-oriented, partly because the data judged is synthetic and partly because these models aim to perform well on benchmark leaderboards. Below, we list the metrics used.

2.3 Metrics

2.3.1 Traditional evaluation metrics

The traditional indicators used to evaluate model generated text, such as Perplexity , Accuracy , Precision , Recall, F1-Score, and AUROC are as follows:

Perplexity (PPL) quantifies how well a model predicts a sample text, calculated as the exponential of the negative average log-likelihood of the text according to the model (1). Accuracy is calculated as the ratio of correctly predicted observations to the total observations (2). Precision measures the accuracy of the positive predictions made by a classification model (3). Recall measures the proportion of actual positives that are correctly identified by a classification model (4). The F1-Score is the harmonic mean of precision and recall, offering a balance between them (5). True Positive Rate (TPR), also known as the Recall, is the proportion of positives correctly identified by the model. False Positive Rate (FPR), is the proportion of negatives incorrectly marked as positives out of all actual negatives (6). AUROC, in binary classification, displays results at various thresholds, reflecting the model's ability to distinguish between positives and negatives to assess its classification performance (7).

$ppl = \mathrm{e}^{(-\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\log \mathrm{P}(\omega_i \mid \omega_{i-1},...,\omega_1))}$ (1)	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ (2)	$Precision = \frac{TP}{TP+FP}$ (3)
$Recall = \frac{TP}{TP+FN}$ (4)	$F_1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$ (5)	$FPR = \frac{FP}{FP+TN}$ (6)
$AUROC = \int_0^1 \frac{TP}{TP+FP} d \frac{FP}{FP+TN}$ (7)		

2.3.2 Alternative evaluation indicators

Recent detection studies have focused on using LLMs (black-box and white-box), with almost no concentrate on language itself. Specific linguistic metrics have been few used. Few studies have achieved detection tasks in a fully white-box manner, which can offer convincing interpretability. Clearly, there remains a lack of research in this specific linguistic aspect of LLMs. To our knowledge, most works on LLMs have not reported experimental results for this task. After research, we found that many studies attempt to enhance the interpretability of LM detection results, such as using log curves, PPL, entropy to explain the results, or using explanatory tools such as SHAP [31]and visualization techniques (t-SNE projection) [33]. It can be foreseen that there is a general desire among researchers for these detection results more explainable.

In addition to these traditional metrics mentioned above, from a statistical perspective, Uniform Information Density (UID) feature [20] capitalizes on the human tendency to use shorter elements for less information and longer elements for more, which aims to maintain a nearly uniform information rate. PHD [43] employs topological data analysis, specifically persistent homology. It analyzes samples' mathematical and topological properties, and transforms them into point clouds in high-dimensional space.

3. Database/ Corpus

The development and utilization of these diverse databases and corpora play a crucial role in advancing the understanding of AIGC detection and enhancing the precision of detection mechanisms.

The ArguGPT[15] corpus is a significant contribution to computational linguistics, which contains a balanced dataset of 4,153 human and 4,038 AI-generated essays from academic tests like TOEFL and GRE. It enhances understanding of linguistic differences between human and AI-generated texts, and also advances AIGC detection technology. Experiments have shown that SVM and RoBERTa-based detectors have achieved over 90% accuracy on this corpus. Guo et al. created HC3[16], the first

Human-ChatGPT Comparison Corpus, which has gathered paired texts from humans and ChatGPT across multiple languages and fields such as computer science and medicine. This dataset supports computational linguistics and ChatBot development. The RoFT database[21] uses a game-like annotation platform to challenge participants in oreder to identify when texts switch from human to AI-generated, which has accumulated over 20,000 annotations across various domains and enhancing assessments of text generation models. The GPABenchmark[40] is specifically designed for assessing ChatGPT's utility in academic writing, with tasks that include generating, and polishing abstracts. GossipCop++ and PolitiFact++[44] datasets focus on fake news detection, it combines high-quality articles written by human and LLM-synthesized news. These datasets provide a robust framework to study detection biases and effectiveness, which are critical for developing and evaluating fake news detectors.

4. Discussion

Essentially, detecting AIGT tasks should be analyzed through mathematical statistical formulas. However, currently there is too much work piling up on black box detection. Currently, it is difficult to detect text generated by black-box model, although the fine-tuned white box language model detector has decent accuracy on specific topics. Detectors, as deep learning models, still inherently exhibit black-box characteristics due to complex neural networks and massive nonlinear data. As mentioned earlier, many efforts are being made to enhance the interpretability of LM detection results, but these "explanations" merely provide insights into which parts of the model affect decision-making. Most end users or developers may still not fully understand the complex layer structure within the model.

To some extent, methods such as external knowledge retrieval, classifier-based detection, consistency checks, and manual reviews can mitigate this issue, but they do not require understanding of the model's internal structure or decision-making processes[45]. These only involve comparing model outputs with known facts. Existing detection benchmarks, such as FELM and HaluEval, provide standards for assessing the factuality and consistency of models, but these methods and benchmarks are still based on black-box approaches. Black-box methods are not acceptable in key areas, despite of practicality in simplifying the detection process, such as healthcare, finance, and law due to a lack of deep understanding of the model's decision-making pathways. In these fields, incorrect information can lead to serious consequences, thus there is a need for more precise diagnostics and solutions to address the root causes of hallucinations.

5. Conclusion

In a word, based on linguistic perspective, we advocate shifting the focus from model-centric to data-centric. AIGT detection, using numerical analysis, not only facilitates the detection of texts generated by open-source and closed-source LMs but also enhances high interpretability. The white-box method provides a transparent view of the model's working principles. This can enable researchers and developers to understand and refine model behavior, thereby ensuring the accuracy and reliability of generated content.

References

[1] Yang, Y., Zhao, N., Ma, T., Yuan, Z., & Deng, C. (2023). 'Paperdemic' during the COVID-19 pandemic. European Journal of Internal Medicine, 108, 111-113.

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

[3] N. Islam, D. Sutradhar, H. Noor, J. T. Raya, M. T. Maisha, and D. M.Farid, "Distinguishing human generated text from chatgpt generated text using machine learning," arXiv preprint arXiv:2306.01761, 2023.

[4] Lalit, D.R., Bhutani, D.P., Verma, D.N., & jain, A. AI-Generated Text Detection: A Review.

[5] Cai, S., & Cui, W. (2023). Evade ChatGPT Detectors via A Single Space. ArXiv, abs/2307.02599.

[6] Krishna, K.; Song, Y.; Karpinska, M.; Wieting, J.; and Iyyer,M. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. arXiv:2303.13408.

[7] Lu, N., Liu, S., He, R., & Tang, K. (2023). Large Language Models can be Guided to Evade AI-Generated Text Detection. ArXiv, abs/2305.10847.

[8] Shi, Z., Wang, Y., Yin, F., Chen, X., Chang, K., & Hsieh, C. (2023). Red Teaming Language Model Detectors with Language Models. Transactions of the Association for Computational Linguistics, 12, 174-189.

[9] Peng, X., Zhou, Y., He, B., Sun, L., & Sun, Y. (2024). Hidding the Ghostwriters: An Adversarial Evaluation of AI-Generated Student Essay Detection. ArXiv, abs/2402.00412.

[10] Deng, C., Zhang, T., He, Z., Xu, Y., Chen, Q., Shi, Y., Zhou, L., Fu, L., Zhang, W., Wang, X., Zhou, C., Lin, Z., & He, J. (2023). K2: A Foundation Language Model for Geoscience Knowledge Understanding and Utilization. Web Search and Data Mining.

[11] Lin, Z., Deng, C., Zhou, L., Zhang, T., Xu, Y., Xu, Y., He, Z., Shi, Y., Dai, B., Song, Y., Zeng, B., Chen, Q., Shi, T., Huang, T., Xu, Y., Wang, S., Fu, L., Zhang, W., He, J., Ma, C., Zhu, Y., Wang, X., & Zhou, C. (2023). GeoGalactica: A Scientific Large Language Model in Geoscience. ArXiv, abs/2401.00434.

[12] Dan, Y., Lei, Z., Gu, Y., Li, Y., Yin, J., Lin, J., Ye, L., Tie, Z., Zhou, Y., Wang, Y., Zhou, A., Zhou, Z., Chen, Q., Zhou, J., He, L., & Qiu, X. (2023). EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education. ArXiv, abs/2308.02773.

[13] He Wei. Opportunities and Challenges Brought by ChatGPT from the Linguistic Perspective[J]. Journal of Ocean University of China (Social Sciences), 2023, (6): 94-103.

[14] Hu, H., Zhang, Z., Huang, W., Lai, J.Y., Li, A., Ma, Y., Huang, J., Zhang, P., & Wang, R. (2023). Revisiting Acceptability Judgements. ArXiv, abs/2305.14091.

[15] Liu, Y.; Zhang, Z.; Zhang, W.; Yue, S.; Zhao, X.; Cheng, X.;Zhang, Y.; and Hu, H. 2023. ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models. arXiv:2304.07666.

[16] Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.;Yue, J.; and Wu, Y. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv:2301.07597.

[17] Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D.F., & Chao, L.S. (2023). A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions. ArXiv, abs/2310.14724.

[18] Rosati, D. (2022). SynSciPass: detecting appropriate uses of scientific text generation. SDP.

[19] He, X., Shen, X., Chen, Z.J., Backes, M., & Zhang, Y. (2023). MGTBench: Benchmarking Machine-Generated Text Detection. ArXiv, abs/2303.14822.

[20] Venkatraman, S., Uchendu, A., & Lee, D. (2023). GPT-who: An Information Density-based Machine-Generated Text Detector. ArXiv, abs/2310.06202.

[21] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In Proceedings of AAAI.

[22] Deng, Z., Gao, H., Miao, Y., & Zhang, H. (2023). Efficient Detection of LLM-generated Texts with a Bayesian Surrogate Model. ArXiv, abs/2305.16617.

[23] Mireshghallah, F., Mattern, J., Gao, S., Shokri, R., & Berg-Kirkpatrick, T. (2023). Smaller Language Models are Better Black-box Machine-Generated Text Detectors. ArXiv, abs/2305.09859.

[24] Aguilar-Canto, F.J., Cardoso-Moreno, M.A., Jiménez, D., & Calvo, H. (2023). GPT-2 versus GPT-3 and Bloom: LLMs for LLMs Generative Text Detection. IberLEF@SEPLN.

[25] OpenAI. Openai ai text classifier, January 2023. URL https://beta.openai.com/ai-text-classifier.

[26] Koike, R., Kaneko, M., & Okazaki, N. (2023). OUTFOX: LLM-generated Essay Detection through In-context Learning with Adversarially Generated Examples. AAAI Conference on Artificial Intelligence.

[27] Edward Tian. Gptzero: an ai detector, 2023. URL https://gptzero.me/.

[28] Guo, Z., & Yu, S. (2023). AuthentiGPT: Detecting Machine-Generated Text via Black-Box Language Models Denoising. ArXiv, abs/2311.07700.

[29] Prillaman, M. (2023). 'ChatGPT detector' catches AI-generated papers with unprecedented accuracy. Nature.

[30] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending Against Neural Fake News. ArXiv, abs/1905.12616.

[31] Mitrović S, Andreoletti D, Ayoub O. ChatGPT or human? detect and explain. explaining decisions of machine learning model for detecting short ChatGPT-generated text. arXiv preprint arXiv. 2023;2301.13852.

[32] Bhattacharjee, A., Kumarage, T., Moraffah, R., & Liu, H. (2023). ConDA: Contrastive Domain Adaptation for AI-generated Text Detection. ArXiv, abs/2309.03992.

[33] Chen, Y., Kang, H., Zhai, V., Li, L., Singh, R., & Raj, B. (2023). Token Prediction as Implicit Classification to Identify LLM-Generated Text. Conference on Empirical Methods in Natural Language Processing.

[34] Tian, Y., Chen, H., Wang, X., Bai, Z., Zhang, Q., Li, R., Xu, C., & Wang, Y. (2023). Multiscale Positive-Unlabeled Detection of AI-Generated Texts. ArXiv, abs/2305.18149.

[35] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. CoRR, abs/2301.11305, 2023. doi: 10.48550/arXiv.2301.11305.

[36] Park, J., Rashid, A., Li, L., & Mitchell, E. (2023). DetectChatGPT: Black-Box Zero-Shot Detection of LLM-Generated Text.

[37] Yang, X., Zhang, K., Chen, H., Petzold, L.R., Wang, W.Y., & Cheng, W. (2023). Zero-Shot Detection of Machine-Generated Codes. ArXiv, abs/2310.05103.

[38] 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, abs/2310.05130.

[39] Wang, P., Li, L., Ren, K., Jiang, B., Zhang, D., & Qiu, X. (2023). SeqXGPT: Sentence-Level AI-Generated Text Detection. Conference on Empirical Methods in Natural Language Processing.

[40] Liu, Z., Yao, Z., Li, F., & Luo, B. (2023). On the Detectability of ChatGPT Content: Benchmarking, Methodology, and Evaluation through the Lens of Academic Writing.

[41] Wu, K., Pang, L., Shen, H., Cheng, X., & Chua, T. (2023). LLMDet: A Third Party Large Language Models Generated Text Detection Tool. Conference on Empirical Methods in Natural Language Processing.

[42] Su, J., Zhuo, T.Y., Wang, D., & Nakov, P. (2023). DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. Conference on Empirical Methods in Natural Language Processing.

[43] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Piontkovskaya, I., Nikolenko, S.I., & Burnaev, E. (2023). Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts. ArXiv, abs/2306.04723.

[44] Su, J., Zhuo, T.Y., Mansurov, J., Wang, D., & Nakov, P. (2023). Fake News Detectors are Biased against Texts Generated by Large Language Models. ArXiv, abs/2309.08674.

[45] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ArXiv, abs/2311.05232.