EasyChair Preprint

№ 15361

On the Gap Between AI-Generated and Human-Written Patent Texts

Zhanhao Xiao, Wei Hu, Yanqiang Wu, Weiqi Chen, Huihui Li and Xiaoyong Liu

November 4, 2024

On the Gap between AI-generated and Human-written Patent Texts

Zhanhao Xiao[1,2], Wei Hu[1], Yanqiang Wu[1], Weiqi Chen[1], Huihui Li[1,2], and Xiaoyong Liu[3,✉]

[1] School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China
[2] Guangdong Provincial Key Laboratory of Intellectual Property and Big Data, Guangdong Polytechnic Normal University, Guangzhou 510665, China
[3] School of Data Science and Engineering, Guangdong Polytechnic Normal University, Guangzhou 510665, China

Abstract. Since the GPT-X models have made progress in generative tasks, a large number of large language models (LLMs) have sprung up. When the powerful features of LLMs have attracted the interest of numerous researchers, their misuse has also become a source of growing concern for human beings. In fact, LLMs have been used to generate fake news, fake academic papers, and fake patent application documents. Detecting whether content is generated by artificial intelligence (AI) has been a significant problem. Unfortunately, to our knowledge, there is currently no existing research focused on AI-generated patent text detection, nor are there any datasets tailored for patents publicly available. In this paper, to explore the differences between AI-generated and human-written patent texts, we generate a set of patent abstract texts by ChatGPT, in Chinese and English, from granted patent claims. Each generated patent abstract text corresponds to its original patent abstract. We analyze the linguistic characteristics of two types of patent texts by various comparison experiments. We anticipate that our work can assist people in identifying the patents generated by AI from the ocean of patents.

Keywords: Artificial Intelligence Generation · Large Language Model · Text Generation Detection · Patent Texts

1 Introduction

Recently, large language models (LLMs) have revolutionized the field of natural language processing research. Artificial intelligence (AI)-generated content technology continues to advance, starting with the progress made with the GPT-X models developed by OpenAI. LLMs can generate texts or pictures based on prompt words or sentences provided by the users. However, the potential misuse of LLMs raises human concerns. For example, they are used to generate fake news headlines and content, which are used to manipulate public opinion or mislead the public to undermine social stability or promote specific political

agendas. In addition, what has attracted particular attention is the abuse of LLMs to generate patent content to deceive patent examination offices to obtain patent grants. This will lead to a consequence: patents may be generated by one sentence without deliberation and the meaningless and low-quality patents may become ubiquitous. Furthermore, the improper use of AI technology will result in intellectual property infringement. In many intellectual property offices, patent applications generated by AI are either required to be disclosed or are prohibited entirely. Thus, such a detector assists examiners in identifying them and streamlines the examination process. Therefore, exploring the gap between AI-generated and human-written patent texts is a necessary research task.

To the best of our knowledge, there is currently no existing research focused on AI-generated patent text detection, nor are there any datasets tailored for patents publicly available. Different from general text content, patent content exhibits unique professional and normative characteristics, which may make patent text detection distinct from generic detection. In this paper, we construct a dataset for patent text generation and detection for the first time. This dataset includes patents from four domains: artificial intelligence, biomedicine, electrical engineering, and machinery manufacturing. Then we analyze the linguistic differences between patent texts generated by LLMs and written by humans. Specifically, we attempt to explore the following questions by experiments:

1. What are the differences in vocabulary features between AI-generated and human-written patent texts?
2. Do AI-generated texts have different features on the part of speech?
3. What differences in dependency distributions exist between AI-generated and human-written patent texts?
4. Do AI-generated and human-written patent texts differ in sentiment polarity?
5. What are the disparities in language model perplexity performance between AI-generated and human-written patent texts?

2 Related Work

2.1 Text Generation Detection Datasets

Recently, researchers have proposed various datasets to study identifying AI-generated texts.

– **CHEAT** [1]: The CHEAT dataset is the most comprehensive publicly available resource for detecting academic content generated by ChatGPT. It consists of 15,000 human-authored abstracts and 35,000 ChatGPT-generated abstracts. The human-authored abstracts are sourced from the IEEE Xplore database and span a broad range of topics in computer science. The ChatGPT-generated abstracts are organized into three categories: i) Generating a 200-word abstract by inputting a paper title and some keywords; ii) Polishing a human-written abstract using ChatGPT through specific prompt templates; iii) Mixing refined abstracts with human-written ones by using randomly constructed masks to determine the sentences to be replaced.

– **HC3** [2] is one of the most known datasets, which includes both human-written and ChatGPT-generated texts. It gathers the responses from humans and ChatGPT to the same questions. The dataset was constructed using a prompt template to input questions into ChatGPT, adjusting the temperature parameters to ensure the generated content aligns with the intended answer, which comprises Chinese and English branches. Specifically, the English branch, HC3-en, includes 58K human answers and 26K ChatGPT answers across 24K questions, primarily sourced from the ELI5 dataset, WikiQA dataset, etc. The Chinese branch, HC3-zh, includes 22K human answers and 17K ChatGPT answers covering more domains, such as medicine, finance, psychology, law, etc. The human answers in HC3-zh are sourced from WebTextQA, BaikeQA, etc.

There are other related datasets including Turing Bench [3], GROVER [4], TweepFake [5], MGTBench [6], ArguGPT [7], DeepfakeTextDetect [8], M4 [9], Scientific-articles Benchmark [10]. Unfortunately, the above datasets do not involve patent texts. However, the majority of existing detection datasets are limited to a specific domain, such as academic papers, financial news, Wikipedia, or certain question-and-answer platforms. Different from the above content, patent content contains more detailed and professional technical descriptions, often uses repetitive terminology, adheres to a specific format, and generally features complex sentence structures. These features may make patent content distinct from general text content. This paper proposes a dataset about patent abstracts to explore the gap between AI-generated and human-written patent texts.

2.2 Text Generation Detection Methods

Existing methods for detecting AI-generated texts can be broadly divided into two categories: metric-based and model-based methods.

Generally, metric-based methods utilize a pre-trained model to extract the distinguishing features of the input texts. For instance, GLTR [11] is a classical tool for detecting text generation based on probability ranking. DetectGPT [12] detects AI-generated texts by examining the curvature of the probability function for a given text. It originates from their finding that AI-generated texts often fall within the negative curvature region of the log probability function, while human-written texts hardly do. Metric-based detectors also includes Log-Likelihood [13], Log-Rank [12], Entropy [11], LRR and NPR [14], etc.

Model-based approaches are typically trained on a corpus containing both human-written and AI-generated texts, enabling the classification model to distinguish the texts generated by AI. For example, Guo et al. [2] proposed a RoBERTa-based detector to distinguish human-written texts from ChatGPT-generated texts. The detector was fine-tuned using the HC3 dataset. The authors provide two ways to train the detector. The first one only leverages the pure answer texts, and the second one leverages the question-answer text pairs to train the model jointly. The experimental results indicate that training with question-and-answer pairs leads to superior performance, potentially due to the richer semantic information encapsulated within these pairs. Additionally, Amrita et

al. [15] explored the effectiveness of ChatGPT as a detector, trained on publicly available datasets TuringBench [3], NeuralNews [16], IMDb[4], and TweepFake [5]. Experimental results indicate that ChatGPT frequently misclassifies AI-generated texts as human-written texts. They indicated that AI-generated texts typically possess greater fluency and consistency, which makes them superficially resemble human-written texts. Other model-based methods also include ConDA [17]. However, these methods may suffer from the issue of generalization.

3 A Novel Dataset for Patent Texts

To explore the difference between AI-generated and human-written patent texts, we constructed a novel dataset called Patent Abstract for Detection (PAD), which contains Chinese and English patent texts covering four domains: artificial intelligence, biomedicine, machinery manufacturing, and electrical engineering.

Our original patent data comes from Google Patents[5]. We chose the granted patents from 2019 to 2021 when LLMs, in particular ChatGPT, are not used universally. We put each patent claim to an LLM, ChatGPT3.5-Turbo, with appropriate temperature parameters to generate an abstract. We then store the abstract text produced by the model alongside its corresponding human-authored patent abstract as paired entries in the dataset. Table 1 shows the quantitative features of the dataset.

Table 1: Details of the PAD dataset

Domain	***English**	***Chinese**
arti	9118	17827
bio	7613	12508
elec	10581	18694
mech	19746	18873
ALL	47058	67902

4 Experiments

Next, we give the experimental analysis performed on our dataset to explore the differences between patent abstracts generated by ChatGPT and by humans.

4.1 Vocabulary Feature Analysis

Next, we analyze the quantitative characteristics of the collected corpus, including the average length (AvgLen), the total number of words (T-words), the number of unique words (U-words), and word density. Word density is defined as the ratio of the unique word number to the total number of words. As shown in Table 2, the average and total scales of patent abstracts generated by ChatGPT are greater than those of human-written abstracts. However, the number of unique words in human-written patent abstracts is slightly lower than that of

[4] https://huggingface.co/datasets/imdb (Accessible on 10 Sep 2024)
[5] https://patents.google.com (Accessible on 10 Sep 2024)

ChatGPT-generated abstracts. Consequently, the word density of human-written patent texts is higher, indicating that humans tend to employ a more diverse vocabulary in their expressions.

Table 2: Comparison of quantitative characteristics in PAD

	*English				*Chinese			
	AvgLen	**U-words**	**T-words**	**Density**	**Avg.len**	**U-words**	**T-words**	**Density**
human	259.1	104421	12197275	**0.0086**	221.3	98298	8531004	**0.0115**
ChatGPT	574.0	112300	270133111	0.0042	335.4	113475	12979049	0.0087

4.2 Part-of-speech Analysis

We utilize the open-source Python library SpaCy[6] to analyze patent texts generated by ChatGPT and written by humans. The occurrence characteristics of different part-of-speech (POS) tags are compared, as presented in Fig.1.

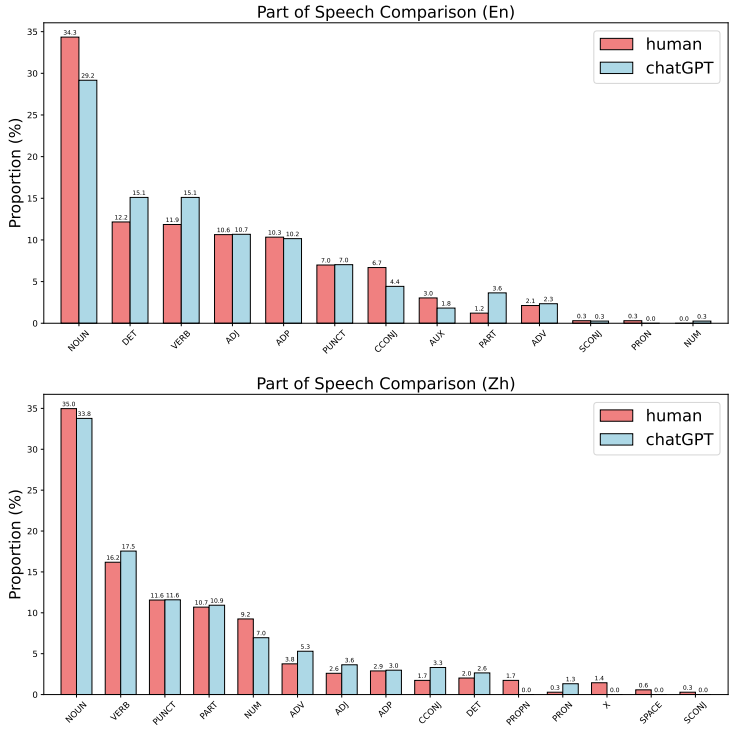


Fig. 1: Comparison of part-of-speech distribution between ChatGPT-generated and human-written patent abstracts (above: English; below: Chinese). The results are sorted descendingly by the ratios of human-written patent abstracts. "X" represents the part of speech not to be specifically classified.

[6] We used the `en_core_web_sm` model for English and the `zh_core_web_sm` model for Chinese.

Both the Chinese and English patent abstracts have high proportions of nouns (NOUN) and verbs (VERB). The dominance of nouns indicates that proper nouns are used to describe technologies. The high proportion of verbs demonstrates the dynamic characteristics of inventions in patents. In the English texts, humans use nouns 17.5% higher than ChatGPT, while in Chinese use 3.6% higher. However, in the English texts, the verb usage proportion of humans is 26.9% lower than ChatGPT, and in Chinese, it is 8% lower than ChatGPT. This is because ChatGPT often describes technical processes in detail, utilizing more verbs to maintain its clarity.

4.3 Dependency Relation Analysis

We utilize dependency relation analysis to reveal the relationships between words within patent text sentences. The experiments were also conducted using the open-source library SpaCy. Fig. 2 shows the results of dependency relations in the English and Chinese patent texts.

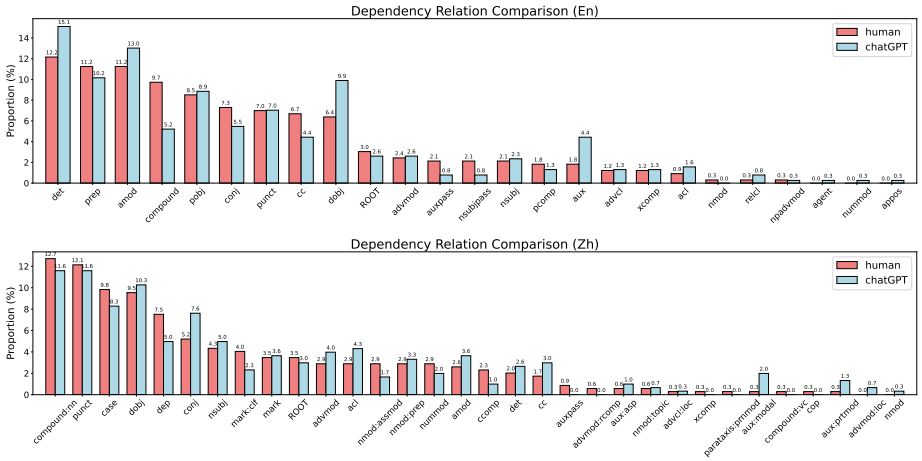


Fig. 2: Comparison of dependency relation ratios between patent abstract texts generated by humans and ChatGPT (above: English; below: Chinese). The results are sorted descendingly by the ratios of human-written texts.

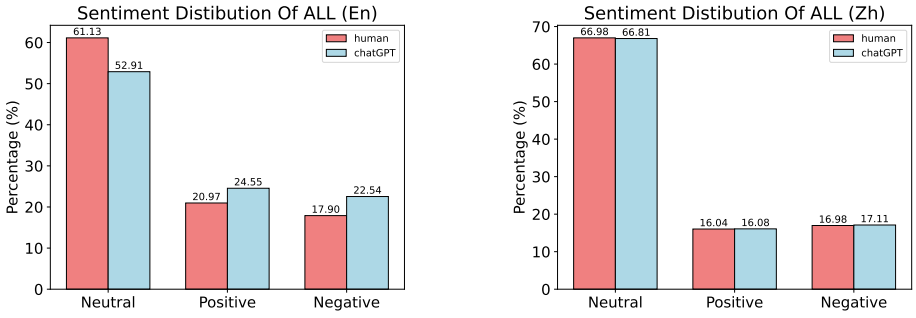
In the English texts, both human-written and ChatGPT-generated texts have high percentages of determiner (det), prepositional modifier (prep), and adjectival modifier (amod) relationships. Each relationship accounts for more than 10%. This is because patent texts must describe technical details and features, which inherently require using more adjectives to modify nouns, using prepositions to modify noun phrases and provide supplementary explanations, and using qualifiers to define technical terms and contexts within the patent text precisely. In particular, humans use 86.5% more compound relationship structures, which indicates that human experts focus more on expression diversity. In contrast, for the direct object (dobj) and auxiliary (aux) relationships, the usage of ChatGPT is 54.7% and 144.4% higher than that of human-written texts. This is because ChatGPT often uses more straightforward sentence structures for clarity.

In the Chinese texts, both abstracts have high percentages of compound noun (compound:nn) and punctuation (punct) relationships, each of which accounts

for more than 10%. Notably, the proportion of the compound noun relationship used by humans is 9.5% higher than ChatGPT. Additionally, humans use case marking (case) relationships 18.1% more frequently and dependent (dep) relationships 50% more frequently than ChatGPT. This indicates that human-written patent texts exhibit more complex sentence structures. However, ChatGPT uses the conjunct (conj) relationship 46.2% higher, and the adjectival clause (acl) relationship 48.3% higher than humans. This is because ChatGPT prefers concise and smooth sentence structures, leading to more conjunctions for parallel sentences. This contrasts with human writing, which frequently employs compound sentences, which results in a stronger inter-sentential correlation and makes the texts more complicated.

4.4 Sentiment Analysis

Generally, human emotions are inherently reflected in natural language. Since ChatGPT-generated texts are trained on human-written data, we aim to explore the emotional differences between human-written and model-generated texts. For sentiment analysis of patent texts, we employed a fine-tuned multilingual model, XLM-RoBERTa-base[7] in the experiments. The comparison results of the sentiment distributions are shown in Fig. 3.



(a) English Sentiment Analysis (b) Chinese Sentiment Analysis

Fig. 3: Proportions of neutral, positive, and negative emotional words

In Fig. 3, we find that the proportion of neutral emotions is the largest for both humans and ChatGPT, which account for more than 50% in both the Chinese and English texts. The results are in line with the objectivity of patent texts. More specifically, patent texts are descriptive which aim to accurately and objectively detail technological inventions. Also, these texts adhere to strict formatting and linguistic standards, which demand neutrality and professionalism in language. Both human-written and ChatGPT-generated patent texts conform to these conventions.

Notably, in the English texts, the proportion of neutral emotions in human-written content is 15.5% higher. However, the proportion of positive emotions is 17.1% lower, and that of negative emotions is 25.9% lower. The reason may

[7] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment (Accessible on 10 September 2024)

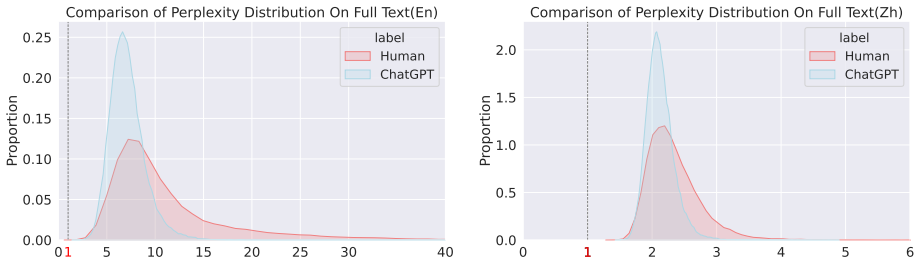
be that humans strictly follow the patent writing specifications, while ChatGPT prefers the overall fluency of expression when generating patent texts.

4.5 Language Model Perplexity Analysis

The perplexity (PPL) is often used as a metric to evaluate the performance of a language model. A lower PPL indicates that the language model is more confident in its predictions and is considered a better model. The training of the language model is carried out on a large-scale text corpus, which can be considered that it has learned some common language patterns and text structures. Therefore, PPL can be used to measure how well the text conforms to common characteristics. We use the open source GPT-2 small[8] (Wenzhong-GPT 2-110M[9] in Chinese) model to calculate the PPLs of the constructed PAD dataset. The perplexity of a sentence is defined as the exponent of the negative average log-likelihood of its words under a language model. Formally,

$$\text{Perplexity}(S) = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i)} \tag{1}$$

where N is the total number of words in the sequence S, $P(w_i)$ is the probability of the i-th word output by a given language model.



(a) PPL Distributions of the English Texts (b) PPL Distributions of the China Texts

Fig. 4: Comparison of perplexity in the English and Chinese texts. The horizontal axis represents the perplexity value, while the vertical axis represents the proportion of different perplexity values.

Fig. 4 illustrates the results of the PPL distributions of human-written and ChatGPT-generated texts. It can be observed that the ChatGPT-generated texts have a relatively lower PPL compared to human-written texts. Also, the PPL values of ChatGPT-generated texts are more concentrated in the lower regions. This is because ChatGPT is good at capturing common patterns and structures in the texts, which leads to ChatGPT generating patent abstracts based on several patterns learned from the training corpora. Against, the writing of humans is uncertain in determining what words and what sentences to follow, which makes the human-written texts individual. Therefore, the human-written texts have higher PPL values and exhibit a long-tailed distribution.

[8] https://huggingface.co/gpt2 (Accessible on 10 September 2024)

[9] https://huggingface.co/IDEA-CCNL/Wenzhong-GPT2-110M (Accessible on 10 September 2024)

5 Discussion

In this paper, we conducted an experimental exploration for several research questions, aiming to understand the differences between patent abstracts generated by an LLM and by humans. From the vocabulary perspective, human-written patent abstracts typically have a larger vocabulary than ChatGPT-generated patent abstracts. This reflects that humans are more creative and utilize a greater variety of synonyms and diverse expressions when writing patent abstracts. Furthermore, humans possess more expert knowledge of the patent domain, while ChatGPT depends on its excellent summarization ability obtained from its training on the vast of corpora.

From the perspectives of parts of speech and dependent relations, humans use more nouns (NOUN), fewer verbs (VERB), fewer determiners (DET), and have more compound relationships (compound, compound:nn), fewer direct object relationships (dobj), and fewer determiner (det) relationships. This is because humans use more complex noun phrases and compound relationships to increase information density. Additionally, the writing specifications of patent texts require humans not to use demonstrative pronouns. On the contrary, ChatGPT focuses more on the fluency and readability of the generated texts and uses more demonstrative pronouns to make the texts shorter. Furthermore, ChatGPT-generated texts contain more verbs and determiners to describe technology details.

Regarding emotional polarity, for English patent texts, humans show more neutral, less positive, and fewer negative emotions, due to the objectivity requirements of patent writing. Whereas, ChatGPT uses different generation strategies, instead of pursuing a neutrality generation strategy.

Finally, from the perspective of perplexity, human-written abstracts have relatively higher perplexity values with a decentralized distribution, which indicates the creativity and individuality of human beings. On the other hand, ChatGPT may generate patent abstracts based on a latent common framework and avoid highly confusing statements to guarantee the smoothness and readability of the generated texts. It is easier for another language model to predict the words given sentence prefixes.

6 Conclusion

In this paper, we explore the differences between LLM-generated and human-written patent abstracts from five perspectives in natural language processing. The experimental results show their quantitative distinctions in these metrics. Through experimental analysis, it was found that the AI-generated and human-written patent abstracts differ in wording preferences, sentence structures, and perplexities, and have a similar distribution of emotional polarity.

Furthermore, the purpose of this exploration is to fill the research gap of lacking an AI generation detection dataset for patent texts. On the other hand, we hope our attempt contributes to building a better AI-generated patent text detector by further taking the found differences into account. Second, we only evaluate the texts generated by ChatGPT as it is a representative of LLMs. It would be interesting to explore the patent texts generated by other LLMs.

References

1. Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. Cheat: A large-scale dataset for detecting ChatGPT-written abstracts. *arXiv preprint arXiv:2304.12008*, 2023.
2. Biyang Guo, Xin Zhang, Ziyuan Wang, et al. Jiang, and Nie. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.
3. Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. TURINGBENCH: A benchmark environment for turing test in the age of neural text generation. In *EMNLP*, pages 2001–2016, 2021.
4. Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. pages 9051–9062, 2019.
5. Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. TweepFake: About detecting deepfake tweets. *PLOS ONE*, 16(5):1–16, 2021.
6. Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. Mgtbench: Benchmarking machine-generated text detection. *arXiv preprint arXiv:2303.14822*, 2023.
7. Yikang Liu, Ziyin Zhang, Wanyang Zhang, et al. Yue, and Zhao. ArguGPT: evaluating, understanding and identifying argumentative essays generated by gpt models. *arXiv preprint arXiv:2304.07666*, 2023.
8. Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. Mage: Machine-generated text detection in the wild. In *ACL*, pages 36–53, 2024.
9. Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, et al. Shelmanov, and Tsvigun. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *EACL*, pages 1369–1407, 2024.
10. Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci, and Georg Groh. Distinguishing fact from fiction: A benchmark dataset for identifying machine-generated scientific papers in the llm era. In *TrustNLP 2023*, pages 190–207, 2023.
11. Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. GLTR: Statistical detection and visualization of generated text. In *ACL*, pages 111–116, 2019.
12. Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *ICML*, pages 24950–24962. PMLR, 2023.
13. 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.
14. Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *EMNLP*, pages 12395–12412, 2023.
15. Amrita Bhattacharjee and Huan Liu. Fighting fire with fire: can chatgpt detect ai-generated text? *ACM SIGKDD Explorations Newsletter*, 25(2):14–21, 2024.
16. Reuben Tan, Bryan Plummer, and Kate Saenko. Detecting cross-modal inconsistency to defend against neural fake news. In *EMNLP*, pages 2081–2106, 2020.
17. Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. Conda: Contrastive domain adaptation for ai-generated text detection. In *in IJCNLP*, pages 598–610, 2023.