CHEAT: A Large-scale Dataset for Detecting ChatGPT-writtEn AbsTracts

Peipeng YuJinan University
ypp865@163.com

Jiahan Chen Jinan University 335486990@qq.com Xuan Feng Jinan University fenffef@163.com Zhihua Xia Jinan University xia_zhihua@163.com

Abstract

The powerful ability of ChatGPT has caused widespread concern in the academic community. Malicious users could synthesize dummy academic content through ChatGPT, which is extremely harmful to academic rigor and originality. The need to develop ChatGPT-written content detection algorithms call for large-scale datasets. In this paper, we initially investigate the possible negative impact of Chat-GPT on academia, and present a large-scale CHatGPT-writtEn AbsTract dataset (CHEAT) to support the development of detection algorithms. In particular, the ChatGPT-written abstract dataset contains 35,304 synthetic abstracts, with Generation, Polish, and Mix as prominent representatives. Based on these data, we perform a thorough analysis of the existing text synthesis detection algorithms. We show that ChatGPT-written abstracts are detectable, while the detection difficulty increases with human involvement. Our dataset is available in https://github.com/botianzhe/CHEAT.

1 Introduction

ChatGPT, a natural language processing tool based on artificial intelligence technology, has attracted widespread attention in recent times. Based on user needs, ChatGPT could complete tasks with high quality, such as coding, translation, thesis writing(Macdonald et al., 2023), and so on. While ChatGPT brings convenience to human life, the potential harm of its synthetic content has gradually emerged(Ufuk, 2023). As an important part of the academic thesis, the abstract has been proven to be brilliantly synthesized by ChatGPT(Gao et al., 2022; Mok, 2023). Malicious researchers could synthesize plausible academic content without practical research, which seriously undermines academic originality.(Thorp, 2023). Meanwhile, although the synthetic content exhibit high levels of authenticity, the academic rigor and the conclusion correctness cannot be guaranteed(Else, 2023). How

Human-written abstract in satellite-bome Terahertz wave ground detection situation, the quantitative estimation of Terahertz wave atmospheria absorption attenuation (oss, especialty with designated satellite position and down-looking angle information, has always been fundamental and a key technology application for various Terahertz communication modes, such as wide band & high-speed network, interestellar communication, satellite-ground station link, statosphere geno-cariff communication, long distance vast data transfer, short range wireless security communication etc. This paper designs an atmospheric absorption loss estimation software on satellite global THz wave ground detection. The realized functions of this software including scene establishment, Besic functions and calculation methods were explained in detail. Finally, the monthly climage calculation restution of stations and calculation methods were explained in detail. Finally, the monthly climage calculation restution of stations are stationary and calculation methods were explained in detail. Finally, the monthly climage calculation restution of stations are stationary and the station and the station and the stationary and the station and the stationary and the stationary and the stationary and the station and the stationary and the stati

ChatGPT-generated abstract

Inis paper discusses the estimation of the atmospheric absorption loss for satellite ground detection at the terahertzwave band. The impact of atmospheric iwaves on the attenuation and absorption of ferahertz-wave signals is, analyzed, and an atmospheric imodeling approach is proposed. By considering the influence of water vapor and other atmospheric components, the proposed method can effectively estimate the atmospheric absorption loss in different weather conditions. To implement the approach, a software tool is developed, which provides an accurate estimation of the atmospheric absorption loss and supports the optimization of satellite ground detection detection systems. The effectiveness of the proposed method is verified through simulations and experiments, which show that it can improve the performance of terahertz-wave band satellite ground detection in various scenarios. Overall, this work provides insights into the atmospheric effects on terahertz-wave signals and presents a practical solution for atmospheric absorption loss estimation that can benefit satellite ground detection applications.

ChatGPT-polished abstract

in satellite-borne Terahertz wave ground detection, accurately estimating the atmospheric absorption attenuation loss Terlahertz waves is crucial for yarious Terahertz communication modes, including wideband and high-speed networks, interstellar communication, satellite-ground station links, stratosphere aefocraft communication, long distance data transfer, and short-range wireless communication. To address this, this paper introduces an atmospheric absorption loss estimation software for satellite global THZ wave ground detection. The software offers key features such as scene establishment, fundamental functions, and calculation methods, all of which are explanced in detail. Lastly, the monthly change calculation results of satellite ground detection in the 0.34 THz working band and 10°-90° downooking angle are presented.

Figure 1: The different distributions of human-written and ChatGPT-written abstracts. The visualization results are obtained by GLTR (Gehrmann et al., 2019).

to detect ChatGPT-written abstracts and ensure the academic originality have been pressing issues that need to be addressed.

As concerns over ChatGPT escalate, there has been a surge of interest in detecting ChatGPT content. The availability of large-scale datasets is an enabling factor for the development of ChatGPT content detection methods. So far, researchers have constructed multiple detection datasets for question answering and querying scenarios(Guo et al., 2023; Mitrović et al., 2023). Some studies also evaluated the quality of ChatGPT-written text with a small amount of data(Huang et al., 2023; Cao et al., 2023; Gao et al., 2022). However, due to the limitations of OpenAI, there is currently no large-scale dataset to support the development of ChatGPT-written abstract detection algorithms. \(^1\).

In this paper, we focus on the practical need to generate large-scale and diverse ChatGPT-written abstracts. We explore different abstract synthesis methods using ChatGPT, including *Generation*, *Polish*, and *Mix*. As shown in Figure 1, ChatGPT-

https://www.nature.com/articles/d41586-023-00553-9

written abstracts are almost indistinguishable from the human-written ones, but there are some differences in semantic distribution. We aim to construct a large-scale dataset to analyze the abnormal distribution patterns in ChatGPT-written abstracts. The contributions are summarized as follows:

- (1) We present a large-scale ChatGPT-written abstract dataset, CHEAT, which is currently the largest dataset available for ChatGPT-written abstract detection.
- (2) We analyze the distribution differences between human-written and ChatGPT-written abstracts. Compared with human-written abstracts, ChatGPT-written abstracts are more informative, but lack of logic and definition precision.
- (3) We conduct an evaluation of current detection algorithms on the CHEAT dataset. The experimental results show that existing schemes lack effectiveness in detecting ChatGPT-written abstracts, and the detection difficulty increases with human involvement.

2 Related Work

In this section, we will introduce the related works of Text synthesis and Text synthesis detection.

2.1 Text Synthesis

ChatGPT, an artificial intelligence chatbot program, is developed based on the Generative Pretrained Transformer (GPT) models. The earliest GPT-1(Radford et al., 2018) was proposed in 2018, which could perform various natural language processing tasks such as semantic inference, question answering, and classification. Although GPT-1 could effectively handle some unknown tasks, its generalization ability was still not sufficient compared to fine-tuned supervised models. After that, Radford et al., Radford et al., 2019) proposed GPT-2 in 2019, which used more network parameters and larger datasets to learn more universal knowledge and achieved the best performance in multiple specific language modeling tasks. Later, Brown et al. (Brown et al., 2020) proposed GPT-3. They employed self-supervised mechanisms to learn general knowledge and performed well in various tasks. However, due to directly learning from large-scale text corpora, GPT-3 was prone to generating erroneous or offensive content, which limited

et al. (Ouyang et al., 2022) designed the Instruct-GPT language model to improve the quality of synthetic content. They combined supervised learning and Reinforcement Learning from Human Feedback(RLHF) to promote the synthetic content better following user intent. The recent ChatGPT furtherly learned to rank the quality of output results, and improved the model's understanding ability successfully. The Instruct Learning and Reinforcement Learning strategies had driven the success of ChatGPT, enabling it to synthesize academic contents that can deceive most detectors.

2.2 Text Synthesis Detection

Existing text synthesis detection algorithms could be roughly divided into traditional detection algorithms and deep learning-based detection algorithms. Traditional detection algorithms analyze abnormal patterns of synthetic text by extracting hand-crafted features, while deep learning-based algorithms train detection models on large-scale datasets to achieve accurate detection of synthetic text. The following provides a detailed introduction to the two types of approaches.

Traditional text synthesis detection Early text synthesis algorithms left obvious synthesis traces and abnormal distribution patterns in statistical features. Sebastian et al., (Gehrmann et al., 2019) developed a visualization tool called GLTR. They integrated multiple detection methods based on statistical features(word order, predicted distribution entropy, and so on.) to identify anomalies in synthetic text. Fröhling et al. (Fröhling and Zubiaga, 2021) found that most synthetic texts were repetitive and lacked purpose, so they extracted statistical features of text style and successfully identified various synthetic text using Random Forests and Support Vector Machines. Lundberg et al. (Levin et al., 2023) generated 50 abstracts using ChatGPT and then analyzed the differences between humanwritten abstracts and ChatGPT-written abstracts using Grammarly. They found that ChatGPT was able to generate more unique words and had fewer grammatical errors. Recently, Guo et al. (Guo et al., 2023) discovered that ChatGPT-written text tends to use more connecting words, providing new ideas for detecting synthetic text.

Deep learning-based text synthesis detectionThe rich information stored in large language models could provide guidance for text synthesis de-

Table 1: The searching keywords used for collecting human-written abstracts.

Natural language processing	Feature extraction	Artificial Intelligence	Knowledge Representation and Reasoning	Internet of Things	
Computational modeling	Labeling	Machine Learning	Expert Systems	Cloud Computing	
Training	Neural networks	Deep Learning	Fuzzy Logic	Cybersecurity	
Supervised learning	Nonlinear systems	Computer Vision	Genetic Algorithms	Data Mining	
Brightness	Convergence	Robotics	Swarm Intelligence	Predictive Analytics	
Estimation	networks	Reinforcement Learning	Big Data Analytics	Decision Support Systems	

tection. Solaiman et al., (Solaiman et al., 2019) fine-tuned the detection model based on RoBERTa and achieved the best performance in the generatedweb detection task. Tay et al. (Tay et al., 2020) found that different language models would leave different defects in the synthesized text. Thus, they trained corresponding detectors for specific language models to achieve high-accuracy detection, but performed poorly in cross-model detection. After that, Ippolito et al., (Ippolito et al., 2020) constructed a novel text synthesis dataset and fine-tuned the BERT classification model, significantly improving the detection accuracy on synthetic texts. Although deep learning algorithms could detect synthetic text well, their interpretability greatly limits their application. The research focus of current deep learning based solutions is to explain the underlying principles behind the detection results.

3 The ChatGPT-written Abstract Dataset

Although some datasets for ChatGPT-written content detection are available, there is no large-scale ChatGPT-written abstract datasets until now. To provide more relevant data for developing detection methods, we constructed the ChatGPT-written abstract (CHEAT) dataset.

3.1 Basic information

The CHEAT dataset is consisted of 15,395 human-written abstracts and 35,304 ChatGPT-written abstracts. The average length of all abstracts is 163.9 and the total vocabulary size is 130,272. The human-written abstracts are searched from IEEE Xplore, a mega repository of scholarly literature. As shown in Table 1, we select 30 keywords to search for matching article abstracts. The collected abstracts all originate from the field of computer science, and cover areas such as natural language processing, computer vision, and machine learning. Among them, abstracts with less than 100 words accounted for 11.6%, abstracts with 100-200 words accounted for 67.7%, and abstracts with more than 200 words accounted for 20.7%. ChatGPT-written

abstracts are then synthesized through the interface provided by OpenAI.

3.2 Abstract Synthesis Methods

The synthetic abstracts in the CHEAT dataset are consisted of Generation abstracts, Polish abstracts, and Mix abstracts. We input the human-written abstracts into ChatGPT (gpt-3.5-turbo) through the OpenAI interface to obtain the corresponding output. Specifically, we obtain synthesized abstracts in three ways:

Generation: ChatGPT is capable of generating plausible abstracts from keywords. Malicious users could use ChatGPT to generate abstracts directly for publication, thus obtaining scientific results without any cost. To detect such synthesis, we create the ChatGPT-Generation dataset to develop detection algorithms. Specifically, we use ChatGPT to output the generated abstract by entering the following command: "Generate a 200-word abstract of the paper in English based on the title and keywords; your answer only needs to include the generated paragraph.", followed by the title and keywords in the human-written data.

Polish: Unlike traditional text polishing algorithms, ChatGPT is able to utilize its rich linguistic knowledge to optimize the original text and enhance its readability. Malicious users are likely to evade paper checking by text polish, causing damage to academic originality(Khalil and Er, 2023). In this paper, we create the ChatGPT-Polish dataset to support the relevant detection algorithm. Specifically, we employ ChatGPT to output a polished abstract by entering the following command: "Polish the following paragraphs in English, your answer just needs to include the polished text.", followed by the human-written abstract.

Mix: Malicious users are likely to mix humanwritten abstracts with polished abstracts to evade detection algorithms. To address this problem, we create a more challenging dataset, ChatGPT-Mix, based on the polished abstracts. Specifically, we first decompose the polished abstracts and humanwritten abstracts according to their semantics, then construct a random mask to determine which sentences need to be replaced, and finally replace the polished abstracts with text from the human-written abstracts to obtain the final mixed abstract. By controlling the number of 1 in the mask, we are able to effectively control the text replacement rate and synthesize mixed abstracts with different detection difficulties.

4 Evaluation and Analysis

ChatGPT's ability to synthesize abstracts is potentially harmful to academic originality and correctness. In this section, we first analyze the linguistic differences between human-written and ChatGPT-written abstracts, and then evaluate the detection performance of existing algorithms on the CHEAT dataset. After that, we explore the judgment basis of the deep learning based detection methods.

4.1 Linguistic Analysis

In this subsection, we evaluate the synthesis quality of ChatGPT-written abstracts from a linguistic perspective, including lexical analysis and dependency analysis.

Lexical analysis In the field of natural language processing, each word could be classified as one of lexical categories. The part-of-speech (POS) tagging task aims to determine the grammatical class of each word in a given sentence. In this part, we use the POS module in NLTK (Bird, 2006) to calculate the lexical distributions of abstract texts in the CHEAT dataset, and sort them by lexical percentage. As shown in Figure 2, we show the statistics of the top ten lexicalities. It can be seen that noun (NOUN) occupies the largest proportion of all lexicalities, while punctuation (PUNCT), verb (VERB), adjointion (ADP), adjective (ADJ), and determiner (DET) occupy most of the others. Comparing human-written and ChatGPT-written abstracts, the following findings could be made:

- (1) The proportions of *NOUN*, *VERB* in ChatGPT-written abstracts are higher than those of human-written abstracts. It can be argued that the rich knowledge contained in ChatGPT can provide a more diverse vocabulary for abstract synthesis and make it more informative.
- (2) The proportions of adposition (ADP), proper noun (PROPN), and auxiliary (AUX) in human-written abstracts are larger than those of ChatGPT-written abstracts. This indicates that humans tend to pay more attention to abstract structure, consis-

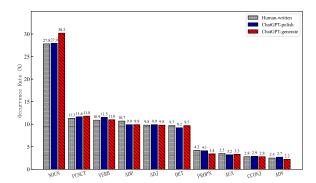


Figure 2: The lexical distribution for human-written, ChatGPT-polished, and ChatGPT-generated abstracts.

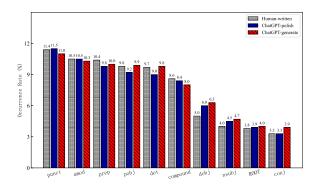


Figure 3: The dependency distribution forhuman-written, ChatGPT-polished, and ChatGPT-generated abstracts.

tency, and logic, while ChatGPT is weaker in these aspects.

Dependency analysis Dependency grammar is an important tool for natural language understanding. It is able to point out the syntactic collocation relations between words. In this paper, we calculate the dependency properties between individual words and attempt to analyze the differences between human-written and ChatGPT-written abstracts. We compute statistical histograms of dependencies in each type of abstracts, and then rank them according to the percentage. As shown in Figure 3, we present the statistics for the top ten dependencies. There is some similarity in the distribution of the human-written abstracts and ChatGPT-written abstracts, but there are also some differences. Specifically, we have the following findings:

- (1) The Adjectival modifier (amod), Prepositional modifier (prep), and Compound modifier (compound) are more used in human-written abstracts. Compared with ChatGPT-written abstracts, human-written abstracts tend to apply more modifiers to define the words precisely.
 - (2) The proportions of the Direct Object (dobj),

Table 2: The detection performance of existing schemes.	. The accuracy and AUC(Area Under the ROC Curve) are
used as metrics.	

	Datasets						
Methods		Generation		Polish		Mix	
		AUC	ACC	AUC	ACC	AUC	
Grover (Zellers et al., 2019)	54.24	56.34	53.33	55.45	50.89	51.71	
Zerogpt(ZeroGPT, 2023)	67.32	78.8	52.71	57.35	50.61	52.59	
OpenAI-detector (Solaiman et al., 2019)	75.97	84.41	54.07	56.17	52.18	55.23	
ChatGPT-detector-roberta (Guo et al., 2023)	75.54	81.91	53.65	47.28	51.92	63.71	
Chatgpt-qa-detector-roberta (Guo et al., 2023)	85.56	97.6	53.53	64.39	51.67	65.28	

Nominal subject (nsubj), and Root words(ROOT) in the human-written abstracts are smaller than those of ChatGPT-written abstracts. This is similar with the distributions in the lexical analysis. ChatGPT could provide more the rich vocabulary for synthesizing abstracts.

4.2 Text Synthesis Detection Evaluation

The distribution differences between ChatGPT-written abstracts and human-written abstracts provide the feasibility of their detection. This subsection evaluate the detection performance of existing algorithms on ChatGPT-written abstracts.

4.2.1 Compared Text Synthesis Detection Methods

We consider five text synthesis detection methods in our experiment. Due to the need to evaluate each method on the CHEAT dataset, we only capture those schemes that have code or the model parameters publicly available.

- **Grover Detector** Grover is a controlled text generation model proposed by Rowan Zellers et al.(Zellers et al., 2019). The Grover detector is trained based on the Grover generation model to determine whether the text is generated by the neural network model.
- **ZeroGPT** ZeroGPT detector(ZeroGPT, 2023) is an online detector, which is trained based on 10 million articles and texts. It is capable of performing machine-generated text detection in multiple languages with high accuracy.
- OpenAI-detector The detector(Solaiman et al., 2019) is officially provided by OpenAI. It employs the original text and the GPT-2 generated text to fine-tune the RoBERTa model to determine whether the text is machine generated or not.

- ChatGPT-detector-roberta Guo et al. (Guo et al., 2023) constructed the HC3 (Human ChatGPT Comparison Corpus) dataset consisting of nearly 40K questions and their corresponding human/ChatGPT answers, and then used these data to fine-tune the RoBERTa model to obtain an accurate synthetic text detector.
- ChatGPT-qa-detector-roberta The correlation between the question and answer is also used to detect the synthetic text. Guo *et al.*(Guo et al., 2023) trained a detection model using Q&A statements in HC3 dataset, and achieved high accuracy in synthetic text detection.

4.2.2 Detection Performance

In this part, we evaluate the performance of existing detection algorithms on our CHEAT dataset. The detection accuracy (ACC) and area under the ROC curve (AUC) are applied to evaluate the effectiveness of the detection algorithms. We first evaluate the detection performance when the algorithm is not trained with our CHEAT dataset. As shown in Table 2, due to the differences of training corpus, existing algorithms have difficulty in detecting ChatGPT-written abstracts. The ROC curve shown in Figure 4 also demonstrates the weakness of existing detection schemes. The Chatgpt-qadetector-roberta could obtain high detection accuracy on the ChatGPT-Generation dataset, but fails on the ChatGPT-Polish and ChatGPT-Mix dataset. Existing schemes are still inefficient in detecting ChatGPT-written abstracts.

After that, we trained multiple models using ChatGPT-Generation, ChatGPT-Polish, and ChatGPT-Mix, respectively. The detection performance of trained models is then evaluated on the corresponding test datasets. The AUC scores are

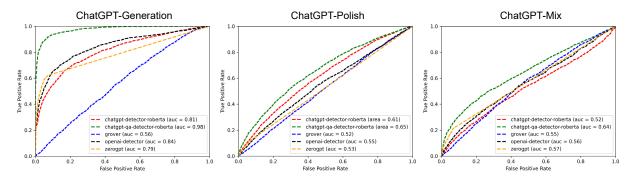


Figure 4: ROC curves of existing detection schemes on three datasets(ChatGPT-Generation, ChatGPT-Polish, and ChatGPT-Mix).

Table 3: The detection performance of different models. The models are trained on the CHEAT training dataset and then evaluated on the corresponding test dataset.

Models	Datasets				
ivioucis	Polish	Generation	Mix		
Distilbert (Sanh et al., 2019)	99.43	100	85.07		
BERT (Kenton et al., 2019)	99.48	100	86.62		
Roberta (Liu et al., 2019)	99.72	100	52.93		
BERT-multilingual (Pires et al., 2019)	99.49	100	60.16		
PubMedBERT (Gu et al., 2021)	99.56	100	87.83		

presented in Table 3. Due to the specialized training on the CHEAT dataset, these models obtain better detection performance on the ChatGPT-written abstracts. In particular, the PubMedBERT, associated with abstract content detection, obtains the best detection performance on our CHEAT dataset. From the results, we believe that the fully generated abstracts are highly feasible to detect. However, similar with the evaluation of existing schemes, the detection difficulty increases with the human involvement. These models could obtain auc scores of 100 on the ChatGPT-Generation dataset while get lower auc score on the ChatGPT-Polish dataset, which are influenced by human writing. Considering the most difficult case, these models obtain the lowest auc score when the human-written abstracts are mixed with ChatGPT-polished abstracts. Detecting ChatGPT-written content with human involvement is still a challenge for existing detection algorithms.

4.2.3 Explainability Analysis

The well-trained detection model could obtain high accuracy, but have difficulty interpreting the output. It is essential for us to understand whether detection models are perceiving particular word patterns to make judgments. In this subsection, we use SHAP (SHapley Additive exPlanations) (Lundberg and

Lee, 2017) to explain the judgments made by detection models. Specifically, SHAP uses the classical Shapley values from game theory to link optimal credit allocation with local explanations. It is able to assign feature importance values to each input word of the detection model. To obtain the model interpretion, we first train a high-accuracy detection model based on the PubMedBERT. Then, we compute the SHAP values between the detection results and the input abstracts. We count the judgment bases for each kind of abstracts, and present the top ten terms of judgment bases in Figure 5. When we attempt to discriminate human-written and ChatGPT-generated abstracts, it can be found that ChatGPT tend to use more special-patterns, such as 'zi', 'iv', 'tunv', and 'gran'. When the human involvement increases, precise words gradually play a more important role in detection tasks. We also visualized the SHAP values for two abstracts in Figure 6. The positive SHAP value indicates a positive impact on classifying the text as ChatGPT-written abstract, and vice versa. It can be seen that detectors could correlate certain words, such as "it holds great potential", "mapping of mobile robots", and "and security areas.", to detect abnormal distributions of ChatGPT-written abstracts. When human involvement increases, words biased towards human-written abstracts would appear, in-

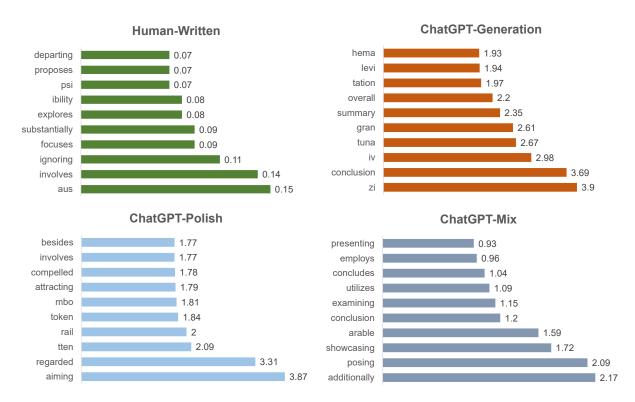


Figure 5: Visualization of SHAP value statistics. The top 10 words ranked by contribution are listed for human-written and ChatGPT-written abstracts.

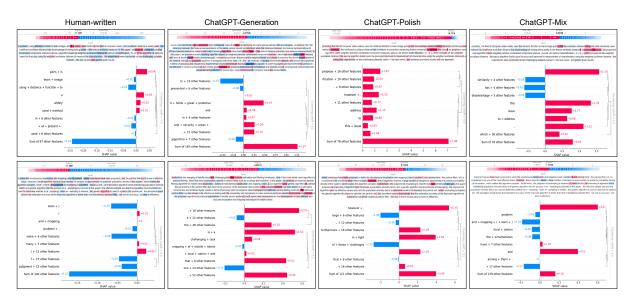


Figure 6: Visualization of SHAP value for single abstract. The sentence semantics and the top 10 words ranked by contribution are calculated for the selected abstract.

creasing the difficulty of text synthesis detection.

5 Discussion & Conclusion

Although the current ChatGPT demonstrates stunning synthesis ability, we demonstrate that the ChatGPT-written abstracts could be detected by a well-trained detector. To train the detection model using domain-specific knowledge, we construct a large-scale CHatGPT-writtEn AbsTract (CHEAT)

dataset, which exceeds the size of all existing publicly available synthetic abstract datasets.

In this paper, we evaluate the performance of existing detection algorithms on the CHEAT dataset. The experimental results in this paper show that the abstracts generated entirely by ChatGPT are detectable, while the synthetic abstracts with human guidance possess detection difficulty, especially when mixed with human-written text. We

need not be too alarmed by the abstracts fully generated by ChatGPT. Rather, we must direct our attention towards ChatGPT-written abstracts with human guidance. We aspire for this dataset to serve as a stepping stone for ChatGPT-written content detection research, particularly in the realm of paper abstracts.

References

- Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.
- Holly Else. 2023. Abstracts written by chatgpt fool scientists. *Nature*, 613(7944):423–423.
- Leon Fröhling and Arkaitz Zubiaga. 2021. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science*, 7:e443.
- Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. 2022. Comparing scientific abstracts generated by chatgpt to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*, pages 2022–12.
- Sebastian Gehrmann, SEAS Harvard, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *ACL* 2019, page 111.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv* preprint arXiv:2301.07597.
- Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and

- limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.
- Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822.
- Jacob Devlin Kenton, Chang Ming-Wei, and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Mohammad Khalil and Erkan Er. 2023. Will chatgpt get you caught? rethinking of plagiarism detection. *arXiv preprint arXiv:2302.04335*.
- Gabriel Levin, Raanan Meyer, Eva Kadoch, and Yoav Brezinov. 2023. Identifying chatgpt-written obgyn abstracts using a simple tool. *American Journal of Obstetrics & Gynecology MFM*, 5(6).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv* preprint arXiv:1907.11692.
- Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Calum Macdonald, Davies Adeloye, Aziz Sheikh, and Igor Rudan. 2023. Can chatgpt draft a research article? an example of population-level vaccine effectiveness analysis. *Journal of Global Health*, 13:01003.
- Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.
- Kimberley Mok. 2023. Chatgpt writes scientific abstracts that can fool experts. https://thenewstack.io/chatgpt-writes-scientific-abstracts-well-enough-to-fool-experts/.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI Tech Report*.

- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv* preprint arXiv:1910.01108.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *OpenAI Tech Report*.
- Yi Tay, Dara Bahri, Che Zheng, Clifford Brunk, Donald Metzler, and Andrew Tomkins. 2020. Reverse engineering configurations of neural text generation models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 275–279.
- H. Holden Thorp. 2023. Chatgpt is fun, but not an author. *Science*, 379(6630):313–313.
- Furkan Ufuk. 2023. The role and limitations of large language models such as chatgpt in clinical settings and medical journalism. *Radiology*, page 230276.
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 9054–9065.
- ZeroGPT. 2023. Ai text detector. https://www.zerogpt.com.