English Dataset

Model	HC3	Translation	Summarization	Paraphrasing	Overall
RoBERTa-HC3	99.40	57.79	58.49	83.03	76.81
RoBERTa-HC3 Plus	98.84	71.36	99.21	96.53	89.93
InstructDGGC	**99.52**	**75.59**	**99.49**	**97.87**	**91.73**

Chinese Dataset

Model	HC3	Translation	Summarization	Paraphrasing	Overall
RoBERTa-HC3	**98.65**	62.68	53.69	59.91	67.87
RoBERTa-HC3 Plus	92.64	61.49	94.25	**92.67**	87.12
InstructDGGC	92.21	**65.19**	**94.26**	92.47	**87.70**

Table 2: Test set accuracy of the model on the HC3 data and our proposed translation, summarization, paraphrasing data.

ter configuration and we train for 4 epochs with a learning rate of $1e^{-4}$ and a batch size of 32.

4.2 Experiment Result and Analysis

As shown in Table 2, we respectively select the best checkpoint according to the results of the validation sets of HC3 and HC3-SI then give the results of the test set. For HC3-SI, we give the score details of each task, involving translation, summarization, and paraphrasing tasks.

First, we focus on the RoBERTa-based model. RoBERTa-HC3 performed poorly on our proposed dataset, indicating that the model struggles to handle semantic-invariant data generated by ChatGPT. We then used HC3 Plus to train RoBERTa. For semantic-invariant data, we observed that, aside from a decline in performance on the Chinese translation data, the model's performance on other datasets improved significantly. We noted that for the translation task, ChatGPT generates sentences that are much more similar to the target sentences in the dataset compared to tasks like summarization and paraphrasing. As a result, detecting AI-generated content in translation data is more challenging for the model, even with a substantial amount of training data. Appendix C provides examples and a more detailed analysis. Additionally, we observed that RoBERTa-HC3, when trained only on HC3 data, demonstrates stronger detection capabilities on HC3 compared to RoBERTa-HC3 Plus. This suggests that training a detector for a specific task is easier than training one that needs to accommodate various types of task data.

Next, we focus on InstructDGGC. As shown in Table 2, we observe a significant improvement in performance on English data compared to RoBERTa-HC3 Plus, demonstrating the strong generalization capabilities of language models fine-tuned on a large number of instructions. However, for Chinese data, aside from a noticeable improvement in translation tasks, the results for other tasks remain largely unchanged. We believe the limited improvement is due to the model's insufficient instruction-following capability, likely stemming from a lack of fine-tuning data for Chinese instructions. Moreover, compared to HC3 Plus, InstructDGGC performs better on translation data, further confirming that it is more robust.

Finally, we focus on the overall score. Compared to RoBERTa-HC3 Plus, InstructDGGC achieved better overall performance, with an improvement of 1.8% for English data and 0.58% for Chinese data. This indicates that fine-tuning instruction models for detecting generated text is a promising method.

5 Conclusion

In this work, we experimentally demonstrate that detecting ChatGPT-generated text becomes more difficult on semantic-invariant tasks, for which we propose a more extensive and comprehensive dataset HC3 Plus. And we use the instruction fine-tuning model to achieve a detector and experimental results show that our detector outperforms the RoBERTa-based method proposed by previous work. We hope that our analysis of semantic-invariant text detection can provide some insights for future work.

6 Limitations

While our experiments show that detecting semantic-invariant text is more challenging for detectors, our dataset only considers the current version of ChatGPT, i.e., `GPT-3.5-Turbo-0301`. As ChatGPT undergoes continuous iterative improvement, our dataset may become outdated compared to the evolving versions of ChatGPT. In future work, we plan to further investigate the impact of these iterations on detection performance.

References

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: statistical detection and visualization of generated text. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 111–116. Association for Computational Linguistics.

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. *CoRR*, abs/2301.07597.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint arXiv:1506.05865*.

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, ACL 2020, Online, July 5-10, 2020*, pages 1808–1822. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

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

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

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.

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

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4762–4772. International Committee on Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

A Related Work

OpenAI released a neural network-based detector to facilitate the detection of outputs of GPT-2 (Radford et al., 2019) models. Another stream of work focuses on zero-shot AI text detection without any

additional training overhead (Mitchell et al., 2023; Ippolito et al., 2020; Gehrmann et al., 2019). After the release of ChatGPT, (Guo et al., 2023) collected a human ChatGPT comparison corpus HC3. Then based on this dataset, they trained detection systems to distinguish whether the text was generated by ChatGPT or humans. However, HC3 mainly focuses on the question-answering task, which neglects the semantic-invariant tasks popular in real-world scenarios. (Sadasivan et al., 2023) found that using a lightweight paraphraser can cause most detectors to fail. Furthermore, (Wang et al., 2022) proposed T*k*-instruct, which performed fine-tuning on a large number of instructions and showed strong generalization ability. To counter the above issues, we propose a new dataset that considers semantic-invariant scenarios. We also train a novel detector based on T*k*-instruct which achieves better detection performance.

B HC3-SI Dataset Description

As shown in Table 3, we give the number of samples in the data set HC3-SI. For Chinese data, the number of samples in the train/val/test set are $42708/4746/22516$. For English data, the number of samples in the train/val/test set The sample sizes are $95745/10641/38142$ respectively. The total sample size of our proposed HC3-SI is $210,000$, nearly twice that of HC3. In addition, our data does not overlap with HC3, and they can be used together.

English Dataset

Dataset	train	val	test
CNN/DailyMail	11160	1240	6000
Xsum	11158	1240	5994
WMT'19 de→en	11077	1231	4264
WMT'19 fr→en	11043	1227	5934
WMT'19 zh→en	11039	1227	5990
WMT'19 ro→en	11035	1227	3960
HC3-Paraphrase-en	29233	3249	6000
Total in English	95745	10641	38142

Chinese Dataset

Dataset	train	val	test
LCSTS	11160	1240	6000
news2016	11023	1225	5918
WMT'19 en→zh	10823	1203	5976
HC3-Paraphrase-zh	9702	1078	4622
Total in Chinese	42708	4746	22516

Table 3: The size of our proposed dataset HC3-SI.

C Sample Description

Figure 2 gives examples of translation, summary generation, and paraphrasing respectively in HC3-SI. Compared with summarization and paraphrasing, the translation task has stronger semantic constraints from the source sentence, which leads to a higher similarity between the text generated by ChatGPT and the text written by humans. Taking the French-to-English sample in Figure 2 as an example, the ChatGPT-generated text has a high overlap of word fragments and many synonymous replacement words compared with the human-written text. Such samples are indiscriminate sources for detectors, which is also verified by our experimental results. For summarization and paraphrasing, although the text generated by ChatGPT and the text written by humans have roughly the same semantics, there are obvious differences in style. Therefore, a well-trained detector will recognize them more easily.