[36] Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970, 2022.

[37] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[38] Bright Xu. Nlp chinese corpus: Large scale chinese corpus for nlp, September 2019.

[39] Yi Yang, Scott Wen-tau Yih, and Chris Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL - Association for Computational Linguistics, September 2015.

[40] Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385, 2019.

A Appendix

A.1 HC3 Dataset Splits Creation

We create 5 and 7 splits for HC3 English and Chinese, respectively. Most of the data come from the publicly available Question-Answering (QA) datasets, where details are listed in the following. For these QA data, we directly input the questions to ChatGPT and collect at least one answer.

We also crawled some wiki concepts and explanations from Wikipedia and BaiduBaike, where explanations are treated as human expert answers and concepts are used to construct the questions, details ref to bellow paragraphs.

For HC3-English, we create five dataset splits:

1. `reddit_eli5`. Sampled from the ELI5 dataset [10].
2. `open_qa`. Sampled from the WikiQA dataset [39].
3. `wiki_csai`. We collected the descriptions of hundreds of computer science-related concepts from Wikipedia[13] as the human experts' answers to questions like "Please explain what is `<concept>`?"
4. `medicine`. Sampled from the Medical Dialog dataset [6].
5. `finance`. Sampled from the FiQA dataset [23], which is built by crawling StackExchange[14] posts under the Investment topic.

For HC3-Chinese, we create seven dataset splits:

1. `open_qa`. Sampled from the WebTextQA and BaikeQA corpus in [38].
2. `baike`. We collected the descriptions of more than a thousand information science-related concepts from BaiduBaike[15] as the human experts' answers to questions like "我有一个计算机相关的问题，请用中文回答，什么是`<concept>`"
3. `nlpcc_dbqa`. Sampled from the NLPCC-DBQA dataset [8].
4. `medicine`. Sampled from the Medical Dialog dataset [6].
5. `finance`. Sampled from the FinanceZhidao dataset [31].

[13] https://www.wikipedia.org/

[14] https://stackexchange.com/

[15] https://baike.baidu.com/

6. `psychology` Sampled from a public Chinese Psychological Question Answering Dataset[16].
7. `law`. Sampled from the LegalQA dataset[17].

A.2 Additional Results

Here we demonstrate the additional results of dependency relations for the Chinese corpus, as is shown in Figure 6. The conclusion is basically consistent with the main paper.

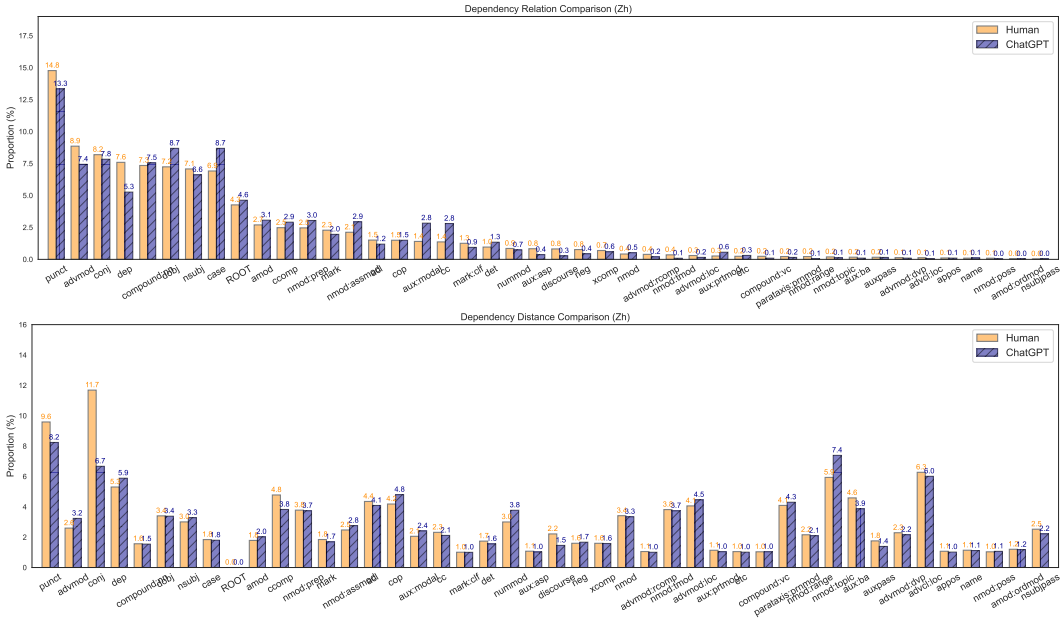

Figure 6: Top-30 dependency relations (upper) and corresponding dependency distances (lower) comparison between human and ChatGPT answers in the HC3-Chinese. Results are sorted by relations proportion of human answers.

Other detailed results, including vocabulary features, sentiment analyses, and dependency parsing results for each data source are all available at our project GitHub repository at `https://github.com/Hello-SimpleAI/chatgpt-comparison-detection`.

A.3 Human Evaluations Examples

For evaluation examples of our human evaluations, please visit our project GitHub repository at `https://github.com/Hello-SimpleAI/chatgpt-comparison-detection`.

[16] https://aistudio.baidu.com/aistudio/datasetdetail/38489
[17] https://github.com/siatnlp/LegalQA