

Lastly, **RoBERTa-based-detector is effective in handling Out-Of-Distribution scenarios**. When compared to the original model, it demonstrates a significant decrease in performance on GLTR’s OOD test datasets, with a drop of up to 28.8% on English datasets(*filtered-full* → *filtered-full* – *filtered-full* → *filtered-sent*) and 45.5% on Chinese datasets(*raw-full* → *raw-full* – *raw-full* → *raw-sent*). However, RoBERTa maintains consistent performance with F1-scores varying by no more than 19%.

5.4.2 How will the indicating words influence the detector?

We first collected a bunch of indicating words for both humans and ChatGPT. For example, ChatGPT’s indicating words (or phrases) include "AI assistant", "I’m sorry to hear that", and "There’re a few steps...", etc. and humans’ indicating words may include "Hmm", "Nope", "My view is", etc. In the filtered version, we remove all sentences in the answers that contain the indicating words for both humans and ChatGPT.

According to Table 4, **removing the indicating words helps the models trained on full-text to perform better across different content granularities**. For example, the RoBERTa-*filter-full* performs significantly better than RoBERTa-*raw-full* in terms of sentence-level and mix-level evaluations, improving more than 3% F1 scores on average. However, **the filtering may slightly hurt the performances of the models trained on sentences**. This may be because the indicating words play a bigger part in the sentence-level text compared with the full text. Removing the indicating words may make some sentences literally unable to be distinguished.

5.4.3 Which granularity is more difficult to detect? Full-text or sentence?

Through the extensive experimental results in Table 5, we conclude that **detecting ChatGPT generated texts is more difficult in a single sentence than in a full text**. This conclusion can be proved by the following two points: First, our results show that both English and Chinese sentence-based detectors (i.e., *raw-sent* and *filtered-sent* versions) achieve satisfactory results w.r.t. the testing task of detecting either ChatGPT generated paragraphs or sentences, whereas the opposite is not true—*raw-full* and *filtered-full* are relatively inferior when detecting ChatGPT generated sentences. In other words, detectors trained on "hard samples" (i.e., sentence corpus) are much easier to solve simple task (i.e., detecting full corpus), while "simple samples" (i.e., full corpus) may be less useful for solving more difficult task (i.e., sentence corpus).

Second, we observe that although both full and sentence corpus are provided in the *raw-mix* and *filtered-mix* versions, it is still more difficult for them to detect single sentences generated by ChatGPT. This is even more obvious for the Chinese corpus, where the F1-score of *raw-mix* trained on the Chinese corpus is 94.09% for testing raw sentence answers, compared to that 97.43% for testing raw full answers. Similar results can be observed for the filtered corpus, where F1-score of *filtered-mix* is 95.61% for testing filtered sentence answers, compared to its F1-score of 97.66% for testing filtered full answers. One possible explanation is that the expression pattern of ChatGPT is more obvious (therefore more easily detected) when paragraphs of text are provided, whereas it is more difficult to detect generated single sentences.

Test → Train ↓	English						Chinese							
	full	raw sent	mix	full	filtered sent	mix	Avg.	full	raw sent	mix	full	filtered sent	mix	Avg.
full-raw	99.82	81.89	84.67	99.72	81.00	84.07	88.53	98.79	83.64	86.32	98.57	82.77	85.85	89.32
sent-raw	99.40	98.43	98.56	99.24	98.47	98.59	98.78	97.76	95.75	96.11	97.68	95.31	95.77	96.40
mix-raw	99.44	98.31	98.47	99.32	98.37	98.51	98.74	97.70	95.68	96.04	97.65	95.27	95.73	96.35
full-filtered	99.82	87.17	89.05	99.79	86.60	88.67	91.85	98.25	91.04	92.30	98.14	91.15	92.48	93.89
sent-filtered	96.97	97.22	97.19	99.09	98.43	98.53	97.91	96.60	92.81	93.47	97.94	95.86	96.26	95.49
mix-filtered	96.28	96.43	96.41	99.45	98.37	98.53	97.58	97.43	94.09	94.68	97.66	95.61	96.01	95.91

Table 5: F1 scores (%) of RoBERTa models at full & sent & mix mode.

5.4.4 Which corpus is more helpful for model training? Full-text, sentence, or mix of the two?

We find that both English and Chinese RoBERTa-based **detectors are more robust when fine-grained corpus data is available in model training**. The sentence-based detectors outperform full-based detectors w.r.t. F1-scores, while the latter can be significantly improved when the sentence corpus is injected in model training, as we observe that mix-based detectors also achieve satisfactory results. For English corpus, *raw-full* only achieves 81.89% F1-score for testing sentence answers, while *raw-sent* is significantly better with 98.43% F1-score, as shown in Table 5. Moreover, the relatively inferior detection performance can be improved by injecting sentence answers into the detector, where we find that *raw-mix* can also obtain significant improvement (with 98.31% F1-score) over the detectors trained on only full answers. Similar conclusions can be acquired for the filtered versions, where both *filtered-sent* and *filtered-mix* significantly outperform *filtered-full* version w.r.t. F1-score, which holds for both English and Chinese corpus.

We indicate that the above conclusions could also hold for other types of detectors like GLTR Test-2 feature-based detectors, as is shown in Table 4. For GLTR Test-2, the average performance of F1-score of *raw-full* and *filtered-full* is 61.74% and 69.47%, respectively, compared to that of *raw-sent* 76.26% and *filtered-sent* 76.41%, where the performance of detectors trained on the mixed corpus is close to the sentence-based versions.

Taking into account the conclusions of the previous paragraph about the detection difficulty between full and sentence answers, we indicate that the fine-grained corpus is helpful for distinguishing ChatGPT generated texts, as it additionally provides guidance and hints in model training for detecting the subtle patterns of ChatGPT hidden in single sentences.

5.4.5 Will a QA-style detector be more effective than a single-text detector?

Table 6 demonstrates the results of both *raw-full* and *filtered-full* models across all test datasets.

On English datasets, the QA model’s F1-scores are superior to that of the single model, except for two *full* test datasets, where it averages 97.48% F1-scores and surpasses single model by 5.63%. There exist some differences in Chinese datasets, where the single model outperforms QA in *raw-full* train dataset. However, the QA model still yields the best evaluation at 94.22%.

In conclusion, **the QA model is generally more effective than the single model and is suitable for filtered scenarios. And the QA training makes models more robust to the sentence inputs.**

Test →	English						Chinese							
	full	raw sent	mix	full	filtered sent	mix	Avg.	full	raw sent	mix	full	filtered sent	mix	Avg.
Train → <i>raw</i> - full														
Single	99.82	81.89	84.67	99.72	81.00	84.07	88.53	98.79	83.64	86.32	98.57	82.77	85.85	89.32
QA	99.84	92.68	93.70	99.75	92.34	93.46	95.30	98.99	80.56	83.85	98.73	80.24	83.89	87.71
Train → <i>filtered</i> - full														
Single	99.82	87.17	89.05	99.79	86.60	88.67	91.85	98.25	91.04	92.30	98.14	91.15	92.48	93.89
QA	99.70	96.14	96.64	99.70	96.07	96.61	97.48	97.29	92.10	93.01	97.18	92.40	93.31	94.22

Table 6: F1 scores (%) of RoBERTa models trained with QA & Single settings.

5.4.6 Which data sources are more difficult for the ChatGPT detectors? and What are the conditions that make it easier to detect ChatGPT?

As shown in Table 7, the evaluation results based on *filtered-full* model are separated by various sources in our HC3 dataset.

On the English datasets, the F1-scores for human answers are slightly higher than those for ChatGPT without any exceptions, regardless of whether RoBERTa or GLTR is used on full-text test datasets. However, the F1-scores for ChatGPT are highly inconsistent on transferring test datasets particularly open-qa dataset with varying performance. **In terms of data resource, reddit-elis5 and finance-en has higher values, while wiki-csai poses a challenge for detectors.**

On the Chinese datasets, the F1-scores of humans and ChatGPT are comparable with no significant difference. This suggests that the difficulty in detecting ChatGPT depends on the data source. **It is observed that open-qa and baike have better performance, whereas the nlpcc-dbqa has lower performance.**

Above all, the evaluations on Chinese dataset show more stability on transferring test dataset compared to the English datasets. Furthermore, it's evident that the F1-scores of ChatGPT are lower than those of human answers, regardless of whether the dataset is English or Chinese. This indicates that **ChatGPT's detector relies more heavily on In-Distribution models.**

Model	Test	F1-hu	F1-ch	F1-hu	F1-ch	F1-hu	F1-ch	F1-hu	F1-ch	F1-hu	F1-ch
English											
Chinese											
RoBERTa	full	99.34	99.28	99.69	99.62	99.53	98.60	100.00	100.00	96.59	96.37
	sent	78.84	85.84	84.06	80.45	70.74	26.78	77.27	93.31	68.91	84.12
GLTR	full	97.50	97.37	98.28	97.96	92.68	82.20	98.22	99.40	95.76	95.72
	sent	46.60	75.26	45.41	61.72	42.01	17.81	38.12	87.05	39.24	76.94

Model	Test	F1-hu	F1-ch								
English											
Chinese											
RoBERTa	full	98.87	97.99	97.78	98.50	98.75	99.33	97.42	95.42	94.61	93.99
	sent	95.00	80.46	93.77	86.23	91.17	93.77	90.10	63.29	86.08	88.88
GLTR	full	86.67	80.42	82.41	88.89	85.75	93.15	77.25	69.78	81.62	77.91
	sent	36.91	32.80	33.99	46.22	36.45	75.21	46.39	27.50	48.10	71.72

Table 7: Human (F1-hu) and ChatGPT (F1-ch) detection F1 scores (%) w.r.t. different data source, models are trained on filtered full text, tested on filtered full and sent. On HC3-Chinese, we omitted the results of *medicine* and *psychology* domains, which are similar to *finance* and *open_qa*, respectively.

6 Conclusion

In this work, we propose the HC3 (Human ChatGPT Comparison Corpus) dataset, which consists of nearly 40K questions and their corresponding human/ChatGPT answers. Based on the HC3 dataset, we conduct extensive studies including human evaluations, linguistic analysis, and content detection experiments. The human evaluations and linguistics analysis provide us insights into the implicit differences between humans and ChatGPT, which motivate our thoughts on LLMs' future directions. The ChatGPT content detection experiments illustrate some important conclusions that can provide beneficial guides to the research and development of AIGC-detection tools. We make all our data, code, and models publicly available to facilitate related research and applications at <https://github.com>Hello-SimpleAI/chatgpt-comparison-detection>.

7 Limitations

Despite our comprehensive analysis of ChatGPT, there are still several limitations in the current paper, which will be considered for improvement in our future work:

1. Despite our efforts in data collection, the amount and range of collected data are still not enough and the data from different sources are unbalanced, due to limited time and resources. To make more accurate linguistic analyses and content detection, more data with different styles, sources, and languages are needed;
2. Currently, all the collected ChatGPT answers are generated **without special prompts**. Therefore, the analysis and conclusions in this paper are built upon ChatGPT's most general style/state. For example, using special prompts such as "Pretending you are Shakespeare..." can generate content that bypasses our detectors or make the conclusions in this paper untenable;