How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection

Biyang Guo[1][†][*], **Xin Zhang**[2][*], **Ziyuan Wang**[1][*], **Minqi Jiang**[1][*], **Jinran Nie**[3][*]
Yuxuan Ding[4], **Jianwei Yue**[5], **Yupeng Wu**[6]
[1]AI Lab, School of Information Management and Engineering
Shanghai University of Finance and Economics
[2]Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen)
[3]School of Information Science, Beijing Language and Culture University
[4]School of Electronic Engineering, Xidian University
[5]School of Computing, Queen's University, [6]Wind Information Co., Ltd

Abstract

The introduction of ChatGPT[2] has garnered widespread attention in both academic and industrial communities. ChatGPT is able to respond effectively to a wide range of human questions, providing fluent and comprehensive answers that significantly surpass previous public chatbots in terms of security and usefulness. On one hand, people are curious about how ChatGPT is able to achieve such strength and how far it is from human experts. On the other hand, people are starting to worry about the potential negative impacts that large language models (LLMs) like ChatGPT could have on society, such as fake news, plagiarism, and social security issues. In this work, we collected tens of thousands of comparison responses from both human experts and ChatGPT, with questions ranging from open-domain, financial, medical, legal, and psychological areas. We call the collected dataset the **H**uman **C**hatGPT **C**omparison **C**orpus (**HC3**). Based on the HC3 dataset, we study the characteristics of ChatGPT's responses, the differences and gaps from human experts, and future directions for LLMs. We conducted comprehensive human evaluations and linguistic analyses of ChatGPT-generated content compared with that of humans, where many interesting results are revealed. After that, we conduct extensive experiments on how to effectively detect whether a certain text is generated by ChatGPT or humans. We build three different detection systems, explore several key factors that influence their effectiveness, and evaluate them in different scenarios. The dataset, code, and models are all publicly available at `https://github.com/Hello-SimpleAI/chatgpt-comparison-detection`.

1 Introduction

Since its dazzling debut in November 2022, OpenAI's ChatGPT has gained huge attention and wide discussion in the natural language processing (NLP) community and many other fields. According to OpenAI, ChatGPT is fine-tuned from the GPT-3.5 series with Reinforcement Learning from Human Feedback (RLHF; [7, 32]), using nearly the same methods as InstructGPT [25], but with slight differences in the data collection setup. The vast amount of knowledge in GPT-3.5 and the meticulous fine-tuning based on human feedback enable ChatGPT to excel at many challenging NLP

---

[*]Equal Contribution.

[†]Project Lead. Corresponding to `guo_biyang@163.com`

[+]Each author has made unique contributions to the project.

[2]Launched by OpenAI in November 2022. `https://chat.openai.com/chat`

tasks, such as translating natural language to code [5], completing the extremely masked text [15] or generating stories given user-defined elements and styles [40], let alone typical NLP tasks like text classification, entity extraction, translation, etc. Furthermore, the carefully collected human-written demonstrations also make ChatGPT able to admit its mistakes, challenge incorrect premises and reject even inappropriate requests, as claimed by OpenAI[3].

The surprisingly strong capabilities of ChatGPT have raised many interests, as well as concerns:

On the one hand, **people are curious about how close is ChatGPT to human experts**. Different from previous LLMs like GPT-3 [4], which usually fails to properly respond to human queries, InstructGPT [25] and the stronger ChatGPT have improved greatly in interactions with humans. Therefore, ChatGPT has great potential to become a daily assistant for general or professional consulting purposes [20, 21]. From the linguistic or NLP perspectives, we are also interested in where are the remaining gaps between ChatGPT and humans and what are their implicit linguistic differences [14, 18].

On the other hand, **people are worried about the potential risks brought by LLMs like ChatGPT**. With the free preview demo of ChatGPT going virus, a large amount of ChatGPT-generated content crowded into all kinds of UGC (User-Generated Content) platforms, threatening the quality and reliability of the platforms. For example, Stack Overflow, the famous programming question-answering website, has temporarily banned ChatGPT-generated content[4], because it believes *"the average rate of getting correct answers from ChatGPT is too low, the posting of answers created by ChatGPT is substantially harmful to the site and to users who are asking and looking for correct answers"*. Many other applications and activities are facing similar issues, such as online exams [33] and medical analysis [20]. Our empirical evaluation of ChatGPT on legal, medical, and financial questions also reveals that potentially harmful or fake information can be generated.

Considering the opaqueness of ChatGPT and the potential social risks associated with model misuse, we make the following contributions to both the academy and society:

1. To facilitate LLM-related research, especially the study on the comparison between humans and LLMs, we collect nearly 40K questions and their corresponding answers from human experts and ChatGPT, covering a wide range of domains (open-domain, computer science, finance, medicine, law, and psychology), named as the **Human ChatGPT Comparison Corpus** (**HC3**) dataset. The HC3 dataset is a valuable resource to analyze the linguistic and stylist characteristics of both humans and ChatGPT, which helps to investigate the future improvement directions for LLMs;

2. We conduct comprehensive **human evaluations** as well as **linguistic analysis** on human/ChatGPT-generated answers, discovering many interesting patterns exhibited by humans and ChatGPT. These findings can help to distinguish whether certain content is generated by LLMs, and also provide insights about where language models should be heading in the future;

3. Based on the HC3 dataset and the analysis, we develop several **ChatGPT detecting models**, targeting different detection scenarios. These detectors show decent performance in our held-out test sets. We also conclude several key factors that are essential to the detector's effectiveness.

4. We **open-source** all the collected comparison corpus, evaluations, and detection models, to facilitate future academic research and online platform regulations on AI-generated content.

2   Human ChatGPT Comparison Corpus (HC3)

ChatGPT is based on the GPT-3.5 series, which is pre-trained on the super-large corpus, consisting of web-crawled text, books, and codes, making it able to respond to all kinds of questions. Therefore, we are curious how will a human (especially an expert) and ChatGPT respond to the same question respectively. Inspired by [1], we also want to evaluate whether ChatGPT can keep honest (not fabricate information or mislead the user), harmless (shouldn't generate harmful or offensive content),

---

[3]https://openai.com/blog/chatgpt/
[4]https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned

HC3-English

	# Questions	# Human Answers	# ChatGPT Answers	Source
All	24322	58546	26903	
reddit_eli5	17112	51336	16660	ELI5 dataset [10]
open_qa	1187	1187	3561	WikiQA dataset [39]
wiki_csai	842	842	842	Crawled Wikipedia (A.1)
medicine	1248	1248	1337	Medical Dialog dataset [6]
finance	3933	3933	4503	FiQA dataset [23]

HC3-Chinese

	# Questions	# Human Answers	# ChatGPT Answers	Source
All	12853	22259	17522	
open_qa	3293	7377	3991	WebTextQA & BaikeQA [38]
baike	4617	4617	4617	Crawled BaiduBaike (A.1)
nlpcc_dbqa	1709	1709	4253	NLPCC-DBQA dataset [8]
medicine	1074	1074	1074	Medical Dialog dataset [6]
finance	689	1572	1983	ChineseNlpCorpus (A.1)
psychology	1099	5220	1099	from Baidu AI Studio (A.1)
law	372	690	505	LegalQA dataset (A.1)

Table 1: Meta-information of the HC3 dataset. The English (resp. Chinese) contains 5 (resp. 7) splits.

and how *helpful* (provide concrete and correct solutions to the user's question) it is compared to human experts.

Taking these into account, we decided to collect a comparison corpus that consists of both human and ChatGPT answers to the same questions. We believe such a comparison corpus can be a valuable and interesting source to study the nature of the language of both humans and language models.

2.1 Human Answers Collection

Inviting human experts to manually write questions and answers is tedious and unaffordable for us to collect a large amount of data, therefore we construct the comparison dataset mainly from two sources:
• Publicly available question-answering datasets, where answers are given by experts in specific domains or the high-voted answers by web users;
• Wiki text. We construct question-answer pairs using the concepts and explanations from wiki sources like Wikipedia[5] and BaiduBaike[6].

The split-data source mapping is shown in Table 1, and please refer to Appendix A.1 for further detailed information.

2.2 ChatGPT Answers Collection

Based on the collected human question-answering datasets, we use ChatGPT to generate answers to these questions. Since the ChatGPT is currently only available through its preview website, we manually input the questions into the input box, and get the answers, with the aid of some automation testing tools. Answers by ChatGPT can be influenced by the chatting history, so we refresh the thread for each question.

To make the answer more aligned with human answers, we add additional instructions to ChatGPT for specific datasets. For example, the human answers from the `reddit-eli5` dataset split are under the context of "Explain like I'm five", therefore we use this context to instruct ChatGPT by adding "Explain like I'm five" at the end of the original question. More detail can be found in the Appendix.

---

[5]`https://www.wikipedia.org/`
[6]`https://baike.baidu.com/`

ChatGPT can generate different answers given the same question in different threads, which is perhaps due to the random sampling in the decoding process. However, we found the differences can be very small, thereby we only collect one answer for most questions.

2.3 Human ChatGPT Comparison Corpus (HC3)

For each question, there can be more than one human/ChatGPT answer, therefore we organize the comparison data using the following format:

```
1  {
2      "question": "Q1",
3      "human_answers": ["A1", "A2"],
4      "chatgpt_answers": ["B1"]
5  }
```

Overall, we collected $24,322$ questions, $58,546$ human answers and $26,903$ ChatGPT answers for the English version, and $12,853$ questions, $22,259$ human answers and $17,522$ ChatGPT answers for the Chinese version. The meta-information of each dataset split is illustrated in Table 1.

3 Human Evaluation & Summarization

In this section, we invite many volunteer testers and conduct extensive human evaluations from different aspects. After the human evaluation, we make our collected comparison corpus available to the volunteers and ask them to manually conclude some characteristics. We then summarize the feedback from the volunteers combined with our observations.

3.1 Human Evaluation

The human evaluation is divided into the **Turing test** and the **Helpfulness Test**. The Turing Test [34] is a test of a machine's ability to exhibit intelligent behavior that is indistinguishable from a human. We invite 17 volunteers, divided into two groups: 8 experts (who are frequent users of ChatGPT) and 9 amateurs (who have never heard of ChatGPT). This is because people who are familiar with ChatGPT may have memorized some patterns exhibited by ChatGPT, helping them to easily distinguish the role.

We designed four types of evaluations, using different query formats or testing groups. We introduce the specific evaluation design and results in the following parts:

$\mathcal{A}$. **Expert Turing Test, Paired Text** (`pair-expert`)
The `pair-expert` test is conducted in the **expert** group. Each tester is required to do a series of tests, each test containing one question and a **pair** of answers (one from humans and another from ChatGPT). The tester needs to determine which answer is generated by ChatGPT.

$\mathcal{B}$. **Expert Turing Test, Single Text** (`single-expert`)
The `single-expert` test is also conducted in the **expert** group. Each tester is required to do a series of tests, each test containing one question and a **single** answer randomly given by humans or ChatGPT. The tester needs to determine whether the answer is generated by ChatGPT.

$\mathcal{C}$. **Amateur Turing Test, Single Text** (`single-amateur`)
The `single-amateur` test is conducted in the **amateur** group. Each tester is required to do a series of tests, each test containing one question and a **single** answer randomly given by humans or ChatGPT. The tester needs to determine whether the answer is generated by ChatGPT.

$\mathcal{D}$. **Helpfulness Test** (`helpfulness`)
We are also curious about how helpful are the answers from ChatGPT compared with humans' answers to one question. Note that helpfulness is a very subjective metric, which can be influenced by many factors, including emotion, tester personality, personal preference, etc. Therefore, simply providing more accurate information or a more detailed analysis may not always lead to a more helpful answer.

The `helpfulness` test is conducted in the **expert** group. Each tester is required to do a series of tests, each containing one question and a **pair** of answers (one from human and another from ChatGPT).

Human Evaluation (En)

	Pair-expert	Single-expert	Single-amateur	Helpfulness
All	0.90	0.81	0.48	0.57
reddit_eli5	0.97	0.94	0.57	0.59
open_qa	0.98	0.78	0.34	0.72
wiki_csai	0.97	0.61	0.39	0.71
medical	0.97	0.97	0.50	0.23
finance	0.79	0.73	0.58	0.60

Human Evaluation (Zh)

	Pair-expert	Single-expert	Single-amateur	Helpfulness
All	0.93	0.86	0.54	0.54
open_qa	1.00	0.92	0.47	0.50
baike	0.76	0.64	0.60	0.60
nlpcc_dbqa	1.00	0.90	0.13	0.63
medicine	0.93	0.93	0.57	0.30
finance	0.86	0.84	0.84	0.75
psychology	1.00	1.00	0.60	0.67
law	1.00	0.77	0.56	0.56

Table 2: Human evaluations of ChatGPT generated answers for both English and Chinese.

Each tester is asked to pretend that the question is proposed by him/herself, and needs to determine which answer is more helpful to him/her.

Settings. We sample around 30 `<question, human_answer, chatgpt_answer>` triplets from each split (i.e., *reddit_eli5*, *wikipedia*, *medical*, etc.) as the samples for the human evaluation. We allocate 2-5 testers for each split and report their average results. For all Turing tests, we report *the proportion that ChatGPT-generated answer is correctly detected* by testers. For the helpfulness test, we report *the proportion that ChatGPT-generated answer is considered to be more helpful*.

Results. Several conclusions can be drawn from the results shown in Table 2. Comparing the results of `pair-expert` and `single-expert`, we can find that **it is easier to distinguish ChatGPT-generated content when providing a comparison pair** than only providing a single answer. Comparing the results of `single-expert` and `single-amateur`, we can find that **the accuracy of experts is much higher than that of amateurs**. The `helpfulness` test gives the proportion of questions that volunteers think the ChatGPT answer is more helpful to them. Surprisingly, results show that **ChatGPT's answers are generally considered to be more helpful than humans' in more than half of questions**, especially for finance and psychology areas. By checking the specific answers in these domains, we find that ChatGPT can usually provide more concrete and specific suggestions. However, ChatGPT performs poorly in terms of helpfulness for the medical domain in both English and Chinese. The ChatGPT often gives lengthy answers to medical consulting in our collected dataset, while human experts may directly give straightforward answers or suggestions, which may partly explain why volunteers consider human answers to be more helpful in the medical domain.

3.2 Human Summarization

After the above evaluations, we open our collected HC3 dataset to the volunteers where they can freely browse the comparison answers from humans and ChatGPT. All dataset splits are allocated to different volunteers, and each volunteer is asked to browse at least 100 groups of comparison data. After that, we ask them to summarize the characteristics of both human answers and ChatGPT answers. Eventually, we received more than 200 feedbacks, and we summarize these findings as follows:

Distinctive Patterns of ChatGPT

(a) **ChatGPT writes in an organized manner, with clear logic**. Without loss of generality, ChatGPT loves to define the core concept in the question. Then it will give out detailed answers step by step and offers a summary at the end, following the deduction and summary structure;

(b) **ChatGPT tends to offer a long and detailed answer.** This is the direct product of the Reinforcement Learning with Human Feedback, i.e. RLHF, and also partly related to the pattern (a) unless you offer a prompt such as "Explain it to me in one sentence";

(c) **ChatGPT shows less bias and harmful information**. ChatGPT is neutral on sensitive topics, barely showing any attitude towards the realm of politics or discriminatory toxic conversations;

(d) **ChatGPT refuses to answer the question out of its knowledge.** For instance, ChatGPT cannot respond to queries that require information after September 2021. Sometimes ChatGPT also refuses to answer what it believes it doesn't know. It is also RLHF's ability to implicitly and automatically determine which information is within the model's knowledge and which is not.

(e) **ChatGPT may fabricate facts.** When answering a question that requires professional knowledge from a particular field, ChatGPT may fabricate facts in order to give an answer, though [25] mentions that InstructGPT model has already shown improvements in truthfulness over GPT-3. For example, in legal questions, ChatGPT may invent some non-existent legal provisions to answer the question. This phenomenon warns us to be extra careful when using ChatGPT for professional consultations. Additionally, when a user poses a question that has no existing answer, ChatGPT may also fabricate facts in order to provide a response.

Many of the conclusions mentioned above like (b),(c),(d) are also discussed in [12] by Fu et al.

Major Differences between Human and ChatGPT

(a) **ChatGPT's responses are generally strictly focused on the given question, whereas humans' are divergent and easily shift to other topics.** In terms of the richness of content, humans are more divergent in different aspects, while ChatGPT prefers focusing on the question itself. Humans can answer the hidden meaning under the question based on their own common sense and knowledge, but the ChatGPT relies on the literal words of the question at hand;

(b) **ChatGPT provides objective answers, while humans prefer subjective expressions.** Generally, ChatGPT generates safer, more balanced, neutral, and informative texts compared to humans. As a result, ChatGPT is excellent at interpreting terminology and concepts. On the other hand, human answers are more specific and include detailed citations from sources based on legal provisions, books, and papers, especially when providing suggestions for medical, legal, and technical problems, etc.;

(c) **ChatGPT's answers are typically formal, meanwhile humans' are more colloquial.** Humans tend to be more succinct with full of oral abbreviations and slang such as "LOL", "TL;DR", "GOAT" etc. Humans also love to apply humor, irony, metaphors, and examples, whereas ChatGPT never uses antiphrasis. Additionally, human communication often includes the "Internet meme" as a way to express themselves in a specific and vivid way;

(d) **ChatGPT expresses less emotion in its responses, while human chooses many punctuation and grammar feature in context to convey their feelings.** Human uses multiple exclamation mark('!'), question mark('?'), ellipsis('...') to express their strong emotion, and use various brackets('(', ')', '[', ']') to explain things. By contrast, ChatGPT likes to use conjunctions and adverbs to convey a logical flow of thought, such as "In general", "on the other hand", "Firstly,..., Secondly,..., Finally" and so on.

Overall, these summarised features indicate that ChatGPT has improved notably in question-answering tasks for a wide range of domains. Compared with humans, we can imagine ChatGPT as a conservative *team* of experts. As a "team", it may lack individuality but can have a more comprehensive and neutral view towards questions.

	English	avg. len.	vocab size	density	Chinese	avg. len.	vocab size	density
human	**All**	142.50	**79157**	**2.33**	**All**	102.27	**75483**	**5.75**
ChatGPT		**198.14**	66622	1.41		**115.3**	45168	3.05
human	*reddit_eli5*	134.21	**55098**	**2.46**	*nlpcc_dbqa*	24.44	10621	**25.43**
ChatGPT		**194.84**	44926	1.38		**78.21**	**11971**	8.96
human	*open_qa*	35.09	9606	**23.06**	*open_qa*	93.68	**40328**	**13.13**
ChatGPT		**131.68**	**16251**	10.40		**150.66**	26451	5.35
human	*wiki_csai*	**229.34**	**15859**	**8.21**	*baike*	**112.25**	**28966**	**5.59**
ChatGPT		208.33	9741	5.55		77.19	14041	3.94
human	*medicine*	92.98	**11847**	**10.42**	*medicine*	92.34	**9855**	**9.94**
ChatGPT		**209.61**	7694	3.00		**165.41**	7211	4.06
human	*finance*	202.07	**25500**	**3.21**	*finance*	80.76	2759	**5.05**
ChatGPT		**226.01**	21411	2.41		**120.84**	**4043**	4.94
human	-	-	-	-	*psychology*	**254.82**	**16160**	**5.77**
ChatGPT	-	-	-	-		164.53	5897	3.26
human	-	-	-	-	*law*	28.77	2093	**19.55**
ChatGPT	-	-	-	-		**143.76**	**3857**	7.21

Table 3: Average answer length, vocabulary size and density comparisons on our corpus.

4 Linguistic Analysis

In this section, we analyze the linguistic features of both humans' and ChatGPT's answers, and try to find some statistical evidence for the characteristics concluded in Section 3.

4.1 Vocabulary Features

In this part, we analyze the vocabulary features of our collected corpus. We are interested in how humans and ChatGPT differ in the choice of words when answering the same set of questions.

Since the number of human/ChatGPT answers is unbalanced, we randomly sample one answer from humans and one answer from ChatGPT during our statistical process. We calculated the following features: **average length** (L), which is the average number of words in each question; **vocab size** (V), the number of unique words used in all answers; we also propose another feature called **density** (D), which is calculated by $D = 100 \times V/(L \times N)$ where N is the number of answers. Density measures how *crowded* different words are used in the text. For example, if we write some articles that add up to 1000 words, but only 100 different words are used, then the density is $100 \times 100/1000 = 10$. The higher the density is, the more different words are used in the same length of text.

In Table 3, we report the vocabulary features for both English and Chinese corpus. Looking at both features of *average length* and *vocab size*, we can see that: **compared to ChatGPT, human answers are relatively shorter, but a larger vocabulary is used.** This phenomenon is particularly obvious in the Chinese open_qa split and the medical splits in both languages, where the average length of ChatGPT is nearly twice longer than that of humans, but the vocab size is significantly smaller.

This phenomenon is also reflected by the *density* factor. The word density of humans is greater than ChatGPT's in **every split**, which further reveals that **humans use a more diverse vocabulary in their expressions**.

4.2 Part-of-Speech & Dependency Analysis

In this part, we compare the occurrences of different part-of-speech (POS) tags and the characteristics of the dependency relations.

4.2.1 Part-of-Speech

Figure 1 illustrates the comparisons between humans and ChatGPT in terms of POS usage. In HC3-English, ChatGPT uses **more** NOUN, VERB, DET, ADJ, AUX, CCONJ and PART words, while using less ADV and PUNCT words.

Figure 1: Part-of-Speech distribution comparison between ChatGPT and human answers. Results are sorted by POS proportion of human answers. The upper figure is for the HC3-English dataset and the lower is for the HC3-Chinese dataset.

A high proportion of nouns (NOUN) often indicates that the text is more argumentative, exhibiting informativeness and objectivity [24]. Accordingly, adposition (ADP) and adjective (ADJ) words also tend to appear more frequently [11]. The frequent co-occurrence of conjunctions (CCONJ) along with nouns, verbs, and adposition words indicates that the structure of the article and the relationships of cause-and-effect, progression, or contrast are clear. The above are also typical characteristics in academic papers or official documents [29]. We believe the RLHF training process has a great influence on ChatGPT's writing style, which partly explains the difference in the POS tags distribution.

4.2.2 Dependency Parsing

Dependency parsing is a technique that analyzes the grammatical structure of a sentence by identifying the dependencies between its words. We parse the answers in the corpus and compare the proportion of different dependency relations and their corresponding dependency distances. Figure 2 shows the comparison between humans and ChatGPT in HC3-English. Due to the limited space, the Chinese version is placed in the Appendix A.2.

The comparison of dependency relations exhibits similar characteristics to that of POS tags, where ChatGPT uses more determination, conjunction, and auxiliary relations. In terms of the dependency distance, ChatGPT has much longer distances for the `punct` and `dep` relations, which is perhaps due to the fact that CharGPT tends to use longer sentences. However, ChatGPT has obviously shorter `conj` relations. According to the analysis of POS tags, ChatGPT usually uses more conjunctions than humans to make the content more logical, this may explain why the `conj` relations of ChatGPT are relatively shorter than humans.

4.3 Sentiment Analysis

Humans are emotional beings, it is natural that our emotions are reflected in our words, to some extent. ChatGPT is learned on large-scale human-generated text, but it is further fine-tuned with human instructions. Therefore we are curious how "emotional" ChatGPT is compared with humans.

We use a multilingual sentiment classification model[7] fine-tuned on Twitter corpus [2] to conduct sentiment analysis for both English and Chinese comparison data. Note that deep learning-based models can be greatly influenced by some indicating words (such as "but" and "sorry" can easily

---

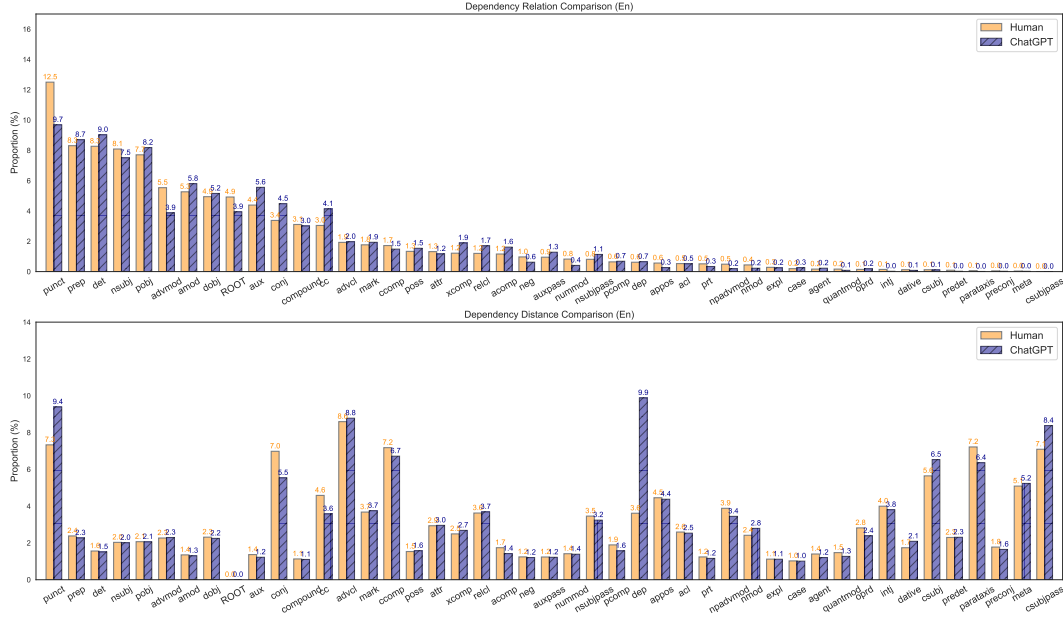[7]https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

Figure 2: Top-30 dependency relations (upper) and corresponding dependency distances (lower) comparison between human and ChatGPT answers in HC3-English. Results are sorted by relations proportion of human answers.

(a) Sentiment distribution of HC3-English

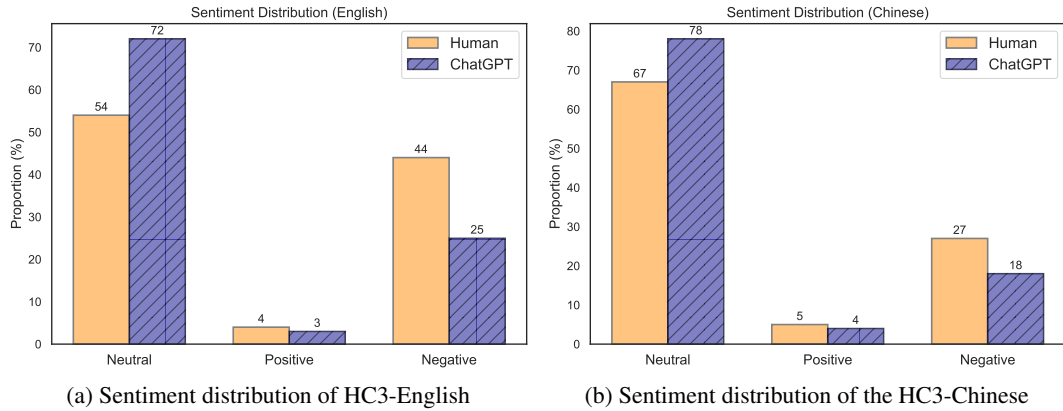(b) Sentiment distribution of the HC3-Chinese

Figure 3: Proportions of three kinds of sentiments (neutral, positive, and negative) in our corpus.

fool the classifier to predict the "negative" label), making the predictions biased [16]. Therefore, the sentiment given by the classifier is only a reference to the true sentiment behind the text.

Figure 3 shows the comparison of the sentiment distribution of humans and ChatGPT. Several findings can be drawn from the results: First, we find that the proportion of neutral emotions is the largest for both humans and ChatGPT, which is in line with our expectations. However, **ChatGPT generally expresses more neutral sentiments than humans**. Then, the proportion of negative emotions is significantly higher than that of positive emotions. Notably, **humans express significantly more negative emotions than ChatGPT**. The proportion of humans' positive emotions is also slightly higher than that of ChatGPT. Overall, ChatGPT is less emotional than humans, though it is not completely emotionless.

4.4 Language Model Perplexity

The perplexity (PPL) is commonly used as a metric for evaluating the performance of language models (LM). It is defined as the exponential of the negative average log-likelihood of the text under

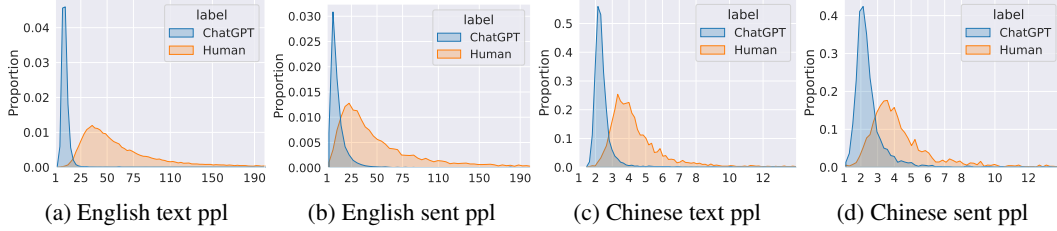| (a) English text ppl | (b) English sent ppl | (c) Chinese text ppl | (d) Chinese sent ppl |

Figure 4: PPL distributions on both English and Chinese data, as well as both text and sentence levels.

the LM. A lower PPL indicates that the language model is more confident in its predictions, and is therefore considered to be a better model. The training of LMs is carried out on large-scale text corpora, it can be considered that it has learned some common language patterns and text structures. Therefore, we can use PPL to measure how well a text conforms to common characteristics.

We use the open-source GPT-2 small[8] (Wenzhong-GPT2-110M[9] for Chinese) model to compute the PPL (both text-level and sentence-level[10] PPLs) of the collected texts. The PPL distributions of text written by humans and text generated by ChatGPT are shown in Figure 4.

It is clearly observed that, regardless of whether it is at the text level or the sentence level, the content generated by ChatGPT has relatively lower PPLs compared to the text written by humans. ChatGPT captured common patterns and structures in the text it was trained on, and is very good at reproducing them. As a result, text generated by ChatGPT have relatively concentrated low PPLs.

Humans have the ability to express themselves in a wide variety of ways, depending on the context, audience, and purpose of the text they are writing. This can include using creative or imaginative elements, such as metaphors, similes, and unique word choices, which can make it more difficult for GPT2 to predict. Therefore, human-written texts have more high-PPL values, and show a long-tailed distribution, as demonstrated in Figure 4.

5 ChatGPT Content Detection

AI-generated content (AIGC) is becoming increasingly prevalent on the internet, and it can be difficult to distinguish it from human-generated content, as shown in our human evaluation (sec 3.1). Therefore, AIGC detectors are needed to help identify and flag content that has been created by a machine, to reduce the potential risks to society caused by improper or malicious use of AI models, and to improve the transparency and accountability of the information that is shared online.

In this section, we conduct several empirical experiments to investigate the ChatGPT content detection systems. Detecting AI-generated content is a widely studied topic [19, 27]. Based on these [30, 13, 27], we establish three different types of detection systems, including machine learning-based and deep learning-based methods, and evaluate them on different granularities and data sources. Detailed results and discussions are provided.

5.1 Methods

Detection of machine-generated text has been gaining popularity as text generation models have advanced in recent years[19, 27]. Here, we implement three representative methods from classic machine learning and deep learning, i.e, a logistic regression model trained on the GLTR Test-2[13] features, a deep classifier for single-text detection and a deep classifier for QA detection. The deep classifiers for both single-text and QA are based on RoBERTa [22], a strong pre-trained Transformer [35] model. In fact, algorithms for OOD detection or anomaly detection [17] can also be applied to develop ChatGPT content detectors, which we leave for future work.

---

[8]https://huggingface.co/gpt2

[9]https://huggingface.co/IDEA-CCNL/Wenzhong-GPT2-110M

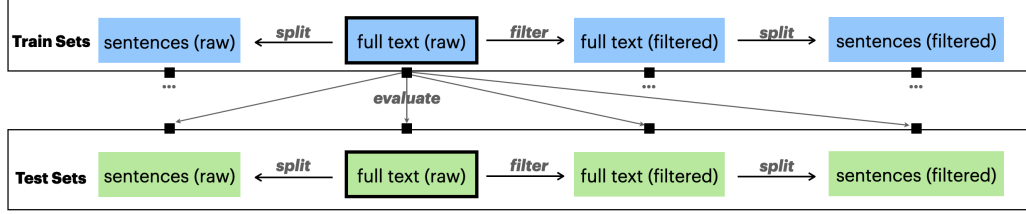[10]For English text, we used NLTK[3] for sentence segmentation (HarvestText for Chinese).

Figure 5: The experiment design for the training and testing of detectors. Different dataset versions are generated through filtering or splitting.

GLTR. [13] studied three tests to compute features of an input text. Their major assumption is that to generate fluent and natural-looking text, most decoding strategies sample high probabilities tokens from the head of the distribution. We select the most powerful Test-2 feature, which is the number of tokens in the Top-10, Top-100, Top-1000, and 1000+ ranks from the LM predicted probability distributions. And then a logistic regression model is trained to finish the classification.

RoBERTa-*sinlge*. A deep classifier based on the pre-trained LM is always a good choice for this kind of text classification problem. It is also investigated in many studies and demo systems [30, 9, 27]. Here we fine-tune the RoBERTa [22] model.

RoBERTa-*QA*. While most content detectors are developed to classify whether a single piece of text is AI-generated, we claim that a detector that supports inputting both a question and an answer can be quite useful, especially for question-answering scenarios. Therefore, we decide to also build a QA version detector. The RoBERTa model supports a text pair input format, where a separating token is used to join a question and its corresponding answer.

5.2 Implementation Details

For the LM used by GLTR, we use gpt2-small [28] for English, and Wenzhong-GPT2-110M released by [36] for Chinese, it is the same with sec. 4.4. For RoBERTa-based deep classifiers, we use `roberta-base`[11] and `hfl/chinese-roberta-wwm-ext`[12] checkpoints for English and Chinese, respectively. All the above models are obtained from huggingface `transformers` [37].

We train the logistic regression model by sklearn [26] on the GLTR Test-2 features from trainset, and search hyper-params following the code of [27]. The RoBERTa-based detectors are trained by the facilities of `transformers`. Specifically, we use the AdamW optimizer, setting batch size to 32 and learning rate to $5e-5$. We finetune models by 1 epoch for English, and 2 epochs for Chinese.

5.3 Experiment Design

The HC3 dataset consists of questions and their corresponding human/ChatGPT answers. We extracted all the `<question, answer>` pairs, and assigned label 0 to pairs with human answers and label 1 to pairs with ChatGPT answers.

Simply using the original answers from humans and ChatGPT to train a binary classifier is the most straightforward way. However, there might be some issues by doing so:

- First, based on the observations in Section 3, both human answers and ChatGPT answers may contain some obvious indicating words that may influence the effectiveness of models;
- Second, users may want to detect whether a single sentence is generated by ChatGPT, instead of the full text. This can be quite difficult for a classifier that is only trained on full texts;
- Third, taking the corresponding question of the answer into account may help the detector to make a more accurate judgment, compared with only considering the answer itself. This

---

[11]https://huggingface.co/roberta-base
[12]https://huggingface.co/hfl/chinese-roberta-wwm-ext

can be widely applied to many QA platforms (like Quora, Stack Overflow, and Zhihu) to find out which answer below a certain question is generated by AI.

Therefore, we design different groups of experiments to study these key questions:
• How will the indicating words influence the detector?
• Is it more challenging for the ChatGPT detectors to detect sentence-level content? Is it harder to train a sentence-level classifier?
• Can the corresponding question help detectors detect the origin of the answer more accurately?

Figure 5 shows how we generate different types of training and testing sets. Specifically, we use the collected raw corpus to construct the first train-test sets (the "full text (raw)" in the figure), which we call the ***raw-full*** version. Then we filter away the indicated words in the text to obtain the ***filtered-full*** version. By splitting the full text into sentences, we obtain the ***raw-sent*** version and the ***filtered-sent*** version. We also combine the full text and the sentences into a mixed version, namely the ***raw-mix*** and ***filtered-mix*** version. Overall, we have six different versions of training and testing sets. Evaluating a model's performance on version B's testing set which is trained on version A's training set can be seen as an out-of-distribution (OOD) generalization evaluation, which is more challenging since it requires the model to be robust when facing sample style changes.

5.4 Results

Following the above experiment design, we conduct comprehensive empirical studies on all kinds of derived corpus. Table 4 shows the test F1 scores.

Test →			English							Chinese				
	full	*raw* sent	mix	full	*filtered* sent	mix	Avg.	full	*raw* sent	mix	full	*filtered* sent	mix	Avg.
Train ↓							RoBERTa							
raw full	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
raw sent	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**
raw mix	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
filtered full	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
filtered sent	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
filtered mix	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
Train ↓							GLTR Test-2							
raw full	98.26	71.58	76.15	98.22	70.19	75.23	81.61	89.61	44.02	53.72	85.89	43.58	53.62	61.74
raw sent	86.26	88.18	87.96	87.72	88.23	88.19	87.76	84.49	71.79	74.01	84.06	70.29	72.90	76.26
raw mix	95.97	86.45	87.81	96.13	86.24	87.73	90.06	86.45	70.85	73.59	84.94	69.14	72.14	76.19
filtered full	98.31	70.91	75.65	98.30	69.48	74.72	81.23	89.46	58.69	64.52	86.51	55.45	62.18	69.47
filtered sent	84.00	88.25	87.71	85.68	88.35	87.99	87.00	84.56	71.85	74.07	84.22	70.59	73.18	76.41
filtered mix	95.36	86.73	87.97	95.60	86.56	87.92	90.02	86.30	71.00	73.70	84.98	69.45	72.40	76.31

Table 4: F1 scores (%) of different models on each testset, average of each language are reported.

5.4.1 Which detector(s) is more useful? ML-based or DL-based? and Why?

According to Table 4, we can derive following conclusions:

Firstly, **the robustness of RoBERTa-based-detector is better than GLTR**. The F1-scores of RoBERTa decrease slightly (1.5-2% in English datasets and 2-3% in Chinese datasets) when sentences are split by comparing the leading diagonal elements in *raw→raw* and *filtered→filtered*. In contrast, the GLTR reduces significantly by over 10% in English datasets, and above 15% in Chinese datasets. Above all, the RoBERTa-based-detector is more robust with anti-interference character. In contrast, the GLTR reduces significantly by over 10% in English datasets, above 15% in Chinese datasets. Above all, the RoBERTa-based-detector is more robust with anti-interference character.

Secondly, **RoBERTa-based-detector is not affected by indicating words.** The F1-scores of RoBERTa only slightly decreased by 0.03% in English *full* dataset, and 0.65% in Chinese *full* dataset, as seen in the minus of relevant leading diagonal elements in *raw→raw* versus *filtered→filtered*. On the contrary, evaluations based on GLTR decrease by up to 3.1% on Chinese datasets, though tiny rise on English datasets, indicating that GLTR is sensitive to indicating words, easily influenced by the patterns of ChatGPT.

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 →	English							Chinese						
Train ↓	*raw*			*filtered*			Avg.	*raw*			*filtered*			Avg.
	full	sent	mix	full	sent	mix		full	sent	mix	full	sent	mix	
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.**

	English							Chinese						
	raw			*filtered*			Avg.	*raw*			*filtered*			Avg.
Test →	full	sent	mix	full	sent	mix		full	sent	mix	full	sent	mix	
	Train → *raw* - 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 → *filtered* - 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-eli5` 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											
		finance		medicine		open_qa		reddit_eli5		wiki_csai	
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
Chinese											
		finance		law		open_qa		nlpcc_dbqa		baike	
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;

3. ChatGPT (perhaps) is mainly trained on English corpus while less on Chinese. Therefore, the conclusions drawn from the HC3-Chinese dataset may not always be precise.

Acknowledgments

References

[1] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[2] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Language Resources and Evaluation Conference*, pages 258–266, Marseille, France, June 2022. European Language Resources Association.

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[6] Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. Meddialog: a large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*, 2020.

[7] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *neural information processing systems*, 2017.

[8] Nan Duan. Overview of the nlpcc-iccpol 2016 shared task: Open domain chinese question answering. In *Natural Language Understanding and Intelligent Applications*, pages 942–948, Cham, 2016. Springer International Publishing.

[9] Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415, 2021.

[10] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: long form question answering. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3558–3567. Association for Computational Linguistics, 2019.

[11] Zhihui Fang. The language demands of science reading in middle school. *International journal of science education*, 28(5):491–520, 2006.

[12] Yao Fu, Hao Peng, and Tushar Khot. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*, Dec 2022.

[13] Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy, July 2019. Association for Computational Linguistics.

[14] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

[15] Biyang Guo, Yeyun Gong, Yelong Shen, Songqiao Han, Hailiang Huang, Nan Duan, and Weizhu Chen. Genius: Sketch-based language model pre-training via extreme and selective masking for text generation and augmentation. *arXiv preprint arXiv:2211.10330*, 2022.

[16] Biyang Guo, Songqiao Han, and Hailiang Huang. Selective text augmentation with word roles for low-resource text classification. *arXiv preprint arXiv:2209.01560*, 2022.

[17] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[18] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*, 2022.

[19] Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[20] Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*, 2022.

[21] Michael R King. The future of ai in medicine: a perspective from a chatbot, 2022.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[23] Macedo Maia, Siegfried Handschuh, Andr'e Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[24] William Nagy and Dianna Townsend. Words as tools: Learning academic vocabulary as language acquisition. *Reading research quarterly*, 47(1):91–108, 2012.

[25] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[27] Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. Deepfake text detection: Limitations and opportunities. In *Proc. of IEEE S&P*, 2023.

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[29] Mary J Schleppegrell. *The language of schooling: A functional linguistics perspective*. Routledge, 2004.

[30] 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.

[31] SophonPlus. Chinesenlpcorpus. `https://github.com/SophonPlus/ChineseNlpCorpus`, 2019.

[32] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. *neural information processing systems*, 2020.

[33] Teo Susnjak. Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*, 2022.

[34] Alan M Turing. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer, 2009.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[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