

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 Ding⁴, **Jianwei Yue**⁵, **Yupeng Wu**⁶

¹AI Lab, School of Information Management and Engineering
Shanghai University of Finance and Economics

²Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen)

³School of Information Science, Beijing Language and Culture University

⁴School of Electronic Engineering, Xidian University

⁵School of Computing, Queen’s University, ⁶Wind Information Co., Ltd

Abstract

The introduction of ChatGPT² 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 **Human ChatGPT Comparison Corpus (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.

²Launched 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 OpenAI³.

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 content⁴, 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	
<i>reddit_elis5</i>	17112	51336	16660	ELI5 dataset [10]
<i>open_qa</i>	1187	1187	3561	WikiQA dataset [39]
<i>wiki_csa</i>	842	842	842	Crawled Wikipedia (A.1)
<i>medicine</i>	1248	1248	1337	Medical Dialog dataset [6]
<i>finance</i>	3933	3933	4503	FiQA dataset [23]

HC3-Chinese

	# Questions	# Human Answers	# ChatGPT Answers	Source
All	12853	22259	17522	
<i>open_qa</i>	3293	7377	3991	WebTextQA & BaikeQA [38]
<i>baike</i>	4617	4617	4617	Crawled BaiduBaike (A.1)
<i>nlpcc_dbqa</i>	1709	1709	4253	NLPCC-DBQA dataset [8]
<i>medicine</i>	1074	1074	1074	Medical Dialog dataset [6]
<i>finance</i>	689	1572	1983	ChineseNlpCorpus (A.1)
<i>psychology</i>	1099	5220	1099	from Baidu AI Studio (A.1)
<i>law</i>	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 Wikipedia⁵ and BaiduBaike⁶.

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-elis5 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/>