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:

```
{
    "question": "Q1",
    "human_answers": ["A1", "A2"],
    "chatgpt_answers": ["B1"]
}
```

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.