	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.

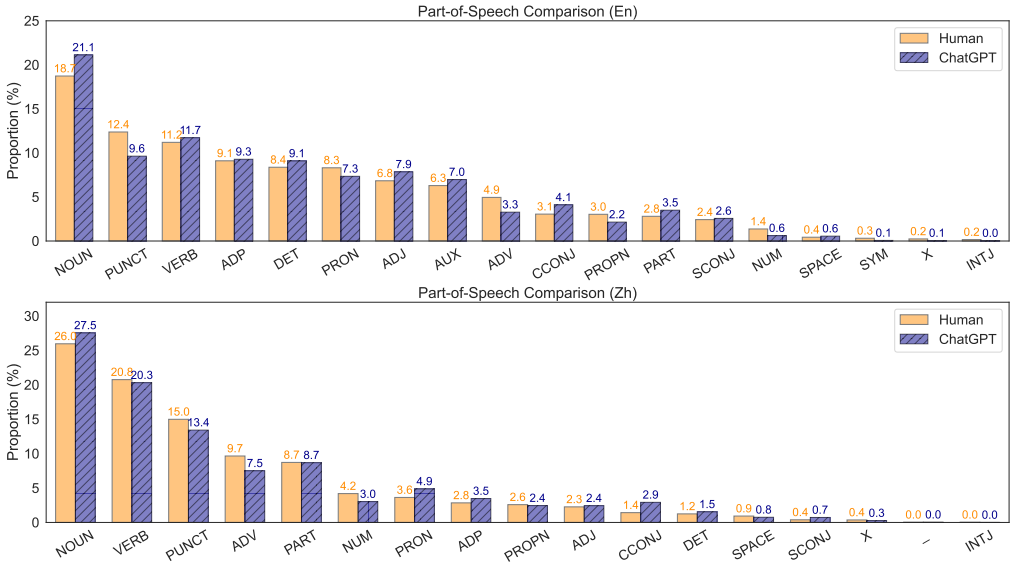


Figure 1: Part-of-Speech distribution comparison between ChatGPT and human answers. Results are sorted by POS proportion of human answers. The upper figure is for the HC3-English dataset and the lower is for the HC3-Chinese dataset.

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

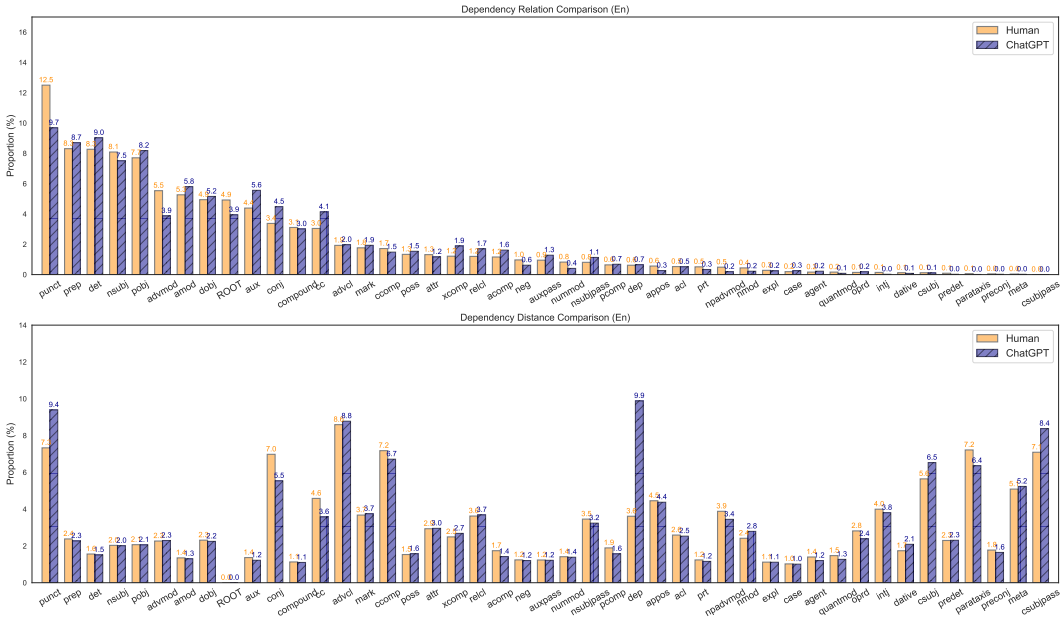
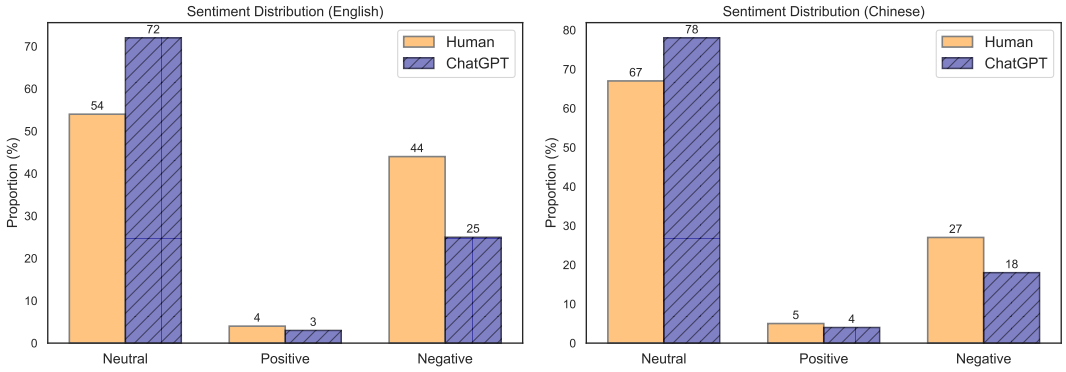


Figure 2: Top-30 dependency relations (upper) and corresponding dependency distances (lower) comparison between human and ChatGPT answers in HC3-English. Results are sorted by relations proportion of human answers.



(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