Can ChatGPT Understand Too?
A Comparative Study on ChatGPT and Fine-tuned BERT

Qihuang Zhong[◇,ℜ]*, **Liang Ding**[ℜ], **Juhua Liu**[◇], **Bo Du**[◇], **Dacheng Tao**[ℜ]
[◇]Wuhan University [ℜ]JD Explore Academy, JD.com Inc.
zhongqihuang@whu.edu.cn, dingliang1@jd.com

Abstract

Recently, ChatGPT has attracted great attention, as it can generate fluent and high-quality responses to human inquiries. Several prior studies have shown that ChatGPT attains remarkable generation ability compared with existing models. However, the quantitative analysis of ChatGPT's understanding ability has been given little attention. In this report, we explore the understanding ability of ChatGPT by evaluating it on the most popular GLUE benchmark, and comparing it with 4 representative fine-tuned BERT-style models. We find that: 1) ChatGPT falls short in handling paraphrase and similarity tasks; 2) ChatGPT outperforms all BERT models on inference tasks by a large margin; 3) ChatGPT achieves comparable performance compared with BERT on sentiment analysis and question-answering tasks. Additionally, several bad cases from inference tasks show the potential limitation of ChatGPT.

1 Introduction

Large language models (LLMs), such as GPT-3 (Brown et al., 2020) and InstructGPT (Ouyang et al., 2022), have swept the natural language processing (NLP) community. Due to their emergent abilities (Wei et al., 2022a), these LLMs can achieve impressive few-shot and zero-shot performance in a variety of NLP tasks. More recently, ChatGPT[1], developed by OpenAI upon InstructGPT (Ouyang et al., 2022), has attracted great attention. Encouragingly, different from prior public chatbots, ChatGPT is able to generate fluent and comprehensive responses to various human inquiries, and even correct inappropriate human questions.

In light of the conventional wisdom that "GPT-style models work well in generation tasks, but perform poorly for understanding tasks, even worse than the base-sized BERT (Devlin et al., 2019)", we wonder whether there is a similar phenomenon in the ChatGPT scenario. For the generation ability of ChatGPT, several prior studies (Jiao et al., 2023; Bang et al., 2023) have shown that ChatGPT can achieve comparable or even better performance than existing LLMs on several generation tasks. However, it is still unclear whether ChatGPT works well on understanding tasks too.

In this report, we provide a systematic study to explore the question: "*can ChatGPT understand too*". This question is answered by evaluating ChatGPT on the authoritative and popular GLUE (Wang et al., 2019) benchmark, spanning 8 understanding tasks, i.e., sentiment analysis, linguistic acceptability, paraphrase, textual similarity, natural language inference, and question answering. For reference, we also compare it with 4 representative BERT-style models. Through a series of experiments and analyses, we find that:

- ChatGPT falls short in handling paraphrase and similarity tasks. Specifically, ChatGPT performs poorly in negative paraphrase and neutral similarity samples, respectively.
- ChatGPT outperforms all BERT-style models on inference tasks by a large margin, indicating its impressive reasoning ability.
- ChatGPT achieves comparable performance compared with BERT-base on sentiment analysis and question answering tasks.
- Despite its good performance on inference tasks, ChatGPT may generate some contradictory or unreasonable responses, which would be its potential limitations.

The remainder of this report is designed as follows. We present the evaluation settings and experimental results in Section 2. In Section 3, we discuss some bad cases in details. Conclusions are described in Section 4.

* Work was done when Qihuang was interning at JD Explore Academy.

[1] https://chat.openai.com

Task	#Pos.	#Neg.	#Neu.	Description	Template Prompt
	Single-Sentence Tasks				
CoLA	25	25	-	acceptablity	For the sentence: "[text]", is the sentence grammarly correct?
SST-2	25	25	-	sentiment	For the sentence: "[text]", is the sentiment in this sentence positive or negative?
	Similarity and Paraphrase Tasks				
MRPC	25	25	-	paraphrase	For the sentence pair "[text_1]" and "[text_2]", do these two sentences have the same semantics?
STS-B	total of 50			similarity	Determine the similarity between the following two sentences: "[text_1]" and "[text_2]". The score should be ranging from 0.0 to 5.0, and can be a decimal.
QQP	25	25	-	paraphrase	For the sentence pair "[text_1]" and "[text_2]", do these two sentences have the same semantics?
	Inference Tasks				
MNLI	25	25	25	NLI	Given the sentence "[text_1]", determine if the following statement is entailed or contradicted or neutral: "[text_2]"
QNLI	25	25	-	QA/NLI	Given the question "[text_1]", determine if the following sentence contains the corresponding answer: "[text_2]"
RTE	25	25	-	NLI	Given the sentence "[text_1]", determine if the following statement is entailed: "[text_2]"

Table 1: Task statistics, descriptions and prompts. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. For ease of illustration, we use "#Pos./#Neg./#Neu." to denote the positive, negative and neutral instances for each task. Considering the limits of ChatGPT, we randomly sample 25 instances for each class from the dev set of each task for evaluation, except for STS-B, where we randomly sample 50 instances from a uniform distribution. In the prompts, `[text]`, `[text_1]` and `[text_2]` are input slots.

2 ChatGPT vs. BERT

In this section, we first introduce the evaluation setting (§2.1), and then present the major results (§2.2). Lastly, some analyses of why ChatGPT performs well or poorly are provided (§2.3).

2.1 Evaluation Setting

Here, we briefly introduce the evaluation setting, including downstream tasks and datasets, baselines, and prompts for ChatGPT.

Tasks and Datasets. Following many prior works (Zhong et al., 2022a, 2023), we use the widely-used GLUE benchmark (Wang et al., 2019) for model evaluation purposes. As one of the most popular NLU benchmarks, GLUE consists of several challenging NLU tasks, including linguistic acceptability (CoLA, Warstadt et al. (2019)), sentiment analysis (SST-2, Socher et al. (2013)), paraphrase (MRPC, Dolan and Brockett (2005)), textual similarity (STS-B, Cer et al. (2017)), question paraphrase (QQP), textual entailment (MNLI, Williams et al. (2018), RTE, Giampiccolo et al. (2007)) and question-answer entailment (QNLI, Rajpurkar et al. (2016)). Considering the limits of testing ChatGPT, we follow Jiao et al. (2023) and randomly sample a subset of the dev set as the evaluation data for each task[2]. Specifically, since most of GLUE tasks are classification tasks (except STS-B which is a regression task), we randomly sample 25 instances for each class from the dev set. For STS-B, we randomly sample 50 instances from a uniform distribution. Table 1 shows the detailed task descriptions and statistics.

For evaluation, we report the performance with Accuracy ("*Acc.*") metric for most tasks, except the Pearson and Spearman correlation ("*Pear./Spea.*") for STS-B, the Matthew correlation ("*Mcc.*") for CoLA, the additional F1 score for MRPC and QQP.

Baselines. We compare ChatGPT (Jan 31 Version) with 4 representative BERT-style models, as the BERT models are commonly used as the baselines to evaluate the understanding ability (Zhong et al., 2022b). Specifically, base-sized/ large-sized BERT and RoBERTa (Liu et al., 2019) are used. All models are fine-tuned on the full training set for each task, where the fine-tuning hyper-parameters

[2]We provide the script and data in https://github.com/WHU-ZQH/ChatGPT-vs.-BERT

Method	CoLA	SST-2	MRPC		STS-B		QQP		MNLI		QNLI	RTE	GLUE
	Mcc.	*Acc.*	*Acc.*	*F1*	*Pear.*	*Spea.*	*Acc.*	*F1*	*m.*	*mm.*	*Acc.*	*Acc.*	avg.
BERT-base	56.4	88.0	90.0	89.8	83.0	81.9	80.0	80.0	82.7	82.7	84.0	70.0	79.2
BERT-large	62.4	96.0	92.0	91.7	88.3	86.8	88.0	88.5	82.7	88.0	90.0	82.0	85.4
RoBERTa-base	61.8	96.0	90.0	90.6	90.2	89.1	84.0	84.0	84.0	88.0	92.0	78.0	84.7
RoBERTa-large	65.3	96.0	92.0	92.0	92.9	91.1	90.0	89.4	88.0	90.7	94.0	84.0	87.8
ChatGPT	56.0	**92.0**	**66.0***	**72.1***	**80.9**	**72.4***	**78.0**	79.3	**89.3***	81.3	84.0	**88.0***	78.7

Table 2: Overall comparison between ChatGPT and fine-tuned BERT-style models on GLUE benchmark. The results in green denote that ChatGPT surpasses the BERT-base model by a clear margin (> 2% (↑) score), while the red results denote ChatGPT under-performs BERT-base (> 2% (↓) score)). More specifically, "*" means that the performance difference between ChatGPT and BERT-base is larger than 10%.

are the same to Zhong et al. (2022c). To estimate the lower bound of ChatGPT's understanding ability, we mainly focus on the comparison between ChatGPT and the basic base-sized BERT.

Prompts for ChatGPT. For each task, we design task-specific prompts for triggering the understanding ability of ChatGPT. Specifically, inspired by (Jiao et al., 2023), we also ask ChatGPT to generate the prompts for each task, by inputting the following human inquiries:

```
provide five concise prompts or
templates that can make you deal with
the [x] task
```

where the [x] is the task slot. Taking the sentiment analysis as an example, we show this process in Figure 1.

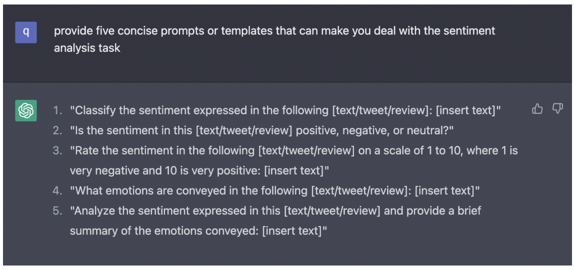

Figure 1: Prompts for sentiment analysis, generated by ChatGPT.

We evaluated ChatGPT on the sentiment analysis task with these five candidate prompts in the preliminary experiments and found a slight performance difference. Thus, for simplicity, we choose one typical prompt for each task and show them in Table 1.

2.2 Main Results

The full results on the GLUE benchmark are shown in Table 2. Overall, ChatGPT can achieve comparable average performance compared with BERT-base (78.7% vs. 79.2%), but still underperforms the other powerful BERT-style models (e.g., RoBERTa-large, 87.8%) by a clear margin. These results show that ChatGPT attains the basic understanding ability, but there is still quite some room for improvement.

Specifically, comparing ChatGPT with BERT-base on specific tasks, we can find that: 1) ChatGPT performs poorly on the paraphrase and similarity tasks, i.e., MRPC and STS-B, where the performance drop is up to 24% score. 2) ChatGPT surpasses all BERT-style models on natural language inference tasks, i.e., MNLI and RTE, indicating its superiority on inference/reasoning. 3) ChatGPT is comparable to BERT-base on the single sentence classification tasks, i.e., sentiment analysis (SST-2) and linguistic acceptability (CoLA), and QA-related tasks, i.e., QNLI.

2.3 Analysis

As seen in Table 2, ChatGPT works well on inference tasks, but falls short in handling paraphrase and similarity tasks. Here, we investigate how ChatGPT works on these special tasks in detail.

Inference Tasks. To have a closer look at why ChatGPT achieves impressive performance on inference tasks, we report the per-class accuracy of ChatGPT and compared models on MNLI and RTE tasks. The results are shown in Table 3. It can be seen that, ChatGPT outperforms BERT-base by a large margin among all settings. Especially, in the class of "entailment", i.e., the premise entails the hypothesis, ChatGPT even surpasses all powerful BERT models by a clear margin. These results continue showing the effective inference ability of ChatGPT, especially reasoning factual input.

Paraphrase Task. Similar to the above analysis, we also report the per-class accuracy of Chat-

Method	MNLI-m			RTE	
	Entailment	Contradiction	Neutral	Entailment	Not_Entailment
BERT-base	88.0	88.0	72.0	76.0	64.0
BERT-large	76.0	92.0	80.0	80.0	84.0
RoBERTa-base	84.0	88.0	80.0	80.0	76.0
RoBERTa-large	84.0	92.0	88.0	92.0	76.0
ChatGPT	92.0* (↑ **4.0**)	96.0* (↑ **8.0**)	80.0 (↑ **8.0**)	96.0* (↑ **20.0**)	80.0 (↑ **16.0**)

Table 3: Per-class accuracy (%) of ChatGPT and BERT-style models on MNLI-m and RTE. The number in parentheses indicates the performance improvement over BERT-base. "*" denotes that ChatGPT outperforms all BERT-style models.

Method	MRPC	
	Entailment	Not_Entailment
BERT-base	88.0	92.0
BERT-large	88.0	96.0
RoBERTa-base	96.0	84.0
RoBERTa-large	92.0	92.0
ChatGPT	88.0 (↓ 0)	**44.0 (↓ 47.0)**

Table 4: Per-class accuracy (%) of ChatGPT and BERT-style models on MRPC. The number in parentheses indicates the performance drops over BERT-base.

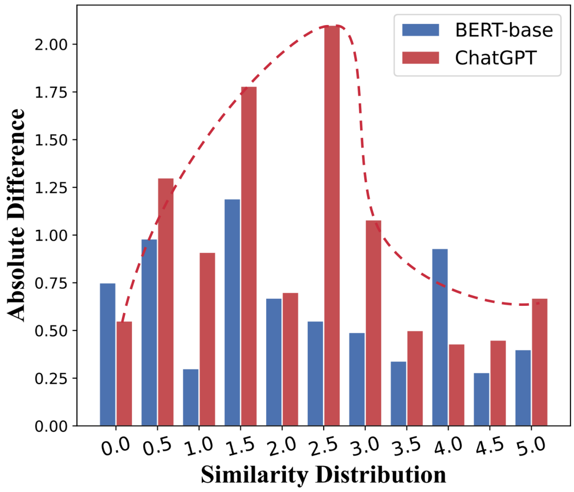

Figure 2: The comparison between BERT-base and ChatGPT on STS-B. The x-axis denotes the similarity distribution of STS-B, and the y-axis denotes the absolute difference between prediction and ground-truth.

GPT and other models on the paraphrase task, i.e., MRPC, in Table 4. Surprisingly, ChatGPT achieves comparable performance compared with BERT-base when evaluating on "entailment" samples, but there is a dramatic performance drop (up to 47% score) in the class of "not_entailment", where the sentences in the pair are not semantically equivalent. This indicates that ChatGPT is not sensitive to the semantic difference between a pair of sentences, which might be related to lack of human feedback on this aspect during model training.

Similarity Task. Since the STS-B is a regression task, we choose some samples from the uniform similarity distribution, ranging from 0 for no meaning overlap to 5 for meaning equivalence, and show the absolute difference between predictions and ground-truths for ChatGPT and BERT-base, respectively. As seen in Figure 2, ChatGPT underperforms BERT-base in most cases, as it generally predicts far from the ground-truths. To be more specific, we can observe that ChatGPT performs worse when the sentences in the pair have a lower similarity (<2.5 score), which is similar to the observation from Table 4. It can also be found that, ChatGPT is difficult to accurately predict the similarity score for a pair of sentences around the decision boundary (around the 2.5 score). One of the reasons is ChatGPT is not fine-tuned on the STS-B task and cannot determine a correct decision boundary.

3 Case Study

In this section, we show some bad cases of ChatGPT to explore its potential limitations, and attempt to explain why ChatGPT falls short in handling the negative samples of paraphrase task.

First, while ChatGPT works well for the inference task, it still fails to make the correct predictions in some cases. As seen in Figure 3, ChatGPT can generate fluent responses to both inquiries due to its powerful generation ability. However, we observe that these responses are somewhat contradictory and even unreasonable. For example, in the upper case, ChatGPT says "`...Jane was hungry and that this was the reason for giving`

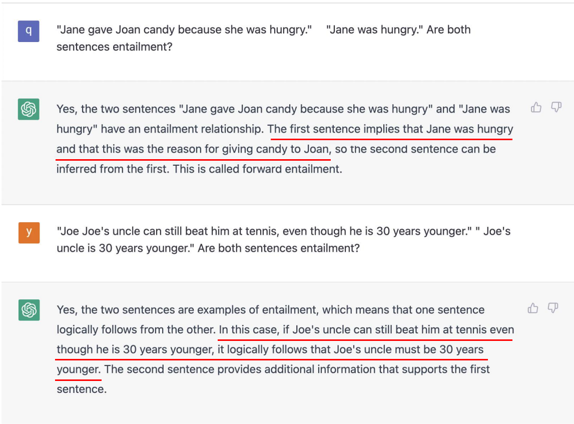

Figure 3: Failures of ChatGPT in inference task. The ground-truth for both cases is "not_entailment", but ChatGPT makes the "entailment" predictions. (Data: 2022.02.09)

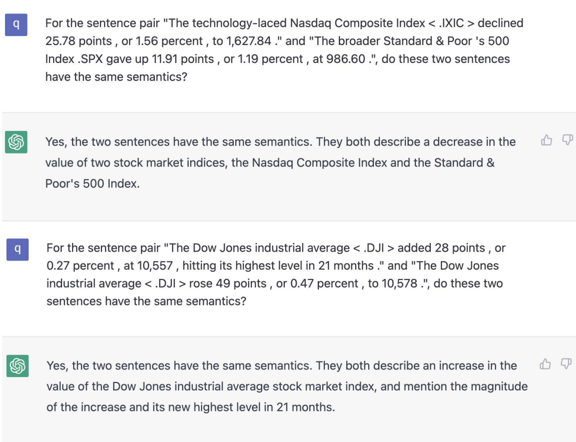

Figure 4: Failures of ChatGPT in paraphrase task. The ground-truth for both cases is "not_entailment", but ChatGPT makes the "entailment" predictions. (Data: 2022.02.09)

`candy to Joan,...`", which is very confusing. If Jane was indeed hungry, Jane would not give candy to Joan, but eat the candy himself (herself). There is a similar phenomenon in the lower case, where ChatGPT answers with confused logic. In general, ChatGPT is able to generate fluent responses following a certain pattern, but appears to have limitations in really reasoning the sentences. One evidence is that ChatGPT even fails to answer the questions, such as the cases in Figure 3, that are easily answered by humans.

On the other hand, some example failures of ChatGPT in the paraphrase task are shown in Figure 4. Both cases are in the class of "not_entailment", i.e., the sentences are not semantically equivalent. ChatGPT thinks the two sentences have the same semantics, as both sentences describe a decrease (increase) in the value, which can be viewed as a coarse-grained semantic similarity. However, we can easily find that the major difference between the two sentences is the value difference, determining the "not_entailment" polarity of these cases. We refer to this value difference as the fine-grained semantic difference. These cases show that such a discrepancy between coarse-grained and fine-grained semantic information might be one of the reasons why ChatGPT struggles with handling negative samples in the paraphrase task. This also indicates that strengthening the ability of ChatGPT to extract fine-grained semantic information would effectively improve its performance on the paraphrase and similarity tasks.

4 Conclusion

In this report, we empirically investigate the understanding ability of ChatGPT on a diversity of natural language understanding tasks. Through a series of quantitative studies, we find that ChatGPT works well on inference tasks, but falls short in handling paraphrase and similarity tasks, especially for the negative instances. Overall, ChatGPT attains a comparable understanding ability compared with fine-tuned BERT-base, but still underperforms the other powerful BERT-style models, such as RoBERTa-large, by a clear margin. We hope our study could facilitate more research on how to address the limitations and improve the understanding performance of ChatGPT, such as leveraging prompt engineering (e.g., chain-of-thought (Wei et al., 2022b) and its variants (Wang et al., 2022; Kojima et al., 2022)) to exploit its capabilities.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*.

Daniel Cer, Mona Diab, Eneko E Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *IWP*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *ACL-PASCAL*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *TACL*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *TMLR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2022a. E2s2: Encoding-enhanced sequence-to-sequence pretraining for language understanding and generation. *arXiv preprint*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2022b. Panda: Prompt transfer meets knowledge distillation for efficient model adaptation. *arXiv preprint*.

Qihuang Zhong, Liang Ding, Keqin Peng, Juhua Liu, Bo Du, Yibing Zhan, and Dacheng Tao. 2023. Bag of tricks for effective language model pretraining and downstream adaptation: A case study on glue. *arXiv preprint*.

Qihuang Zhong, Liang Ding, Li Shen, Peng Mi, Juhua Liu, Bo Du, and Dacheng Tao. 2022c. Improving sharpness-aware minimization with fisher mask for better generalization on language models. *arXiv preprint*.