Evaluating ChatGPT: Strengths and Limitations in NLP Problem Solving

Yanyi Wu*

Department of Mathematics, Brandeis University, Waltham, MA, United States

*Corresponding author: yanyiwu@brandeis.edu

Abstract. This paper critically analyzes ChatGPT's problem-solving performance on a range of natural language processing (NLP) tasks. Using a comparative methodology, it compares ChatGPT's performance with that of its predecessor, GPT-3.5, in seven different domains: summarization, named entity recognition, arithmetic, natural language inference, symbolic and logical reasoning, question answering, conversation, and arithmetic. The process entails a methodical assessment highlighting ChatGPT's replies' quantitative and qualitative elements. The findings show that although ChatGPT performs very well on math and question-answering tasks, it struggles with summarization and commonsense reasoning. The conversation sheds light on the subtleties of these findings while considering the applications and development of AI. The article concludes that although ChatGPT is a significant progress in natural language processing, its uneven problem-solving performance highlights the need for continued development and optimization of artificial intelligence models. This study aids in understanding the state and promise of AI-driven language models in challenging problem-solving situations.

Keywords: ChatGPT; GPT-3.5; natural language processing; zero-shot learning.

1. Introduction

Within these two years, more complex and adaptable language models have been created because of the rapid development of artificial intelligence (AI) and natural language processing (NLP) [1]. By utilizing patterns and structures collected in previously gathered data, a kind of pre-trained transformer for generative processing called "ChatGPT" is created to produce new data. When data scientists examine the connection between data science and ChatGPT in more detail, they see that data science is a valuable framework essential to ChatGPT's development. Big data analytics is embodied by ChatGPT, which uses advanced data science approaches to handle and learn from large datasets. In this sense, machine learning (ML) and natural language processing—two subfields of data science—play a critical role in gathering and storing enormous volumes of textual data and processing complex patterns to understand and emulate human language. They strive to give ChatGPT the fantastic capacity to provide replies that are both coherent and contextually appropriate [2]. Large language models (LLM), one of the newest machine learning (ML) models, have been associated with pioneering AI-powered language analysis. LLMs are well-developed AI programs based on a great deal of textual training. These models can produce language similar to that of a human being and carry out various language tasks, including sentiment analysis, question answering, etc. These models, including the "Generative Pre-trained Transformer (GPT)" and the "Bidirectional Encoder Representations from Transformers (BERT)" and their chat-based program (ChatGPT), have received much attention recently and produced previously unattainable degrees of precision in the broader scope of language analysis [3].

Recently, the NLP community has expressed an amount of interest in OpenAI's ChatGPT LLM. Similar to InstructGPT, ChatGPT was developed by "using reinforcement learning from human feedback (RLHF) to train the GPT-3.5 series model" [4]. RLHF consists of three main parts: using supervised learning to train a language model, collecting human preference-based comparison data to create a reward model, and using reinforcement learning to optimize the language model about the reward model [5]. ChatGPT has demonstrated remarkable capabilities during RLHF training,

including producing high-quality replies to human input, rejecting improper queries, and adjusting earlier mistakes in light of subsequent interactions [6].

The integration of computer methods with linguistics, which attempts to close the gap between machine comprehension and human communication, is at the core of natural language processing (NLP). With the advancement of machine learning, NLP has significantly increased in power, especially when it comes to more advanced models for learning, like neural networks. These algorithms can pick up on intricate linguistic subtleties and patterns because they have been trained on enormous databases of human language. As a result, NLP applications—which now include sentiment analysis, chatbots like ChatGPT, speech recognition, and language translation—have grown more complex. Noticeably, one well-known use of NLP is sentiment analysis, which is the process of examining text to ascertain the sentiment that lies behind it. It involves identifying neutral, negative, or positive emotions in spoken or written words. This method is frequently applied to comprehend customer perspectives in reviews, feedback, and social media. Businesses use sentiment analysis to track brand reputation and determine what the public thinks about their goods and services. Sentiment analysis systems can correctly detect and categorize the sentiments expressed in large volumes of unstructured text data by utilizing machine learning techniques.

In this paper, the purpose is to evaluate the performance of ChatGPT in dealing with NLP tasks in different ways, which include arithmetic tasks, commonsense and symbolic and logical reasoning tasks, natural language inference, question answering, dialogue, summarization, named entity recognition, and sentiment analysis. The evaluation's two goals were to: (i) find out if ChatGPT could handle these problems and it is accuracy in solving them; and (ii) compare ChatGPT's ability to solve problems with other machine intelligence like GPT-3.5. The overall goals of this study were to evaluate ChatGPT's potential as an intelligent tool for problem-solving in different datasets and investigate the highest degree it can achieve [7].

2. Methods

2.1. Seven Representative Tasks

In this experiment, researchers evaluate seven representative tasks: arithmetic (including MultiArith, GSM8K, Addsub, AQuA, SingleEq, and SVAMP), commonsense (including CSQA, strategyQA, and COPA), symbolic (including Letter Concatenation Operation and Coin Tossing, logical (Date understanding and tracking shuffled objects), natural language interface (RTE and CB), question answering (BoolQ), dialogue (Mutual), summarization (SAMSum), named entity recognition (CoNLL03), and sentiment analysis (SST2) [8]. These datasets have four types of reasoning tasks: arithmetic, commonsense, symbolic, and logical reasoning. Table 1 displays the data from several databases. If the labels are available for assessment, it automatically utilizes the test split for all datasets. For CommonsenseQA and COPA, a validation split is used. Researchers employ the question-only set (open-domain setting) from the BIG-bench collaboration [9] for StrategyQA by Wei [10].

2.2. Zero-shot Learning

Before entering the methodology parts, zero-shot learning needs to be introduced. Solving unknown tasks without labeled training examples is the goal of zero-shot learning. This presents a significant obstacle because models usually rely on enormous volumes of training data. Zero-shot learning approaches previously used can be roughly categorized into two types: (i) model-based approaches for learning a model for unseen samples directly [11] and (ii) concrete-example-based approaches that seek to "obtain labeled instances for unseen tasks in order to enhance model learning" [12]. LLMs are better at zero-shot learning, as evidenced by more recent research [13]. The most current innovation in LLMs is ChatGPT, which has demonstrated incredible aptitude in several dialogue-related areas. In this study, researchers take a step further and investigate ChatGPT's zero-shot learning capabilities on several tasks other than discussion [14].

As stated above, the primary comparison is between GPT-3.5 and ChatGPT regarding their zero-shot learning performance on various tasks. In order to solve the test problem, researchers use the function $Y = f(P, X)$ as the model for the mapping between these variables. Given an instruction P and test question X as input, this function produces a target text Y. For zero-shot-CoT in this experiment, researchers employ the identical two-stage prompting as Kojima et al. to produce the general procedures and then request the final output [15].

2.3. Chain-of-Thought Prompting

Before responding to the questions, LLMs are prompted to develop intermediate reasoning stages using the Chain-of-Thought (CoT) method [10]. Kojima et al. [15] used self-generated rationales in zero-shot CoT to show that LLMs are reasonable zero-shot reasoners. By introducing visual elements into CoT reasoning, Zhang et al. have more recently presented multimodal-CoT [16]. The model with less than one billion parameters surpassed human performance on the ScienceQA datasets, exceeding GPT-3.5 by 16% [17].

3. Results and Discussion

3.1. Arithmetic Reasoning

Table 1 displays ChatGPT and GPT-3.5 performance accuracy on six arithmetic reasoning datasets, with the best result either with or without chain-of-thought (CoT). In five of the six datasets without CoT, ChatGPT outperforms GPT-3.5 better, indicating its superior arithmetic reasoning skills. In Figure 1, there is a concrete instance of the performance of ChatGPT and GPT-3. ChatGPT correctly interprets the changes in the question, yielding the correct response of "74 lives." However, GPT-3.5 yields an incorrect result due to an elementary calculation error, although its tracking is correct. In sum, ChatGPT performs better than GPT-3.5 in arithmetic reasoning tasks.

Q: Wendy had 43 lives when she was playing a video game. She lost eight lives at a challenging section of the game. How many lives would she have if she received 39 more on the following level?
ChatGPT Response: 74 lives
GPT-3.5 Response: 73 lives

Fig. 1 An easy math question where ChatGPT get correct, but GPT-3.5 does not.

Table 1. Accuracy Performance of ChatGPT and GPT-3.5

Model	MultiArith	GSM8K	AddSub	AQUA-RAT	Single Eq	SVAMP	Letter
ChatGPT	95.8	78.9	88.6	53.5	91.5	77.5	70.2
GPT-3.5	83.7	59.5	87.3	40.6	86.4	73.6	54.4
Fine-tuning	96.2	63.1	93.9	45.3	93.1	79	99.4
	Coin	Date	Object	CSQA	StrategyQA	COPA	RTE
ChatGPT	65.8	72.6	58.7	73.7	61.1	82	85.9
GPT-3.5	97.8	77	39.7	74.9	61.1	93	80.1
Fine-tuning	100	65.3	23.9	82.3	77.8	95	95.8
	CB	BoolQ	Mutual	SAMsum	CoNLL03	SST2	
ChatGPT	89.3	87.3	76.2	31	53.2	93.7	
GPT-3.5	83.9	84.7	75.2	32.4	53.5	88.8	
Fine-tuning	100	91.2	93.5	40.5	94.6	97.5	

3.2. Commonsense, Symbolic and Logical Reasoning

Table 2 presents the zero-shot performance of ChatGPT with and without CoT and the few-shot performance of other models, including GPT-3.5 below.

First, applying CoT might not necessarily result in better performance in an experiment requiring commonsense thinking. The investigation of Kojima et al. indicates that the CoT technique often produces flexible and acceptable rationales in commonsense reasoning problems, but the final prediction needs to be corrected. The findings suggest that more detailed prior information may be needed for commonsense reasoning tasks, while this problem may be lessened by reducing model size [15]. Second, in contrast to mathematical, logical datasets, the performance of ChatGPT is worse than GPT-3.5 with or without CoT under every circumstance, suggesting that GPT-3.5 has superior matching skills [16, 17].

Table 2. Accuracy performance of ChatGPT compared to GPT-3.5 in commonsense

	CSQA		StrategyQA		COPA	
	No CoT	CoT	No CoT	CoT	No CoT	CoT
Zero-shot performance						
ChatGPT	75	71.5	61.1	61.1	93	85
Few-shot performance						
GPT-3.5 variants	82	79.9	68.6	77.8	95	-

3.3. Natural Language Inference

Table 1 displays the outcomes of several models on the RTE and CB natural language inference tasks. From the RTE and CB part, ChatGPT performs significantly better in the zero-shot situation than GPT-3.5, illustrating ChatGPT's better zero-shot sentence relation inference capability.

3.4. Question Answering

Table 1 shows that ChatGPT outperforms GPT-3.5 by almost 3% in the BoolQ datasets, similar to the previous findings. As Clark et al. demonstrated, answering the BoolQ question sets necessitates challenging "entailment-like" reasoning. As a result, ChatGPT is more capable of handling activities that require thinking [18].

The accuracy performance of GPT-3.5 and ChatGPT is displayed in Table 3. ChatGPT performs much better in the "Yes" class than GPT-3.5, suggesting it prefers to handle factual data. Additionally, ChatGPT usually creates ambiguous replies, such as "It is unclear," even though we expect it to output "Yes" or "No" via job instructions. Since sometimes ChatGPT does not output the exact answer to meet the requirement, ChatGPT's performance is even worse than previously.

Table 3. In each “Yes” and “No” situation, the performance of ChatGPT and GPT-3.5 on BoolQ datasets

	ChatGPT	GPT-3.5
Yes	88.9(+7.8)	81.1
No	84.6(-6.0)	90.6

3.5. Dialogue

As anticipated, ChatGPT outperforms GPT-3.5, which aligns with ChatGPT's remarkable discussion capability that the community has already noted.

Figure 2 provides a specific illustration of a situation where GPT-3.5 is having trouble, and ChatGPT provides accurate replies. ChatGPT can provide pertinent information while reasoning more effectively in the current situation. This demonstrates ChatGPT's superior thinking power yet again.

Q: f: Father, will you please give Jerry and me some cash? M: Why? f: We would want to purchase rock concert tickets. We'll give it back to you in a few days. M: What is the cost of these tickets? f: $18.00 per unit. M: Eighteen dollars? It is not a tiny amount. Which option is the right one? Respond Options: (A) f: in actuality, a total of $18. (B) f: Indeed, the total cost of two tickets was $44. (C) f: the total is $36. Nonetheless, there is a $2 discount for two tickets, for a total of $36. (D) f: indeed, a total of $36.
A: For items A through D, the response is
ChatGPT Response: D
GPT-3.5 Response: C

Fig. 2 A concrete example ChatGPT answers correctly, butGPT-3.5 does no.

3.6. Summarization

Table 1 reports the scores of GPT-3.5 and ChatGPT on the SAMSum task for the summarization part. Surprisingly, from the table, GPT-3.5 outperforms ChatGPT in every measure. This exciting phenomenon needs deep investigation.

3.7. Named Entity Recognition

Table 1 reports the best result of the zero-shot performance of ChatGPT and GPT-3.5 in CoNLL03 datasets for named entity recognition. Table 4 shows that the overall performance of ChatGPT and GPT-3.5 differ little. Compared to fine-tuning results, both fail to achieve a satisfactory result since the accuracy is significantly different. Especially for Misc, ChatGPT, and GPT-3.5 underperform the fine-tuned result, indicating both need to improve at dealing with named entity recognition tasks.

Table 4. Different models of CoNLL03 datasets. “Org” is an organization, and “Misc” is a miscellaneous entity.

	ChatGPT	GPT-3.5	Fine-tuned
All	53.2	53.5	94.6
Misc	4.1	4.8	83
Org	51.4	50	91.9

3.8. Sentiment Analysis

From Table 1, compared to GPT-3.5, ChatGPT performs significantly better. Table 5 calculates the per-class accuracy of both positive and negative models to investigate why ChatGPT performs better than GPT-3.5. It is evident that Chat-GPT's performance varies among courses. On negative samples, it performs significantly better than GPT-3.5, indicating its ability to detect negative instances within a text, but on positively labeled data, the data is similar. However, both underperform the fine-tuned methods. The reason might be that it judges a positive or negative stance to a neutral circumstance [19] or does not provide an exact yes or no answer instead of claiming “unclear.”

Table 5. Different models on per-class accuracy on SST2 datasets

	ChatGPT	GPT-3.5	Fine-tuned
All	93.7	88.8	97.5
Positive	90.8	88.1	-
Negative	96.7	89.5	-

4. Conclusion

The zero-shot learning abilities of ChatGPT on various representative tasks, including different datasets, are tested in this experiment. Comprehensive study and experimental findings showed ChatGPT's performance in many NLP tasks and its present drawbacks. The data shows that ChatGPT is good at solving arithmetic questions, and its ability can even be compared to fine-tuning results. However, ChatGPT still needs to be improved to deal with summarization and commonsense tasks. Compared to fine-tuning results, the performance of ChatGPT on commonsense, natural language inference, symbolic tasks, question-answering, dialogue, summarization, named entity recognition, and sentiment analysis all underperform from fine-tuning one even at a large margin, indicating that ChatGPT still has a large extent to improve the model. This work may inspire future research, such as using ChatGPT's reasoning and conversational skills in NLP tasks and correcting generalist models' shortcomings in situations where they are currently unsatisfactory.

References

[1] Ray, Partha Pratim. ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope. Internet of Things and Cyber-Physical Systems, 2023, 121–154.

[2] Molly, Ruby. How ChatGPT Works: The Model Behind the Bot. Towards Data Science, Working paper, 2023.

[3] Devlin, Jacob, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Cornell University, 2018.

[4] Christiano, Paul F, et al. Deep Reinforcement Learning From Human Preferences. Neural Information Processing Systems, 2017, 4299–4307.

[5] Ouyang Long, et al. Training Language Models to Follow Instructions With Human Feedback. Working paper, 2022.

[6] Guo Biyang, et al. How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. Cornell University, 2023.

[7] Orrù Graziella, et al. Human-like Problem-solving Abilities in Large Language Models Using ChatGPT. Frontiers in Artificial Intelligence, 2023.

[8] Qin Chengwei, et al. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? arXiv Cornell University, 2023.

[9] BIG-bench collaboration BIG-bench collaboration. 2021. Beyond the imitation game: Measuring and extrapolating the capabilities of language models, 2022. https://github.com/google/BIG-bench.

[10] Wei Jason, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Cornell University, 2022.

[11] Xiaolong Wang, Yufei Ye, Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 6857–6866.

[12] Zhang Li, et al. Learning a Deep Embedding Model for Zero-Shot Learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2021–2030.

[13] Brown T B, et al. Language Models Are Few-Shot Learners. Cornell University, 2020.

[14] Qin Chengwei, et al. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? Cornell University, 2023.

[15] Kojima Takeshi, et al. Large Language Models Are Zero-Shot Reasoners. Cornell University, 2022.

[16] Zhang Zhuosheng, et al. Multimodal Chain-of-Thought Reasoning in Language Models. Cornell University, 2023.

[17] Lü Ping, et al. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. Cornell University, 2022.

[18] Clark Christopher, et al. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. Cornell University, 2019.

[19] Kocoń Jan, et al. ChatGPT: Jack of All Trades, Master of None. Information Fusion, 2023.