Why Does ChatGPT Fall Short in Providing Truthful Answers?

Shen Zheng* **Jie Huang*** **Kevin Chen-Chuan Chang**
University of Illinois at Urbana-Champaign, USA
{shenz2, jeffhj, kcchang}@illinois.edu

Abstract

Recent advancements in large language models, such as ChatGPT, have demonstrated significant potential to impact various aspects of human life. However, ChatGPT still faces challenges in providing reliable and accurate answers to user questions. To better understand the model's particular weaknesses in providing truthful answers, we embark an in-depth exploration of open-domain question answering. Specifically, we undertake a detailed examination of ChatGPT's failures, categorized into: *comprehension*,*factuality*, *specificity*, and *inference*. We further pinpoint *factuality* as the most contributing failure and identify two critical abilities associated with factuality: *knowledge memorization* and *knowledge recall*. Through experiments focusing on factuality, we propose several potential enhancement strategies. Our findings suggest that augmenting the model with granular external knowledge and cues for knowledge recall can enhance the model's factuality in answering questions.

1 Introduction

ChatGPT/GPT-4 [OpenAI, 2022, 2023] has gained substantial recognition for its practical applications, providing useful and informative responses to a wide range of queries. Recent studies have conducted comprehensive technical evaluations of ChatGPT on numerous NLP tasks, demonstrating that ChatGPT outperforms other models across various tasks [Bubeck et al., 2023, Bang et al., 2023, Jiao et al., 2023, Qin et al., 2023].

However, in spite of the impressive capabilities exhibited by ChatGPT, researchers have highlighted some challenges of ChatGPT, such as its inability to perform reliable reasoning [Bang et al., 2023], translate low-resource languages effectively [Jiao et al., 2023], solve complex mathematical problems [Frieder et al., 2023], and provide accurate information [Bang et al., 2023]. While these shortcomings are documented, the specific limitations of ChatGPT that contribute to these challenges are not entirely clear in the existing literature. Taking question answering as a representative example: Is the model's failure due to its inability to reason or a lack of knowledge to answer the question? Is the issue a result of insufficient knowledge, or does the model struggle to recall the internal knowledge with the question? Is the difficulty in recalling knowledge the root cause, or does the model have trouble understanding the question's context or intent?

In this study, we delve into an in-depth exploration of the weakness of ChatGPT in the context of complex open-domain question answering, as this task aligns closely with users' everyday search demands and requires extensive knowledge as well as robust understanding and reasoning capabilities. Our goal is to identify common failure modes of ChatGPT in providing truthful answers, pinpoint the specific abilities in which ChatGPT is deficient that contribute to these failures, and consider potential strategies for improvement.

*Equal contribution.

ICBINB Workshop at NeurIPS 2023.

To this end, we first employ a thematic analysis to analyze instances of ChatGPT's failures and categorize them into four primary error types: *comprehension error*, *factuality error*, *specificity error*, and *inference error*. We then pinpoint the *factuality* deficiency as the primary failure and identify *knowledge memorization* and *knowledge recall* as critical abilities for answering questions with *factuality*. Furthermore, we propose several potential strategies to help mitigate these deficiencies. Our results indicate that ChatGPT's *factuality* can be enhanced by supplying granular external knowledge and cues for knowledge recall. Our findings provide practical insights for developing more reliable question answering systems.

2 Related Work

A substantial body of research has been conducted on examining various aspects of ChatGPT, including its general evaluation [Bang et al., 2023, Qin et al., 2023, Kocoń et al., 2023], understanding abilities [Zhong et al., 2023], mathematical abilities [Frieder et al., 2023], bug fixing performance [Sobania et al., 2023], out-of-distribution (OOD) behaviors [Wang et al., 2023], translation behaviors [Jiao et al., 2023], and question answering performance [Guo et al., 2023, Tan et al., 2023]. Although ChatGPT showcased compelling performance, the research community has surfaced several issues concerning its reasoning [Borji, 2023, Bang et al., 2023], factual accuracy [Bang et al., 2023], solving complex mathematical problems [Frieder et al., 2023, Borji, 2023] and ethical implications [Zhuo et al., 2023, Borji, 2023, Ray, 2023]. However, these studies predominantly concentrate on the categorization and identification of common problems, with limited in-depth investigation into the underlying deficiencies that contribute to the failures. In this work, we identify common failures in question answering scenarios, delve into the fundamental ability shortcomings that lead to these errors, and proposes potential strategies to mitigate these failures based on our experimental insights.

3 Models and Datasets

We focus on complex open-domain question answering, using two widely-used benchmark datasets: HotpotQA [Yang et al., 2018] and BoolQ [Clark et al., 2019], both of which use Wikipedia as their knowledge source. We selected 200 questions from HotpotQA for analyzing the errors made by ChatGPTs. For factuality evaluation, we sampled an additional 500 questions from HotpotQA and 1000 questions from BoolQ. We evaluated the performances of both GPT-3.5 and GPT-4 using these datasets. To generate responses from GPT-3.5 and GPT-4, we utilized the public OpenAI API.[1]

4 ChatGPT's Failures

4.1 Thematic Analysis

We examined the model's responses to 200 HotpotQA samples using thematic analysis [Braun and Clarke, 2012], a method for identifying patterns or "themes" within data. The process starts by extracting preliminary "codes" from data, which are later assembled into broader themes. For the purpose of ensuring a rigorous and comprehensive thematic analysis, a two-annotator approach was employed. We asked two independent annotators, both proficient in the subject and experienced in qualitative analysis to independently review the data set. Initially, they worked separately to identify the codes, analyze, and report patterns (themes) within the data. Upon completion of their independent analyses, the annotators convened in a collaborative session to compare, discuss, and reconcile any discrepancies in their identified themes. After the inter-annotator discussion, we grouped them several themes and validated these themes with an extra 40 questions. We confirm a complete representation of errors can be achieved through these themes.

To quantify the degree of consistency between the annotators, we employed Cohen's Kappa κ [McHugh, 2012]. This statistical measure is used to evaluate the reliability of agreement between two raters beyond what is expected by chance. The formula for Cohen's Kappa is given by:

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)}$$

[1] `https://openai.com/blog/openai-api`. We use the model gpt-3.5-0301 and gpt-4-0314.

Table 1: Examples of four types of errors (text in brackets represents the groundtruth, text in blue indicates the entities compared and their information, text in red indicates where the model (GPT-3.5) makes mistakes).

Type	Question and GroundTruth	Responses
Comprehension	What is a rare breed of dog that was derived as a variant of Rat Terrier, Shiloh Shepherd dog or American Hairless Terrier? *(American Hairless Terrier)*	One rare breed of dog that was derived as a variant of Rat Terrier, Shiloh Shepherd dog or American Hairless Terrier is the Teddy Roosevelt Terrier.
Factuality	Are the New Orleans Outfall Canals the same length as the Augusta Canal? *(yes)*	No... the Orleans Canal is approximately 22 miles (35 kilometers) long ... The Augusta Canal ... is approximately 7 miles (11.3 kilometers) long
Specificity	1. What profession do Kçbç Abe and Agatha Christie share? *(playwright)* 2. What genre do Superheaven and Oceansize belong to? *(rock)*	1. Author, ... 2. Superheaven and Oceansize are not of the same genre. Superheaven ... play a style of alternative rock ... Oceansize ... was a British progressive rock band ...
Inference	1. Which band has more members, Muse or The Raconteurs? *(The Raconteurs)* 2. Which is currently more valuable, Temagami-Lorrain Mine or Meadowbank Gold Mine? *(Meadowbank Gold Mine)*	1. Muse has more members than The Raconteurs. Muse is a British rock band with three members ... The Raconteurs ... Raconteurs has four members. 2. I cannot provide the current valuation ... Meadowbank Gold Mine... was producing gold at a rate of approximately 220,000 ounces per year. On the other hand, Temagami-Lorrain Mine is a historic iron mine located in Ontario, Canada that has been inactive for many years.

In this formula, p_o represents the observed proportion of agreement, p_e represents the expected proportion of agreement.

In our study, the computed inter-coder agreement was 0.8394, which suggests a high degree of consensus between the annotators. We finalize the four identified themes as: problem comprehension and intent, factual correctness, specificity level, and reasoning, referred to as *comprehension error*, *factuality error*, *specificity error*, and *inference error*.

Comprehension errors refer to the failure in comprehending the question context and intention. In our experiment, the model demonstrates proficiency in comprehending the question, but it exhibits shortcomings when faced with questions containing grammar mistakes or ambiguity. For instance, the question shown in the Comprehension row of Table 1 presents a challenge to the model due to an incorrect interrogative pronoun "what", which should be "which". Consequently, the model fails to recognize that the question is seeking a selection between the two items marked in blue and instead misinterprets it as a choice among all last three items.

A *factuality error* occurs when a model lacks the necessary supporting facts to produce an accurate answer [Petroni et al., 2019, Lee et al., 2023]. This may be due to the model's lacking knowledge of a particular entity, attribute, or event. The example in the Factuality row of Table 1 shows a mistake when the model has the incorrect knowledge about the length of two canals. While this type of error is straightforward, it accounts for a majority of errors in the model.

Specificity errors occur when the model fails to answer a problem at the appropriate level of specificity [Adiwardana et al., 2020, Huang et al., 2023a]. This can manifest as the model providing an answer that is too general or too specific. For instance, in the specificity row of Table 1, for the first question, the correct answer is playwright, but if the model answers with author, it is not specific enough. Similarly, in the second question, the ground truth is rock, but the model predicts Superheaven as alternative rock and Oceansize as progressive rock, and claims that they do not share the same genre. It is also classified as specificity error.

An *inference error* occurs when a model possesses the necessary knowledge to answer a question, but fails to reason with the facts effectively to arrive at the correct answer [Huang and Chang, 2023, Wei et al., 2023]. For instance, for the first question in the inference row of Table 1, the model may know that Muse has three members and The Raconteurs has four members, but still incorrectly claims that Muse has more members. Additionally, the model may fail to make predictions based on

Table 2: Number of failures under different settings.

	#Correct	#Wrong	Comprehension	Factuality	Specificity	Inference
GPT-3.5	130	70	11	46	6	7
GPT-3.5+evi	186	14	1	6	2	5
GPT-4	149	51	6	37	3	5

Table 3: Examples of questions and responses on factuality related abilities (The second column demonstrates the erroneous question and response, and third column illustrates our prompt to test the two abilities. Text in blue indicates the model's responses, and text in red is where the model make a mistake).

Abilities	Erroneous Responses	Test Conversation
Memorization	–> Are the New Orleans Outfall Canals the same length as the Augusta Canal?: –> No...the Orleans Canal is approximately 22 miles (35 kilometers) ...The Augusta Canal...is approximately 7 miles (11.3 kilometers) long	–> User (asking essential knowledge#1): What's the length of New Orleans Outfall Canals? –> Model: The total length of the New Orleans Outfall Canals is approximately 22.5 miles (36.2 kilometers). –> User (asking essential knowledge#2): What's the length of Augusta Canal? –> Model: The Augusta Canal is approximately 7 miles (11.3 kilometers) long.
Recall	–> Chang Ucchin was born in korea during a time that ended with the conclusion of what? –> I'm sorry, but I need more information to answer your question accurately.	–> User (asking essential knowledge#1): When was Chang Ucchin born in korea? –> Model: Chang Ucchin ... was born on March 26, 1891... –> User (asking original question): Chang Ucchin was born in korea during a time that ended with the conclusion of what? –> Model: Chang Ucchin was born in Korea during a time when Korea was under Japanese rule, ...ended with the conclusion of World War II in 1945.

commonsense. In the second question, although holding the knowledge that the Meadowbank Gold Mine is still producing gold and the Temagami-Lorrain Mine has been inactive for years, the model still fails to deduce that the former one is currently more valuable due to its ongoing production.

4.2 Results

We counted the frequency of errors across the four categories. We used GPT-3.5 as the baseline model by feeding it plain questions. To study the effect of providing external evidence and to investigate GPT-4's improvement, we also explored providing questions with gold evidence (accurate piece of information that provides a definitive answer to a question) to the GPT-3.5 model (GPT-3.5+evi) and plain questions to the GPT-4 model. The results are summarized in Table 2.

Based on our experiments, we made the following observations: 1) Nearly half of the failures are due to *factuality error*, followed by *inference error*, *comprehension error* and *specificity error*. 2) Providing evidence not only addresses *factuality* but also significantly mitigates *comprehension* and *specificity errors*. 3) The GPT-4 model demonstrates some improvements compared to GPT-3.5, particularly in addressing *comprehension* and *specificity errors*. However, *factuality* is only marginally improved. Our finding underlines *factuality* as the primary concern in open-domain QA for its dominance among errors, its impact on other error types, and GPT-4's inadequate improvement in this area.

5 Abilities Behind Factuality

As we pinpoint in Section 4.2 that *factuality* is the most critical concern in open-domain QA, we focus on investigating factuality. Drawing from cognitive science research on human memory organization in QA [leh, 1980], we treat *factuality error* as a cognitive failure in retrieving relevant knowledge from LLMs' memory for a question.

We define *essential knowledge* p with respect to a question q as *the knowledge indispensable for answering question* q and identify two key abilities for the knowledge retrieval process:

Definition 1 *(Knowledge Memorization) There exist an appropriate prompt* s *which, when fed into the model, will result in the essential knowledge* p*.*

Definition 2 *(Knowledge Recall) Given the question* q *as the prompt, the model is able to output the memorized essential knowledge* p*.*

Focusing on these two abilities, we conducted experiments with questions the model couldn't answer due to lack of knowledge. To test knowledge memorization, we rephrased essential knowledge as a question. E.g., in Table 3, we evaluated the model's memorization by asking about canal lengths. To further evaluate knowledge recall, we re-asked the original question in the same conversation. If the model answered accurately in this setting but not when only asked the original question, it indicated a recall issue. In the recall row of Table 3, the model correctly answers the query about Chang's birth event after a preceding related question, but fails without this context.

We find that 6 out of 46 errors stem from the recall issues, while the rest come from memorization. Our novel insight that distinguishes knowledge memorization and recall offers a new perspective for addressing knowledge-related problems.

6 Improving Factuality for QA

6.1 Settings

In our previous experiment, we demonstrated that a factuality error can stem from knowledge memorization or recall. To investigate how to mitigate it, we conducted experiments using the HotpotQA and BoolQ datasets. We used the plain question configuration with the GPT-3.5 model as our baseline. Since the GPT-3.5 and GPT-4 share similar underlying architectures and training processes, we experimented only on the former to draw observations. The prompts of our experiments are shown in Appendix A. For evaluation, we employed partial match [Mavi et al., 2022], which examines whether the ground truth is a substring of the prediction.

6.1.1 Knowledge Memorization

The limitations of models in memorizing facts has steered research towards retrieval-augmented language models, such as those augmented with external corpora [Izacard et al., 2022, Shi et al., 2023, Huang et al., 2023b] or search engines [Lazaridou et al., 2022, Komeili et al., 2022]. However, noise and non-essential information can compromise retrieval, e.g., notable methods retrieve the whole web pages from Bing search as knowledge [Komeili et al., 2022]. Hence, we refer to the retrieved knowledge as context information and define *granularity* as the length ratio of the context information to the essential knowledge. We investigate the role of *granularity* in performance across four different granularity settings.

- **Sentence level.** Directly provide external knowledge at the sentence level.
- **Passage level.** We offer gold evidence sentences along with other sentences related to the entities.
- **Section level.** We supply the Wikipedia section containing the gold evidence sentences.

6.1.2 Knowledge Recall

To mitigate knowledge recall issues, we considered the knowledge recall process as retrieving the values (essential knowledge) in LLMs' memory with the given keys (the plain question by default), and tested whether supplying entity-related keys aids the knowledge recall process. Based on the keys provided, we proposed the following settings:

- **Complete entity name.** We give the model complete Wikipedia names of the core entities in the question, e.g., for the film "Samson and Deliah", we provide "Samson and Deliah (1984 film)".
- **Definition sentences.** We provide the entity's initial Wikipedia sentences as the definition or background in addition to the entity names, ensuring no essential knowledge is present.

Table 4: Factuality experiments on HotpotQA and BoolQ.

	HotpotQA	BoolQ
Plain Question	0.58	0.71
External - Sentence	0.73	0.869
External - Passage	0.718	-
External - Section	0.603	0.77
Recall - Complete Entity Name	0.597	0.755
Recall - Wiki Background	0.646	0.789
Recall - Random Sentences	0.55	-

- **Random relevant sentences.** We also provide other random sentences from the entity's Wikipedia page along with entity names, again avoiding essential knowledge.

6.2 Observations and Implications

Finding I. Finer granularity of external knowledge yields better results. Table 4 (top) shows that external knowledge incorporation boosts performance, and effectiveness is influenced by knowledge granularity. Including essential knowledge with other sentences affects (although only slightly) results, while using the whole Wikipedia section with evidence greatly reduces the performance gain. This suggests that performance decreases with coarser input knowledge granularity.

Finding II. Providing relevant keys aids in recalling essential knowledge. Table 4 (bottom) indicates that supplying the model with keys can enhance accuracy. Supplying the complete entity names improves the performance, while providing entity background or definition sentences further aids knowledge recall, even without the essential knowledge. However, random sentences from the entities' Wikipedia page do not improve the performance, but instead decrease it.

Based on these findings, we explore strategies to boost the factuality in question answering from a LLM research perspective.

- **Provide external context with fine granularity as evidence to help memorize essential knowledge.** Although LLMs possess a vast amount of implicit knowledge, it is evident that there is still a significant amount of knowledge that is difficult to cover during training or challenging to recall during inference. Therefore, building an IR system to retrieve knowledge with finer granularity based on the question would be helpful according to our findings. Some attempts in this direction have been observed in systems such as New Bing[2], Bard[3], and ChatGPT plugins[4].
- **Provide descriptions of entities as keys to help recall essential knowledge.** In our analysis, we observe that sometimes ChatGPT indeed memorize the essential knowledge to solve the question, but cannot recall the knowledge with the question. A relevant finding is highlighted in the study by Huang et al. [2022], where it is observed that while language models can memorize a substantial number of email addresses, they struggle to associate specific email addresses with corresponding individual names. Improving the recall capabilities could be an intriguing direction for developing more powerful language models.

7 Conclusion

Our study explores ChatGPT's common errors in the context of truthful open-domain question answering, identifies four error types and pinpoints *factuality error* as the most critical error. We further define *essential knowledge*, and examine two crucial abilities *knowledge memorization* and *knowledge recall* associated with factuality. We study the impacts of evidence granularity on *knowledge memorization* and providing relevant keys on *knowledge recall*. We finally propose methods to improve ChatGPT's factuality in QA, contributing to the understanding of factuality and offering insights to enhance QA systems and language models, promoting more reliable LLMs.

[2] `https://www.bing.com/new`

[3] `https://bard.google.com`

[4] `https://openai.com/blog/chatgpt-plugins`

References

The process of question answering - a computer simulation of cognition. *American Journal of Computational Linguistics*, 6(3-4), 1980.

D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, et al. Towards a human-like open-domain chatbot. *ArXiv preprint*, abs/2001.09977, 2020.

Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023.

A. Borji. A categorical archive of chatgpt failures, 2023.

V. Braun and V. Clarke. *Thematic analysis.* American Psychological Association, 2012.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712, 2023.

C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300.

S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner. Mathematical capabilities of chatgpt, 2023.

B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023.

J. Huang and K. C.-C. Chang. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 2023.

J. Huang, H. Shao, and K. C.-C. Chang. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

J. Huang, K. C.-C. Chang, J. Xiong, and W.-m. Hwu. Can language models be specific? how? In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 2023a.

J. Huang, W. Ping, P. Xu, M. Shoeybi, K. C.-C. Chang, and B. Catanzaro. Raven: In-context learning with retrieval augmented encoder-decoder language models. *arXiv preprint arXiv:2308.07922*, 2023b.

G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv*, 2208, 2022.

W. Jiao, W. Wang, J. tse Huang, X. Wang, and Z. Tu. Is chatgpt a good translator? yes with gpt-4 as the engine, 2023.

J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, A. Kocoń, B. Koptyra, W. Mieleszczenko-Kowszewicz, P. Miłkowski, M. Oleksy, M. Piasecki, Łukasz Radliński, K. Wojtasik, S. Woźniak, and P. Kazienko. Chatgpt: Jack of all trades, master of none, 2023.

M. Komeili, K. Shuster, and J. Weston. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.579.

A. Lazaridou, E. Gribovskaya, W. Stokowiec, and N. Grigorev. Internet-augmented language models through few-shot prompting for open-domain question answering, 2022.

N. Lee, W. Ping, P. Xu, M. Patwary, P. Fung, M. Shoeybi, and B. Catanzaro. Factuality enhanced language models for open-ended text generation, 2023.

V. Mavi, A. Jangra, and A. Jatowt. A survey on multi-hop question answering and generation, 2022.

M. McHugh. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82, 2012. doi: 10.11613/BM.2012.031.

OpenAI. Chatgpt: Optimizing language models for dialogue. *OpenAI*, 2022.

OpenAI. Gpt-4 technical report, 2023.

F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250.

C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang. Is chatgpt a general-purpose natural language processing task solver?, 2023.

P. P. Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 2023.

W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, and W.-t. Yih. Replug: Retrieval-augmented black-box language models. *ArXiv preprint*, abs/2301.12652, 2023.

D. Sobania, M. Briesch, C. Hanna, and J. Petke. An analysis of the automatic bug fixing performance of chatgpt, 2023.

Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, and G. Qi. Evaluation of chatgpt as a question answering system for answering complex questions, 2023.

J. Wang, X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, H. Huang, W. Ye, X. Geng, B. Jiao, Y. Zhang, and X. Xie. On the robustness of chatgpt: An adversarial and out-of-distribution perspective, 2023.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259.

Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert, 2023.

T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing. Exploring ai ethics of chatgpt: A diagnostic analysis, 2023.

A Prompts

A.1 Knowledge Memorization

The prompts for all the settings in knowledge memorization are:

```
Use the following knowledge [entity1:  evidence1, entity2:  evidence2, ...],
answer the question:  [question].
```

A.2 Knowledge Recall

The prompt for complete entity name setting is:

```
Use the knowledge about [entity1, entity2, ...], answer the question:
[question].
```

The prompts for the other two settings (definition sentences and random relevant sentences) are:

```
Use the knowledge about [entity1, entity2, ...], and with the following
background knowledge [entity1:  evidence1, entity2:  evidence2, ...],
answer the question:  [question].
```