International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 11 | Nov - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

An Experiment Based Evaluation of Logical Reasoning Abilities of GPT-3.5 and GPT -4

Ansh Tiwari[1], Ashmit Dubey[2], Mahesh Kr Tiwari[3] , Rinku Raheja[4]

[1]National Post Graduate College, Department of Computer Science
[2] National Post Graduate College, Department of Computer Science
[3] National Post Graduate College, Department of Computer Science
[4] National Post Graduate College, Department of Computer Science

---

Abstract

Employing logical logic capability is a comprehensive natural language understanding bid. With the release of Generative Pretrained Transformer 4(GPT- 4), stressed as" advanced" at logic tasks, we're eager to learn the GPT- 4 performance on colorful logical logic tasks. This report analyses multiple logical logic datasets, with popular marks like LogiQA and ReClor, and recently- released datasets like AR- LSAT. We test themulti-choice reading appreciation and natural language conclusion tasks with marks taking logical logic. We further construct a logical logic out-ofdistribution dataset to probe the robustness of ChatGPT and GPT- 4. A comparison in terms of performance has also been been made between ChatGPT and GPT-4 . Trial results show that ChatGPT performs significantly better than the RoBERTa forfeiture- tuning system on utmost logical logic marks. With early access to the GPT- 4 API we're suitable to conduct violent trials on the GPT- 4 model. The results show GPT- 4 yields indeed higher performance on utmost logical logic datasets. Among marks, ChatGPT and GPT- 4 do fairly well on well- known datasets like LogiQA and ReClor. still, the performance drops significantly when handling recently released and out- of- distribution datasets. Logical logic remains grueling for ChatGPT and GPT- 4, especially on outof- distribution and natural language conclusion datasets. We release the prompt- style logical logic datasets as a standard suite and name it LogiEval.

Keywords: Generative Pretrained Transformer 4 (GPT-4), Natural Language Understanding (NLU), Multi-choice Reading Comprehension, Out-of-distribution Dataset, RoBERTa Fine-tuning.

1 Introduction

Logical logic is essential to mortal intelligence, and incorporating logical logic capacities into natural language understanding(NLU) systems has been an active exploration interest from the morning of artificial intelligence(Cresswell, 1973)(Kowalski, 1979)(Iwanska ´, 1993). Experimenters have been exploring colorful approaches to achieve this thing, including rule- grounded styles, emblematic systems(MacCartney and Manning, 2007a), fine-tuning large language models(Wang et al., 2018), and combining both neural and emblematic approaches(Li and Srikumar, 2019). In the traditional logical and semantic approach, computational linguists developed emblematic systems exercising First- Order- sense(FOL) or Natural sense(MacCartney and Manning, 2007a) to attack abecedarian conclusion tasks. Rule- grounded models struggle to unravel problems like the RTE challenge(Dagan et al., 2005) with hand- drafted rules and theorem provers. Formal sense logic espoused by early experimenters came up with emblematic systems and hand- drafted rules, where knowledge was represented explicitly using formal sense or other emblematic representations. With rules, the systems can reuse deduction operations. still, these approaches face challenges in handling nebulosity and scalability. They're brittle when dealing with real- world natural language

data. The period of neural network models sees the rise of large- scale NLI datasets as popular marks. For illustration, the SNLI(Bowman et al., 2015) and theMulti-genre NLI(MNLI)(Williams et al., 2018a) datasets are created through crowdsourcing, featuring an immense data size and broad content. They beget the development of models with better representation capacities and come the go- to standard for natural language understanding exploration. The giant vault in model performance comes with the arrival of Motor- grounded(Vaswani et al., 2017) language models like BERT(Devlin et al., 2018) when the training schemes of similar models enable them to pierce colossal unlabelled corpora. As a result, erecting language models with trillions of parameters come pos- sible(Brown et al., 2020)(Raffel et al., 2019). The paradigm ofpre-training and fine- tuning has since come the dominant result to textual conclusion tasks. Experimenters fine- tune language models on task-specific datasets afterpre-training models on massive textbook corpora. Largepre-trained language models(LMs) achieve beyond- human performances on popular NLI and MRC marks, prompting for further sophisticated marks in textual conclusion. NLP exploration on logical logic regains instigation with the recent releases of a sprinkle of datasets, specially LogiQA and Reclor. The datasets are collected from logical logic examinations similar as Chinese Civil menial Examinations and Law School Admission Test(LSAT). These tests are challenging indeed for humans and are golden- labeled data with good quality. Logical logic is exploited in numerous probing tasks over largePre-trained Language Models(PLMs) and downstream tasks like question- answering and dialogue systems. PLMs perform inadequately compared to traditional marks. Despite the progress made so far, achieving mortal- suchlike logical logic capabilities in NLU systems remains a grueling task. GenerativePre-trained Motor 4(GPT- 4)(OpenAI, 2023), as well as ChatGPT, is a newlyreleased language model developed by OpenAI, designed to understand and inducemulti-modal contents. GPT- 4 is promoted to retain indeed more important capabilities in tasks that bear logical logic. Logical logic is essential to mortal intelligence, enabling us to draw conclusions, make prognostications, and break problems grounded on given information. Incorporating logical logic into language models like GPT- 4 can revise natural language understanding(NLU) systems, making them more accurate, robust, and able of understanding complex information in natural language. The evaluation of ChatGPT and GPT- 4 for logical logic tasks explores their performance on several logical logic marks, detailing the strengths and limitations of ChatGPT and GPT- 4 in this sphere. former ChatGPT evaluation indicates that ChatGPT performs better on deducible logic than inductive bones

(Bang et al., 2023) and that ChatGPT performs better on factual inputs(Qin et al., 2023). After the release of GPT- 4, we will bandy two tasks for assessing its logical logic capacitiesmulti-choice reading appreciation and natural language conclusion. Both tasks are logic-heavy and serve as a playground for testing models' logic capacities. Several logical logic datasets have been released under the two tasks. These marks proved to be delicate to break for PLMs. We hope this report will exfoliate further light on the logical logic capability of ChatGPT and GPT- 4. Our benefactions are as follows:

1. We test ChatGPT and GPT- 4 on two logicalreasoning tasksmulti-choice reading appreciation and natural language conclusion. We conduct trials on multiple logical logic marks to dissect the logical logic capability of ChatGPT and GPT- 4.

2. We introduce LogiEval, a test suite for testing prompt- grounded large language models. We release the standard at https// github.com/csitfun/LogiEval

2. Our trials show that both ChatGPTand GPT- 4 are good at working well- known logical logic reading appreciation marks but struggle at handling out- of- distribution datasets. Their performances on natural language conclusion tasks taking logical logic still need perfecting.

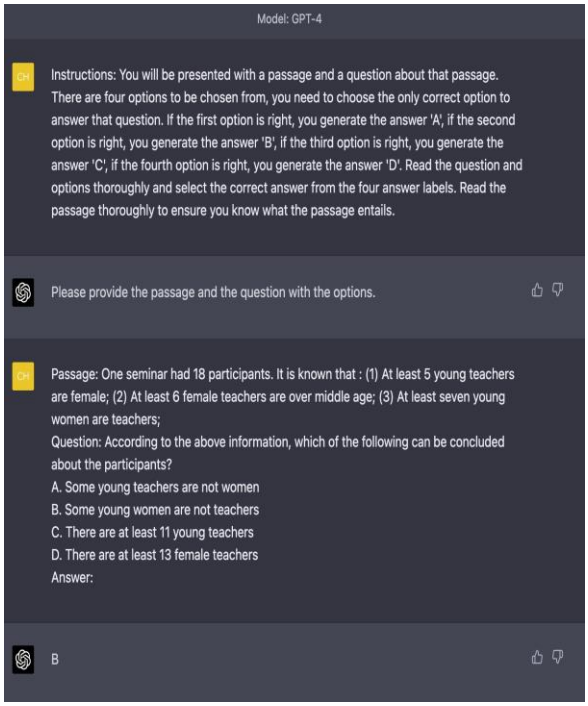

Figure 1: Instruction and question format for logical reading comprehension tasks.

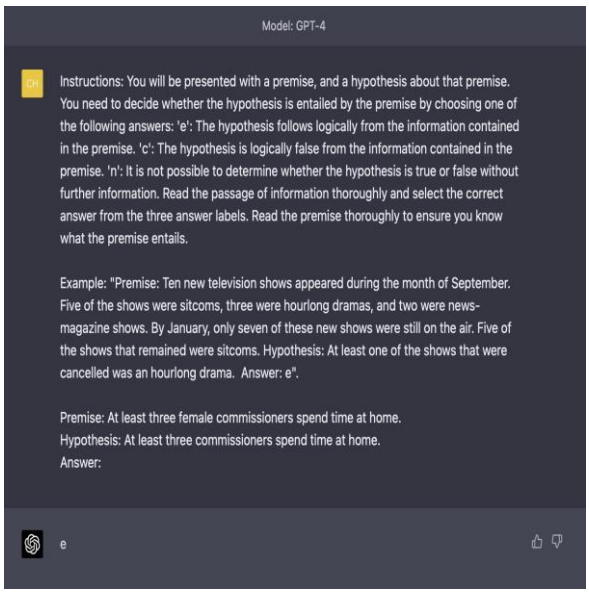

Figure 2: Instruction and question format for natural language inference tasks.

2 Evaluation Settings

We considermulti-choice reading appreciation and natural language conclusion tasks for our evaluation. Multi-choice reading appreciation is heavily tested on large language models for these tasks frequently have easily- formed and high- quality datasets. On the other hand, the natural language conclusion task is a abecedarian task for assessing logic capacities.

The datasets and the evaluation settings are handed as follows:

2.1 Datasets

2.1.1 Multi-choice Reading Appreciation

Machine reading is a popular task in NLP. In the typical multiple- choice task setting, given a passage and a question, a system is asked to elect the most applicable answer from a list of seeker answers.

LogiQA is a dataset specifically designed formulti-choice question- answering tasks that involve logical logic. The dataset is espoused from the Chinese Civil Service Examination, restated into English, and released in Chinese and English. The dataset has been streamlined to the 2.0 interpretation, where the data size has been enlarged. We choose the test sets of both the Chinese and English performances for our evaluation. Figure 3 shows an illustration from the LogiQA 2.0 test set.

ReClor is another logical logic dataset designed for reading appreciation tasks taking logical logic. It collects question- answering exemplifications from the LSAT examinations 2, which are targeted to testing mortal logical logic capacities. We use the development set for our testing because the test set does n't include gold markers.

Dataset	LogiQA 2.0 test	LogiQA 2.0 zh test	ReClor dev	AR-LSAT test	LogiQA 2.0 ood
Size	1572	1594	500	230	1354
Human avg.	86.00	88.00	63.00	56.00	83.00
human ceiling	95.00	96.00	100.00	91.00	99.00
RoBERTa	48.76	35.64	55.01	23.14	33.22
ChatGPT (API)	52.37	53.18	57.38	20.42	38.44
GPT-4 (Chat UI)	75.26(73/97)	51.76 (44/85)	92.00 (92/100)	18.27 (19/104)	48.21(54/112)
GPT-4 (API)	72.25	70.56	87.20	33.48	58.49

represents the out-of-distribution data of LogiQA 2.0.

Table 1:ChatGPT and GPT-4 performance on the Logical multi-choice machine reading comprehension task
(accuracy %). “LogiQA 2.0 zh test” refers to the test set of the LogiQA 2.0 Chinese version. “LogiQA 2.0 ood”

2.1.2 Natural Language Inference

Natural language conclusion is the task of deciding the logical relationship between a thesis and a premise. The typical scheme is a textbook bracket where the model needs to choose one from three markers entailment, contradiction, and neutral. ConTRoL(Liu et al., 2020) is an NLI dataset that further investigates contextual logic under the NLI frame. It has 36.2 of premise, hypothesis and label under the order of logical logic. Figure 4 shows an illustration from the ConTRoL dataset. MED(Yanaka et al., 2019b) and HELP(Yanaka et al., 2019a) are two NLI datasets fastening on monotonicity logic, which is an essential conception in Natural sense(MacCartney and Manning, 2007b). The datasets are generated through monotonicity rules and only probe monotonicity-related conclusion specifically. analogous to

the NLI section of our dataset, MED and HELP probe introductory sense marvels in natural language, which is monotonicity in particular. For the HELP dataset, we use the train set for our evaluation. Then's an illustration from the HELP dataset represents the eschewal- of- distribution data of LogiQA 2.0.

Premise: Tom said that neither parents had ever been to Boston.

Hypothesis: Tom said that neither one of his parents had ever been to Boston.

Label: Entailment

ConjNLI (Saha et al., 2020) is a challenging stress test for NLI over conjunctive sentences, where the premise differs from the hypothesis by having conjuncts being removed, added, or replaced. Logical reasoning about conjunctions is heavily tested in ConjNLI. Premise-hypothesis pairs are created automatically by applying conjunct operations on collected conjunctive sentences. Here is an example from the ConjNLI dataset:

Premise: In Quebec, an allophone is a resident, usually an immigrant, whose mother tongue or home language is neither French nor English.

Hypothesis: In Quebec, an allophone is a resident, usually an immigrant, whose mother tongue or home language is not French.
Label: Entailment

TaxiNLI (Joshi et al., 2020) is an NLI dataset re-annotated on the MNLI (Williams et al., 2018b) dataset with fine-grained category labels. The annotation includes logical categories like connectives, mathematical, and deduction. Notice that TaxiNLI is a subset of the MNLI dataset, so we include the MNLI dataset for our comparison as a traditional NLI benchmark. Here is an example from the TaxiNLI dataset:

Premise: and that you’re very much right but the jury may or may not see it that way so you get a little anticipate you know anxious there and go well you know.

Hypothesis: Even if you’re correct, I think the jury would pick up on that.

Label: Contradiction

2.1.3 Out- of- distribution Data

AR- LSAT(Wang et al., 2022) is a new dataset of logical logic questions from the Law School Admission Test. Released in 2022, it has 2064 questions, each describing a logic game belonging to three dominant types(1) ordering game,(2) grouping game, and(3) assignment game. It's noticed that each question has five options rather than four. Figure 5 shows an illustration from the AR- LSAT test set.

Besides, we construct a LogiQA 2.0 outof- distribution dataset, which incorporates the recently released Chinese Civil menial test from 2022. The test set is a collection of logical logic tests designed by experts from 2022 onwards.

Dataset	ConTRoL test	ConjNLI test	HELP	MED	TaxiNLI test	MNLI dev
Size	805	623	35891	5382	10071	9815
Human avg.	87.00	89.00	81.00	91.00	97.00	98.00
Human ceiling	94.00	100.00	95.00	99.00	100.00	100.00
RoBERTa	48.76	38.94	39.47	46.83	49.91	90.02
ChatGPT (API)	58.45	47.03	42.13	55.02	57.30	55.40
GPT-4 (Chat UI)	58.18(64/110)	61.00 (61/100)	53.33 (56/105)	75.79 (72/95)	75.47(80/106)	68.00 (68/100)
GPT-4 (API)	56.40	72.71	46.01	89.42	60.08	64.08

Table 2: ChatGPT and GPT-4 performance on the natural language inference task (accuracy %).

2.2 Experiment Setting

We take RoBERTa- base(Liu et al., 2019) as our birth model. Following a fine- tuning scheme, we use Huggingface's RoBERTa- base model as ourpre-trained language model. RoBERTa- base is trained on the training set for 5 ages for each dataset. We also set up a mortal birth by reporting the average and ceiling performance of mortal testees. For ChatGPT and GPT- 4, we follow an instruction-prompt scheme for both Natural Language Inference andmulti-choice reading appreciation tasks. Figure 1 shows the instruction formulti-choice reading appreciation tasks. After probing the styles of promptdesigning for logic tasks, we find that there are substantially three types of prompt designing for NLI tasks, specifying the markers(entailment, neutral or contradiction)(Qin et al., 2023), specifying the logic system(induction or deduction)(Bang et al., 2023), and chain- of- study logic(Kojima et al., 2023) which will be specified in the coming chapter. Among these, specifying the markers system suits our purpose for utmost of our NLI datasets are 3- marker bracket tasks. therefore, we prompt GPT with the 3 three possible connections between the thesis and conclusion, entailment, contradiction and neutral, every time we ask a question to GPT. The instruction we use for themulti-choice reading appreciation task is in Appendix A. For assessing ChatGPT, We use the Eval frame handed by OpenAI, a suite for assessing OpenAI models and an open- source registry of marks. The model we choose is " gpt- 3.5 turbo "(interpretation March 23, 2023). piecemeal from task structure, we offer an in- environment illustration to each API call to guarantee controlled affair. GPT- 4 has been limited access to subscribe druggies from March 14, 2023. We're granted early access to GPT- 4 API by incorporating requests to the OpenAI Eval depository. So we're suitable to use the GPT- 4 API and the OpenAI Eval frame. The model we use is " dereliction- gpt- 4 "(interpretation March 14, 2023). We also use the GPT- 4 Chat UI to conduct our GPT- 4 trials and farther analyses with two OpenAI Plus accounts.

3 Results

3.1 Experiment results on themulti-choice reading appreciation tasks

Table 1 shows the results of themulti-choice reading appreciation datasets.

3.1.1 The performance of ChatGPT

ChatGPT shows a performance increase compared to the birth model on several long- standing marks. The delicacy of the LogiQA 2.0 test set is 53.37, nearly 4 points advanced than the RoBERTa base model. The performance gap between ChatGPT and RoBERTa is salient when testing on the Chinese interpretation of LogiQA 2.0, which indicates the performance thickness of ChatGPT in both Chinese and English languages. ChatGPT yields the stylish performance on the ReClor dataset with an delicacy of 57.38, compared with RoBERTa's 55.01 delicacy. still, ChatGPT gests a huge performance drop on out- ofdistribution datasets. On the AR- LSAT test set, the delicacy is only 20.42, lower than the performance of RoBERTa base. On LogiQA 2.0 ood, the performance is 38.44, still lower than that of RoBERTa base. From the trials above, ChatGPT performs well on well- known Logical logic like LogiQA and ReClor. The delicacy of ChatGPT surpasses fine- tuning styles by a small periphery. still, when tested on the recently re- leased dataset, videlicet AR- LSAT, and on LogiQA out- of- distribution dataset, the performance declined significantly. Despite its limitations, ChatGPT still represents a significant advancement in natural language understanding and demonstrates the eventuality of language models to reason logically.

3.1.2 The performance of GPT- 4

GPT- 4 performs remarkably better than ChatGPT when doing primer tests on LogiQA and ReClor. On the LogiQA 2.0 test set(1572 cases), GPT4 yields an delicacy of 72.25. On the Chinese interpretation of the LogiQA 2.0 test set(1594 cases), the delicacy is 70.56, which is analogous to the performance on the English interpretation. On the ReClor dev set(500 cases, ReClor does n't include gold markers on its test), GPT- 4 reaches an 87.20 delicacy which is the loftiest score among all three models. still, when tested on the AR- LSAT test set(230 intances), GPT- 4 performs unexpectedly worse with only a 33.48 delicacy. The test result on LogiQA 2.0 ood data(1354 cases) shows that GPT- 4 gets 58.49 correctness, which is significantly lower than that on the LogiQA 2.0 test set. nonetheless, the performance is still the loftiest among all three models. We'll not haste to the conclusion, but it's safe to say that GPT- 4's performance drop on out- of- distribution datasets is conspicuous. For comparison, the GPT- 4 Chat UI results are also handed, where we manually test a sprinkle of data cases.

3.2 Experiment results on the natural language inference task

Trial results on the natural language conclusion task Table 2 shows the results on the natural language conclusion datasets.

3.2.1 The performance of ChatGPT

ChatGPT performs better than the RoBERTa model on the logical logic NLI datasets we test. On the ConTRoL test set, the delicacy is 58.45, advanced than the RoBERTa- base model by nearly 10 percent. On the ConjNLI test set, ChatGPT yields 47.03 delicacy, which outperforms RoBERTa by around 9 percent. On the HELP dataset, ChatGPT

gets a 42.31 delicacy, around 3 points advanced than that of RoBERTa. On the MED dataset, ChatGPT gives 55.02 delicacy, nearly 9 percent advanced than that of RoBERTa. On the TaxiNLI test set, ChatGPT gives 57.30 delicacy, over 7 percent advanced than that of RoBERTa. For comparison, ChatGPT gives 55.40 delicacy on the MNLI dev set, which is significantly lower than that of RoBERTa, which indicates that ChatGPT is n't optimized for answering three-labeled natural language conclusion questions. Since it's noticed that ChatGPT is n't good at following NLI task instructions, we give an in- environment illustration to help the model induce task markers, as shown in Figure 2. Overall, the results show that ChatGPT surpasses OK - tuned RoBERTa by only a small periphery for logical logic NLI datasets. The performance of GPT- 4 We test GPT- 4's performance on logical logic NLI datasets. On the ConTRoL test set(805 cases), GPT- 4 performs slightly lower than ChatGPT, yielding a 56.40 delicacy. The performance of GPT- 4 on the ConjNLI test(623 cases) and the MED test(5382 cases) is significantly better, with an delicacy of 72.71 and , independently. still, on the HELP(35891 cases) and TaxiNLI test(10071 cases), the performance of GPT- 4 is slightly better than that of RoBERTa and ChatGPT, with an delicacy of 46.01 and 60.08, independently. The GPT- 4 performance on the MNLI dev set is slightly better than on the TaxiNLI test, and yields 64.08 delicacy, which indicates logical logic adds further challenges to the GPT- 4 model. We also include the testing results with GPT- 4 Chat UI and around 100 data cases for each NLI dataset. The results on the six NLI datasets indicate that GPT- 4 does n't perform largely on logical logic natural language conclusion compared to multichoice reading appreciation. We also notice that GPT- 4 can not affair markers steadily indeed though the instruction is handed in the natural language conclusion task script. From this, we infer that GPT- 4 is n't good at following the instruction for the natural language conclusion task, though it's well- trained to follow the instruction for the multichoice reading appreciation task.

4 Analysis

The trial results show that ChatGPT and GPT- 4 surpass RoBERTa on utmost logical logic marks, including popular marks like LogiQA and ReClor and less- known datasets like AR- LSAT. still, the performance drop on out- of-distribution datasets is conspicuous for both GPT models, indicating they struggle to handle new and strange data. therefore, we conduct more case studies with the GPT- 4 chat UI and farther dissect its capacities.

4.1 Answer and Reason

For GPT- 4 homemade tests, we record the answers GPT- 4 gives and the logic for the answer. Figure 6 gives an illustration of GPT- 4's answer and logic. In this illustration, GPT- 4 did it rightly. From the paragraph's inconsistency between the analogous drunk driving rate both with and without drunk driving checks, and the claim that the strict checks lower the drunk driving rate, GPT- 4 chooses a fact prior to the contemporary situation that the drunk driving rate used to be high before strict checks, to break this contradiction. In our assessment of GPT- 4 on the logiQA dataset, we audited the first 10 crimes made by the model. Four were distributed as logical crimes, similar as affirming the question, negating the thesis, and soliciting the question. Three were linked as compass crimes, including attributing predicates to incorrect subjects or assigning characters to the wrong objects. The remaining three crimes fall in the incapability to resolve semantic nebulosity, wherein GPT- 4 named a simply good response when asked for an optimal bone

. still, this limited sample of crimes does n't number the conclusion that GPT- 4 is unskillful to handle logic questions, as there are also cases where it directly identifies the correct answer. GPT- 4's occasional selection of wrong answers suggests that farther examination is necessary to determine whether some features in the questions may spark similar incorrect choices. In- environment literacy In this section, we test the in- environment literacy capability of GPT- 4. We observe that GPT- 4 is prone to affair more correct answers within the same discussion window after roughly eight discussion rounds; GPT- 4's delicacy increases after seeing further exemplifications. During this

procedure, no feedback is handed to the discussion. To illustrate this miracle, we conduct an trial on the LogiQA 2.0 ood data and the ConTRoL dataset, each representing a typical test case for the task ofmulti-choice reading appreciation and natural language conclusion. We aimlessly elect 20 cases from each dataset for the following testing. Flash back that GPT- 4's performance is n't competitive on these two datasets. We first test the 20 cases from the same discussion window; also, we test each case of 20 in a new discussion window. The testing results are shown in Table 3 For the LogiQA 2.0 ood dataset, GPT- 4 yields 9 correct answers when the 20 exemplifications are in the same discussion window. still, without the environment, the number of correct answers drops to 5. For the ConTRoL dataset, we find that GPT- 4 answers 13 questions rightly with the environment, and it drops to 7 without the environment. excursus C shows an illustration where GPT- 4 answers the question rightly inside the environment while does n't give the correct answer in a new discussion window. Chain- of- study Persuading Chain- of- study(Hut) egging is explored by numerous experimenters and shows promising results on complexmulti-step logic tasks(Kojima et al., 2023). This section explores zero- shot Hut egging for GPT- 4 on logical logic datasets. The trial is conducted on the LogiQA 2.0 ood data. We choose the same 112 cases as we do primer tests with GPT- 4, which is shown in Table 1. For this round, we add the prompt" Let's think step by step" to the instruction. By adding this prompt, GPT- 4 generates longer logic textbooks illustrating the logic way. We collect the final answer for each Hut logic process and get 61 correct answers out of 112 questions, which is advanced than the former trial without Hut egging . Overall, the evaluation of the logical logic capability of ChatGPT and GPT- 4 highlights the significance of developing further sophisticated marks in textual conclusion to ameliorate NLU systems' logical logic capacities further. The results also suggest that there's still room for enhancement in language models' logical logic capacities, particularly when handling out- of- distribution datasets. Experimenters need to continue developing further sophisticated marks in textual conclusion to ameliorate NLU systems' logical logic capacities further. Exploring new approaches to training language models that can more handle out- of- distribution datasets and other challenges associated with real- world operations is important.

Dataset	LogiQA 2.0 ood	ConTRoL
# instances	20	20
in context	45.00 (9/20)	65.00 (13/20)
w/o context	25.00 (5/20)	35.00 (7/20)

Table 3: GPT-4 performance with/without context.

5 Conclusion

In conclusion, our analysis highlights the significant strides made by ChatGPT and GPT-4 in handling logical reasoning tasks, particularly in familiar datasets like LogiQA and ReClor. Both models consistently outperform baseline systems in multi-choice reading comprehension, demonstrating their advanced reasoning capabilities. However, the models struggle with out-of-distribution datasets, such as AR-LSAT, revealing a key limitation in adapting to novel logical constructs. Despite notable advancements, challenges persist in achieving human-like logical reasoning in natural language understanding, especially for inference tasks under unfamiliar conditions. This study underscores the importance of developing more sophisticated benchmarks and training methods to enhance language models' robustness and reasoning depth. Future research should focus on refining these models' adaptability to diverse logical structures, ultimately paving the way for more resilient and contextually aware AI systems.

References:

1. Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.
2. Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
3. Candlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.
4. Maxwell John Cresswell. 1973. *Logics and languages (1st ed.)*. Routledge.
5. Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proc. of ICML*.
6. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*.
7. Lucja Iwanska. 1993. Logical reasoning in natural lan-´ guage: It is all about knowledge. *Minds and Machines*.
8. Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. Taxinli: Taking a ride up the NLU hill. *CoRR*.
9. Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.
10. Robert Kowalski. 1979. *Logic for problem solving*, volume 7. Ediciones Díaz de Santos.
11. Tao Li and Vivek Srikumar. 2019. Augmenting neural networks with first-order logic. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 292–302, Florence, Italy. Association for Computational Linguistics.
12. Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2020. Natural language inference in context - investigating contextual reasoning over long texts. *CoRR*.
13. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*.
14. Bill MacCartney and Christopher D. Manning. 2007a. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.
15. Bill MacCartney and Christopher D Manning. 2007b. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.
16. OpenAI. 2023. Gpt-4 technical report.
17. Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver?
18. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*.
19. Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. Conjnli: Natural language inference over conjunctive sentences. *CoRR*.
20. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob
21. Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz
22. Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*.
23. Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018.
24. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

25. Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. 2022. From lsat: The progress and challenges of complex reasoning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
26. Adina Williams, Nikita Nangia, and Samuel Bowman. 2018a. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of NAACL*.
27. Adina Williams, Nikita Nangia, and Samuel Bowman. 2018b. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*
28. *1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
29. Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, , and Johan Bos. 2019a. Help: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM2019)*.
30. Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. Can neural networks understand monotonicity reasoning?