GLoRE: Evaluating Logical Reasoning of Large Language Models

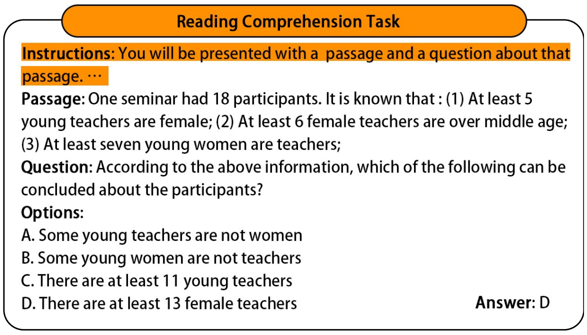
Anonymous ACL submission

Abstract

Recently, large language models (LLMs), including notable models such as GPT-4 and burgeoning community models, have showcased significant general language understanding abilities. However, there has been a scarcity of attempts to assess the logical reasoning capacities of these LLMs, an essential facet of natural language understanding. To encourage further investigation in this area, we introduce GLoRE, a meticulously assembled **G**eneral **Lo**gical **R**easoning **E**valuation benchmark comprised of 12 datasets that span three different types of tasks. Our experimental results show that compared to the performance of human and supervised fine-tuning, the logical reasoning capabilities of open LLM models necessitate additional improvement; ChatGPT and GPT-4 show a strong capability of logical reasoning, with GPT-4 surpassing ChatGPT by a large margin. We propose a self-consistency probing method to enhance the accuracy of ChatGPT and a fine-tuned method to boost the performance of an open LLM. We release the datasets and evaluation programs to facilitate future research.

1 Introduction

Large Language Models (LLMs) (OpenAI, 2023; Touvron et al., 2023; Anil et al., 2023) are increasingly being aligned with real-world tasks (Bubeck et al., 2023; Ouyang et al., 2022; Qin et al., 2023; Chung et al., 2022), demonstrating advanced capabilities in handling complex reasoning tasks and showing significant adaptability and versatility across various applications, from simple everyday tasks to specialized domains such as coding, mathematics, law, medicine, and finance (Li et al., 2022; Frieder et al., 2023; Choi et al., 2023; Kung et al., 2023; Wu et al., 2023b). Previous work has shown pre-trained models' proficiency in natural language understanding tasks (Goyal et al., 2023; Zhong et al., 2023a). However, studies also reveal areas of deficiency (Kocoń et al., 2023; Wang et al., 2023), calling into question the overall reasoning capabilities of these models (Chalmers, 2023).



Reading Comprehension Task

Instructions: You will be presented with a passage and a question about that passage. ···

Passage: One seminar had 18 participants. It is known that : (1) At least 5 young teachers are female; (2) At least 6 female teachers are over middle age; (3) At least seven young women are teachers;

Question: According to the above information, which of the following can be concluded about the participants?

Options:
A. Some young teachers are not women
B. Some young women are not teachers
C. There are at least 11 young teachers
D. There are at least 13 female teachers

Answer: D

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

Logical reasoning is fundamental to human intelligence, and natural language-based logical reasoning has remained a vibrant research interest since the inception of artificial intelligence (Cresswell, 1973; Kowalski, 1979; Iwańska, 1993; Liu et al., 2020b; Yu et al., 2020). Figure 1 represents a showcase of testing logical reasoning in reading comprehension. To successfully respond to such logical reasoning questions, LLMs typically need to engage in multi-step, algorithmic, symbolic, and compositional reasoning (Liu et al., 2020b). Thus, logical reasoning serves as a suitable testbed for evaluating the abilities of LLMs to process complex information in natural language accurately, robustly, and logically.

To this end, we present a General Logical Reasoning Evaluation (GLoRE) benchmark, evaluating instruction-tuned LLMs for LLM logical reasoning tasks on several logical reasoning datasets, detailing the strengths and limitations of LLMs in this domain. Similar to GLUE (Wang et al., 2018) and Super-GLUE (Wang et al., 2019) for natural language understanding, GLoRE assembles a range of different datasets that evaluates logical reasoning. Specifically, we consider three types of logical reasoning tasks, including Multi-choice Reading Comprehension (Lai et al., 2017), Natural Language Inference (NLI) (Dagan et al., 2005), and True-or-

False (Yes-or-No) Questions (Clark et al., 2019). The three task formats cover a broad spectrum of logical reasoning phenomena, where high-quality logical reasoning datasets were released and remain challenging for pre-trained language models before LLM (Huang and Chang, 2023; Clark et al., 2019; Koreeda and Manning, 2021). Overall, GLoRE covers 12 datasets with 72,848 instances in total.

Using GLoRE, we evaluate the logical reasoning ability of both powerful commercial models like GPT-4, and popular open-sourced models like the ones based on LLaMA (Touvron et al., 2023), Falcon (Almazrouei et al., 2023) and Mistral (Jiang et al., 2024), testing their instruction-following and problem-solving abilities for logical reasoning tasks. Results show that commercial LLMs outperform open-source LLMs and pre-trained LMs by a large margin on zero-shot settings, with GPT-4 drawing close to human performances on specific datasets. However, the performance of GPT 4 and other models does not remain stable across the board, with significant variations between different datasets, which can indicate their sensitivity to data distributions. The susceptibility of models to variations in data distribution is further confirmed by observations that both in-context learning and supervised fine-tuning predominantly enhance the performance of Large Language Models (LLMs) across specific test distributions. This demonstrates their robust learning ability. Interestingly, Chain-of-Thought reasoning can be helpful to logical reasoning, as indicated by prior work (Kojima et al., 2023; Chen et al., 2023; Saparov and He, 2022; Yang et al., 2022), but only to a very limited extent, which suggests that it might take effect mostly by offering relatively superficial patterns. Our results show both promises and challenges – on the one hand, LLMs show the potential to give solid performances and learn effectively on logical reasoning datasets; on the other hand, they show much sensitivity to the data distribution, and therefore, the robustness needs further enhancement.

To our knowledge, GLoRE is the first instruction-prompt evaluation suite for logical reasoning, and we are the first to evaluate LLMs' complex logical reasoning abilities comprehensively. We release our benchmark at `https://anonymous.com`.

2 Related Work

Logical Reasoning with Natural Language. Tapping into logical reasoning capabilities represents a holistic endeavour in natural language understanding (NLU). A variety of methods have been explored to realize this objective, including symbolic systems (Mccarthy, 2002; Poole et al., 1987; MacCartney and Manning, 2007a), fine-tuning of language models (Wang et al., 2018; Huang et al., 2021; Xu et al., 2022; Liu et al., 2023b), and hybrid approaches combining neural and symbolic elements (Li and Srikumar, 2019; Saha et al., 2020; Sanyal et al., 2022).

The recent introduction of evaluation datasets, notably LogiQA (Liu et al., 2020b) and Reclor (Yu et al., 2020), has reinvigorated the focus on logical reasoning in NLP research. Logical reasoning is now leveraged in numerous probing tasks over large Pre-trained Language Models (PLMs) and applied to downstream tasks such as question-answering and dialogue systems (Shi et al., 2021; Beygi et al., 2022). Despite these advancements, the aspiration to emulate human-like logical reasoning capabilities within NLU systems remains a significant challenge for traditional models (Liu et al., 2020b; Huang and Chang, 2023). In this study, our goal is not only to quantitatively evaluate the capability of Large Language Models (LLMs) in addressing the previously mentioned challenge but also to underscore the significance of our work in providing a validated platform for enhancing various reasoning methods with our data.

LLM Reasoning Evaluation. Despite progress in evaluating LLMs for specific reasoning tasks like arithmetic (Qin et al., 2023) and commonsense (Bang et al., 2023), a yawning gap exists in comprehensively assessing their logical reasoning. While LLMs excel at specific tasks like arithmetic reasoning (Qin et al., 2023), they face challenges in complex areas like multi-step reasoning (Fu et al., 2023) and abstract scenarios (Gendron et al., 2023). ChatGPT exhibits strengths in chat-specific reasoning and some commonsense domains (Bang et al., 2023; Ott et al., 2023), but struggles with tasks requiring longer chains of inference (Bang et al., 2023). Other LLMs like FLAN-T5 (Chung et al., 2022), LLaMA (Touvron et al., 2023), and PaLM (Anil et al., 2023) show potential in general deductive reasoning (Saparov et al., 2023), while InstructGPT and Codex excel in specialized domains like medical reasoning (Liévin et al., 2022). Despite these advances, limitations in data bias (Orrù et al., 2023), and complex reasoning tasks necessitate further research and optimization to fully unlock the reasoning potential of LLMs (Wu et al., 2023c).

The Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021) evaluates the capabilities of large language models in various domains, ranging from the foundational areas of knowledge like mathematics and history to highly specialized fields such as law and ethics. However, compared to the GLoRE benchmark, MMLU does not introduce logical reasoning data into the scope, making it incapable of testing complex logical reasoning tasks.

Big-Bench Hard (BBH) (Suzgun et al., 2022) isolates 23 most challenging tasks from BIG-Bench (bench authors, 2023). These tasks comprise general language understanding, arithmetic and algorithmic reasoning, and logical deduction. However, in comparison to our benchmark, the data size of the logical reasoning section in BBH is very small. HumanEval (Chen et al., 2021) serves as a hand-written evaluation set for coding. The programming problems included are designed to assess language comprehension, reasoning, algorithms, and simple mathematics. While similar to logical reasoning in that code generation necessitates complex reasoning skills, GLoRE differs in presenting logical reasoning problems via natural language prompts.

ARB (Sawada et al., 2023) is a benchmark for advanced reasoning over multiple fields like mathematics, physics, biology, chemistry, and law. Similar to GLoRE, it introduces a challenging subset of math and physics problems that require advanced symbolic reasoning. However, the benchmark constraints its problem on the above subjects with domain knowledge, not general logical reasoning questions, which is the focus of GLoRE.

3 The GLoRE Dataset

Dataset	**Size**	**Target**
LogiQA 2.0 test	1,572	4-way multi-choice
LogiQA 2.0 zh test	1,594	4-way multi-choice
ReClor dev	500	4-way multi-choice
AR-LSAT test	230	5-way multi-choice
LogiQA22	1,354	4-way multi-choice
ConTRoL	805	E, C, N
HELP	35,891	E, C, N
TaxiNLI test	10,071	E, C, N
NaN-NLI	259	E, C, N
FraCas	346	Yes, No, Neutral
RuleTaker dev	10,068	Yes, No
ProofWriter dev	10,158	Yes, No

Table 1: Data statistics. (“E” refers to “entailment”; “C” refers to “contradiction”; “N” refers to “neutral”.)

As mentioned in the introduction, GLoRE contains three NLU tasks: Multi-choice Reading Comprehension, NLI, and Yes-or-No. First, Multi-choice reading comprehension (Lai et al., 2017) is essential in verbal reasoning tests, which cover abundant high-quality logical reasoning problems in the wild. Second, Unlike multi-choice reading comprehension, NLI (Dagan et al., 2005) is more general and centric on entailment relations in a simpler task format, which is a fundamental task for evaluating reasoning abilities (Poliak et al., 2018; Demszky et al., 2018). Third, the Yes-or-No reasoning task (Clark et al., 2019) is a combination of question-answering and textual entailment, which can serve as a playground for testing models’ reasoning abilities (Clark et al., 2020; Tafjord et al., 2021). The data statistics are shown in Table 1.

3.1 Multi-choice Reading Comprehension (MRC)

Within the standard multi-choice reading comprehension (MRC) task setting, a system is presented with a passage and a question, and the objective is to choose the most suitable answer from a set of candidate responses. An example of logical MRC can be seen in Figure 1. Particularly, GLoRE contains five such datasets:

LogiQA (Liu et al., 2020b) is a logical MRC dataset derived from the Chinese Civil Service Examination, translated into English, and made available in both Chinese and English versions. Figure 3 in Appendix A illustrates an example. We adopt the second version of LogiQA (Liu et al., 2023a) and use both the English (**LogiQA 2.0**) and Chinese (**LogiQA 2.0 zh**) test sets for our evaluation.

ReClor (Yu et al., 2020) comprises question-answering examples from the LSAT exams designed to assess human logical reasoning abilities. We use the development set for our testing as the test set does not provide gold labels.

AR-LSAT (Wang et al., 2022) is a dataset of analytical reasoning questions from the Law School Admission Test. Each question contains five options rather than four. An example from the AR-LSAT test set can be found in Figure 4 in Appendix A.

LogiQA22 is collected and processed according to the LogiQA 2.0 format after ChatGPT was released. It incorporates the newly released Chinese Civil Servant Exams from 2022, which are not included in the original LogiQA dataset.

3.2 Natural Language Inference (NLI)

NLI is the task of determining the logical relationship between a hypothesis and a premise. The typical scheme involves text classification, where the model selects one of three labels: *entailment*, *contradiction*, and *neutral*. An logical NLI example is shown in Figure 5.

ConTRoL (Liu et al., 2020a) is an NLI dataset that offers an in-depth examination of contextual reasoning within the NLI framework. Figure 5 in Appendix A displays an example of ConTRoL. Approximately 36.2% of premise-hypothesis pairs fall under the category of logical reasoning in this dataset. We choose the logical reasoning portion for our evaluation.

HELP (Yanaka et al., 2019) is an NLI dataset emphasizing monotonicity reasoning, a crucial concept in Natural Logic (MacCartney and Manning, 2007b). An example from the HELP dataset can be seen in Figure 6 in Appendix A. We use the training set for our evaluation.

TaxiNLI (Joshi et al., 2020) is an NLI dataset that has been re-annotated based on MNLI (Williams et al., 2018), with categories include logical categories such as connectives, mathematical reasoning, and deduction. An example from the TaxiNLI dataset can be found in Figure 7 in Appendix A.

NaN-NLI (Truong et al., 2022) is a test suite designed to probe the capabilities of NLP models in capturing sub-clausal negation. An example from the NaN-NLI dataset is depicted in Figure 8 in Appendix A. The successful handling of sub-clausal negation can be seen as a strong indicator of a model's logical reasoning capacity.

3.3 True-or-False (Yes-or-No) Questions (TF)

The **FraCaS** test suite (Pulman, 1996), converted to RTE style by MacCartney and Manning (2007a), presents complex entailment problems involving multi-premised contexts. The original FraCas dataset is a three-way classification ("Yes", "No", "Don't know") task. The ability to determine entailment relationships in this context is closely tied to logical reasoning. Figure 9 in Appendix A illustrates an example. We convert the "Don't know" label into a single "Neutral" token.

The **RuleTaker** (Clark et al., 2020) dataset is a synthetic creation designed to examine the reasoning ability of transformer models (Vaswani et al., 2017) over natural language rules. This task explicitly targets logical reasoning by asking models to reason over a set of rules and facts to generate true-or-false responses as output. An example from the RuleTaker dataset is shown in Figure 10 in Appendix A.

The **ProofWriter** (Tafjord et al., 2021) dataset generates sets of facts and rules, each followed by questions, which can be proven true or false using proofs of various depths. Figure 11 in Appendix A presents an example from the ProofWriter dataset.

4 Evaluation Methodology

We consider seven logic reasoning evaluation scenarios for open-sourced LLMs and closed API-based or UI-based models such as ChatGPT and GPT-4, which include *zero-shot evaluation*, *few-shot and Chain-of-Thought evaluation*, *instruction tuning evaluation*.

Zero-shot Evaluation In this setup, the task input is transposed into a prompt via templates, and the gold label is verbalized (Liu et al., 2021b). The LLMs need to generate the verbalized gold answer. Prior research indicated that ChatGPT could underperform in question-answering scenarios if the instructions were not appropriately optimized (Zhong et al., 2023b). Consequently, we investigated different zero-shot prompting methods to enhance the performance of the tested models. The instructions differ slightly for different datasets, according to their target outputs. The finalized instructions for the three types of tasks are integrated into GLoRE.

Few-shot Evaluation LLMs are capable of achieving efficient in-context learning (Dong et al., 2023), where different numbers of context examples and in-context demonstration methods (Liu et al., 2021a) can be used. In this study, we randomly sampled a few instances (1 for 1-shot, 2 for 2-shot, and 5 for 5-shot) from each dataset to conduct few-shot experiments respectively. For each sampled instance, we append it to the beginning of the existing prompt. For the experiment, we use the same model configuration as in the zero-shot scenario.

Instruction Tuning An appealing benefit of open-sourced LLMs, such as LLaMA, lies in their amenability to task-specific fine-tuning (Wu et al., 2023a). This feature allows us to optimize their performance more precisely, offering a distinct edge over their closed counterparts. We consider an evaluation method by fine-tuning the open-sourced LLM model using instruction-tuning, providing specific instructions to address distinct tasks.

We converted a specific logic reasoning training set into the instruction-prompting framework as shown in Appendix B. This process entailed reforming the dataset such that each instance was paired with a clear, directive instruction, an input, and a target output. We then fine-tuned an open-sourced LLM with this transformed training dataset and the fine-tuning process. After instruction-tuning, we evaluate the model performance on the specific test set for the training task and the zero-shot performance on the other logic reasoning tasks to examine its cross-task generalization ability.

Chain-of-Thought Evaluation It has been shown that Chain-of-Thought (CoT) can improve the math (Imani et al., 2023; Chen et al., 2022) and logic (Ling et al., 2023) capabilities of LLMs. We explore zero-shot CoT prompting (Kojima et al., 2023) on logical reasoning datasets.

5 Results

5.1 Evaluated Models

We adopted **RoBERTa-base** (Liu et al., 2019) as a baseline, fine-tuning it on the training set over five epochs for each dataset. The community models selected for comparison include FALCON-40B-INSTRUCT (Almazrouei et al., 2023) LLAMA-30B-SUPERCOT (Touvron et al., 2023) and MIXTRAL-8X7B, both of which are highly-regarded open language model representations (LLMs) available on the HuggingFace Hub.[1]

Both **ChatGPT** and **GPT-4** are evaluated with the OpenAI Evaluation framework[2], a comprehensive tool designed for the evaluation of OpenAI models. The specific versions of the models assessed are labeled as "gpt-3.5-turbo-0301" for ChatGPT and "gpt-4-0314" for GPT-4, respectively. Moreover, we engage the GPT-4 Chat UI to conduct a series of case studies on GPT-4. These examinations probe into the model's in-context learning abilities and chain-of-thought reasoning capabilities, by using two OpenAI Plus accounts.

All experiments were executed on 40G VRAM A100 GPUs based on the HuggingFace transformers library. Our evaluation metrics consisted of classification accuracy scores. Additionally, we utilized reported accuracies for datasets where human performance data was available and recorded both the average and peak performance of human participants to establish a human baseline. For the LogiQA22 dataset, we engaged five co-authors as test subjects and computed their accuracy based on 150 test examples.

5.2 Main Results

Zero-shot Results Table 2 outlines the primary zero-shot evaluation results. The first block presents both the average and maximum human performance. Notably, with the exception of the ReClor and AR-LSAT tasks, humans achieve an average accuracy exceeding 80%. On ReClor and AR-LSAT, the averaged human performance is 63.00% and 56.00%, respectively, showing the challenge of these LSAT tasks. The human ceiling performance is close to 100%, showcasing human proficiency in logical reasoning tasks.

The second block details the supervised fine-tuning results of RoBERTa-base, a model containing only 125M parameters. RoBERTa-base achieves accuracy rates of 48.76% and 33.22% on LogiQA 2.0 and LogiQA22, respectively. The overall performance of RoBERTa-base lags behind average human performance, suggesting that supervised models may struggle to learn logical reasoning. Moreover, the model's performance on MRC tasks is lower than on NLI and TF tasks, which can be because of more output ambiguities (multi-choice vs. three-way or Yes/No). On the NaN-NLI dataset, RoBERTa yields 90.02% accuracy, the best performance reaching the human level. This might be because NaN-NLI is a negation data converted from sentence-level NLI datasets by rules. Fine-tuned RoBERTa is able to learn superficial artifacts from the data. While ProofWriter requires complex reasoning skills, RoBERTa-base's superior performance (55.92%) on this task suggests its potential to tackle specific types of logical reasoning tasks.

The third block presents the zero-shot results for LLaMA, Falcon, and Mixtral. The average performance across all tasks is strikingly similar for LLaMA and Falcon (32.34% for LLaMA and 32.28% for Falcon), suggesting that LLaMA-30B's logical reasoning capabilities are comparable to those of Falcon 40B. However, both LLaMA and Falcon fall short of RoBERTa-base's performance on nearly all task types, with the notable exception of RT for Falcon. Specifically, the accuracy results on the MRC tasks for LLaMA and Falcon are approximately 20%, a figure which is even lower than expected from a random guess in a 4-way classi-

---

[1] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[2] https://github.com/openai/evals

Task	MRC					NLI				TF			Average
Dataset	LQ	LQ zh	RC	AL	LQ22	CT	HL	TN	NN	FC	RT	PW	
Human avg.	86.00	88.00	63.00	56.00	83.00	87.00	81.00	97.00	94.00	92.00	84.00	82.00	82.75
Human Ceiling	95.00	96.00	100.00	91.00	99.00	94.00	95.00	100.00	100.00	97.00	95.00	93.00	96.25
RoBERTa	48.76	35.64	55.01	30.90	33.22	48.76	39.47	49.91	**90.02**	32.01	53.50	55.92	47.76
LLaMA	19.31	26.35	17.81	17.98	18.41	24.10	32.26	41.91	47.29	40.00	48.89	53.78	32.34
Falcon	23.21	19.77	26.77	12.70	17.33	16.13	28.49	44.66	53.31	35.57	56.11	53.33	32.28
Mixtral-8x7B	45.29	36.81	48.92	41.40	38.97	50.84	33.27	40.86	50.13	32.08	46.84	44.80	42.52
ChatGPT	52.37	53.18	57.38	51.49	38.44	**58.45**	42.13	57.30	56.59	49.13	54.74	53.95	52.10
GPT-4	**72.25**	**70.56**	**87.20**	**73.12**	**58.49**	56.40	**46.01**	**60.08**	76.74	**75.35**	**60.19**	**59.66**	**66.34**

Table 2: LLMs' performance on the GLoRE benchmark. *LQ*: LogiQA 2.0, *RC*: ReClor, *AL*: AR-LSAT, *CT*: ConTRoL, *HL*: HELP, *TN*: TaxiNLI, *NN*: NaN-NLI, *FC*: FraCas, *RT*: RuleTaker, *PW*: ProofWriter. All results are in %, the best ones are in **bold**, and the second best ones are in underline.

Types	ChatGPT	GPT-4	LLaMA	Falcon
Categorical reasoning	83.83% (389/464)	95.04% (441/464)	22.84% (106/464)	20.91% (97/464)
Sufficient condition reasoning	44.99% (175/389)	63.75% (248/389)	20.82% (81/389)	20.56% (80/389)
Necessary condition reasoning	37.46% (124/331)	60.73% (201/331)	19.64% (65/331)	25.38% (84/331)
Conjunctive reasoning	26.79% (75/280)	35.00% (98/280)	7.86% (22/280)	12.86% (36/280)
Disjunctive reasoning	15.75% (60/381)	27.03% (103/381)	7.87% (30/381)	17.85% (68/381)

Table 3: LLMs' performance across reasoning types (accuracy %).

fication. These findings indicate that instruction-tuned LLMs face challenges with logical reasoning tasks without incorporating specific in-context demonstrations. Furthermore, we observe a smaller performance gap between LogiQA and LogiQA22 for these models compared to RoBERTa, implying that without specific in-domain tuning, their performance remains relatively stable and is not significantly impacted by the presence of test data distribution. Mixtral-8x7b, on the other hand, shows a significant performance increase compared to the other two open models, indicating the efficiency of a mixture-of-expert model.

The fourth block provides the zero-shot results of ChatGPT and GPT-4. Both models, particularly GPT-4, exceed RoBERTa-base in several MRC benchmarks. However, we observed a significant performance drop on LogiQA22. For instance, GPT-4's accuracy on LogiQA22 dropped to 58.49% compared to a solid 72.25% on LogiQA 2.0, indicating that these models are sensitive to data distribution, while struggle with unfamiliar data distributions. In NLI tasks and true-or-false questions, ChatGPT and GPT-4 showed notable improvements over the fine-tuned RoBERTa across most datasets. Specifically, ChatGPT exhibited the best performance with 58.45% accuracy on the ConTRoL dataset, surpassing GPT-4. Again, GPT4's performance varies across datasets for NLI, showing sensitivity to data distribution.

Model	0-shot	1-shot	2-shot	5-shot
LLaMA	32.34	32.89	35.03	39.62
Falcon	32.28	33.14	33.76	35.72
ChatGPT	52.10	55.85	57.43	60.32
GPT-4	66.34	70.31	71.44	75.83

Table 4: Average accuracies on GLoRE few-shot evaluation.

The results of TF questions are similar. Intuitively, the underlying logical rules are consistent across different datasets, but the data distributions are different. If a model makes use of correct rationales, it should give consistent levels of performance across distributions. Our observations in Table 2 contradict the above, which shows that the model rationale is not the same as the human rationale.

Results Across Tasks and Reasoning Types In our experiments, we evaluated the performance of the LLMs on three types of tasks. We found that the performance of models varied significantly across tasks and reasoning types. Table 2 lists out the detailed scores.

In zero-shot scenarios, the open-source models falcon-40b-instruct and llama-30b-supercot performed significantly below RoBERTa and human baselines on machine reading comprehension and natural language inference tasks, with the exception of binary classification problems, where the performance gap is not salient. Specifically, ChatGPT exemplifies similar performance to the two open-source models, indicating their incapability on TF questions. However, ChatGPT and GPT-4 showed improved performance compared to RoBERTa, even in zero-shot conditions. In particular, GPT-4 performed close to or even surpassed the human level on datasets such as ReClor.

Overall, GPT-4 and ChatGPT models show remarkable capability in tackling some logical MRC datasets. The performance is not as competitive when facing the NLI and TF tasks (NLI and TF are three-way or two-way classification tasks; however, most of the accuracies are even lower). Apart from that, we observed a significant performance drop in newly cultivated data for these commercial models, a trend not mirrored by the open-source models. The shift in data distribution might contribute to the performance drop of the intensive instruction-tuned models.

5.3 The Effect of In-Domain Training

The above experiments show that the performances of LLMs are sensitive to the data distribution. Even though the underlying reasoning principles are the same, LLM performance varies significantly across datasets. This suggests that LLMs might not reason using the correct rationale, but rely on superficial features. To further investigate the influence of data distribution, we consider training on datasets where LLMs perform weakly – using in-context learning for commercial LLMs and supervised fine-tuning for open-source LLMs.

Few-shot Results for GPT-4 Few-shot learning aims to educate models on the data distribution with as few instances as possible. The few-shot evaluation tests the efficiency of models to solve similar problems. Evaluation results are shown in Table 4. With the increase of in-context examples, the accuracy of each tested model on the GLoRE benchmark increases. The models we tested all show in-context learning abilities on the logic reasoning benchmark. Among them, GPT-4 witnesses the highest performance gain with over 9 percent accuracy boost on the 5-shot scenario compared to zero-shot.

Instruction-tuned LLaMA We conducted instruction tuning (Section 4) with the LogiQA 2.0 training set using LLaMA-7b. The fine-tuning process, spanning 2 epochs, leveraged the computational capabilities of 2 A100 GPUs. The results of this experiment are illustrated in Table 5. First, post fine-tuning with Alpaca's instructions, a substantial improvement in performance was observed across all tasks, underscoring the effectiveness of instruction-tuning. As Alpaca's instructions were not task-specific for logical reasoning tasks, the improvements can be largely attributed to the model's enhanced general instruction comprehension capabilities. Second, our tuned LLaMA-7B model markedly outperformed the baseline LLaMA-7B model and Alpaca. On LogiQA 2.0, the accuracy is improved from 18.04% to 52.74%, achieving a performance higher than the fine-tuned RoBERTa-base result (48.76%). Although the instruction-tuning only uses the LogiQA 2.0 training dataset, the tuned model can generalize the logic reasoning ability to the other datasets. For instance, on LogiQA 2.0 zh, the performance is boosted from 19.06% to 31.18%, while on ReClor, the fine-tuned model achieved 55.20% accuracy, outperforming Alpaca by 32.82 points. These results demonstrate that instruction-finetuning can improve the zero-shot logic reasoning performance via transfer learning. Moreover, the instruction-tuned model's performance on LogiQA22 (35.16%) even surpassed that of the RoBERTa-based classification model (33.22%), demonstrating the potential benefits of generalization using instruction-tuning.

Dataset	**7b-base**	**Alpaca**	**7b-tuned**
LogiQA 2.0 test	18.04	22.99	**52.74**
LogiQA 2.0 zh test	19.06	22.54	**31.18**
ReClor dev	15.83	22.38	**55.20**
AR-LSAT test	13.91	13.16	**21.43**
LogiQA22	20.25	21.16	**35.16**

Table 5: Fine-tune LLaMA on the LogiQA dataset (accuracy %). "7b-base" is the base model of LLaMA-7b; "Alpaca" is an instruction-tuned LLaMA-7b with GPT-4 Alpaca data; "7b-tuned" is our fine-tuned LLaMA-7b on the LogiQA 2.0 training set. All results are in %.

Model	**w/o CoT**	**w/ CoT**
LLaMA	32.34	35.05
Falcon	32.28	34.98
ChatGPT	52.10	55.75
GPT-4	66.34	68,47

Table 6: Chain-of-Thought evaluation on GLoRE. All results are in %.

	CoT correct	**CoT wrong**
w/o CoT correct	65.00	1.33
w/o CoT wrong	3.50	30.21

Table 7: The confusion matrix for GPT-4 results on the LogiQA22 data with/without CoT. All results are in %.

Model	Coherence	Completeness	Correctness	Relevance
LLaMA	3.38	3.53	3.00	4.50
Falcon	3.21	3.44	3.15	4.50
ChatGPT	4.00	4.81	3.76	4.72
GPT-4	4.52	4.81	4.51	4.89

Table 8: Human evaluation of CoT generations.

5.4 Chain-of-Thought Prompting

It has been shown that Chain-of-Thought prompting can give stronger performances for reasoning (Wei et al., 2023; Kojima et al., 2023). One advantage of Chain-of-Thoguht reasoning is that it increases the interpretability, where we an gain understanding of the reasoning steps. Table 6 shows the results on GLoRE with/without CoT. Apart from that, we calculate the confusion matrix of GPT-4 results in Table 7. All models experience a performance gain with the CoT prompting, ranging from 2 to 3 percent. The confusion matrix further illustrates the significance of performance elevation with CoT prompting.

Manual Evaluation and Case Study We further evaluate the reasoning processes by LLMs, and the results are shown in Table 8. The human evaluation is conducted on 100 data instances randomly selected from the benchmark. The objective is to assess the model's capability to produce logically coherent reasoning pathways leading up to the final answer, rather than solely the correctness of the outcome.

The four dimensions we include in our evaluation metrics are detailed as follows:

a. Coherence: Measure the logical consistency in the reasoning process. Are there any jumps in logic or contradictory statements?

b. Completeness: Does the model cover all aspects of the question? Is every step in the reasoning process explained?

c. Correctness: Beyond the final answer, are the intermediate conclusions accurate?

d. Relevance: Is the content of the reasoning pertinent to the question at hand? Are there any unrelated digressions?

We adopt a 5-point Likert scale for each metric:

1 = Poor, 2 = Below Average, 3 = Average, 4 = Above Average, 5 = Excellent.

It can be seen that the models give relatively low scores on the coherence and correctness of the reasoning chains. Surprisingly, some 11% of incorrect reasoning chains can lead to correct outputs, as an example shown in Figure 2. This further shows that LLM might not rely on exact reasoning chains for deriving the conclusion, but might make use of superficial features in the chain instead. The results indicate the need for further enhancing the causal nature of LLM reasoning.

MRC: Correct Answer, Wrong Inference

Passage: At a job fair of the company, the administration department, human resources department and office of the company each plan to recruit a staff member. Three graduates from the Chinese Department, the History Department and the philosophy Department came to apply for these three different positions. The recruitment information shows that the graduates of the history department are older than the graduates of the office, the graduates of the philosophy department and the graduates of the human resources department wear similar colors, and the graduates of the human resources department are younger than those of the Chinese department.
Question: If the above argument is true, which of the following options can be inferred?
Options:
A.Graduates of philosophy Department are older than graduates of the History Department.
B.Graduates of the Chinese department are older than those of the philosophy department.
C.Graduates from the History Department apply for the administration department.
D.Graduates of the Chinese Department apply for office.
Answer: B
Reason: The passage states that the graduates of the human resources department are younger than those of the Chinese department. It also mentions that the graduates of the philosophy department and the graduates of the human resources department wear similar colors. Since the graduates of the human resources department are younger than those of the Chinese department, we can infer that the graduates of the Chinese department are older than those of the philosophy department. The other options are not directly supported by the information provided in the passage.

Figure 2: GPT-4 responses with correct answer yet wrong inference.

We further elaborated on two specific case studies in Appendix C. These case studies provide detailed examples of how the models responded to specific prompts and where GPT-4 made the right and wrong predictions and rationales.

6 Conclusion

We assembled GLoRE, a comprehensible dataset for evaluating the logical reasoning ability of ChatGPT, GPT-4, and other strong open-source LLMs on multiple logical reasoning tasks. Our results show that ChatGPT and GPT-4 outperform the traditional fine-tuning method on most logical reasoning benchmarks. In contrast, community models are weak on GLoRE, while instruction-tuning on similar data increases the models' performance. Finally, supervised fine-tuning, in-context learning, and voting techniques all lead to stronger results. Both quantitative and qualitative evaluation suggest that existing LLMs may rely on relatively superficial patterns in solving logical reasoning tasks, and research on enhancing the underlying inference mechanism can be useful for addressing such issues.

Limitatins

While the GLoRE benchmark provides valuable insights into the logical reasoning capabilities of large language models (LLMs), there are several limitations to consider:

Dataset Bias The effectiveness of evaluating logical reasoning in LLMs heavily relies on the quality and diversity of the datasets used. Biases present in the training data may impact the generalizability of the results and the model's performance on real-world scenarios.

Task Specificity The logical reasoning tasks included in the GLoRE benchmark may not cover the full spectrum of reasoning abilities required for comprehensive natural language understanding. Certain types of reasoning, such as causal reasoning or temporal reasoning, may not be adequately addressed in the current evaluation framework.

Scalability As LLMs continue to grow in size and complexity, scalability issues may arise in evaluating their logical reasoning abilities. The computational resources required for training and testing these models on increasingly complex tasks could be a limiting factor.

Addressing these limitations and exploring avenues for further research will be essential to enhance the robustness and applicability of logical reasoning evaluations in large language models.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

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 arXiv:2302.04023*.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Sajjad Beygi, Maryam Fazel-Zarandi, Alessandra Cervone, Prakash Krishnan, and Siddhartha Reddy Jonnalagadda. 2022. Logical reasoning for task oriented dialogue systems.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

David J. Chalmers. 2023. Could a large language model be conscious?

Jialin Chen, Zhuosheng Zhang, and Hai Zhao. 2023. Modeling hierarchical reasoning chains by linking discourse units and key phrases for reading comprehension. *arXiv preprint arXiv:2306.12069*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. Chatgpt goes to law school. *Available at SSRN*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proc. of IJCAI*.

Maxwell John Cresswell. 1973. *Logics and languages (1st ed.)*. Routledge.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *MLCW*.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. 2023. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint arXiv:2305.17306*.

Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2023. Large language models are not abstract reasoners. *arXiv preprint arXiv:2305.19555*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey.

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. Dagn: Discourse-aware graph network for logical reasoning. *arXiv preprint arXiv:2103.14349*.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Lucja Iwańska. 1993. Logical reasoning in natural language: It is all about knowledge. *Minds and Machines*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. Taxinli: Taking a ride up the NLU hill. *CoRR*.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, page 101861.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert Kowalski. 1979. *Logic for problem solving*, volume 7. Ediciones Díaz de Santos.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. 2017. RACE: Large-scale Reading Comprehension dataset from Examinations. In *EMNLP*.

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.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. *arXiv preprint arXiv:2306.03872*.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2020a. Natural language inference in context - investigating contextual reasoning over long texts. *CoRR*.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023a. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023b. Logicot: Logical chain-of-thought instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2908–2921.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3?

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020b. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *CoRR*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

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*.

Bill MacCartney and Christopher D. Manning. 2007a. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Bill MacCartney and Christopher D Manning. 2007b. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

John Mccarthy. 2002. Programs with common sense. 1.

OpenAI. 2023. Gpt-4 technical report.

Graziella Orrù, Andrea Piarulli, Ciro Conversano, and Angelo Gemignani. 2023. Human-like problem-solving abilities in large language models using chatgpt. *Frontiers in Artificial Intelligence*, 6:1199350.

Simon Ott, Konstantin Hebenstreit, Valentin Liévin, Christoffer Egeberg Hother, Milad Moradi, Maximilian Mayrhauser, Robert Praas, Ole Winther, and Matthias Samwald. 2023. Thoughtsource: A central hub for large language model reasoning data. *arXiv preprint arXiv:2301.11596*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

David Poole, Randy Goebel, and Romas Aleliunas. 1987. *Theorist: A Logical Reasoning System for Defaults and Diagnosis*, pages 331–352. Springer New York, New York, NY.

Stephen G. Pulman. 1996. Using the framework.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver?

Swarnadeep Saha, Sayan Ghosh, Shashank Srivastava, and Mohit Bansal. 2020. Prover: Proof generation for interpretable reasoning over rules.

Soumya Sanyal, Harman Singh, and Xiang Ren. 2022. Fairr: Faithful and robust deductive reasoning over natural language.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using ood examples. *arXiv preprint arXiv:2305.15269*.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. Arb: Advanced reasoning benchmark for large language models.

Jihao Shi, Xiao Ding, Li Du, Ting Liu, and Bing Qin. 2021. Neural natural logic inference for interpretable question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3673–3684, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Thinh Hung Truong, Yulia Otmakhova, Timothy Baldwin, Trevor Cohn, Jey Han Lau, and Karin Verspoor. 2022. Not another negation benchmark: The NaN-NLI test suite for sub-clausal negation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 883–894, Online only. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. 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*.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxin Jiao, Yue Zhang, and Xing Xie. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023b. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023c. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*.

Fangzhi Xu, Jun Liu, Qika Lin, Yudai Pan, and Lingling Zhang. 2022. Logiformer. In *Proceedings of the 45th*

International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, , and Johan Bos. 2019. 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)*.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2022. Language models as inductive reasoners.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *Proc. of ICLR*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023a. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023b. Agieval: A human-centric benchmark for evaluating foundation models.

A Datasets Examples

We illustrate data examples mentioned in Section 3 here.

Figure 3 is an example from the LogiQA 2.0 test set. In this example, investigators want to certify the connection between astrological signs and personality. However, the volunteers who attended the program were biased because introverted people are less likely to attend such investigations. This fact flaws the conclusion of the investigation.

B Instructions and Prompts for the Three Tasks

The instructions differ slightly for different datasets, according to their target outputs. **Instruction and Prompt for the Multi-Choice Reading Comprehension Task** `Instructions: You will be presented with a passage and a question about that passage. There are four options to be chosen from, you need to choose the only correct option to answer that question. If the first option is right, you generate the answer ‘A’, if the second option is right, you generate the answer ‘B’, if the third option is right, you generate the answer ‘C’, if the fourth option is right, you generate the answer ‘D’, if the fifth option is right, you generate the answer ‘E’. Read the question and options thoroughly`

Passage: For a television program about astrology, investigators went into the street and found twenty volunteers born under the sign of Gemini who were willing to be interviewed on the program and to take a personality test. The test confirmed the investigators’ personal impressions that each of the volunteers was more sociable and extroverted than people are on average. This modest investigation thus supports the claim that one’s astrological birth sign influences one’s personality.

Question: Which one of the following, if true, indicates the most serious flaw in the method used by the investigators?

A. People born under astrological signs other than Gemini have been judged by astrologers to be much less sociable than those born under Gemini.

B. There is not likely to be a greater proportion of people born under the sign of Gemini on the street than in the population as a whole.

C. People who are not sociable and extroverted are not likely to agree to participate in such an investigation.

D. The personal impressions the investigators first formed of other people have tended to be confirmed by the investigators’ later experience of those people.

Figure 3: A multi-choice reading comprehension example from the LogiQA 2.0 dataset.

Context: A loading dock consists of exactly six bays numbered 1 through 6 consecutively from one side of the dock to the other. Each bay is holding a different one of exactly six types of cargo fuel, grain, livestock, machinery, produce, or textiles. The following apply: The bay holding grain has a higher number than the bay holding livestock. The bay holding livestock has a higher number than the bay holding textiles. The bay holding produce has a higher number than the bay holding fuel. The bay holding textiles is next to the bay holding produce.

Question: Which one of the following CANNOT be the type of cargo held in bay 4?

A. "grain"

B. "livestock"

C. "machinery"

D. "produce"

E. "textiles"

Answer: A

Figure 4: An example from the AR-LSAT dataset.

Premise: Ten new television shows appeared during the month of September. Five of the shows were sitcoms, three were hourlong dramas, and two were news-magazine shows. By January, only seven of these new shows were still on the air. Five of the shows that remained were sitcoms.

Hypothesis: At least one of the shows that were cancelled was an hourlong drama.

Label: Entailment

Figure 5: An NLI example from the ConTRoL dataset.

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

Figure 6: An NLI example from the HELP dataset.

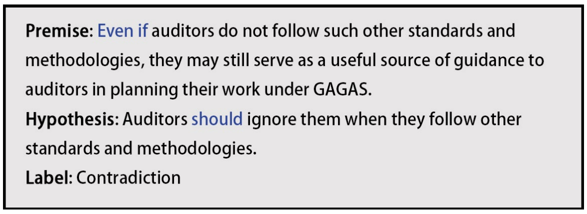

Premise: Even if auditors do not follow such other standards and methodologies, they may still serve as a useful source of guidance to auditors in planning their work under GAGAS.
Hypothesis: Auditors should ignore them when they follow other standards and methodologies.
Label: Contradiction

Figure 7: An NLI example from the TaxiNLI dataset.

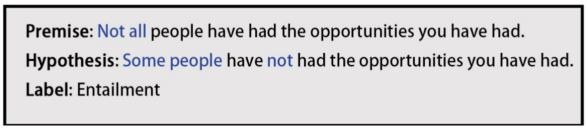

Premise: Not all people have had the opportunities you have had.
Hypothesis: Some people have not had the opportunities you have had.
Label: Entailment

Figure 8: An NLI example from the NAN-NLI dataset.

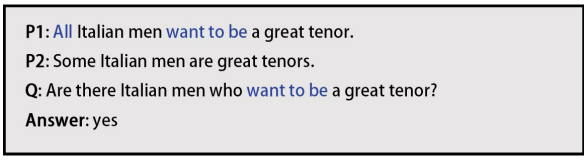

P1: All Italian men want to be a great tenor.
P2: Some Italian men are great tenors.
Q: Are there Italian men who want to be a great tenor?
Answer: yes

Figure 9: An example from the FraCaS dataset.

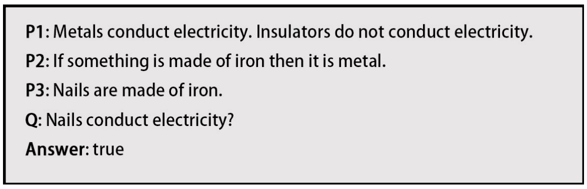

P1: Metals conduct electricity. Insulators do not conduct electricity.
P2: If something is made of iron then it is metal.
P3: Nails are made of iron.
Q: Nails conduct electricity?
Answer: true

Figure 10: An example from the RuleTaker dataset.

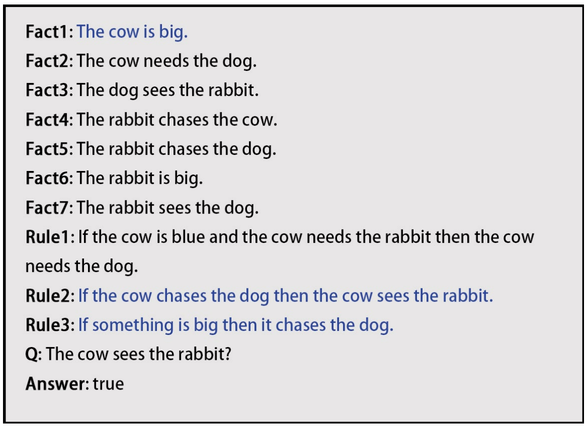

Fact1: The cow is big.
Fact2: The cow needs the dog.
Fact3: The dog sees the rabbit.
Fact4: The rabbit chases the cow.
Fact5: The rabbit chases the dog.
Fact6: The rabbit is big.
Fact7: The rabbit sees the dog.
Rule1: If the cow is blue and the cow needs the rabbit then the cow needs the dog.
Rule2: If the cow chases the dog then the cow sees the rabbit.
Rule3: If something is big then it chases the dog.
Q: The cow sees the rabbit?
Answer: true

Figure 11: An example from the ProofWriter dataset.

and select the correct answer from the four answer labels. Read the passage thoroughly to ensure you know what the passage entails.

Instruction and Prompt for the True-or-False Question Answering Task

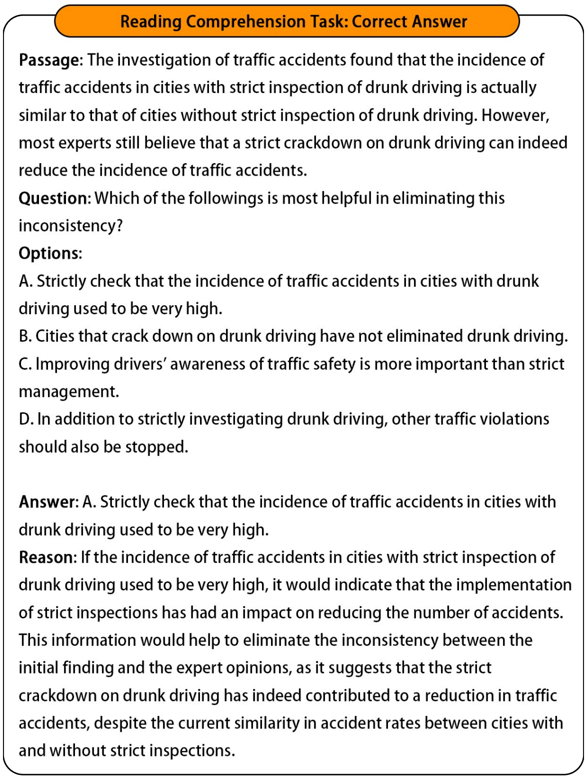
Instructions: You will be presented with a premise and a hypothesis about that premise. You need to decide whether the hypothesis is entailed by the premise by choosing one of the following answers: ‘E’: The hypothesis follows logically from the information contained in the premise. ‘C’: The hypothesis is logically false from the information contained in the premise. ‘N’: It is not possible to determine whether the hypothesis is true or false without further information. Read the passage of information thoroughly and select the correct answer from the three answer labels. Read the premise thoroughly to ensure you know what the premise entails.

Instruction and Prompt for the Natural Language Inference Task

Instructions: You will be presented with a set of facts and rules as premises, and a hypothesis about it. You need to decide whether the hypothesis is entailed by the premise by choosing one of the following answers: ’Yes’: The hypothesis follows logically from the information contained in the premise. ’No’: The hypothesis is logically false from the information contained in the premise. ’Neutral’: It is not possible to determine whether the hypothesis is true or false without further information. Read the passage of information thoroughly and select the correct answer from the three answer labels. Read the premise thoroughly to ensure you know what the premise entails.

C Case Study

Figure 12 gives an example of GPT-4’s answer and reasoning, which GPT-4 answered correctly. According to the passage, experts believe that a strict inspection of drunk driving can reduce traffic accidents, despite the fact that the incidence rate is similar for cities with and without strict checks. To resolve this inconsistency of experts' belief and prove the claim that strict checks lower the incidence of traffic accidents, GPT-4 chooses the new evidence that the drunk driving rate used to be high before strict checks, which can solve this contradiction. Figure 13 gives an example where GPT-4 answered incorrectly. The passage talks about whether humans originated from Africa. The conclusion would be challenged if we find older fossils in other parts of the world. GPT-4 does not have a thorough understanding of the problem. It only focuses on establishing the chronological order of events. The failure to attend to the location of the fossils leads to the wrong answer.



Reading Comprehension Task: Correct Answer

Passage: The investigation of traffic accidents found that the incidence of traffic accidents in cities with strict inspection of drunk driving is actually similar to that of cities without strict inspection of drunk driving. However, most experts still believe that a strict crackdown on drunk driving can indeed reduce the incidence of traffic accidents.

Question: Which of the followings is most helpful in eliminating this inconsistency?

Options:

A. Strictly check that the incidence of traffic accidents in cities with drunk driving used to be very high.

B. Cities that crack down on drunk driving have not eliminated drunk driving.

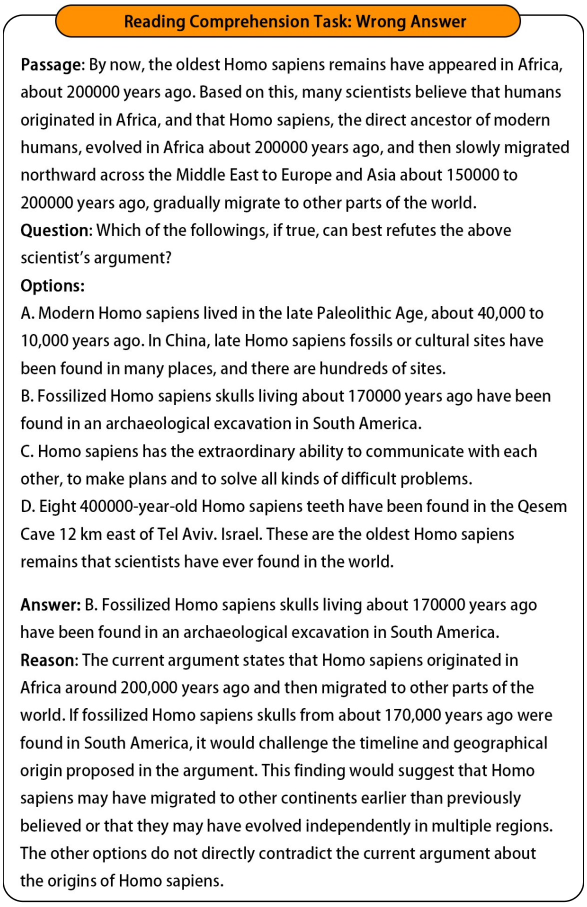
C. Improving drivers' awareness of traffic safety is more important than strict management.

D. In addition to strictly investigating drunk driving, other traffic violations should also be stopped.

Answer: A. Strictly check that the incidence of traffic accidents in cities with drunk driving used to be very high.

Reason: If the incidence of traffic accidents in cities with strict inspection of drunk driving used to be very high, it would indicate that the implementation of strict inspections has had an impact on reducing the number of accidents. This information would help to eliminate the inconsistency between the initial finding and the expert opinions, as it suggests that the strict crackdown on drunk driving has indeed contributed to a reduction in traffic accidents, despite the current similarity in accident rates between cities with and without strict inspections.

Figure 12: An GPT-4 generation example where the model answers the question correctly.



Reading Comprehension Task: Wrong Answer

Passage: By now, the oldest Homo sapiens remains have appeared in Africa, about 200000 years ago. Based on this, many scientists believe that humans originated in Africa, and that Homo sapiens, the direct ancestor of modern humans, evolved in Africa about 200000 years ago, and then slowly migrated northward across the Middle East to Europe and Asia about 150000 to 200000 years ago, gradually migrate to other parts of the world.

Question: Which of the followings, if true, can best refutes the above scientist's argument?

Options:

A. Modern Homo sapiens lived in the late Paleolithic Age, about 40,000 to 10,000 years ago. In China, late Homo sapiens fossils or cultural sites have been found in many places, and there are hundreds of sites.

B. Fossilized Homo sapiens skulls living about 170000 years ago have been found in an archaeological excavation in South America.

C. Homo sapiens has the extraordinary ability to communicate with each other, to make plans and to solve all kinds of difficult problems.

D. Eight 400000-year-old Homo sapiens teeth have been found in the Qesem Cave 12 km east of Tel Aviv. Israel. These are the oldest Homo sapiens remains that scientists have ever found in the world.

Answer: B. Fossilized Homo sapiens skulls living about 170000 years ago have been found in an archaeological excavation in South America.

Reason: The current argument states that Homo sapiens originated in Africa around 200,000 years ago and then migrated to other parts of the world. If fossilized Homo sapiens skulls from about 170,000 years ago were found in South America, it would challenge the timeline and geographical origin proposed in the argument. This finding would suggest that Homo sapiens may have migrated to other continents earlier than previously believed or that they may have evolved independently in multiple regions. The other options do not directly contradict the current argument about the origins of Homo sapiens.

Figure 13: An GPT-4 generation example where the model answers the question incorrectly.