Appendix for "Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond"

Fangzhi Xu*, Qika Lin*, *Member, IEEE,* Jiawei Han, Tianzhe Zhao,
Jun Liu, *Senior Member, IEEE,* Erik Cambria, *Fellow, IEEE*

Abstract—This supplementary material accompanies the paper titled "*Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond*". The following sections firstly provide comprehensive information about the seven LLMs analyzed in this study. Additionally, this file includes a detailed description of the evaluation methodology, prompt engineering techniques, NeuLR dataset, and supplementary analyses. Finally, the case studies for each dataset are presented. Together, these sections provide a comprehensive and detailed account of the methods and results of the study.

1 Details of Large Language Models

In this study, we select three previously representative LLMs (i.e., text-davinci-003, ChatGPT and BARD) and four up-to-date LLMs (i.e., LLaMA3.1-Chat, Mistral-Instruct-v0.3, Claude-3 and GPT-4) for evaluation and analysis. The specifics are itemized in Table 1, which includes the affiliation, usage charge, pre-trained data, and model size of each LLM.

TABLE 1: Details of the selected LLMs. *Affi.* is short for *Affiliation*. *Charge* represents the charges for 1M tokens in the format of (input charge / output charge). *Data* is the latest time of the utilized training data. *B* in the last column represents the one billion.

Model	Affi.	Charge	Data	Size
text-davinci-003	Open-AI	20$	Sep. 2021	-
ChatGPT	Open-AI	2$	Jun. 2021	-
BARD	Google	-	Not report	540B
LLaMA3.1-Chat	Meta	-	2023	8B
Mistral-Inst-v0.3	Mistral	-	2023	7B
Claude-3	Anthropic	3$/15$	2023	-
GPT-4	Open-AI	30$/60$	2023	-

2 Details of Evaluations

This section supplements some details of the evaluation.

2.1 Evaluated Datasets

Considering the huge annotation load, we make a balance between the comprehensiveness and annotation cost. We try to keep the full data of each dataset to evaluate ChatGPT, except for some large-scale datasets, i.e., RuleTaker, α-NLI, α-NLG, AbductiveRules, D*-Ab and LogiQA2NLI. We limit the number of evaluation samples in these datasets to 1000, 1000, 1000, 1000, 1000 and 600 respectively. For text-davinci-003 and BARD, we randomly sample 100 examples for each dataset. It is still more comprehensive than previous works, that only utilize dozens of samples for testing (e.g., 30 samples for each dataset).

Notably, as for the supplemented experiments on the four latest LLMs (LLaMA3.1-Chat, Mistral-Instruct-v0.3, Claude-3 and GPT-4), we randomly sample 100 cases for each dataset. It is expected to maintain a fair comparison.

For the mixed-form reasoning datasets, we calculate the fact length distributions of four datasets for reference. It is observed that most samples have 1-2 or 3-4 facts for reasoning.

TABLE 2: Fact length distributions (%) of mixed-form datasets.

Model	Length				
	1-2	3-4	5-6	7-8	$\geq$ 9
ReClor	28.20	61.60	9.00	1.00	0.20
LogiQA	76.65	13.36	7.22	1.84	0.92
LogiQA 2.0	39.40	45.80	12.60	1.60	0.60
LogiQA2NLI	34.94	43.56	16.95	3.07	1.46

2.2 Evaluation Modes

As mentioned in the main paper, we introduce four metrics to evaluate the LLM performances, i.e., *Answer Correctness*, *Explain Correctness*, *Explain Completeness* and *Explain Redundancy*. Also, we employ five metrics to attribute the error cases. Although these metrics are subjective, their standards are clear and well-defined.

As for the three previously representative LLMs, their evaluations are conducted through manual annotation. In detail, we output the LLM answers through API calling and hire ten graduate students (all majoring in Natural Language Processing) to annotate the answers based on our metric definitions.

As for the four trending LLMs, their evaluations are under the **automatic mode**. Specifically, we utilize GPT-4o model to act as the LLM judge. The definitions of the metrics are fed to the LLM judge as instructions, and it outputs the evaluation scores the same as the human annotators.

3 Supplementary Results on Latest LLMs

First of all, we conduct the pilot experiments to demonstrate the consistency between the manual labeling and GPT-4o. The analysis is conducted under the four evaluation metrics (i.e., answer correctness, explanation correctness, explanation completeness, and explanation redundancy) and the two error categories (i.e., evidence selection and reasoning errors). Fig. 1 shows the results.

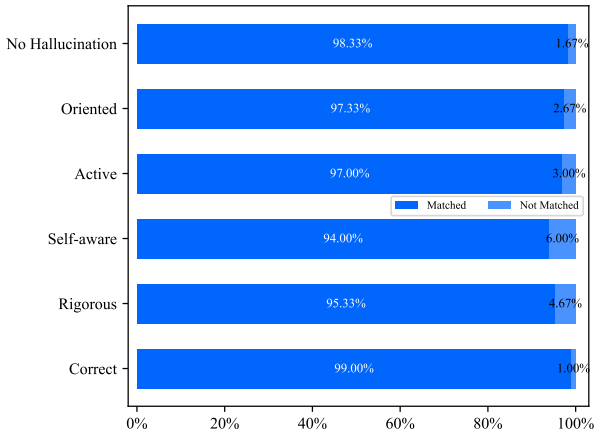


Fig. 1: The evaluation matches between LLM judge and human annotations on the 300 samples.

It illustrates that LLM-as-judge can well align with human performance in our evaluation settings.

Next, we implement the **automatic evaluations** on the four trending LLMs (i.e., LLaMA3.1-Chat-8B, Mistral-Instruct-v0.2-7B, Claude-3, and GPT-4). Fig. 2 and 3 provide the respective results on the evaluation of six dimensions.

4 Prompt Engineering

In the implementation, we consider both the zero-shot and few-shot settings. To obtain appropriate model outputs, we construct the prompts, shown in Table 5 of Appendix.

5 NeuLR dataset

In this section, we will provide more details on how NeuLR is constructed and what is the form of the prompt.

5.1 Construction of NeuLR

According to the explanations in the main paper, current logical reasoning may fail to keep the neutral content for evaluation. Therefore, we propose the new dataset NeuLR to benchmark the pure logical reasoning capability of LLMs. The data is sourced from [1][9]. Detailedly, we transform some of the important content (e.g., entities and properties) into random strings. In NeuLR, we neutralize three types of words, i.e., name, species and property. We replace each type of words with the concatenation of a prefix (i.e., NP, SP, ADP) and six random characters (combination of uppercase and lowercase letters and numbers). Table 4 illustrates examples of the replacement process.

TABLE 4: The construction of neutral content. *Example* column presents some example words. *Prefix* represents the prefix for each type of words. *6-character* represents examples of randomly generated combination of characters.

Type	Example	Prefix	6-character
Name	Bob/Lily	NP	e.g., uF52pT
Species	Dog/Sheep	SP	e.g., 7gfO2k
Property	Big/Red/Smart	ADP	e.g., PT01mx

5.2 Prompt Forms

Different from previous evaluations, we consider adding descriptions of content neutrality in the prompt forms. Specifically, we replace the previous task description prefix with the following form (take abductive reasoning for illustration):

- This is a neutral-content abductive reasoning task. The strings starting with SP represent species. The strings starting with NP represent names. The strings starting with ADP represent property. Given a context and a fact, it is required to generate a short missing fact.

Other parts of the prompts are the same as the previous form.

Further, we consider both zero-shot setting, few-shot setting and chain-of-thought strategy in NeuLR. The zero-shot and few-shot prompt forms are same as the forms in TABLE 5. For chain-of-thought prompt, we utilize the explanation provided with the dataset to help generate the reasoning chains. Next, we will offer examples for each reasoning setting.

- Deductive Reasoning

[chain-of-thought]: There is one example of deductive reasoning: The facts are: SP1Ggwz1 are afraid of SPU4g85d. NPdmbdKB is a SP1Ggwz1. SPArMPn0 are afraid of SPU4g85d. SPU4g85d are afraid of SPArMPn0. NPd0WTGi is a SP1Ggwz1. NP1cTSnm is a SPU4g85d. SPFURodYs are afraid of SP1Ggwz1. NPpEQFEK is a SPU4g85d. The question is: What is NPpEQFEK afraid of? Because NPpEQFEK is a SPU4g85d and SPU4g85d are afraid of SPArMPn0. Therefore, the answer is: SPArMPn0.

- Inductive Reasoning

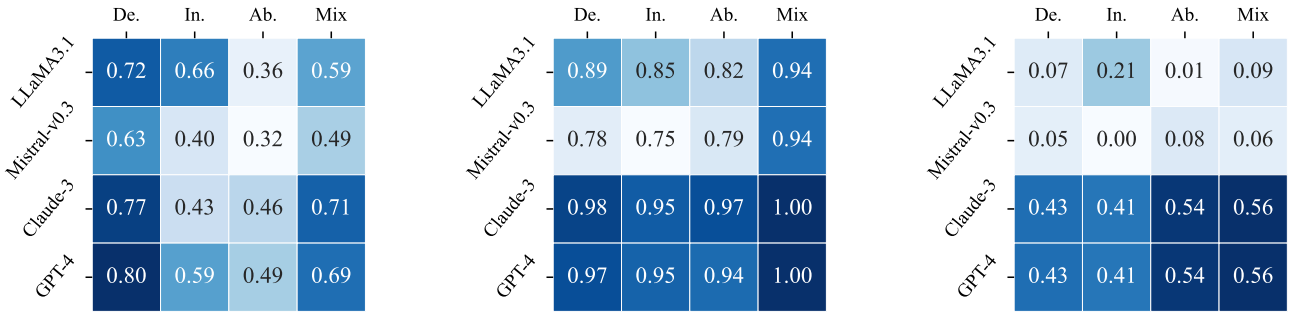
[chain-of-thought]: There is one example of inductive reasoning: The facts are: NP16WBQQ is a SPQFyx7i. NPerfGLN is a SP4Vge77. NP0V3buK is a SPGbXv1C. NPU9kOFg is a SPQFyx7i. NPDHuODI is a SP4Vge77. NPerfGLN is ADPHQuIP9. NP0V3buK is ADPmBU2ts. NP16WBQQ is ADP0rI59W. NPDHuODI is ADPHQuIP9.The question is: What property is NPU9kOFg? Because NPU9kOFg is a SPQFyx7i, NP16WBQQ is a SPQFyx7i and NP16WBQQ is ADP0rI59W. Therefore, the answer is: ADP0rI59W.

- Abductive Reasoning

[chain-of-thought]: ... Based on rule2 and tripleM, we can derive a new fact. Combine it with triple4. Based on rule1, we can derive a new fact. Combine it with rule1. In this way, we can finally derive the given fact...

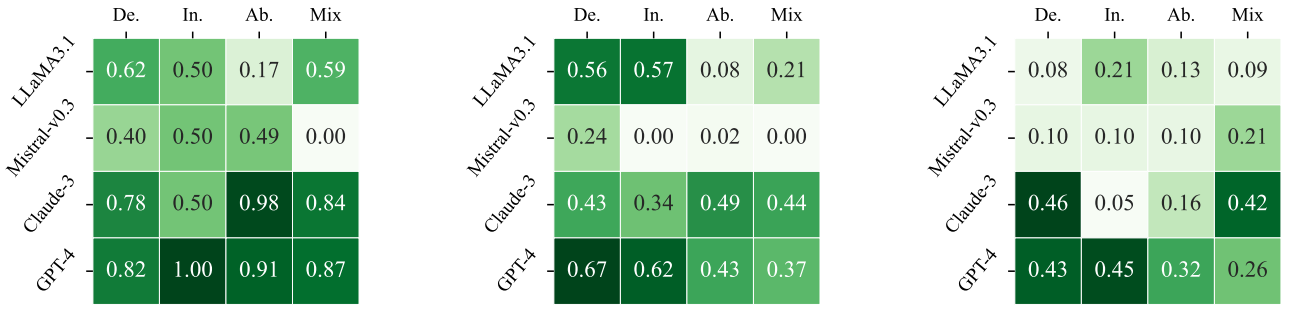
6 Supplementary Analysis

We also provide the performances of text-davinci-003 and BARD with different number of hops in Figure 4. In the



(a) Correctness. (b) Rigorousness. (c) Self-awareness.

Fig. 2: Heatmap results for the correctness, rigorousness, and self-awareness metrics.



(a) Active. (b) Oriented. (c) No hallucination.

Fig. 3: Heatmap results for the activity, orientation and no-hallucination of LLMs.

deductive setting of text-davinci-003, model performance drops with the number of hops increasing. But when the hop number is over five, it witnesses slight gains in performance, which illustrates that text-davinci-003 has the potential to conduct multi-hop reasoning in the deductive reasoning setting. However, in the inductive and abductive settings, the performances of text-davinci-003 decrease sharply when the number of hops increases. Especially, it fails all the cases when the hop number is over six in inductive reasoning, and also when the hop number is greater than one in abductive reasoning. It is inferior to ChatGPT.

For BARD, the situation is quite different. In deductive reasoning, the performance of BARD increases with the hop number adds. Especially, when the hop number is over five, the accuracy reaches 100% without reducing the rigor of reasoning. In inductive reasoning, the performance of BARD also drops at first, but it keeps stable and witnesses obvious gains when the hop number is over six. It demonstrates that BARD is better at conducting inductive reasoning and processing multi-hop scenarios compared with text-davinci-003 and ChatGPT. In abductive reasoning, BARD struggles a lot, inferior to ChatGPT but is better than text-davinci-003.

In all, in the face of complex multi-hop scenarios, LLMs still have much room for improvement. From the results, they do relatively well in deductive reasoning settings. But they are far from good in the inductive and abductive settings, which can also inspire future research on it.

7 Case Studies

We show one reasoning case for each dataset in Table 5-19, where the context and question as well as output of 0-shot ChatGPT, 1-shot ChatGPT, 3-shot ChatGPT, 0-shot Davinci-003 and 0-shot BARD are displayed. We also provided the annotated information about the answer correctness: ✅, explain correctness: ✅, explain completeness: ✅, explain redundancy: ✅, evidence wrong selection: ☑, hallucination: ☑, no reasoning: ☑, perspective mistake: ☑, and process mistake: ☑. (the last five indicators are annotated when the explain explanation is false).

References

[1] J. Weston, A. Bordes, S. Chopra, and T. Mikolov, "Towards ai-complete question answering: A set of prerequisite toy tasks," in *4th International Conference on Learning Representations (ICLR)*, 2016.

[2] B. Dalvi, P. Jansen, O. Tafjord, Z. Xie, H. Smith, L. Pipatanangkura, and P. Clark, "Explaining answers with entailment trees," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2021, pp. 7358–7370.

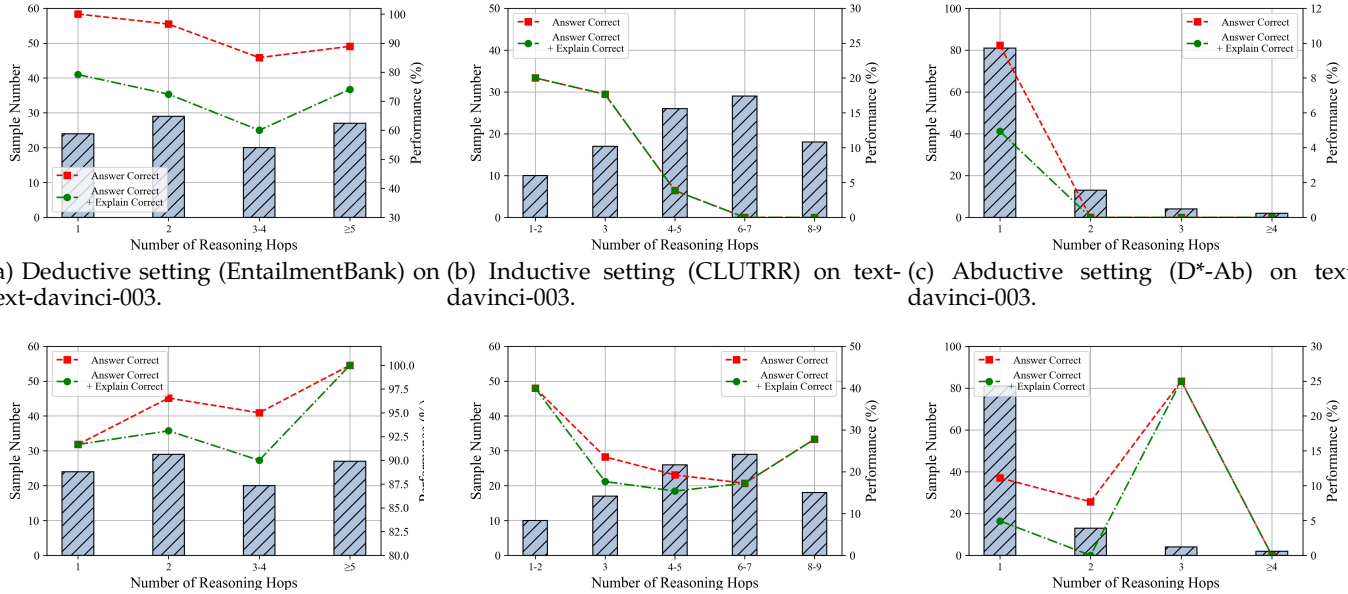
[3] P. Clark, O. Tafjord, and K. Richardson, "Transformers as soft reasoners over language," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 2020, pp. 3882–3890.

[4] S. Han, H. Schoelkopf, Y. Zhao, Z. Qi, M. Riddell, L. Benson, L. Sun, E. Zubova, Y. Qiao, M. Burtell, D. Peng, J. Fan, Y. Liu, B. Wong, M. Sailor, A. Ni, L. Nan, J. Kasai, T. Yu, R. Zhang, S. R. Joty, A. R. Fabbri, W. Kryscinski, X. V. Lin, C. Xiong, and D. Radev, "FOLIO: natural language reasoning with first-order logic," *CoRR*, vol. abs/2209.00840, 2022.

[5] A. Talmor, O. Tafjord, P. Clark, Y. Goldberg, and J. Berant, "Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

TABLE 3: Evaluated Datasets. *Gen.* distinguishes whether the predicted answer is generated text or classified labels. *Explain* denotes whether the explanation is required in the task. # Davinci, # ChatGPT and # BARD columns represent the number of evaluation samples of each dataset for the three LLMs.

Categories	Dataset	Source	Gen.	Explain	# Davinci	# ChatGPT	# BARD
Deductive	**bAbI-15**	[1]	✓	✓	100	1,000	100
	EntailmentBank	[2]	✓		100	340	100
	RuleTaker	[3]		✓	100	1,000	100
	FOLIO	[4]			100	204	100
	Leap-Of-Thought	[5]			100	1,289	100
Inductive	**bAbI-16**	[1]	✓		100	1,000	100
	CLUTRR	[6]	✓		100	1,146	100
Abductive	**α-NLI**	[7]			100	1,000	100
	α-NLG	[7]	✓		100	1,000	100
	AbductiveRules	[8]	✓		100	1,000	100
	D*-Ab	[9]	✓	✓	100	1,000	100
mixed-form	**ReClor**	[10]			100	500	100
	LogiQA	[11]			100	651	100
	LogiQA 2.0	[12]			100	500	100
	LogiQA2NLI	[12]			100	600	100



(a) Deductive setting (EntailmentBank) on text-davinci-003. (b) Inductive setting (CLUTRR) on text-davinci-003. (c) Abductive setting (D*-Ab) on text-davinci-003.

(d) Deductive setting (EntailmentBank) on BARD. (e) Inductive setting (CLUTRR) on BARD. (f) Abductive setting (D*-Ab) on BARD.

Fig. 4: The performances of text-davinci-003 and BARD under different number of hops. Comparison of deductive, inductive and abductive reasoning settings.

[6] K. Sinha, S. Sodhani, J. Dong, J. Pineau, and W. L. Hamilton, "CLUTRR: A diagnostic benchmark for inductive reasoning from text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019, pp. 4505–4514.

[7] C. Bhagavatula, R. L. Bras, C. Malaviya, K. Sakaguchi, A. Holtzman, H. Rashkin, D. Downey, W. Yih, and Y. Choi, "Abductive commonsense reasoning," in *8th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2020.

[8] N. Young, Q. Bao, J. Bensemann, and M. Witbrock, "Abductionrules: Training transformers to explain unexpected inputs," in *Findings of the Association for Computational Linguistics*. Association for Computational Linguistics, 2022, pp. 218–227.

[9] O. Tafjord, B. Dalvi, and P. Clark, "Proofwriter: Generating implications, proofs, and abductive statements over natural language," in *Findings of the Association for Computational Linguistics*. Association for Computational Linguistics, 2021, pp. 3621–3634.

[10] W. Yu, Z. Jiang, Y. Dong, and J. Feng, "Reclor: A reading comprehension dataset requiring logical reasoning," in *8th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2020.

[11] J. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, and Y. Zhang, "Logiqa: A challenge dataset for machine reading comprehension with logical reasoning," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*. ijcai.org, 2020, pp. 3622–3628.

[12] H. Liu, R. Ning, Z. Teng, J. Liu, Q. Zhou, and Y. Zhang, "Evaluating the logical reasoning ability of chatgpt and GPT-4," *CoRR*, vol. abs/2304.03439, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.03439

TABLE 5: Prompt Engineering.

Dataset	Prompt of zero-shot	Prompt of k-shot
Deductive Reasoning		
bAbI (task 15)	[zero-shot prompt]: Given facts: [Context]. Based on the given facts above, answer the following question using deductive reasoning and give simple explanations. The question is: [Question]	There are [k] examples of deductive reasoning: Given facts: [Context] The question is: [Question] The answer is: [Answer] (display k samples) [zero-shot prompt]
EntailmentBank	[zero-shot prompt]: Given facts: [Context]. [Question]. Please answer the question in one sentence using deductive reasoning. And give simple explanations.	There are [k] examples of deductive reasoning: Given facts: [Context] The question is: [Question] The answer is: [Answer] (display k samples) [zero-shot prompt]
RuleTaker	[zero-shot prompt]: Given facts: [Context]. Based on the given facts above, determine whether the following statement is true using deductive reasoning and give simple explanations. The statement is: [Statement].	There are [k] examples of deductive reasoning: Given facts: [Context] The statement is: [Statement] The answer is: [Answer] (display k samples) [zero-shot prompt]
FOLIO	[zero-shot prompt]: Given facts: [Context]. Based on the given facts above, determine whether the following statement is true, false, or uncertain using deductive reasoning and give simple explanations. The statement is: [Statement].	There are [k] examples of deductive reasoning: Given facts: [Context] The statement is: [Statement] The answer is: [Answer] (display k samples) [zero-shot prompt]
Leap-Of-Thought	[zero-shot prompt]: Given facts: [Context]. Based on the given facts above, determine whether the following statement is true using deductive reasoning and give simple explanations. The statement is: [Statement].	There are [k] examples of deductive reasoning: Given facts: [Context] The statement is: [Statement] The answer is: [Answer] (display k samples) [zero-shot prompt]
Inductive Reasoning		
bAbI-16	[zero-shot prompt]: Given facts: [context]. Based on the given facts above, answer the following question using inductive reasoning and give simple explanations. The question is: [Question].	There are [k] examples of inductive reasoning: Given facts: [Context] The question is: [Question] The answer is: [Answer] (display k samples) [zero-shot prompt]
CLUTRR	[zero-shot prompt]: Given facts: [context]. [Question]. Please answer the question in one sentence using inductive reasoning. And give simple explanations.	There are [k] examples of inductive reasoning: Given facts: [Context] The question is: [Question] The answer is: [Answer] (display k samples) [zero-shot prompt]
Abductive Reasoning		
α-NLI	[task description]: Given a context, the abductive reasoning task is to choose the more likely explanation from a given pair of hypotheses choices. And give simple explanations. [zero-shot prompt]: The context is: [Context]. The hypothesis choices are: A. [Option A]. B. [Option B].	[task description] Next, I will give you [k] example(s) for test. The context is [Context]. The hypothesis choice are: A. [Option A]. B. [Option B]. The correct choice is: [Label]. Next, I will give you an example for test. [zero-shot prompt]
α-NLG	[task description]: Given a context, the abductive reasoning task is to generate a valid and short hypothesis. [zero-shot prompt]: The context is: [Context]. Please generate a short hypothesis for the context and give simple explanations.	[task description] Next, I will give you [k] example(s) for test. The context is [Context]. The correct answer is [Label]. Next, I will give you an example for test. [zero-shot prompt]

AbductiveRules	[task description]: Given a context and an observation, the abductive reasoning task is to generate a valid and short explanation. [zero-shot prompt]: The context is: [Context]. The observation is: [Observation]. Please generate a short explanation for the given context and observation.	[task description] Next, I will give you [k] example(s) for test. The context is: [Context]. The observation is: [Observation]. The explanation is: [Explanation]. Next, I will give you an example for test. [zero-shot prompt]
D*-Ab	[task description]: Given a context and a fact, the abductive reasoning task is to generate a short missing fact. [zero-shot prompt]: The context is: [Context+Rule]. The fact is: [Fact]. Please generate a short missing fact for the given context and fact. And give simple explanations.	[task description] Next, I will give you [k] example(s) for test. The context is: [Context+Rule]. The observation is: [Observation]. The explanation is: [Explanation]. Next, I will give you an example for test. [zero-shot prompt]
mixed-form Reasoning		
ReClor	[task description]: This is a Mahcine Reading Comprehension task, given the context and question, you are required to choose the correct answer from the answer set and give explanations. [zero-shot prompt]: The context is: [Context]. The question is: [Question]. [Option A]. [Option B]. [Option C]. [Option D]. Please give the correct answer and simple explanations.	[task description] Next, I will give you [k] example(s) for test. The context is: [Context]. The question is: [Question]. The correct choice is: [Label]. Next, I will give you an example for test. [zero-shot prompt]
LogiQA	[task description]: This is a Mahcine Reading Comprehension task, given the context and question, you are required to choose the correct answer from the answer set and give explanations. [zero-shot prompt]: The context is: [Context]. The question is: [Question]. [Option A]. [Option B]. [Option C]. [Option D]. Please give the correct answer and simple explanations.	[task description] Next, I will give you [k] example(s) for test. The context is: [Context]. The question is: [Question]. The correct choice is: [Label]. Next, I will give you an example for test. [zero-shot prompt]
LogiQA 2.0	[task description]: This is a Mahcine Reading Comprehension task, given the context and question, you are required to choose the correct answer from the answer set and give explanations. [zero-shot prompt]: The context is: [Context]. The question is: [Question]. [Option A]. [Option B]. [Option C]. [Option D]. Please give the correct answer and simple explanations.	[task description] Next, I will give you [k] example(s) for test. The context is: [Context]. The question is: [Question]. The correct choice is: [Label]. Next, I will give you an example for test. [zero-shot prompt]
LogiQA2NLI	[task description]: This is a Natural Language Inference task. Please tell whether the premise and conclusion entailed or NOT entailed. And give simple explanations. [zero-shot prompt]: The premise is: [Premise]. The conclusion is: [Conclusion]. Please give the correct answer and simple explanations.	[task description] Next, I will give you [k] example(s) for test. The premise is: [Premise]. The context is: [Context]. The correct answer is: [Label]. Next, I will give you an example for test. [zero-shot prompt]

TABLE 6: A reasoning case in bAbI-15.

Question, answer and explanation	Context: 1 Sheep are afraid of wolves. 2 Mice are afraid of wolves. 3 Winona is a sheep. 4 Gertrude is a sheep. 5 Wolves are afraid of sheep. 6 Emily is a sheep. 7 Cats are afraid of sheep. 8 Jessica is a sheep. Question: What is jessica afraid of? Answer: Wolf Explanation: IsSheep(jessica) (8)$\xrightarrow{1}$ Afraid(jessica, wolf)
ChatGPT of 0-shot	Jessica is afraid of nothing, as it is not specified in the given facts that she has any known natural predator or fear. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 1-shot	Jessica is a sheep and sheep are afraid of wolves. Therefore, it can be deduced that Jessica is afraid of wolves. Answer and Explain: ✓ ✓ ✓ ✗
ChatGPT of 3-shot	Sheep are afraid of wolves and Jessica is a sheep. Answer and Explain: ✗ ✗ ✓ ✗ Explain Error Type: ☒ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	Jessica is likely afraid of wolves, since wolves are a common source of fear that other animals, including sheep, mice, and cats, have in common. Wolves are known to be predators of sheep, so Jessica being a sheep, would likely be scared of wolves. Answer and Explain: ✓ ✗ ✓ ✓ Explain Error Type: ☑ ☑ ☒ ☑ ☒
Bard of 0-shot	Jessica is a sheep. Sheep are afraid of wolves. Therefore, Jessica is afraid of wolves. Here is the deductive reasoning: 1. Sheep are afraid of wolves. 2. Jessica is a sheep. 3. Therefore, Jessica is afraid of wolves. The first premise is supported by the fact that sheep are prey animals and wolves are predators. Wolves are known to prey on sheep, and sheep have evolved to be afraid of wolves as a way of self-preservation. The second premise is supported by the fact that Jessica is a sheep. This is a given fact in the prompt. The third premise follows logically from the first two premises. If sheep are afraid of wolves, and Jessica is a sheep, then Jessica must be afraid of wolves. Therefore, the conclusion is that Jessica is afraid of wolves. Answer and Explain: ✓ ✓ ✓ ✓

TABLE 7: A reasoning case in EntailmentBank.

Question, answer and explanation	Context: 1 pollination is when pollinating animals / wind / water carry pollen from one flower to another flower. 2 if something is a part of a process then that something is required for that process. 3 helping something has a positive impact on that something. 4 if something is required for a process then that something positively impacts that process. 5 an insect is a pollinating animal. Question: life cycle of some plants includes the process of pollination. Pollination is helped by? Answer: wind, water, and insects Explanation: sent1 & sent5→int1: pollination is when insects / wind / water carry pollen from one flower to another; int1 & sent2→int2: insects / wind / water carrying pollen is required for pollination; int2 & sent4→int3: insects / wind / water have a positive impact on pollination by carrying pollen; int3 & sent3→hypothesis
ChatGPT of 0-shot	Pollination in the life cycle of some plants is helped by pollinating animals, such as insects, which are required for the process of pollination and have a positive impact on it. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 1-shot	Pollination in plants is helped by pollinating animals such as insects, which carry pollen from one flower to another as a necessary part of the process, positively impacting the plants' life cycle. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 3-shot	Pollinating animals such as insects, which are required for the process of pollination to occur. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	Pollination is helped by pollinating animals, such as insects, carrying pollen from one flower to another. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Bard of 0-shot	Sure, here is the answer to the question in one sentence using deductive reasoning: Pollination is helped by pollinating animals, such as insects. Here is the simple explanation: Pollination is the process of transferring pollen from the male part of a flower to the female part of a flower. Pollinating animals, such as insects, help pollination by carrying pollen from one flower to another. Without pollination, plants cannot reproduce and produce seeds. Therefore, pollinating animals help the life cycle of some plants by helping them to reproduce. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☒ ☑ ☒ ☒

TABLE 8: A reasoning case in RuleTaker.

Question, answer and explanation	Context: 1 If the current runs through the bell then the bell is ringing. 2 The circuit includes the battery. 3 If the circuit does not include the switch then the circuit is complete. 4 The circuit includes the bell. 5 If the circuit includes the switch and the switch is on then the circuit is complete. 6 If the current runs through the light bulb then the light bulb is glowing. 7 If the wire is plastic then the wire is not conducting. 8 The circuit includes the switch. 9 If the circuit includes the battery and the battery is not flat then the circuit is powered. 10 If the current runs through the circuit and the circuit includes the bell then the current runs through the bell. 11 The wire is metal. 12 If the wire is metal then the wire is conducting. 13 If the current runs through the circuit and the circuit includes the light bulb then the current runs through the light bulb. 14 If the current runs through the radio then the radio is playing. 15 If the current runs through the circuit and the circuit includes the radio then the current runs through the radio. 16 If the circuit is powered and the circuit is complete and the wire is conducting then the current runs through the circuit. Statement: The current does not run through the circuit. Question: Determine whether the statement is true using deductive reasoning and give explanations. Answer: True
ChatGPT of 0-shot	The circuit is powered, complete, and the wire is conducting. Therefore, we can use deductive reasoning to determine that the statement "The current does not run through the circuit" is false based on the given facts. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 1-shot	If the battery is not flat and the circuit includes the battery, then the circuit is powered: "If the circuit includes the battery and the battery is not flat then the circuit is powered." If the circuit is powered and the wire is metal (which is conducting), then the current runs through the circuit: "If the circuit is powered and the circuit is complete and the wire is conducting then the current runs through the circuit." However, it is stated that the wire is plastic and if the wire is plastic, then it is not conducting: "If the wire is plastic then the wire is not conducting." Therefore, the statement "The current does not run through the circuit" is true. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 3-shot	This statement is false based on the given facts. If the circuit is powered, complete, and the wire is conducting, then the current runs through the circuit, as stated in the first premise. Hence, we can't conclude that the current doesn't run through the circuit. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	Pollination is helped by pollinating animals, such as insects, carrying pollen from one flower to another. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Bard of 0-shot	Sure, here is the answer to the question in one sentence using deductive reasoning: Pollination is helped by pollinating animals, such as insects. Here is the simple explanation: Pollination is the process of transferring pollen from the male part of a flower to the female part of a flower. Pollinating animals, such as insects, help pollination by carrying pollen from one flower to another. Without pollination, plants cannot reproduce and produce seeds. Therefore, pollinating animals help the life cycle of some plants by helping them to reproduce. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☒ ☑ ☒ ☒

TABLE 9: A reasoning case in FOLIO.

Question, answer and explanation	Context: 1 All growth companies' stocks are volatile. 2 If the stock price is volatile, then it is not suitable for a retirement fund. 3 Some companies' stocks are growth companies' stocks. 4 All mature companies' stocks are suitable for a retirement fund. 5 KO is a mature company's stock. Statement: KO is a company stock and a growth company's stock. Question: Determine whether the statement is true, false, or uncertain using deductive reasoning and give explanations. Answer: False Explanation: MatureCompanies'Stocks(KO) (5)$\xrightarrow{4}$ SuitableForRetirementRund(KO) $\xrightarrow{2r}\neg$PriceVolatile(KO)$\xrightarrow{1r}$ $\neg$GrowthCompanies'Stocks(KO) (contradictory)
ChatGPT of 0-shot	A mature company's stock cannot also be a growth company's stock since these are two distinct categories. The given facts say that some companies' stocks are growth companies' stocks and all mature companies' stocks are suitable for a retirement fund. Additionally, KO is specifically identified as a mature company's stock. Therefore, the statement "KO is a company stock and a growth company's stock" is false. Answer and Explain: ✓ ✗ ✗ ✓ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 1-shot	We know from the given facts that KO is "a mature company's stock," but there is no information to suggest that it is also a growth company's stock. Therefore, the statement "KO is a company stock and a growth company's stock" is uncertain based on the given facts. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 3-shot	The given facts state that KO is a mature company's stock. However, it is not stated whether KO is also a growth company's stock. Therefore, the statement "KO is a company stock and a growth company's stock" is uncertain. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	False. While it is true that KO is a company stock, it is not true that it is a growth company's stock. Growth companies' stocks are volatile, whereas the fact given states that mature companies' stocks are suitable for retirement funds. Since KO is a mature company's stock, it follows that it is not a growth company's stock. Answer and Explain: ✓ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Bard of 0-shot	The statement "KO is a company stock and a growth company's stock" is false. Explanation: 1. All growth companies' stocks are volatile. This means that the stock price of a growth company is likely to fluctuate more than the stock price of a mature company. 2. If the stock price is volatile, then it is not suitable for a retirement fund. This is because retirement funds are designed to provide long-term growth, and volatile stocks are more likely to lose value over time. 3. Some companies' stocks are growth companies' stocks. This means that some companies are growing at a faster rate than others. 4. All mature companies' stocks are suitable for a retirement fund. This is because mature companies are less likely to experience sudden changes in their stock price. 5. KO is a mature company's stock. This means that KO is a company that is not growing at a rapid pace. Based on the above facts, we can conclude that KO is not a growth company's stock. This is because KO is a mature company, and mature companies' stocks are not volatile. Therefore, the statement "KO is a company stock and a growth company's stock" is false. Answer and Explain: ✓ ✓ ✗ ✗

TABLE 10: A reasoning case in Leap-Of-Thought.

Question, answer and explanation	Context: 1 A bird is not capable of express feelings. 2 A quack-quack is not capable of suffering dehydration. 3 A privet is capable of suffering dehydration. 4 A music is capable of express feelings. Statement: A jazz is capable of express feelings. Question: Determine whether the statement is true using deductive reasoning and give explanations. Answer: True Explanation: IsMusic(jazz) (common sense)$\xrightarrow{4}$ ExpressFeeling(jazz)
ChatGPT of 0-shot	The statement "A jazz is capable of expressing feelings" can be inferred with a certain level of probability based on the given facts. The fact that "A music is capable of expressing feelings" suggests that jazz, which is a genre of music, is likely to be capable of expressing feelings as well. The fact that "A bird is not capable of expressing feelings" and "A quack-quack is not capable of suffering dehydration" are logically irrelevant to the statement about whether jazz is capable of expressing feelings. The fact that "A privet is capable of suffering dehydration" is also irrelevant, as it does not provide any information or indication one way or the other about the statement. Therefore, based on the given facts, the statement about jazz being capable of expressing feelings can be inferred with a certain level of probability and is likely true. Answer and Explain:
ChatGPT of 1-shot	The statement "A jazz is capable of express feelings" is true using deductive reasoning based on the given facts. Answer and Explain: Explain Error Type:
ChatGPT of 3-shot	This is true, as the given fact "a music is capable of express feelings" includes the subcategory of music that is jazz. Therefore, we can logically deduce that jazz music is also capable of expressing feelings. Answer and Explain:
Davinci-003 of 0-shot	True. Although humans usually associate the sound of a quack-quack with birds, the quack-quack mentioned here is not a bird and is not capable of expressing feelings. Additionally, a privet is capable of suffering dehydration but not expressing feelings. Therefore, the only species that has been mentioned in the facts that is capable of expressing feelings is the music. Although a jazz is not explicitly mentioned in the facts, it would also be categorically considered to be a type of music, so it can logically be assumed that a jazz is capable of expressing feelings. Answer and Explain:
Bard of 0-shot	The statement "A jazz is capable of express feelings" is not true. The fact that a music is capable of expressing feelings does not mean that a jazz is capable of expressing feelings. A jazz is a type of music, but it is not the only type of music. There are many other types of music, and not all of them are capable of expressing feelings. For example, classical music is often seen as being more formal and less emotional than jazz. Therefore, it is not possible to say that all music is capable of expressing feelings, simply because one type of music is capable of doing so. In order to determine whether a jazz is capable of expressing feelings, we would need more information about the specific jazz in question. We would need to know what the jazz is about, who wrote it, and how it was performed. Once we have this information, we can then use deductive reasoning to determine whether the jazz is capable of expressing feelings. For example, if the jazz is about a personal experience that the author had, then it is more likely that the jazz will be able to express feelings. This is because the author will have put their own emotions into the jazz, and this will be reflected in the music. However, if the jazz is about a more general topic, such as the beauty of nature, then it is less likely that the jazz will be able to express feelings. This is because the author will not have put their own emotions into the jazz, and the music will therefore be more objective. Ultimately, whether or not a jazz is capable of expressing feelings is a matter of opinion. There is no scientific way to determine whether a piece of music is capable of expressing feelings. However, by using deductive reasoning, we can gain a better understanding of the potential for a jazz to express feelings. Answer and Explain: Explain Error Type:

TABLE 11: A reasoning case in bAbI-16.

Question, answer and explanation	Context: 1 Greg is a rhino. 2 Brian is a swan. 3 Bernhard is a frog. 4 Brian is green. 5 Greg is gray. 6 Bernhard is yellow. 7 Julius is a lion. 8 Julius is yellow. 9 Lily is a rhino. Question: What color is Lily? Answer: Gray Explanation: IsRhino(Greg) (1), IsGray(Greg) (5)→ IsGray(rhino), IsRhino(Lily) (9)→IsGray(Lily)
ChatGPT of 0-shot	Based on the given facts and inductive reasoning, it is likely that Lily's color is not definitively determined. Although Lily is described as a rhino, there is no information given that directly pertains to Lily's color. Additionally, there is no other animal descriptions that could be used to make an inference regarding Lily's color. Therefore, based on the given facts, it is not possible to make an accurate determination of Lily's color using inductive reasoning. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 1-shot	Lily's color is likely to be unknown as there is no information provided about her color among the other animals. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 3-shot	Brian, Greg, and Bernhard, which do not directly provide clues about what color Lily may be. Therefore, we cannot be certain about her color based on the information given alone. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	Based on the given facts, we can use inductive reasoning to conclude that Lily is gray. This is because all of the other rhinos in the given facts are gray, and there is no evidence to suggest that Lily is any different. In fact, the only thing that distinguishes Lily from the other rhinos is her name, which is a flower that is typically white or yellow. However, we can safely assume that Lily's name is not a reference to her color, as there are many other animals and plants that share the same name. Therefore, the most likely conclusion is that Lily is gray, just like the other rhinos. Here is a more detailed explanation of the inductive reasoning process used to reach this conclusion: 1. We are given the following facts: * Greg is a rhino. * Brian is a swan. * Bernhard is a frog. * Brian is green. * Greg is gray. * Bernhard is yellow. * Julius is a lion. * Julius is yellow. * Lily is a rhino. 2. We notice that all of the rhinos in the given facts are gray. 3. We conclude that there is a strong likelihood that Lily is also gray, as there is no evidence to suggest that she is any different from the other rhinos. 4. We can safely assume that Lily's name is not a reference to her color, as there are many other animals and plants that share the same name. 5. Therefore, the most likely conclusion is that Lily is gray. Answer and Explain: ✓ ✓ ✓ ✓
Bard of 0-shot	Based on inductive reasoning, it is likely that Lily is gray, as both Greg and Julius are gray and they are both rhinos, like Lily. Since the only other colors mentioned among the animals are green and yellow, and none of the rhinos have these colors, it is reasonable to assume that Lily is gray. Answer and Explain: ✓ ✗ ✓ ✓ Explain Error Type: ☒ ☑ ☒ ☑ ☒

TABLE 12: A reasoning case in CLUTRR.

Question, answer and explanation	Context: 1 [Chuck] took his daughter [Sara] to the park that afternoon to play. 2 [Judy] took her sister, [Maxine], out to dinner for her birthday. 3 [Chuck] took his sister [Judy] out to lunch after learning that she got accepted into her first choice for university. 4 [Maxine] asked her daughter, [Joyce], if she would like to go to a movie with her on Saturday night. 5 [Maxine] and her daughter [Joyce] went to see a movie yesterday and then got ice cream afterwards. 6 [Reynaldo]'s wife, [Maxine], was thrilled to learn she was pregnant. 7 [Naomi] went shoe shopping with her sister [Sara]. Question: What is the kinship between Reynaldo and Naomi? Answer: niece Explanation: [(Reynaldo, wife, Maxine) 6, (Maxine, daughter, Joyce) 5] → (Reynaldo, daughter, Joyce) i1 [(Maxine, sister, Judy) 2, (Judy, brother, Chuck) 3] → (Maxine, brother, Chuck) i2 [(Joyce, mother, Maxine) 4, (Maxine, brother, Chuck) i2] → (Joyce, uncle, Chuck) i3 [(Reynaldo, daughter, Joyce) i1, (Joyce, uncle, Chuck) i4] → (Reynaldo, brother, Chuck) i4 [(Chuck, daughter, Sara) 1, (Sara, sister, Naomi) 7] → (Chuck, daughter, Naomi) i5 [(Reynaldo, brother, Chuck) i4, (Chuck, daughter, Naomi) i5] → (Reynaldo, niece, Naomi)
ChatGPT of 0-shot	There is no direct or indirect relation mentioned between Reynaldo and Naomi in the given facts. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 1-shot	Unclear, as there is not enough information about Naomi's relation to Reynaldo in the given facts. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
ChatGPT of 3-shot	The given facts do not provide any information about the kinship between Reynaldo and Naomi, so the question cannot be answered using inductive reasoning. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☑ ☒ ☒
Davinci-003 of 0-shot	Reynaldo and Naomi have no direct kinship; they are both related to Maxine, as Reynaldo is her husband and Naomi is Sara's sister. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☑ ☑ ☒ ☒
Bard of 0-shot	Reynaldo and Naomi are siblings-in-law. Inductive reasoning is a type of reasoning that uses specific examples to make a general conclusion. In this case, we can use the following specific examples to make the general conclusion that Reynaldo and Naomi are siblings-in-law: Maxine is the wife of Reynaldo. Sara is the daughter of Maxine. Naomi is the sister of Sara. Therefore, Naomi is the sister-in-law of Reynaldo. Here are some simple explanations for each of the specific examples: Maxine is the wife of Reynaldo because we are told that Reynaldo's wife is Maxine. Sara is the daughter of Maxine because we are told that Maxine has a daughter named Sara. Naomi is the sister of Sara because we are told that Naomi went shoe shopping with her sister Sara. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☑ ☑ ☒ ☒

TABLE 13: A reasoning case in α-NLI.

Question, answer and explanation	Context: When I was 19, I enlisted with the USMC. When I got back from the hospital, the USMC sent me home. Question: Choose the more likely explanation from following two options: A. I got injured the last day during a training event. B. I was shot in action but didn't regret it. Answer: B
ChatGPT of 0-shot	The more likely explanation is hypothesis choice A: I got injured the last day during a training event. This is because the context mentions the narrator enlisting with the USMC and being sent home after getting back from the hospital, suggesting a military-related scenario. Hypothesis choice A of the narrator getting injured during a training event aligns with these details and provides a probable explanation for why they may have been sent home, as injuries sustained during training may prevent soldiers from fulfilling duties and may require medical attention. Hypothesis choice B of the narrator being shot in action doesn't fit with the context, as there is no indication of any mention of being in combat or any explanation for how this would have been related to the narrator's hospitalization or being sent home, and it doesn't provide an explanation for what may have motivated the narrator's military-related interests or preferences. Answer and Explain: ✗ ✗ ✓ ✗ Explain Error Type: ☒ ☒ ☒ ☒ ☑
ChatGPT of 1-shot	The correct choice is A. The context suggests that the person was enlisted in the USMC and got sent home from the hospital, indicating that they likely got injured during a training event. Choice B suggests an unlikely scenario where the person got shot in action, which cannot be inferred from the context and contradicts the fact that the person did not regret it. Answer and Explain: ✗ ✗ ✓ ✗ Explain Error Type: ☒ ☒ ☒ ☒ ☑
ChatGPT of 3-shot	The correct choice is: A. The context suggests that the narrator enlisted with the USMC and was sent home from the hospital. Hypothesis A, where the narrator got injured the last day during a training event, is the more likely explanation because it fits with the idea of being sent home from the hospital after being injured in the line of duty. Hypothesis B, where the narrator was shot in action but didn't regret it, does not make sense as a plausible explanation for being sent home from the hospital expressed in the context. Answer and Explain: ✗ ✗ ✓ ✗ Explain Error Type: ☒ ☒ ☒ ☒ ☑
Davinci-003 of 0-shot	The more likely explanation is A. I got injured the last day during a training event. This could have been caused by some sort of accident with a weapon or a physical training exercise gone wrong. It is more likely that I would have been injured on the last day of service, as opposed to being shot in action, which would be less likely in the US Marine Corps. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☒ ☒ ☒ ☒ ☑
Bard of 0-shot	The more likely explanation is A. I got injured the last day during a training event. This is because the context states that you enlisted with the USMC at 19, and that you were sent home from the hospital after being injured. This suggests that the injury occurred during training, and that it was serious enough to warrant your discharge from the military. Hypothesis B is less likely because it is not supported by the context. The context does not mention anything about you being shot in action, or about you not regretting it. Additionally, it is not clear why the USMC would send you home if you were not injured. Here is a simple explanation for why you might have been injured during training: You were involved in a training accident. You overexerted yourself and suffered an injury. You were injured by another Marine during training. It is important to remember that these are just possible explanations, and we cannot know for sure why you were injured without more information. Answer and Explain: ✗ ✗ ✓ ✗ Explain Error Type: ☒ ☒ ☒ ☒ ☑

TABLE 14: A reasoning case in α-NLG.

Question, answer and explanation	Context: Joy's mom sent her in the donut store for a dozen donuts. She was happy as she left with 13 donuts instead of 12. Question: Generate a short hypothesis for the context and give simple explanations. Answer: Joy got an extra donut.
ChatGPT of 0-shot	Hypothesis: The fact that Joy's mom sent her to the donut store for a dozen donuts, and Joy left with 13 instead of 12 might have been a result of the potential personal or environmental factors related to perception, generosity, or habit. Explanation: The fact that Joy's mom sent her to the donut store for a dozen donuts suggests that it might have some potential personal or environmental factors related to perception, generosity, or habit. Having personal or environmental factors can involve various types of cognitive or practical factors, such as accuracy, generosity, or routine, and can reflect one's sense of responsibility or automaticity. Thus, the fact that Joy's mom sent her to the donut store for a dozen donuts might have been a result of the potential personal or environmental factors related to perception, generosity, or habit. Additionally, the fact that Joy left with 13 donuts instead of 12 implies that there might have been some additional potential cognitive or contextual factors related to accuracy, availability, or preference. Having cognitive or contextual factors can involve various types of perception or action, such as estimation, surplus, or choice, and can reflect one's sense of intention or opportunity. Furthermore, the fact that Joy was happy as she left with 13 donuts instead of 12 suggests that there might have been some potential affective or symbolic factors related to value, satisfaction, or normativity. Having affective or symbolic factors can involve various types of appraisal or intent, such as pleasure, gratitude, or conformity, and can reflect one's sense of social or moral obligation. Therefore, the hypothesis suggests that the fact that Joy's mom sent her to the donut store for a dozen donuts, and Joy left with 13 instead of 12 might have been a result of the potential personal or environmental factors related to perception, generosity, or habit, as well as the potential cognitive or contextual factors related to accuracy, availability, or preference. Answer and Explain: ✗ ✗ ✓ ✗ Explain Error Type: ☒ ☒ ☒ ☒ ☑
ChatGPT of 1-shot	Hypothesis: The reason why Joy left the donut store with 13 instead of 12 donuts may be due to a generous gesture from the store clerk, a promotional offer that allowed her to receive an extra donut for free, a mistake made by the store clerk, or Joy's persuasive negotiation or effective communication skills. Explanation: The fact that Joy left the donut store with 13 instead of 12 donuts suggests that there may have been an unexpected addition or alteration to her order. One possibility is that the store clerk may have generously offered Joy an additional donut as a kind gesture or to show appreciation for her business. Another possibility is that there may have been a promotional offer where customers who purchased a dozen donuts received an additional donut for free. Alternatively, a mistake may have been made by the store clerk when fulfilling Joy's order, resulting in an extra donut being added to her purchase. Finally, Joy may have effectively persuaded or negotiated with the store clerk to receive an extra donut, perhaps by demonstrating loyal patronage or simply asking for an additional donut. These factors may have contributed to Joy's pleasant surprise at leaving the donut store with an extra donut, and highlight the importance of customer service, promotions, careful order fulfillment, and communication skills in the food service industry. Therefore, the hypothesis proposes that the reason why Joy left the donut store with 13 instead of 12 donuts may be due to a generous gesture from the store clerk, a promotional offer that allowed her to receive an extra donut for free, a mistake made by the store clerk, or Joy's persuasive negotiation or effective communication skills. Answer and Explain: ✓ ✓ ✓ ✗
ChatGPT of 3-shot	Hypothesis: The donut shop may have made an error and accidentally given Joy an extra donut, which she was pleased to receive. Alternatively, the donut store may have had a special promotion where customers receive an extra donut for free, which Joy was not aware of but happily benefited from. This scenario highlights the potential for unexpected surprises and bonuses in our daily lives. Answer and Explain: ✓ ✓ ✓ ✗
Davinci-003 of 0-shot	Hypothesis: The donut store owner gave Joy an extra donut for free. Explanation: The extra donut could indicate that the store owner gave Joy an extra donut for free as a gesture of goodwill or appreciation. This could be due to Joy being loyal customer at the store, or due to her being a child and the store owner having sympathy for her. Answer and Explain: ✓ ✓ ✓ ✗

Bard of 0-shot	Sure, here are some possible hypotheses for why Joy was happy when she left the donut store with 13 donuts instead of 12: The donut store was giving away a free donut with every dozen purchased. The donut store made a mistake and gave Joy an extra donut. Joy was able to negotiate with the donut store owner to get an extra donut. Joy is a very good friend or family member and the donut store owner wanted to give her a little extra something. Here are some simple explanations for each hypothesis: The donut store was giving away a free donut with every dozen purchased because they were trying to promote their business or because they had too many donuts on hand. The donut store made a mistake and gave Joy an extra donut because they were busy or because they were not paying attention. Joy was able to negotiate with the donut store owner to get an extra donut because she was polite and persuasive. Joy is a very good friend or family member and the donut store owner wanted to give her a little extra something because they appreciated her business or because they were close friends or family. It is impossible to know for sure which hypothesis is correct, but all of them are possible explanations for why Joy was happy when she left the donut store with 13 donuts instead of 12. Answer and Explain:

TABLE 15: A reasoning case in AbductiveRules.

Question, answer and explanation	Context: 1 Gary is strong. 2 Things that are tiny, are small, and are short, are thin. 3 All people that are quiet, are poor, and are smart, are clever. 4 Gary is big. 5 If something is short, is tiny, and is nice, it is also thin. 6 If a person is bad, is sad, and is strong, that person is rough. 7 Things that are strong, are huge, and are big, are high. 8 People that are bad, are short, and are strong, are also clever. 9 Dave is poor. 10 Dave is imperfect. 11 Charlie is quiet. 12 Gary is heavy. 13 Anne is tiny. 14 All things that are huge, are small, and are big, are also high. 15 All things that are quiet, are huge, and are poor, are also thin. 16 Anne is little. 17 Charlie is nice. 18 Dave is bad. 19 If a person is tiny, is nice, and is sad, that person is high. 20 If a person is quiet, is nice, and is smart, that person is clever. 21 All things that are small, are big, and are smart, are rough. 22 Anne is small. 23 People that are bad, are poor, and are sad, are also rough. Question: Please generate a short explanation for the given context and observation. The observation is: Dave is rough. Answer: Dave is sad. Explanation: 9, $18 \xrightarrow{23}$ Dave is sad.
ChatGPT of 0-shot	Based on the given context, individuals who possess traits such as badness, sadness, and strength are typically perceived as rough. Dave has been described as poor and imperfect which would suggest a degree of hardship and inadequacy, and the observation confirms that he is perceived as rough. Therefore, it can be inferred that Dave's perceived roughness is a result of his perceived badness, sadness, and poor state in accordance with the given context. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 1-shot	The context indicates that being bad, sad, and strong is associated with being rough, and being poor and imperfect are also negative qualities. The observation that Dave is rough suggests that he may possess these negative traits or exhibit aggressive behavior. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 3-shot	The given observation contradicts the context, as it states that Dave is rough whereas the context indicates that he is poor and imperfect. Therefore, there is no valid explanation for this observation within the given context. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☒ ☒ ☒ ☒
Davinci-003 of 0-shot	Dave is rough because he is bad, poor, and strong, which meets the criteria of people that are bad, poor, and strong being rough. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☒ ☒ ☒ ☒
Bard of 0-shot	Sure, here is a short explanation for the given context and observation: Dave is rough because he is bad, sad, and strong. He is also poor and imperfect. This suggests that he is also quiet and clever. However, he is also tiny and small. This suggests that he is also weak. Here is a more detailed explanation: Dave is bad. This is evident from the fact that he is described as "bad" and "rough". Dave is sad. This is evident from the fact that he is described as "sad" and "imperfect". Dave is strong. This is evident from the fact that he is described as "strong" and "huge". Dave is poor. This is evident from the fact that he is described as "poor" and "imperfect". Dave is imperfect. This is evident from the fact that he is described as "sad" and "imperfect". Dave is quiet. This is evident from the fact that he is described as "quiet" and "clever". Dave is clever. This is evident from the fact that he is described as "clever" and "strong" Dave is tiny. This is evident from the fact that he is described as "tiny" and "small". Dave is small. This is evident from the fact that he is described as "small" and "weak". Dave's badness, sadness, strength, poverty, imperfection, quietness, cleverness, tininess, and smallness suggest that he is a complex and contradictory person. He is capable of great things, but he is also capable of great evil. He is a force to be reckoned with, and he will not be easily defeated. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☑ ☒ ☑ ☒

TABLE 16: A reasoning case in D*-Ab.

Question, answer and explanation	Context: Fact1: The bear chases the cat. Fact2: The bear is big. Fact3: The bear is cold. Fact4: The bear likes the squirrel. Fact5: The bear likes the tiger. Fact6: The bear sees the squirrel. Fact7: The cat is big. Fact8: The cat is cold. Fact9: The cat is young. Fact10: The cat likes the bear. Fact11: The cat likes the tiger. Fact12: The cat sees the bear. Fact13: The cat sees the tiger. Fact14: The squirrel is young. Fact15: The tiger chases the bear. Fact16: The tiger likes the squirrel. Rule1: If the tiger is young then the tiger likes the squirrel. Rule2: If something is round then it likes the bear. Rule3: If something likes the bear then it is cold. Rule4: If the tiger likes the bear and the bear sees the squirrel then the squirrel is round. Rule5: If something likes the squirrel then it is round. Rule6: If something chases the bear and the bear likes the cat then the cat likes the squirrel. Question: Please generate a short missing fact for the given context and fact. And give simple explanations. The fact is: The cat likes the squirrel. Answer: The bear likes the cat. Explanation: Fact15 $\xrightarrow{Rule6}$ The bear likes the cat.
ChatGPT of 0-shot	Missing fact: The squirrel is round. Explanation: According to the given rules, if the tiger likes the bear and the bear sees the squirrel, then the squirrel is round. As the tiger likes the squirrel in the given context, we can conclude that the tiger likes the bear, and since the bear likes the squirrel, we can infer that the squirrel is round. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 1-shot	The fact that the cat likes the squirrel suggests a potential behavioral or social relationship between the two animals, such as companionship or mutual interest in resources. The missing fact that Anne is young could imply a new or additional character to the context. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 3-shot	According to the context, the bear likes the squirrel, and the cat likes the bear, which could suggest a predator-prey relationship between the cat and the squirrel. Moreover, given the fact that the cat likes the squirrel, it is possible that the cat may try to protect the squirrel from the bear's chase. Additionally, the context suggests that. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☒ ☒ ☑ ☒
Davinci-003 of 0-shot	The missing fact is: The tiger is young. Explanation: The rule1 states that if the tiger is young then the tiger likes the squirrel. The fact16 states that the tiger likes the squirrel, so the tiger must also be young. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☒ ☒ ☑ ☒
Bard of 0-shot	Sure, here is a short missing fact for the given context and fact: The squirrel sees the cat. This is the missing fact because it is the only fact that can explain all of the given facts and rules. For example, Rule 4 says that if the tiger likes the bear and the bear sees the squirrel then the squirrel is round. However, we know that the squirrel is round, so the squirrel must also see the bear. Here are some simple explanations for the missing fact: The squirrel is friends with the cat. The squirrel is related to the cat. The squirrel is a food source for the cat. I hope this helps! Let me know if you have any other questions. However, I think there is a mistake in the given context. The fact that "The cat likes the squirrel" contradicts the fact that "The cat chases the squirrel". If the cat likes the squirrel, then it would not chase the squirrel. Therefore, one of these facts must be incorrect. Answer and Explain: ✗ ✗ ✗ ✓ Explain Error Type: ☑ ☑ ☒ ☑ ☒

TABLE 17: A reasoning case in ReClor.

Question, answer and explanation	Context: A science class stored one selection of various fruits at 30 degrees Celsius, a similar selection in similar conditions at 20 degrees, and another similar selection in similar conditions at 10 degrees. Because the fruits stored at 20 degrees stayed fresh longer than those stored at 30 degrees, and those stored at 10 degrees stayed fresh longest, the class concluded that the cooler the temperature at which these varieties of fruits are stored, the longer they will stay fresh. Question: The class's reasoning is flawed in that the class A. neglected to offer any explanation for the results it discovered B. assumed without proof that its thermometer was reliable C. generalized too readily from the fruits it tested to fruits it did not test D. too readily extrapolated from a narrow range of temperatures to the entire range of temperatures. Answer: D
ChatGPT of 0-shot	The correct answer is D. Option A is not relevant to the class's reasoning, as it still made an attempt to draw a conclusion from its results. Option B is a possible flaw, but it is not stated in the argument whether the thermometer was reliable or not, so it cannot be the correct answer. Option C is a possible flaw, but it is not stated whether the class attempted to generalize its results to fruits it did not test, so it cannot be the correct answer. Option D is the correct answer, as the class too readily made a generalization from a narrow range of temperatures (10 degrees, 20 degrees, and 30 degrees) to the entire range of temperatures at which fruits can be stored, which is likely to be much broader. Answer and Explain: ✓ ✓ ✓ ✗
ChatGPT of 1-shot	The correct answer is C. The class's reasoning is flawed in that it generalized too readily from the fruits it tested to fruits it did not test. The experiment only tested a specific selection of fruits and did not test all varieties of fruits, so it is not necessarily true that cooler temperatures will always keep fruits fresh longer. Option A is incorrect because the class did offer an explanation for the results they discovered. Option B is not relevant to the flaw in the class's reasoning. Option D is also not relevant, as the class did not extrapolate from a narrow range of temperatures to the entire range of temperatures. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☒ ☒ ☒ ☑ ☒
ChatGPT of 3-shot	The correct answer is D. The class's reasoning is flawed in that they too readily extrapolated from a narrow range of temperatures to the entire range of temperatures. The class only tested three temperatures, and did not consider the possibility of a threshold beyond which colder temperatures may not be beneficial for all fruits. Therefore, it is too hasty to conclude that the cooler the temperature, the longer the fruits will stay freshbased on such a limited range of temperatures. Answer and Explain: ✓ ✓ ✗ ✗
Davinci-003 of 0-shot	The correct answer is C. The class's reasoning is flawed in that the class generalized too readily from the fruits it tested to fruits it did not test. The class was only testing three types of fruit at three different temperatures, which does not truly demonstrate how all different varieties of fruits can react differently to varying temperatures. Thus, the class should not make a hasty generalization about how all fruits would respond to different temperatures. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☒ ☑ ☒ ☑ ☒
Bard of 0-shot	The correct answer is D. Too readily extrapolated from a narrow range of temperatures to the entire range of temperatures. The class only tested fruits at three temperatures: 30 degrees Celsius, 20 degrees Celsius, and 10 degrees Celsius. They concluded that the cooler the temperature, the longer the fruits will stay fresh. However, this conclusion is based on a very small sample size. It is possible that the relationship between temperature and freshness is not linear. For example, it is possible that fruits will stay fresh for the same amount of time at 10 degrees Celsius and 5 degrees Celsius. The class would need to test fruits at a wider range of temperatures to confirm their conclusion. The other options are not correct. Option A is incorrect because the class did offer an explanation for their results: they said that the cooler the temperature, the longer the fruits will stay fresh. Option B is incorrect because the class did not assume that their thermometer was unreliable. Option C is incorrect because the class did not generalize to fruits they did not test. They only tested fruits at three temperatures, and they only made a conclusion about those three temperatures. Answer and Explain: ✓ ✓ ✓ ✗

TABLE 18: A reasoning case in LogiQA.

Question, answer and explanation	Context: People often use microwave ovens to heat food.Some people think that microwave oven heating changes the molecular structure of food, producing molecules that the human body cannot recognize.These strange molecular formulas are unacceptable to the human body, and some are toxic and may even cause cancer. Therefore, people or animals who often eat microwave foods will have serious physiological changes in their bodies, which will cause serious health problems. Question: Which of the following can best question the above view? A. Microwave heating will not cause more nutrient loss than other cooking methods B. China's microwave oven production standards are consistent with international standards and EU standards C. It is also common for developed countries to use microwave ovens D. The microwave only heats the water molecules in the food, the food does not change chemically. Answer: D
ChatGPT of 0-shot	The answer is D. The given argument suggests that microwave oven heating changes the molecular structure of food, producing molecules that the human body cannot recognize and causing serious health problems. Option D provides an alternative explanation for the process of microwave heating, which states that the microwave only heats the water molecules in the food, and the food does not change chemically. Therefore, option D best questions the view presented in the argument. Option A is not relevant as it talks about nutrient loss rather than the effects of microwave heating on food. Option B and option C are not relevant as they provide information about the production standards of microwave ovens and their usage in different countries, but do not impact the argument presented in the context. Answer and Explain: ✓ ✓ ✓ ✗
ChatGPT of 1-shot	The correct choice is: D. The argument suggests that the use of microwave ovens to heat food changes the molecular structure of the food, producing molecules that the human body cannot recognize, which can lead to serious health problems. Option D challenges this argument by stating that the microwave only heats the water molecules in the food and does not change the chemical structure of the food. This suggests that the argument may be based on a misunderstanding of how microwaves work. Option A, B, and C are not relevant to the argument or do not challenge the premise that microwaves can cause health problems. Answer and Explain: ✓ ✓ ✓ ✗
ChatGPT of 3-shot	The correct choice is: D. The microwave only heats the water molecules in the food, the food does not change chemically, can best question the above view. It suggests that the claim about microwave ovens changing the molecular structure of food is not accurate and provides a scientific explanation for how microwave heating works, which does not involve the production of strange or toxic molecules. Answer and Explain: ✓ ✓ ✗ ✗
Davinci-003 of 0-shot	The best answer is B. China's microwave oven production standards are consistent with international standards and EU standards. This answer questions the view presented in the context that microwave oven heating changes the molecular structure of food, since it implies that safety regulations surrounding the production of microwave ovens would never be put in place if this were true. Therefore, this disproves the view that microwave oven heating produces molecules that the human body cannot recognize, which could be toxic and may even cause cancer. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☒ ☒ ☒ ☑ ☒
Bard of 0-shot	The correct answer is D. The microwave only heats the water molecules in the food, the food does not change chemically. The context states that some people think that microwave oven heating changes the molecular structure of food, producing molecules that the human body cannot recognize. However, this is not true. Microwaves only heat the water molecules in food, and the food does not change chemically. This is why microwaved food is safe to eat. The other options are not necessarily correct. Option A is incorrect, as there is some evidence that microwave heating can cause nutrient loss. Option B is irrelevant, as it does not address the question of whether or not microwave oven heating changes the molecular structure of food. Option C is also irrelevant, as it does not address the question of whether or not microwaved food is safe to eat. Here are some additional information about microwaves and their effects on food: Microwaves are a form of non-ionizing radiation, which means that they do not have enough energy to damage DNA. Microwaves heat food by causing water molecules to vibrate. This vibration produces heat, which cooks the food. Microwaves do not change the chemical structure of food. This means that microwaved food is safe to eat. Some studies have shown that microwave heating can cause nutrient loss. However, these studies have been small and the results have been mixed. More research is needed to determine the extent of nutrient loss from microwaved food. Overall, there is no evidence that microwave oven heating is harmful to human health. Microwaved food is safe to eat. Answer and Explain: ✓ ✗ ✓ ✓ Explain Error Type: ☒ ☑ ☒ ☑ ☒

TABLE 19: A reasoning case in LogiQA 2.0.

Question, answer and explanation	Context: In a track and field qualifier, Zhang Qiang surpassed Li Jin, and Song Zhi's results were better than Wang Ping, but not as good as Ma Zheng. Question: Which of the following options can be inferred? A. Ma Zheng's grades are better than Zhang Qiang. B. Li Jin's results surpass Wang Ping. C. Zhang Qiang's performance is better than Song Zhi. D. Among the five people, Wang Ping ranked third at most. Answer: D
ChatGPT of 0-shot	If Zhang Qiang surpassed Li Jin and Song Zhi's results were better than Wang Ping, but not as good as Ma Zheng, it can be inferred that Zhang Qiang's performance is better than Song Zhi. While the relative positions of Zhang Qiang, Li Jin, Wang Ping, and Ma Zheng are known, there is no information to suggest that Ma Zheng's grades are better than Zhang Qiang (option A) or that Li Jin's results surpass Wang Ping (option B). Option D introduces a ranking that is not directly supported by the information in the passage. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 1-shot	The correct answer is D. From the given information, we know that Zhang Qiang's result was better than Li Jin's, and Song Zhi's result was better than Wang Ping's but not as good as Ma Zheng's. Therefore, we can infer that Ma Zheng's result was better than all the other athletes, including Zhang Qiang, and Li Jin's result was worse than both Zhang Qiang's and Wang Ping's. However, we cannot determine the exact ranking of Wang Ping among the five athletes with the given information, so option D is the correct choice as it cannot be inferred from the given context. Answer and Explain: ✓ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 3-shot	The correct answer is C. The passage states that Zhang Qiang surpassed Li Jin and that Song Zhi's results were better than Wang Ping but not as good as Ma Zheng. Therefore, it can be inferred that Zhang Qiang's performance is better than Song Zhi's since he surpassed Li Jin. Choice A cannot be inferred because Ma Zheng's performance isn't compared to Zhang Qiang's. Choice B also cannot be inferred because there's no comparison between Li Jin's and Wang Ping's performance. Choice D cannot be inferred because there's no information to suggest that Wang Ping ranked third at most. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
Davinci-003 of 0-shot	The correct answer is C. Zhang Qiangś performance is better than Song Zhi. This can be inferred from the given context, as the context stated that "Zhang Qiang surpassed Li Jin", which means Zhang Qiang is better than Li Jin, and "Song Zhiś results were better than Wang Ping", which means Song Zhi is better than Wang Ping, therefore Zhang Qiang is better than Song Zhi. Therefore, option C is the correct choice. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
Bard of 0-shot	The correct answer is D. Among the five people, Wang Ping ranked third at most. The passage states that Zhang Qiang surpassed Li Jin, and Song Zhi's results were better than Wang Ping, but not as good as Ma Zheng. This information can be organized into the following table: — Athlete — Results — — Zhang Qiang — Better than Li Jin — — Song Zhi — Better than Wang Ping — — Ma Zheng — Better than Song Zhi — From this table, we can see that Wang Ping's results are not as good as Ma Zheng's, so Wang Ping ranked third at most. The other answer choices are not supported by the passage. Answer choice A is not supported because the passage does not say that Ma Zheng's grades are better than Zhang Qiang's. Answer choice B is not supported because the passage does not say that Li Jin's results surpass Wang Ping's. Answer choice C is not supported because the passage does not say that Zhang Qiang's performance is better than Song Zhi's. Answer and Explain: ✓ ✓ ✓ ✗

TABLE 20: A reasoning case in LogiQA2NLI.

Question, answer and explanation	Context: The premis is: Researcher: People who participate in opinion surveys often give answers they believe the opinion surveyor expects to hear, and it is for this reason that some opinion surveys do not reflect the actual views of those being surveyed. However, in well-constructed surveys, the questions are worded so as to provide respondents with no indication of which answers the surveyor might expect. So if a survey is well constructed, survey respondents' desire to meet surveyors' expectations has no effect on the survey's results. The conclusion is: the answers of opinion-survey respondents can be influenced by beliefs about the surveyor's expectations even if those beliefs are unfounded. Question: Determine whether the premise and conclusion entailed or NOT entailed. Answer: entailed
ChatGPT of 0-shot	The conclusion is not logically entailed by the given premise. The premise discusses the potential for opinion survey respondents to give answers they believe the surveyor expects to hear, but notes that well-constructed surveys prevent this by wording questions in a way that does not indicate expected answers. It then concludes that if a survey is well constructed, respondents' desire to meet surveyors' expectations has no effect on the survey's results. Therefore, it does not support the conclusion that the answers of opinion-survey respondents can be influenced by beliefs about the surveyor's expectations even if those beliefs are unfounded. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 1-shot	The conclusion is not entailed by the given premise. While the premise discusses how well-constructed surveys are worded so as to provide respondents with no indication of which answers the surveyor might expect, and how this can mitigate respondents' desires to meet surveyors' expectations, it does not provide information to support or refute the claim that the answers of opinion-survey respondents can be influenced by beliefs about the surveyor's expectations even if those beliefs are unfounded. Therefore, the conclusion cannot be reliably inferred from the given premise. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
ChatGPT of 3-shot	Not entailed. The conclusion about the answers of opinion-survey respondents being influenced by beliefs about the surveyor's expectations cannot be logically derived from the given premise. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
Davinci-003 of 0-shot	NOT entailed. The premise is implying that because the survey is well constructed with no indication of what answers the surveyor might expect, the survey respondents' answers will not be impacted by any beliefs they have of the surveyor's expectations. The conclusion is suggesting that even if those beliefs are unfounded, the responses will still be influenced by them. This goes against the premise as the premise has stated that well-constructed surveys will give no indication of the surveyor's expectations, and therefore, their beliefs of the surveyor's expectations will have no influence. Answer and Explain: ✗ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒
Bard of 0-shot	The conclusion is entailed by the premise. The premise states that people who participate in opinion surveys often give answers they believe the opinion surveyor expects to hear, even if those beliefs are unfounded. It also states that well-constructed surveys are worded so as to provide respondents with no indication of which answers the surveyor might expect. This means that even if a survey is well-constructed, respondents may still be influenced by beliefs about the surveyor's expectations. This is because people are often motivated to please others, and they may believe that giving the answer they think the surveyor expects will make them look good. For example, a respondent may believe that a surveyor expects them to support a certain candidate in an election. Even if the survey question is not leading, the respondent may still give the answer they think the surveyor expects, even if they do not actually support that candidate. This is why it is important for surveyors to be aware of the potential for respondents to be influenced by beliefs about their expectations. Surveyors can try to minimize this effect by using clear and unbiased language, and by explaining to respondents that there are no right or wrong answers. Answer and Explain: ✓ ✗ ✗ ✗ Explain Error Type: ☑ ☒ ☒ ☑ ☒