Analysis of LLM's "Spurious" Correct Answers Using Evidence Information of Multi-hop QA Datasets

Ai Ishii[1], **Naoya Inoue**[2,1], **Hisami Suzuki**[1], **Satoshi Sekine**[1]
[1]RIKEN AIP, Tokyo, Japan,
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
ai.ishii@riken.jp, naoya-i@jaist.ac.jp, hisami.suzuki@a.riken.jp, satoshi.sekine@riken.jp

Abstract

Recent LLMs show an impressive accuracy on one of the hallmark tasks of language understanding, namely Question Answering (QA). However, it is not clear if the correct answers provided by LLMs are actually grounded on the correct knowledge related to the question. In this paper, we use multi-hop QA datasets to evaluate the accuracy of the knowledge LLMs use to answer questions, and show that as much as 31% of the correct answers by the LLMs are in fact spurious, i.e., the knowledge LLMs used to ground the answer is wrong while the answer is correct. We present an analysis of these spurious correct answers by GPT-4 using three datasets in two languages, while suggesting future pathways to correct the grounding information using existing external knowledge bases.

1 Introduction

Question Answering (QA) is one of the hallmark tasks that evaluate language understanding capabilities of NLP systems. We are currently witnessing the flourishment of highly capable large language models (LLMs) that solve this complex task, requiring both knowledge and inference skills, with an impressive accuracy (Bang et al., 2023). On the other hand, it has been shown that LLMs can generate content that contradicts facts (Bang et al., 2023; Ji et al., 2023), and several verification results have been reported regarding the evaluation of LLMs' internal knowledge and whether LLMs can provide answers based on facts (Wang et al., 2023; Manakul et al., 2023; Lin et al., 2022; Zheng et al., 2023; Pezeshkpour, 2023).

At this point, it is not clear exactly to what extent such LLMs possess the knowledge needed to solve QA problems and how accurately they perform inference to leverage that knowledge. How often do LLMs rely on "hallucinated" knowledge during inference? Can these hallucinations be remedied by structured knowledge bases (KBs) carefully crafted by humans? Previous studies have reported that correct answers are often obtained despite errors in the reasoning path that LLMs output to solve QA (Bao et al., 2024; Sprague et al., 2024; Nguyen et al., 2024; Ishii et al., 2024). Ishii et al. (2024) shows the specific error patterns by question type in such cases and the possibility of complementing errors with KBs using JEMHopQA dataset[1], which has evidence information in the form of triples, but their analysis is limited to one dataset in Japanese.

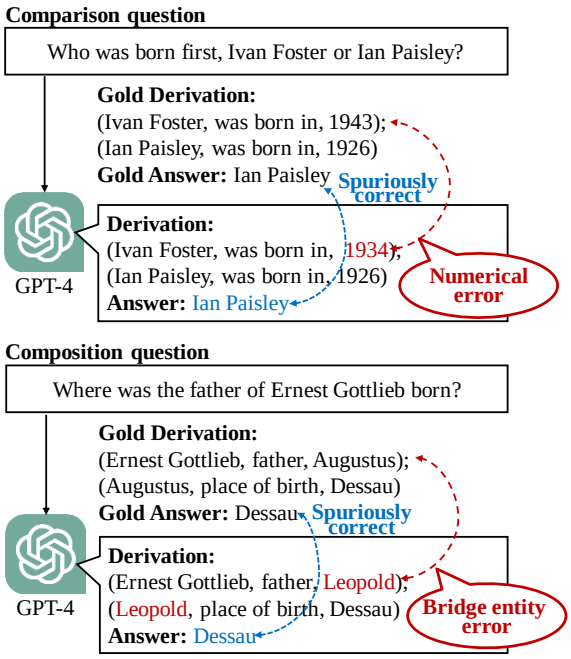

Figure 1: Examples of spurious correct answers. Red text indicates where the model (GPT-4) makes mistakes, blue text indicates where the model's answer is correct. See Appendix A for other types of errors.

In this paper, we focus on investigating how such "spurious" correct answers by LLMs occur more deeply in other datasets and languages. We use three datasets from two languages – HotPotQA (Yang et al., 2018) with $\mathcal{R}^4\mathcal{C}$ (Inoue et al.,

[1]https://github.com/aiishii/JEMHopQA

Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024), pages 24–34

	HotPot	2Wiki	JEMHop
#Avg. question	16.50	11.87	30.71
#Avg. answer	3.42	2.30	4.32
#Avg. derivations	2.50 (3.00)	2.37 (2.42)	2.04 (2.07)

Table 1: Question and answer lengths and number of derivation triples of each dataset. The #Avg. question and #Avg. answer in HotPot and 2Wiki are the average number of tokens, that in JEMHop is the average number of characters, and the number in parentheses in #Avg. derivations is the average number of derivations in each original dataset.

Question type	HotPot	2Wiki	JEMHop
Comparison	19%	27%	61%
Composition	80%	55%	39%
Bridge-comparison	1%	18%	0%

Table 2: Distribution of question types.

2020) and 2WikiMultiHopQA (Ho et al., 2020) for English, and JEMHopQA for Japanese. These three datasets present the task of outputting the knowledge (derivation) that serves as the evidence for the answer in the form of derivation triples (as in Fig. 1), so they can be used directly to measure the spuriousness of correct answers in QA. In addition, we investigate the extent to which gold derivation triples in each dataset are covered by existing KBs, suggesting that hallucinatory knowledge can be corrected by combining LLMs with such KBs.

2 Analysis Method

2.1 Datasets

In this analysis, we use questions, answers, and supporting evidence from widely used HotPotQA, 2WikiMultiHopQA, and JEMHopQA, which are Wikipedia-based multi-hop QA datasets. We use $\mathcal{R}^4\mathcal{C}$ for derivation triples of HotPotQA, and randomly extract 100 instances from the development set as **HotPot**. We randomly extract 100 instances from the 2WikiMultiHopQA development set as **2Wiki** and use all 120 instances from the JEMHopQA development set as **JEMHop**. Table 1 summarizes the details of these datasets, where the average number of derivation triples are roughly the same across them.

In these datasets, questions comprise of three different types[2]: (i) Comparison questions, where the two derivation triples have the same relation, as in at the top of Fig. 1; (ii) Composition questions, where two derivation triples share a "bridge" entity, as in the example at the bottom of Fig. 1 where "Augustus" serves as the bridge; (iii) Bridge-comparison, which combines a composition with a comparison, where a comparison is made after finding the bridge entity, e.g., "Which film has the director who is older, Aardram or Land and Freedom?". The distribution of these three types of questions is shown in Table 2.

As Ishii et al. (2024) reports that comparison questions (numerical comparisons in particular) are more susceptible to spurious correct answers, we created additional datasets that include the samples of such questions in our study. The number of numerical comparison questions differs considerably across our dataset (HotPot: 4%, 2Wiki: 17%, JEMHop: 28%), so we created focused datasets consisting only of numerical comparisons by taking 30 samples from the development set of the datasets, resulting in **HotPot_NC**, **2Wiki_NC**, and **JEMHop_NC**. We also extracted 30 multi-hop QA instances that compare numerical values from DROP (Dua et al., 2019), a widely used QA dataset that requires mathematical operations, and use them as the analysis set **DROP_NC**.

The supporting evidence in each dataset is in the form of triples representing a semi-structured relationship (e.g., "date of birth") between a subject entity ("Ivan Foster") and an object entity ("1943"), as shown in Fig. 1. The questions are those that require multi-hop reasoning, and each question-answer pair is accompanied by two or more derivation steps. The task of evaluating LLMs using each dataset is, given a question Q, (i) to predict the answer A, and (ii) to generate a derivation D that justifies A.

2.2 Evaluation Metrics

Answers For HotPot and 2Wiki, we use exact match (EM) and partial match, F_1 score measuring the average overlap between gold and predicted answers. For JEMHop, we use similarity match (SM) score based on the Levenshtein distance.

Derivations To account for differences in the structure of the triples and to measure semantic matches, the authors manually evaluated derivation triplets. Even if predicted derivation has a different surface form from the gold derivation, it is consid-

[2]Although 2WikiMultiHopQA has an "Inference" type, we consider it a subtype of composition in this paper.

	Answer EM / F_1 or SM (%)		
	HotPot	**2Wiki**	**JEMHop**
Zero-shot	38.7/45.3	23.7/28.8	51.7/52.5
5-shot	39.7/49.9	34.3/39.8	56.1/57.8
CoT 5-shot	**41.5/50.9**	**48.3/56.4**	**62.8/64.5**
Comparison	71.1/78.9	86.4/86.4	81.3/81.3
Composition	35.0/44.9	22.4/36.4	34.0/38.3
Brg-comparison	0.0/0.0	70.4/72.2	-/-

Table 3: Results of GPT-4 with different prompts.

ered correct if the information contained is correct, in the form of a triple, and sufficient to answer the question.

Note that each dataset provides evaluation scripts for both answers and derivation triples, but we use these scripts to evaluate answers only and rely on human evaluation for derivation triples.

2.3 Evaluation Setup Using GPT-4

We use `gpt-4-0613` model via OpenAI API with the prompt for the Chain-of-Thought (CoT) (Wei et al., 2022) 5-shot setting as a method of eliciting the derivation triples that the model uses to infer. More specifically, the CoT 5-shot prompt consists of an instruction to provide a CoT reasoning path, along with 5 few-shot samples. To ensure that the setting of the CoT 5-shot prompt to output the inference path at the same time as the answer does not affect the accuracy of the answer, we also use zero-shot (ask a question only) and non-CoT 5-shot (include 5 random samples from the training set) prompts (see examples in Appendix B).

Based on the results of preliminary experiments, we use temperature parameters of 0.1, 0.2, and 0.0 for HotPot, 2Wiki, and JEMHop, respectively. The maximum token limit is set to 32 for the zero-shot and 5-shot prompts, and to 256 for the CoT prompt. Due to the sampling-based decoding of GPT-4 API, we run each experiment three times and report the average of all runs.

3 Results and Discussion

3.1 How well can GPT-4 answer multi-hop questions correctly?

In Table 3, the first three rows show the results for the answers in the zero-shot, 5-shot, and CoT 5-shot settings for each dataset. In all datasets, the 5-shot setting performed better than the zero-shot setting, and the CoT 5-shot setting achieved the highest accuracy. These results confirm that the CoT 5-shot prompt setting, which outputs the derivation triples simultaneously with the answer, does not affect the accuracy of the answers.

The last two rows show that composition questions are significantly harder to answer correctly than comparison questions in all datasets. A major factor for this large difference is suspected to be that in comparison questions, the two subject entities are explicitly mentioned in the question and the answers tend to be binary (choosing one of the two entities), while in composition question, a bridge entity is implicit and must be identified, and the answers for these questions tend not to be binary (an entity as an answer). Bridge-comparison questions fell in the middle as this tasks for a binary answer while needing to identify a bridge entity.

3.2 When do spurious correct answers occur?

Table 4. shows the performance of GPT-4, where we present the results in a matrix along both answer correctness and derivation triple correctness. Cases where the derivation triples were considered correct even though they differed from the gold derivation triples in this evaluation are described in detail in Appendix C. We found that only 0-1% cases had an error in inference (Answer is F and Derivation is T); the remaining cases had errors in derivation (i.e., hallucination). As shown in the table, spurious correct answers (Answer is T and Derivation is F) comprise 18% of all cases (which is 31% of the correctly answered cases) in 2Wiki and 15.8% (which is 25% of the correctly answered cases) in JEMHopQA, showing that they occur also quite frequently in English. More than 90% of these spurious correct answers occur in comparison questions and bridge-comparison; they occur less frequently in HotPot because there are fewer comparison questions.

The question type that generated spurious correct answers most frequently (38% on 2Wiki and 68% on JEMHop) was questions comparing numerical values or dates (see detail in Appendix A). Therefore, we also manually classified the correctness of the derived triples and answers for the numerical comparison questions, adding the evaluation of HotPot_NC, 2Wiki_NC, JEMHop_NC and DROP_NC (see in §2.1) in the CoT 5-shot setting[4].

[3]Note that this classification table does not include the formatting errors that occurred in two cases in JEMHop and one case in HotPot_NC as derivation triple errors, so the total does not add up to 100%.

[4]As DROP lacks evidence information, few-shot examples

		Derivation Triples					
		HotPot		2Wiki		JEMHop	
		T	F	T	F	T	F
Answer	T	50.0%	**8.0%**	39.0%	**18.0%**	47.5%	**15.8%**
	F	1.0%	41.0%	0.0%	43.0%	0.8%	34.1%

Table 4: Classification of right (T) and wrong (F) of answers and derived triples[3].

		Derivation Triples							
		HotPot_NC		2Wiki_NC		JEMHop_NC		DROP_NC	
		T	F	T	F	T	F	T	F
Answer	T	73.3%	**16.7%**	40.0%	**46.7%**	50.0%	**36.7%**	46.7%	**36.7%**
	F	0.0%	6.7%	0.0%	13.3%	0.0%	13.3%	0.0%	16.7%

Table 5: Classification of right (T) and wrong (F) of answers and derived triples of numerical comparison questions[3].

The results are in Table 5.

In this table, we find that as much as 36-46% of the answers were spuriously correct in 3 of the 4 datasets – with the exception of HotPot_NC, where the rate of spurious correct answers remained lower at 16.7%. While it was not obvious to us why HotPot_NC behaved differently, we could see why spurious correctness happens often in numerical comparison: they occur when the relative order of numbers or dates are not affected even when there is an error in derivation triples. This is also observed when we analyzed the results of bridge comparison questions in 2Wiki – out of 14 correct answers of this type, 10 were in fact spurious in the same manner as the numerical comparison questions: there was an error in the identification of bridge entity (identifying a wrong person), but the relative order of the dates required for the answer was unaffected. In order for the answers to be spuriously correct in this way, the error margin for the numbers/dates in the grounding knowledge must be small enough so as not to impact the relative order. Exactly how "wrong" or "close" GPT-4 is when it comes to the numerical aspect of the grounding information deserves further investigation; we leave this for future work.

3.3 Can External KBs Remedy Spurious Correct Answers?

GPT-4 "hallucinated" wrong derivation triples in 50-60% in each dataset as a whole. We investigated whether this knowledge hallucination can be fixed by using external KBs.

For this, we used two existing KBs on Wikipedia, namely Wikidata (Vrandečić and Krötzsch, 2014) and Shinra[5] (Sekine et al., 2019). The latter extracts attribute-value pairs from Japanese Wikipedia articles and structures them according to the ENE (Sekine, 2008) categories in Sekine et al. (2020); this is used for JEMHopQA only as it is in Japanese. Also, in 2Wiki, all hallucinated derivation triples can be found by Wikidata as the questions of 2WikiMultihopQA derive from the knowledge triples in Wikidata. Therefore, we studied the extent to which gold derivation triples in each dataset are covered by external KBs for HotPot and JEMHop only. Knowledge representation in these KBs is compatible with the derivation triples used in our task, allowing for a straightforward application.

In Table 6, the first three columns show the coverage of derivation triples of each dataset for GPT-4, Wikidata, and GPT-4 combined with Wikidata. We assume that the derivation triples generated by GPT-4 in answering the questions are GPT-4's internal knowledge and estimate GPT-4's coverage by calculating how well GPT-4's internal knowledge covers the gold derivation triples in each dataset. As a multi-hop question requires two or more triples to answer, a partial coverage statistic is also given. We see that GPT-4 provides complete evidence for 51% and 48% of HotPot and JEMHop questions respectively, but if combined with Wikidata, it can cover up to 59% and 63% respectively. The last three columns show the coverage of derivation triples of Shinra, GPT-4 combined Shinra and GPT-4 combined with both KBs. GPT-4 and both KBs seem to complement each other well: GPT-4 combined with both KBs achieves 81.7% of coverage, up by

of CoT-5shot are created using the same data as 2Wiki.

[5] http://shinra-project.info/

Dataset	Coverage	GPT-4	Wikidata (W)	GPT-4+W	Shinra (S)	GPT-4+S	GPT-4+W+S
HotPot ($\mathcal{R}^4\mathcal{C}$)	Full	51.0%	31.0%	59.0%	-	-	-
	Partial	17.0%	51.0%	41.0%	-	-	-
	None	32.0%	18.0%	0.0%	-	-	-
JEMHop	Full	48.3%	29.2%	63.3%	50.0%	78.3%	81.7%
	Partial	23.3%	28.3%	26.7%	29.2%	15.0%	13.3%
	None	28.3%	42.5%	10.0%	20.8%	7.5%	5.0%

Table 6: Coverage of derivation steps in the test set by existing KBs and GPT-4.

31% as compared with GPT-4 alone (48.3%). This indicates that a further improvement in multi-hop QA task is possible by combining LLM with existing KBs, a fruitful direction for future research.

4 Conclusions

In this paper, we presented the evaluation of GPT-4 on multi-hop QA in three datasets in English and Japanese, focusing on how the answers are/are not grounded on the knowledge internal to the model. The results show that almost all of the incorrect answers are due to knowledge hallucination, and that even when the answer is correct, up to 31% of them (40% in numerical comparison questions) are in fact spurious. We also showed that the knowledge GPT-4 uses for grounding is complementary with external KBs, indicating a future direction of integrating them for solving multi-hop questions. Our analysis is based on the assumption that the derivation triples generated by the LLM are reasoning of the LLM, but we hope to clarify whether this assumption is correct in the future.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP20269633 and 19K20332. The authors would like to thank the anonymous reviewers for their insightful feedback.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Guangsheng Bao, Hongbo Zhang, Linyi Yang, Cunxiang Wang, and Yue Zhang. 2024. Llms with chain-of-thought are non-causal reasoners. *Preprint*, arXiv:2402.16048.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

Ai Ishii, Naoya Inoue, Hisami Suzuki, and Satoshi Sekine. 2024. JEMHopQA: Dataset for Japanese explainable multi-hop question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9515–9525, Torino, Italia. ELRA and ICCL.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human

falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. *Preprint*, arXiv:2402.11199.

Pouya Pezeshkpour. 2023. Measuring and modifying factual knowledge in large language models. *arXiv preprint arXiv:2306.06264*.

Satoshi Sekine. 2008. Extended named entity ontology with attribute information. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Satoshi Sekine, Maya Ando, Akio Kobayashi, and Aska Sumida. 2020. Updated Extended Named Entity Definitions and Japanese Wikipedia Classification Data 2019. In *Proceedings of the 26th Conference on Natural Language Processing in Japan (NLP 2020)*.

Satoshi Sekine, Akio Kobayashi, and Kouta Nakayama. 2019. Shinra: Structuring wikipedia by collaborative contribution. In *Conference on Automated Knowledge Base Construction*.

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Cunxiang Wang, Sirui Cheng, Zhikun Xu, Bowen Ding, Yidong Wang, and Yue Zhang. 2023. Evaluating open question answering evaluation. *CoRR*, abs/2305.12421.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in answering questions faithfully? *arXiv preprint arXiv:2304.10513*.

A Detailed Types of Spurious Correct Answers

Table 7 shows the percentage of spurious correct answers by question type in each dataset. They mainly appeared in comparison and bridge-comparison questions, with numerical comparison being the most frequent (38% in 2Wiki comparison , 50% in 2Wiki bridge-comparison, 68% in JEMHop).

Spurious correct answers of comparison questions. Table 8 shows examples of spurious correct answers in comparison questions. In "numerical comparison", the relative order of dates (e.g., "1212" vs "1248") or values (e.g., "1.5" vs "2.0") in GPT-4's derivations matched the gold, despite incorrect date or values. In "shared predicate", the answer condition (e.g., whether authors are the same in both entities) was unaffected, despite different authors ("Meka Tanaka" vs "Oreko Tachibana") in GPT-4's and gold derivations.

Spurious correct answers of composition questions. Table 9 shows examples where the answer was correct despite incorrect bridge entities. In one case, different princes were from the same family and birthplace. In others, the bridge entity was unspecified or non-existent, suggesting the model knew the answer in advance. For example, GPT-4 correctly answered "World War II" for when a facility was established, despite using a non-existent bridge entity.

Spurious correct answers of brige-comparison questions. Table 10 shows examples in "numerical comparison" and "shared predicate" types. The answer was unaffected despite wrong bridge entities, as the relative order of dates (e.g., "1936" vs "1956") or conditions like directors' countries remained unchanged.

B Example of Prompts for GPT-4

The following are examples of the three types of prompts we used in our experiments:

	HotPot	2Wiki	JEMHop
Comparison			
Numerical comparison	16.7%	38.9%	68.4%
Entity selection	16.7%	0.0%	10.5%
Shared predicate	0.0%	0.0%	15.8%
Composition			
Entity or value answer	66.7%	5.6%	5.3%
Bridge-comparison			
Numerical comparison	0.0%	50.0%	-
Shared predicate	0.0%	5.6%	-

Table 7: Types of spurious correct answers by question type. Each percentage is the number of spurious correct answer cases in HoPot (6 cases), 2Wiki (18 cases) and JEMHop (19 cases).

1. **Zero-shot**: ask a question only, as in:

   Output your answers to the following questions.
   Answers should be brief noun phrases or "yes/no" answers.:
   Which film came out first, 3 Dots or Dying God? =>

2. **5-shot**: include 5 random samples from the training set as few-shot examples, as in:

   Output your answers to the following questions, referring to the examples.
   Answers should be brief noun phrases or "yes/no" answers.:
   When was the director of film Antarjali Jatra born? => 24 July 1950
   Who died later, Bob Dispirito or John Wilton? => Bob Dispirito
   (...3 more examples)
   Which film came out first, 3 Dots or Dying God? =>

3. **Chain-of-Thought (CoT) 5-shot**: add an instruction to provide a CoT reasoning path, along with 5 few-shot samples.

   Output your answers and rationale to the following questions in the form of examples.
   Answers should be brief noun phrases or "yes/no" answers.:
   When was the director of film Antarjali Jatra born? => (Antarjali Jatra, director, Goutam Ghose);(Goutam Ghose, date of birth, 24 July 1950) => 24 July 1950
   Who died later, Bob Dispirito or John Wilton? => (Bob DiSpirito, date of death, December 21, 2015);(John Wilton, date of death, 10 May 1981) => Bob Dispirito
   (...3 more examples)
   Which film came out first, 3 Dots or Dying God? =>

C Detailed Manual Evaluation of Derivations

In the manual evaluation of the derivations output by GPT-4, even if the derivations did not exactly match the gold derivations, they were considered correct if they were in the form of triples and provided sufficient information to derive the answer from the question. The specific cases considered correct are as follows:

i Differences in wording (tense, synonymous verbs or nouns, presence or absence of modifiers).

ii Differences in granularity of information (geographic, temporal, etc. units).

iii Differences in type of information.

iv Differences in the amount of information contained in a triple (cases where multiple triples of information in the gold are combined into one in the pred (GPT-4 output) and vice versa).

v Differences in how triples are formed (the subject and object of the triple are opposite, or part of the object of the gold triple is included in the relation of the pred triple, etc.).

Examples for each pattern are shown in Table 11.

Question Type	Example
Numerical comparison	**Question:** Which occured first, the Battle of Las Navas de Tolosa or king Fernando III gave a new fuero to the city? **Gold derivation:** ("Battle of Las Navas de Tolosa", "start time", "July 16, 1212"); ("Giving of new fuero by Fernando III", "start time", "1219") **Gold answer:** Battle of Las Navas de Tolosa **GPT-4's derivation:** ("Battle of Las Navas de Tolosa", "start time", "July 16, 1212"); ("Giving of new fuero by Fernando III", "start time", "1248") **GPT-4's answer:** Battle of Las Navas de Tolosa
Numerical comparison	**Question:** Which start has a higher absolute magnitude, A-type star or B9-type star? **Gold derivation:** ("A-type star", "absolute magnitude", "0.2") ; ("B9-type star", "absolute magnitude", "0.4") **Gold answer:** B9-type **GPT-4's derivation:** ("A-type star", "absolute magnitude", "1.5") ; ("B9-type star", "absolute magnitude", "2.0") **GPT-4's answer:** B9-type star
Shared predicate	**Question:** Are Ai Yazawa the author of both "A" and "Promise Cinderella"? **Gold derivation:** ("A", "author", "Ai Yazawa") ; ("Promise Cinderella", "author", "Oreko Tachibana") **Gold answer:** No **GPT-4's derivation:** ("A", "author", "Ai Yazawa") ; ("Promise Cinderella", "author", "Meka Tanaka") **GPT-4's answer:** No

Table 8: Examples of spurious correct answers in comparison questions. Red text indicates where there was an error in the derivation, blue text indicates that the answer is correct.

Error type	Example
Bridge entity is wrong	**Question:** Where was the father of Ernest Gottlieb, Prince Of Anhalt-Plötzkau born? **Gold derivation:** (“Ernest Gottlieb, Prince of Anhalt-Plötzkau”, “father”, “Augustus, Prince of Anhalt-Plötzkau”); (“Augustus, Prince of Anhalt-Plötzkau”, “place of birth”, “Dessau”) **Gold answer:** Dessau **GPT-4’s derivation:** (“Ernest Gottlieb, Prince of Anhalt-Plötzkau”, “father”, “Leopold, Duke of Anhalt-Dessau”); (“Leopold, Duke of Anhalt-Dessau”, “place of birth”, “Dessau”) **GPT-4’s answer:** Dessau
Bridge entity is not identified	**Question:** A 1946 musical comedy starred a British actor who lived in what country throughout his adult life? **Gold derivation:** (“Two Sisters from Boston”, “is”, “a 1946 musical comedy film”); (“Two Sisters from Boston”, “stars”, “Peter Lawford”); (“Peter Lawford”, “is”, “a British actor”); (“Peter Lawford”, “lived throughout adult life in”, “the United States”) **Gold answer:** United States **GPT-4’s derivation:** (“A 1946 musical comedy”, “starred”, “a British actor”); (“The British actor”, “lived in, “the United States throughout his adult life”) **GPT-4’s answer:** The United States
Bridge entity is fictitious	**Question:** The facility where Hideki Tojo died and died was established after what war? **Gold derivation:** (“Hideki Tojo”, “Place of death”, “Sugamo Prison”); (“Sugamo Prison”, “War that led to its establishment”, “World War II”) **Gold answer:** World War II **GPT-4’s derivation:** (“Hideki Tojo”,“facility where he died and died”, “Suginami Ward Hirozawa Hospital”); (“Suginami Ward Hirozawa Hospital”,“when established”,“post-World War II”) **GPT-4’s answer:** World War II

Table 9: Examples of spurious correct answers in composition questions. Red text indicates where there was an error in the derivation, blue text indicates that the answer is correct.

Question Type	Example
Numerical comparison	**Question:** Which film has the director who is older, Aardram or Land And Freedom? **Gold derivation:** (“Aardram”, “director”, “Suresh Unnithan”); (“Suresh Unnithan”, “date of birth”, “30 July 1956”); (“Land and Freedom”, “director”, “Ken Loach”); (“Ken Loach”, “date of birth”, “17 June 1936”) **Gold answer:** Land And Freedom **GPT-4’s derivation:** (“Aardram”, “director”, “Sibi Malayil”) ; (“Sibi Malayil”, “date of birth”, “2 May 1956”) ; (“Land And Freedom”, “director”, “Ken Loach”) ; (“Ken Loach”, “date of birth”, “17 June 1936”) **GPT-4’s answer:** Land And Freedom
Shared predicate	**Question:** Are the directors of films Penelope (1966 Film) and Sioux Blood both from the same country? **Gold derivation:** (“Penelope (1966 film”) ,“director”, “Arthur Hiller”) ; (“Arthur Hiller”, “country of citizenship”, “Canadian”) ; (“Sioux Blood”, “director”, “John Waters”) ; (“John Waters (director born 1893”) ,“country of citizenship”, “American”) **Gold answer:** No **GPT-4’s derivation:** (“Penelope”, “director”, “Arthur Hiller”) ; (“Arthur Hiller”, “country of birth”, “Canada”) ; (“Sioux Blood”, “director”, “John Ford”) ; (“John Ford”, “country of birth”, “United States”) **GPT-4’s answer:** No

Table 10: Examples of spurious correct answers in bridge-comparison questions. Red text indicates where there was an error in the derivation, blue text indicates that the answer is correct.

Pattern	Derivation examples	
	gold	pred
(i) wording	(Kingdom of the Isles, covered a total land area of, over 8300 km2) (Michaël Llodra, gained victory over, Juan Martín del Potro)	(The Isles, covers, a total land area of over 8300 km2) (Michaël Llodra, defeated, Juan Martín del Potro)
(ii) granularity	(Great Neck School District, is in, the town of North Hempstead, Nassau County, New York, United States) (Disney Magazine, is published quarterly from, December 1965 to April 2005) (Dirk Nowitzki, was born, June 19, 1978)	(Great Neck School District, is located in, Great Neck, New York); (Disney Magazine, ceased publication in,2005) (Dirk Nowitzki, was born in, 1978)
(iii) type	(Shinjo-city, city tree, Cherry tree) (Avengers: Infinity War, previous film, Avengers: Age of Ultron)	(Shinjo-city, city tree, exist) (Avengers: Infinity War, position of the work, third film in the Avengers series)
(iv) information per one triple	(Modest Mouse, was formed in, Issaquah); (Issaquah, is in, Washington) (Finish What Ya Started, features Sammy Hagar, on a rhythm guitar)	(Modest Mouse, formed in, Issaquah, Washington) ("Finish What Ya Started", is a song from, OU812); (OU812, features, Sammy Hagar); (Sammy Hagar,plays,guitar)
(v) form	(Lantern Waste, is the place where, Lucy Pevensie and Mr. Tumnus meet) (The Spiderwick Chronicles (film), follows the adventures on a family as they discover, magical creatures)	(Lucy Pevensie and Mr. Tumnus, meet at, Lantern Waste) (The Spiderwick Chronicles, is about, a New England family who discover magical creatures around their estate)

Table 11: Examples of derivatives that were considered correct.