Lost in the Distance: Large Language Models Struggle to Capture Long-Distance Relational Knowledge

Meiyun Wang*[1], Takeshi Kojima†[1], Yusuke Iwasawa[1], and Yutaka Matsuo[1]

[1]The University of Tokyo

Abstract

Large language models (LLMs) have demonstrated impressive capabilities in handling long contexts, but challenges remain in capturing relational knowledge spread far apart within text. Connecting long-distance knowledge is important for solving tasks as the context length increases: imagine reading a lengthy detective novel where seemingly trivial information introduced early on often becomes essential during the climactic reveal of the culprit. In this study, we expose the "Lost in the Distance" phenomenon, where LLM performance of capturing the relational knowledge degrades significantly when the relational knowledge is separated by noise, i.e., unrelated sentences to solve a task. Specifically, we design an experiment in which we insert artificial noise between two related elements and observe model performance as the distance between them increases. Our findings show that while LLMs can handle edge noise with little impact, their ability to reason about distant relationships declines sharply as the intervening noise grows. These findings are consistent in both forward-looking prediction and backward-looking prediction settings. We validate this across various models (GPT-4, Gemini-1.5-pro, GPT-4o-mini, Gemini-1.5-flash, Claude-3.5-Sonnet) and tasks (causal reasoning and knowledge extraction). These results reveal a significant limitation in how LLMs process relational knowledge over long contexts. We release our code and data to support further research.[1]

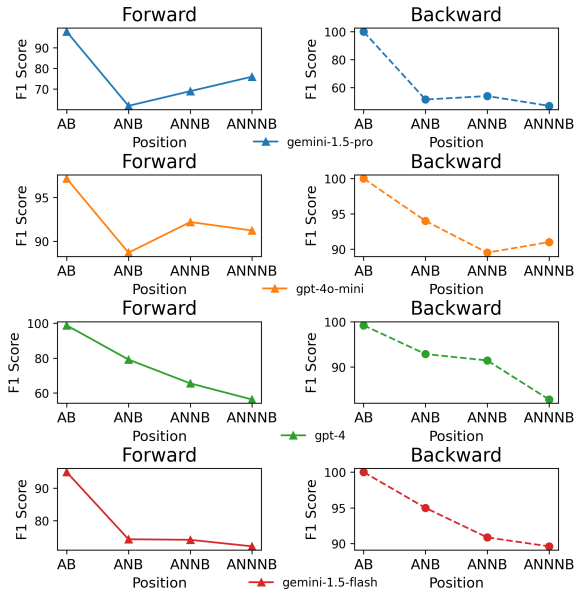

Figure 1: Lost in the Distance: (Left) Forward-looking setting to assess whether LLMs can capture recent knowledge based on distant past relational knowledge. (Right) Backward-looking setting to assess whether LLMs can capture past knowledge based on recent relational knowledge. See Section 3 for details and Appendix A for additional evaluation metrics.

1 Introduction

Recent large language models (LLMs) have demonstrated a remarkable ability to solve long and complex tasks as their capacity to handle longer context lengths has increased significantly. However, as context length grows, LLMs struggle more with filtering out irrelevant information, making it harder to focus on the main task. Liu et al. (2024) identified the "Lost in the Middle" phenomenon, where model performance drops significantly when crucial information is located in the middle of a long context, compared to when it appears at the beginning or end. Similarly, Kamradt (2023) highlighted LLMs' vulnerability in retrieving relevant information from long contexts with their "Needle in a Haystack Test." However, these findings mainly focus on document-level tasks, mimicking the retrieval-augmented generation setup.

This study brings attention to a new challenge

*omiun20@g.ecc.u-tokyo.ac.jp, t.kojima@weblab.t.u-tokyo.ac.jp, iwasawa@weblab.t.u-tokyo.ac.jp, matsuo@weblab.t.u-tokyo.ac.jp

†Corresponding Author

[1]https://github.com/Kirawang23/Lost_in_the_Distance

Findings of the Association for Computational Linguistics: NAACL 2025, pages 4536–4544

System Prompt
Write a high-quality answer to the given question using only the exact words or phrases from the text.
Note that the Text may contain irrelevant information (noise).
Return only the answer by writing 'Answer: XXX'.

User Prompt
Text: (Noise Text1) The first person is Xavier Pendleton. The second person is Uriah Hawthorne. (Noise Text2)
The first person is the champion of the quantum chess world championship, defeating opponents in multiple dimensions.
The second person is the skilled puppeteer who brought the magical world of "Enchanted Strings" to life. (Noise Text3)
Question: What is the name of the skilled puppeteer who brought the magical world of "Enchanted Strings" to life?
Answer:

Table 1: An example prompt for the Name2Description task with the backward-looking prediction setting.

for LLMs when processing long contexts: ***"Can LLMs effectively capture relational knowledge that appears far apart within a context?"*** This ability is crucial, as it affects performance on various downstream tasks such as causal reasoning and knowledge extraction, where the relevant information may be much shorter and hidden within noisy long contexts. To illustrate, imagine reading a lengthy detective novel where seemingly trivial information introduced early on often becomes essential during the climactic reveal of the culprit.

Our experiments show that even state-of-the-art LLMs still struggle to handle relational knowledge when located far apart within a context. Specifically, given artificial noise (N1, N2, N3) and relational knowledge (A and B), we define a context as a sequence {N1, A, N2, B, N3} and ask LLMs to predict B (or A) based on a question about A (or B). The results on GPT-4, Gemini-1.5-pro, GPT-4o-mini, and Gemini-1.5-flash, Claude-3.5-Sonnet in the zero-shot in-context learning (ICL) setting demonstrate that performance significantly drops as the length of the in-between noise (N2) increases. We refer to the phenomenon as ***"Lost in the Distance."*** We observe this phenomenon in both forward-looking prediction and backward-looking prediction settings.

Interestingly, when there is no distance between A and B (i.e., N2 = 0), adding noise at the edges (N1, N3 > 0) alleviates the performance degradation compared to adding in-between noise. This indicates that the location of the noise plays a significant role in performance. We conducted additional experiments with various noise types (e.g., novel excerpts, pre-training corpus samples, and random words) and task types (causal reasoning and knowledge extraction), all showing similar trends.

2 Related Work

Recent LLMs have significantly extended their context lengths to enhance their LLM capabilities in handling complex tasks. For example, early GPT models supported 512 tokens (Radford, 2018), but the latest GPT-4 and GPT-4o-mini models now accommodate up to 8,192 and 128,000 tokens, respectively[2]. Similarly, Gemini-1.5-pro and Gemini-1.5-flash models offer context lengths of 2 million and 1 million tokens, respectively[3].

However, as context lengths grow, research suggests that LLMs may struggle with irrelevant or noisy information, which can interfere with problem-solving. Liu et al. (2024) introduced the "lost in the middle" effect, where LLMs struggle to retrieve information located in the middle of long contexts. Likewise, Kamradt (2023) highlighted the difficulty of retrieving specific details from vast amounts of text in their "Needle in a Haystack Test." Furthermore, Peysakhovich and Lerer (2023) found that as context length increases, LLMs tend to disproportionately prioritize more recent information, known as "recency bias."

While these studies focus on retrieving information from a single point within long contexts, our work addresses a more complex challenge: identifying relationships between two distinct points in the text—essentially, "finding two needles in a haystack." Previous research, such as Levy et al. (2024) and Shi et al. (2023), has shown that inserting irrelevant noise between key elements of a task can significantly impair multi-step reasoning. However, these studies did not explore the specific impact of intermediate noise on retrieving atomic-level knowledge relations, which may serve as the foundation for multi-step reasoning. In contrast, our study directly measures how well LLMs

[2] https://platform.openai.com/docs/models
[3] https://ai.google.dev/gemini-api/docs/models/gemini

can recall atomic-level factual information across long distances in both forward and backward directions, offering new insights into how noise affects various tasks and advancing our understanding of multi-step reasoning.

3 Experiment

3.1 Setting

Model Our experiments use the following four LLMs via API: Claude-3.5-Sonnet (claude-3-5-sonnet-20240620), GPT-4 (gpt-4-2023-06-13), Gemini-1.5-Pro (gemini-1.5-pro-002), GPT-4o-mini (gpt-4o-mini-2024-07-18), and Gemini-1.5-Flash (gemini-1.5-flash-002). We set greedy decoding (temperature set to zero) and limit the maximum output tokens to 100 for rigid predictions.

Noise We select three types of noise: novel, random words, and pre-training corpus. For the novel noise, we pick the top 10 popular ebooks from Project Gutenberg (Gutenberg). For the random words, we randomly select words from WordNet (Miller, 1995). For the pre-training corpus, we use FineWeb (Penedo et al., 2024) and RedPajama (Computer, 2023).

Task Our tasks include "Name2Description", "Parent2Child", and "Cause2Effect". "Name2Description" consists of relations between a fictitious name and his/her fictitious description (Berglund et al., 2023). "Parent2Child" includes relations between a fictitious parent's name and his/her fictitious child's name (Berglund et al., 2023), and "Cause2Effect" contains relations between a causal phrase and its consequent effect phrase (Du et al., 2022).

Prompt We randomly select noise tokens from various noise sources, split them into three chunks, and place them in different positions. N1, N2, and N3 represent the noise chunks in different positions (1: before A, 2: between A and B, 3: after B). Given noise tokens {N1, N2, N3} and relational knowledge A and B, our experiments define a text in context as a sequence of {N1, A, N2, B, N3}. We use LLMs to predict B based on a question about A (forward prediction) and to predict A based on a question about B (backward prediction). We analyze knowledge recall performance by varying the length of N1, N2 and N3. Unless otherwise stated, our experiments use the Gemini-1.5-pro model, the Name2Description task, and

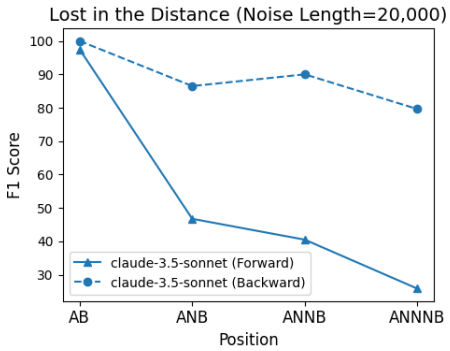

Figure 2: Lost in the Distance with extra-long noise (20,000 tokens for ANNNB) for Claude-3.5-Sonnet .

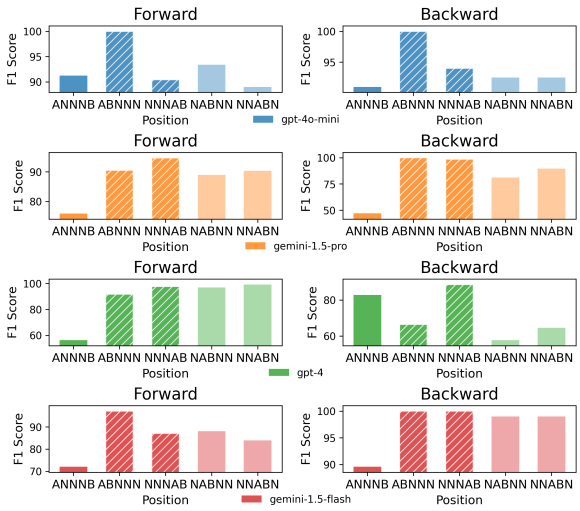

Figure 3: No Distance, Less Degradation.

texts from novels, with a maximum of 3,000 noise tokens as the default setting, and 200 samples tested per task. Table 1 provides an example prompt for the Name2Description task experiment. See Appendix B for the other task settings.

Evaluation We let LLMs generate an answer text given the prompt in a zero-shot setting. We use the F1 score for the evaluation metric instead of the exact match to allow for minor perturbations in the response. Specifically, we calculate the F1 score by taking the harmonic mean of precision and recall, where precision = (# of matched words between prediction and ground truth) / (# of predicted words) and recall = (# of matched words) / (# of ground truth words).

3.2 Result

Lost in the Distance We conducted experiments by varying the length of the intermediate noise (N2) to values of 0, 1000, 2000, and 3000. In the im-

plementation, we split all noise tokens into three chunks of equal length, each referred to as N, and concatenate them to form N2. Figures 1 present the results. The results show that as the intermediate noise increases, the performance of relational knowledge recall decreases significantly. Although different models exhibit varying degrees of performance decline due to their inherent architectural differences and required parameter adjustments, all models consistently follow a downward trend. We hypothesize that the intermediate noise tokens, which are irrelevant information, interfere with the proper attention between knowledge A and B, leading to confusion in the knowledge connection. We refer to this phenomenon as "Lost in the Distance."

Figures 2 show the performance of the state-of-the-art Claude-3.5-Sonnet model with different noise lengths (N2). We set the maximum length to 20,000 tokens, split it into three chunks of equal length, and concatenate them to generate N2. The results indicate that as N2 increases from 6,000 to 20,000 tokens, even the newest LLM struggles to retrieve relational knowledge, providing strong evidence of the "lost in the distance" phenomenon.

No Distance, Less Degradation To complement our findings, we compared the results when noise is placed between the knowledge pair (i.e., {A N2 B}) and when it is not, across three configurations: {A B N3}, {N1 A B}, and {N1 A B N3}. In these three configurations, there is no distance between A and B (i.e., N2 = 0), but with the same amount of total noise added at the edges. We concatenate the noise chunks to generate different noise tokens in different positions (e.g., in the case of {A B N3}, three noise chunks are used to form N3). Figures 3 present the F1 scores for both forward and backward predictions. The leftmost bar corresponds to the "Lost in the Distance" phenomenon, while the rightmost two bars represent the "Lost in the Middle" phenomenon. The relatively lower performance of the leftmost bar suggests that while intermediate noise significantly affects performance, noise at the beginning and end has a much smaller impact. This effect is also influenced by the characteristics of the noise. To further investigate this, we designed a task where the context is more closely tied to the noise. Specifically, we extracted character names from each noise chunk of the novel and used these names directly as questions and answers. For each question, we paired the names with randomly assigned identity information, prompting

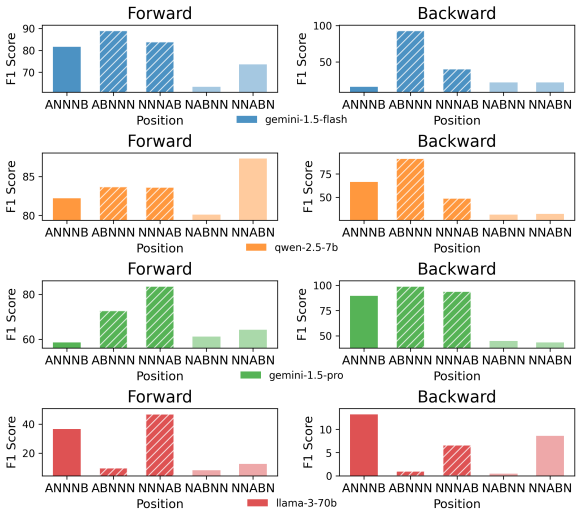

Figure 4: No Distance, Less Degradation: Effect of Context-Related Noise.

the model to identify either the identity associated with a given name or the name corresponding to a specific identity. As the character names appear at different positions within the noise, they introduce varying levels of interference to the model's reasoning. We evaluated this task on several models, including Gemini-1.5-flash, Qwen-2.5-7B, Llama-3-70B, and Gemini-1.5-pro. Figures 4 show that the "Lost in the Distance" phenomenon remains observable even when the noise is more contextually relevant. However, because the interfering elements in the noise appear at different positions, the specific performance varies. Notably, the Llama model generally exhibits lower accuracy, and its output sentences often contain long sequences of noise that directly incorporate the query, leading to differences between its calculated F1 scores and the actual performance of the other models. Another finding is that "Lost in the Middle" phenomenon becomes more significant when the noise is more contextually relevant. Note that "Lost in the Middle" and "Lost in the Distance" are orthogonal phenomena so that they can independently coexist with each other. See Appendix B for further details on this task.

Ablation Study Across Tasks and Noises We validate the "Lost in the Distance" effect across different tasks, as shown in Figures 5. We consistently observe an overall decline in performance as intermediate noise increases, though to varying extents. One interesting observation is that in the Cause2Effect task, performance declines more

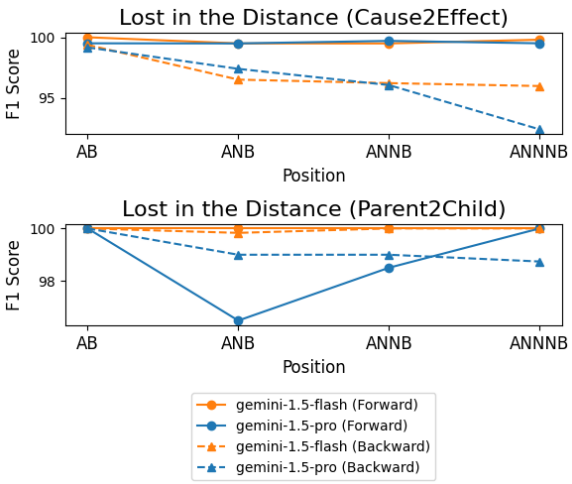

Figure 5: Lost in the Distance (Different Tasks)

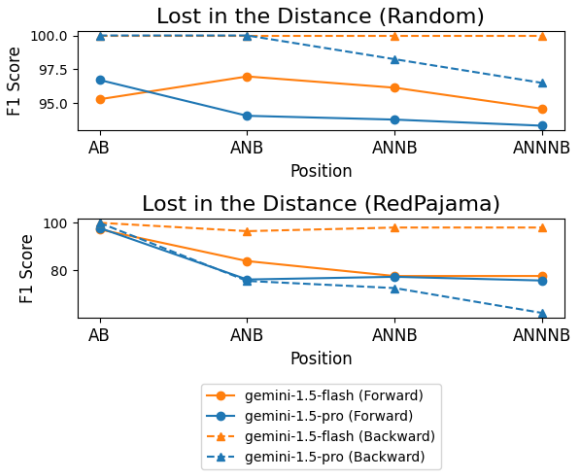

Figure 6: Lost in the Distance (Different Noises)

sharply in backward prediction than in forward prediction. As Berglund et al. (2023) have pointed out, the result demonstrates that LLMs inherently struggle with retrieving information in a backward manner, which is called the "reversal curse." However, we did not observe a clear trend in the Parent2Child task. Nearly all models performed well, possibly because the novel-related noise had a limited impact on this task. In the future, we will explore more diverse tasks to evaluate how different types of noise affect various tasks. Figures 6 show the results under different noise conditions. We can easily observe the "Lost in the Distance" effect across various noise types. The noise types for RedPajama show clear declines as noise increases. In addition, the difficulty of the noise also influences performance. For example, the random noise type contains only random words without any logical structure, making it easier for LLMs to identify as noise. On the other hand, the novel noise type poses greater challenges as it contains a more logical structure and can mislead the model into making incorrect decisions.

4 Conclusion

This study has highlighted that current LLMs capable of handling long contexts still struggle to connect distant relational knowledge both in forward and backward prediction, which we refer to as "lost in the distance." We hope these findings will draw more attention to the importance of distant relational knowledge, leading to the development of new benchmarks and proposals for performance improvement in future work.

5 Limitation

We limited the noise experiments to approximately 4,000 tokens due to budget constraints. However, considering that recent LLMs are expanding context windows beyond our experiment size (see Section 2), evaluating with longer noise contexts is an important direction for future research.

In this study, we created artificial tasks that require long-distance relational knowledge by inserting unrelated texts as noise within a context, as there are no suitable existing benchmarks for measuring such performance. Therefore, future work needs to develop tasks that assess more realistic long-distance knowledge relationships within consistent, long texts.

While this research has identified a new phenomenon, "Lost in the Distance," it is also crucial for future work to propose methods to address problems related to this phenomenon.

References

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446.

Project Gutenberg. Project gutenberg.

Greg Kamradt. 2023. The needle in a haystack test.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *Preprint*, arXiv:2310.01427.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

A Evaluation Metrics

We report Precision, Recall, F1 Score, and Exact Match metrics to demonstrate the impact of the "lost in the distance" phenomenon across different models, as presented in Table 2. While the models show varying levels of performance degradation, a consistent downward trend is observed across all metrics.

B Template Types

Table 3 and 4 show the templates for each task (Parent2Child and Cause2Effect). To make the tasks solvable only when the LLMs pay attention to both A and B, we provide two sets of relationships in the text and randomize their order within the context of each task. We set a question based on the randomly chosen knowledge from each of the two pairs. Table 5 presents the template for the Name2Description task with context-related noise. We collected character names from various novels and used them as both questions and answers for testing.

C Case Study: Output Examples

Table 6 and 7 describes failure and successful output examples from Gemini-1.5-pro model.

Model	Direction	Position	Precision	Recall	F1 Score	Exact Match
gemini-1.5-flash	Forward	AB	97.12	94.34	94.90	93.5
gemini-1.5-flash	Forward	ANB	76.57	73.70	74.25	67.5
gemini-1.5-flash	Forward	ANNB	75.76	74.06	74.08	67.5
gemini-1.5-flash	Forward	ANNNB	73.87	71.89	72.07	67.5
gemini-1.5-flash	Backward	AB	100.00	100.00	100.00	100.0
gemini-1.5-flash	Backward	ANB	95.00	95.00	95.00	95.0
gemini-1.5-flash	Backward	ANNB	90.74	91.25	90.86	90.5
gemini-1.5-flash	Backward	ANNNB	89.39	90.50	89.61	89.0
gemini-1.5-pro	Forward	AB	98.00	97.53	97.69	97.5
gemini-1.5-pro	Forward	ANB	67.03	61.35	61.90	55.0
gemini-1.5-pro	Forward	ANNB	73.08	68.56	68.98	62.5
gemini-1.5-pro	Forward	ANNNB	80.08	75.18	75.97	70.0
gemini-1.5-pro	Backward	AB	100.00	100.00	100.00	100.0
gemini-1.5-pro	Backward	ANB	51.50	51.50	51.50	51.5
gemini-1.5-pro	Backward	ANNB	54.00	54.00	54.00	54.0
gemini-1.5-pro	Backward	ANNNB	47.00	47.00	47.00	47.0
gpt-4	Forward	AB	97.98	100.00	98.80	93.5
gpt-4	Forward	ANB	78.86	83.41	79.21	54.0
gpt-4	Forward	ANNB	69.04	68.03	65.55	39.5
gpt-4	Forward	ANNNB	62.24	56.65	56.32	38.0
gpt-4	Backward	AB	99.12	100.00	99.21	99.0
gpt-4	Backward	ANB	92.04	99.50	92.85	91.0
gpt-4	Backward	ANNB	90.58	98.50	91.43	89.5
gpt-4	Backward	ANNNB	80.91	97.50	82.78	78.5
gpt-4o-mini	Forward	AB	99.50	96.46	97.16	95.0
gpt-4o-mini	Forward	ANB	88.74	89.04	88.69	86.5
gpt-4o-mini	Forward	ANNB	92.33	92.30	92.19	90.5
gpt-4o-mini	Forward	ANNNB	91.43	91.40	91.22	89.0
gpt-4o-mini	Backward	AB	100.00	100.00	100.00	100.0
gpt-4o-mini	Backward	ANB	94.00	94.00	94.00	94.0
gpt-4o-mini	Backward	ANNB	89.50	89.50	89.50	89.5
gpt-4o-mini	Backward	ANNNB	91.00	91.00	91.00	91.0

Table 2: Experimental Results for Lost in the Distance (as shown in Figure 1)

System Prompt
Write a high-quality answer to the given question using only the exact words or phrases from the text.
Note that the Text may contain irrelevant information (noise).
Return only the answer by writing 'Answer: XXX'.

User Prompt
Text: (Noise Text1) The second person is Leonard Bertram Oldman. The first person is Josephine Miller. (Noise Text2) The first person's child is Sienna Miller. The second person's child is Gary Oldman. (Noise Text3)
Question: What is the name of Leonard Bertram Oldman's child?
Answer:

Table 3: An example prompt for Parent2Child task with the forward-looking prediction setting.

System Prompt
Write a high-quality answer to the given question using only the exact words or phrases from the text.
Note that the Text may contain irrelevant information (noise).
Return only the answer by writing 'Answer: XXX'.

User Prompt
Text: (Noise Text1) The second event is "strong interactions occur between gluons". The first event is "he wanted to eat biscuits". (Noise Text2) The second event's effect is "these carrier particles were close to each other". The first event's effect is "his mother made quick bread made with a leavening agent such as baking powder for him". (Noise Text3)
Question: What is the effect of "strong interactions occur between gluons"?
Answer:

Table 4: An example prompt for Cause2Effect task with the forward-looking prediction setting.

System Prompt
Write a high-quality answer to the given question using only the exact words or phrases from the text.
Note that the Text may contain irrelevant information (noise).
Return only the answer by writing 'Answer: XXX'.

User Prompt
Text: The second person is Sherlock Holmes. The first person is James Windibank ... Her boots I didn't observe. She had small round, hanging gold earrings, and a general air of being fairly well-to-do in a vulgar, comfortable, easy-going way." Sherlock Holmes clapped his hands softly together and chuckled... Finally, the second person became the master mixologist who crafted the legendary nebula cocktail. The first person became the brilliant strategist who led the victorious cyberspace defense force.
Question: What is the name of the master mixologist who crafted the legendary nebula cocktail?
Answer:

Table 5: An example prompt for Name2Description task with context-related noise in the backward-looking prediction setting.

System Prompt
Write a high-quality answer to the given question using only the exact words or phrases from the text.
Note that the Text may contain irrelevant information (noise).
Return only the answer by writing 'Answer: XXX'.

User Prompt
Text: The second person is Harrison Ashford. The first person is Nola Westbrook. (...Hasten, been or a that Roger its this (of to matter, one down bosom, ... her which well it dispense donations of brother all her itself the emphasis scarlet God will Chillingworth, by and upon as and so be) The second person is the groundbreaking geneticist who developed the cure for the common cold. The first person is the virtuoso violinist who debuted the symphony of the stars.
Question: What is the name of the groundbreaking geneticist who developed the cure for the common cold?
Answer: Roger Chillingworth. ✗ (Correct answer is Harrison Ashford)

Table 6: A failure example by Gemini-1.5-pro for Name2Description task (Noise type: Novel, Noise position: ANNNB) in the backward-looking prediction setting.

System Prompt
Write a high-quality answer to the given question using only the exact words or phrases from the text.
Note that the Text may contain irrelevant information (noise).
Return only the answer by writing 'Answer: XXX'.

User Prompt
Text: The second event is "the breeder feeds pelicans a lot of food every day." The first event is "the physicist gave an electron to a single charged ion." The first event's effect is "he found that neutralization occured." The second event's effect is "pelicans is very healthy."
Question: What is the cause of " he found that neutralization occured "?
Answer: the physicist gave an electron to a single charged ion ✓

Table 7: A successful example by Gemini-1.5-pro for Cause2Effect task (Without any noise: AB) in the backward-looking prediction setting.