In-Context Learning May Not Elicit Trustworthy Reasoning: A-Not-B Errors in Pretrained Language Models

Pengrui Han [*1,2] **Peiyang Song** [*3] **Haofei Yu** [1] **Jiaxuan You** [1]

Abstract

Recent advancements in artificial intelligence (AI) have led to the development of highly capable large language models (LLMs) demonstrating significant human-like abilities. Yet these pretrained LLMs are often vulnerable to interesting cognitive biases. In this work, we study the A-Not-B error – a developmental stage for human infants, characterized by the persistence of previously rewarded behavior despite changed conditions that warrant even trivial adaptation. Our investigation reveals that LLMs, akin to human infants, erroneously apply past successful responses to slightly altered contexts. Employing various reasoning tasks, we demonstrate that LLMs are susceptible to the A-Not-B error. Notably, smaller models exhibit heightened vulnerability, mirroring the developmental trajectory of human infants. Models pretrained with extensive, high-quality data show significant resilience, highlighting the importance of internal knowledge quality, similar to how rich experiences enhance human cognitive abilities. Furthermore, increasing the number of examples before a context change leads to more pronounced failures, highlighting that LLMs are fundamentally pattern-driven and may falter with minor, non-erroneous changes merely in patterns. We open source all code and results under a permissive MIT license, to encourage reproduction and further research exploration [1].

1. Introduction

In the field of cognitive science, there is a classic cognitive phenomenon called the A-Not-B error (Popick et al., 2011; Smith & Gasser, 2005; Vorms, 2012). In a typical A-Not-B task, an infant repeatedly retrieves an object from one location (Location A) but continues to search for it in the same location even after seeing it being moved into a new location (Location B). The persistence of this error and its eventual resolution reflect significant developmental milestones in cognitive abilities of human infants. It is a key indication of the balance between working memory and long-term memory (Diamond, 1998; Cuevas & Bell, 2010). The resolution of A-not-B errors marks a stage closely tied to the emergence of self-locomotion, where an infant finally develops representations of spatio-temporal relationships, objects, space, and self (Smith & Gasser, 2005).

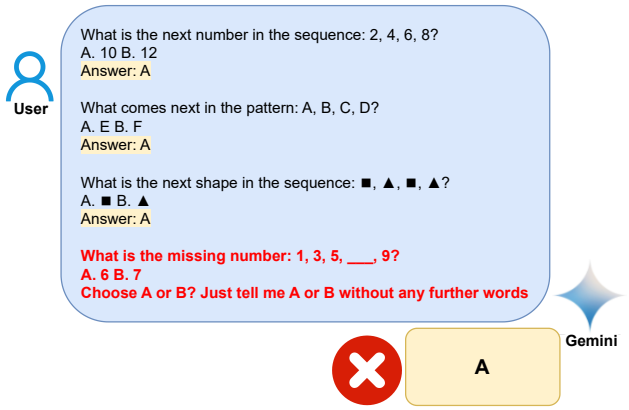

Figure 1. **A-Not-B Adversarial Few-Shot Prompts Mislead Gemini on Simple Questions.** This figure presents an example prompt that tricks advanced model Gemini on simple questions by consistently providing examples with the answer A. The experiments were conducted on June 12, with the note that future updates might lead to different results.

Recent AI advancements, especially with Large Language Models (LLMs) (Saravanan et al., 2023), have significantly impacted many sectors (Feng et al., 2024). These models exhibit not only remarkable human-like cognitive abilities (Ruan et al., 2023; Huang et al., 2022; Han et al., 2024; Zhang et al., 2023; Song et al., 2024; Street et al., 2024), such as reasoning (Wei et al., 2022; Yao et al., 2023; Cai et al., 2023), but also demonstrate great potential to operate within real-world contexts like humans. Through approaches such as in-context learning (ICL) (Xie et al., 2021; Min et al., 2022) and prompt engineering (Giray, 2023),

[*]Equal contribution [1]University of Illinois Urbana-Champaign [2]Caleton College [3]California Institute of Technology. Correspondence to: Pengrui Han <barryhan@carleton.edu>, Peiyang Song <psong@caltech.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]https://github.com/Peiyang-Song/LLM-A-Not-B-Errors

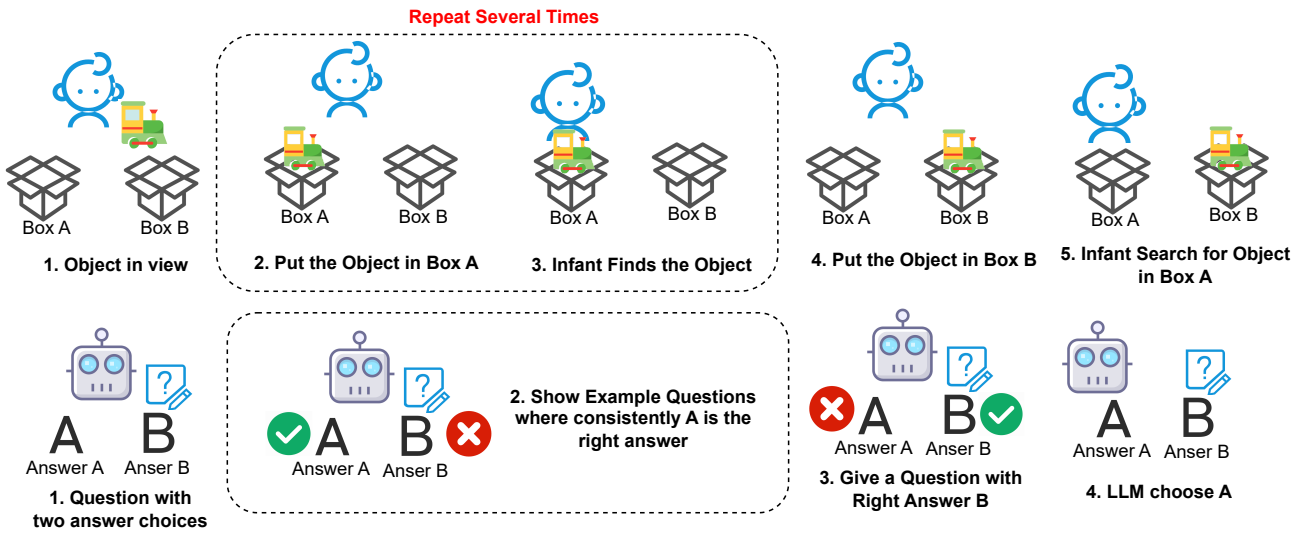

Figure 2. **Illustration of A-not-B Task Peformance in Infants and LLMs.** This figure demonstrates the typical A-not-B error using a two-location (Box A and Box B) setup. The first sequences showcase an infant's repeated actions: placing an object in Box A, observing it being moved to Box B, yet continuing to search in Box A. This depicts the cognitive phenomenon where prior experience overrides current visual cues. On the bottom sequence, the figure analogously presents a scenario in which LLMs are misled by a consistent answer pattern, illustrating the A-not-B type scenarios for LLMs, where these models fail to adapt to changed circumstances.

LLMs can quickly learn and understand from input contexts, reacting with their internal knowledge and capabilities. This is very similar to how humans assimilate information from their environment and apply their knowledge and reasoning skills to navigate and respond to various real-world situations.

However, in scenarios similar to the A-not-B error, where models are shown examples with consistent answers as A and then presented with a new question requiring a different answer B, as illustrated in Figure 1, we see that advanced LLMs like Gemini fail to answer even the simplest questions correctly. This is surprising as even a child with just elementary school mathematical knowledge would not be misled by such an easy trick. The fact that LLMs are highly susceptible to this A-not-B scenarios is an interesting and significant failure case. It reflects LLMs' inability to reason in a trustworthy and consistent manner and casts doubt on whether LLMs actually possess knowledge despite being performant on certain benchmarks.

Therefore, in this work, we continue from the surprising qualitative examples and conduct comprehensive experiments across different reasoning tasks, and prove the generality of this failure case of LLMs. Most importantly, we find that:

- LLMs are frequently significantly misled by A-Not-B style adversarial prompts;
- Smaller models are more susceptible to input changes, paralleling the vulnerability seen in human infants whose cognitive development is still in progress;
- Models pretrained with more extensive and higher quality data are significantly more resilient, highlighting the quality of internal knowledge is crucial when interacting with contexts, akin to rich experiences enhance human cognitive ability;
- Increasing example shots will lead to more failure cases, shedding light on the fact that LLMs are inherently pattern-driven and can fail due to even minor pattern changes that contain no incorrect information.

2. Experiments

2.1. Experiments Setup

Experiment Motivations. Motivated by the original A-not-B experiment in cognitive science, we replicate the scenario with LLMs shown in Figure 2. Before delving into our datasets and detailed experiment setups, we discuss the correspondence between core elements in our designed experiment and those in the original A-not-B experiment. In the A-not-B experiment, the infant constantly observes the placement of "a certain object" to be in the same "Location A". In our experiment, this is parallel to LLMs observing the placement of "the ground truth answer to a MCQ question" to be the same "Option A". Similar to how the infant then observes the placement of the ball in a different "Location B", LLMs are then tasked with a similar style MCQ from the same domain whose ground truth answer is supposed to be "Option B". The infant or the LLMs then chooses

A-Not-B Error		Arithmetic MathQA			Commonsense CommonsenseQA			Causal Winogrande			Scientific SciQ		
Models	# of Shots	Original	A-not-B	Change	Original	A-not-B	Change	Original	A-not-B	Change	Original	A-not-B	Change
Llama3_70B	3	32%	36%	↑12.5%	84%	88%	↑4.8%	32%	36%	↑12.5%	96%	100%	↑4.2%
	5	36%	42%	↑16.7%	86%	86%	↓0.0%	76%	80%	↑5.3%	98%	100%	↑2.0%
	10	36%	32%	↓11.1%	86%	88%	↑2.3%	78%	80%	↑2.6%	100%	100%	↓0.0%
	25	28%	24%	↓14.3%	92%	90%	↓2.2%	86%	78%	↓9.3%	100%	100%	↓0.0%
Llama3_8B	3	62%	22%	↓64.5%	86%	74%	↓14.0%	46%	64%	↑39.1%	96%	90%	↓6.2%
	5	42%	14%	↓66.7%	92%	82%	↓10.9%	54%	76%	↑40.7%	94%	94%	↓0.0%
	10	32%	8%	↓75.0%	94%	82%	↓12.8%	50%	52%	↑4.0%	98%	96%	↓2.0%
	25	36%	6%	↓83.3%	92%	62%	↓32.6%	50%	28%	↓44.0%	96%	90%	↓6.2%
Qwen1.5_72B	3	66%	68%	↑3.0%	96%	96%	↓0.0%	80%	82%	↑2.5%	96%	98%	↑2.1%
	5	56%	50%	↓10.7%	94%	92%	↓2.1%	80%	82%	↑2.5%	94%	96%	↑2.1%
	10	56%	44%	↓21.4%	94%	92%	↓2.1%	74%	76%	↑2.7%	92%	96%	↑4.3%
	25	50%	28%	↓44.0%	94%	92%	↓2.1%	82%	82%	↓0%	92%	94%	↑2.2%
Qwen1.5_7B	3	64%	88%	↑37.5%	84%	86%	↑2.4%	60%	70%	↑16.7%	96%	94%	↓2.1%
	5	72%	90%	↑25.0%	88%	90%	↓2.3%	64%	80%	↑25.0%	94%	92%	↓2.1%
	10	76%	92%	↑21.1%	92%	94%	↑2.2%	86%	82%	↓4.7%	94%	92%	↓2.1%
	25	86%	92%	↑7.0%	98%	96%	↓2.6%	98%	96%	↓2.3%	96%	94%	↓2.1%

Table 1. **Main Result: LLMs are misled by A-Not-B style adversarial prompts.** This table presents the results for all four models across four different reasoning tasks. Accuracy drops are denoted in blue, while accuracy increases are shown in red, both indicating that the LLMs are influenced by the A-not-B style adversarial prompts.

the wrong answer "A". This is a cognitive error because an infant would know well that the ball is in "Location B" if without the previous demonstrations of "ball discovered in "Location A". This corresponds to how LLMs *are* able to choose the correct answer (Option B) if it had not seen the previous MCQs with Option A as the correct answers. That is, LLMs *do* have the capability to identify the correct answers for some MCQs but fail to do so in this mimiced A-not-B scenario. This easily translates to an overall accuracy drop.

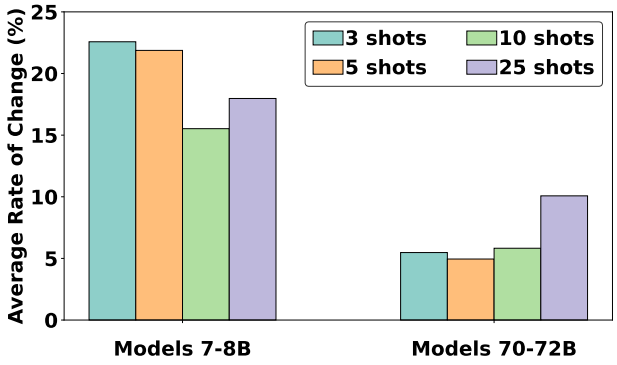

Figure 3. **Smaller models are more vulnerable to A-not-B style adversarial prompts.** Average performance variation across different numbers of few-shot examples for large and small models on four reasoning tasks are shown.

Datasets. Few-shot prompting has been widely adopted for various reasoning tasks of LLMs. To investigate whether LLMs are susceptible to A-not-B errors during few-shot prompting, we choose four representative Question-Answering (QA) datasets, each for a particular category of reasoning task. Specifically, we choose the MathQA dataset (Amini et al., 2019) for arithmetic reasoning, the CommonsenseQA dataset (Talmor et al., 2019) for commonsense reasoning, the winogrande dataset (Sakaguchi et al., 2019) for causal reasoning, and the SciQ dataset (Welbl et al., 2017) for scientific reasoning. All the four datasets consist of multiple-choice questions (MCQs). We preprocess the datasets to split a QA sample into three parts: the question, the choices, and the ground truth answer. Then we modify the QA samples so that each sample has only two choices left, with one of them being the ground truth answer and the other being an incorrect answer. The scenario thus loyally resembles the original setting where A-not-B errors were first observed, in that there are two possible answers and only one is correct. Next, we discuss the detailed settings of experiments.

Experiment Settings. With the modified datasets, we conduct experiments for each reasoning task. We test the models' performance on the modified datasets in two different settings: the original setting and the adversarial setting. In

the original setting, we construct a prompt in the standard few-shot paradigm – we first provide n QA examples, and then ask one last question for the model to answer. We then check if the model's answer agrees with the ground truth. The adversarial setting differs from the original setting mainly in the few shots provided before the question. As shown in Figure 2, in the adversarial setting, we reorder the options so that the answers for all the examples we provide are the first one (Choice A). Then for the final question being asked, we similarly reorder the options so that the ground truth is the second option (Choice B).

With the qualitative examples we reported from state-of-the-art closed-source models such as GPT-4 (OpenAI, 2024) and Gemini (Team, 2024) in Figure 1, we proceed to the larger-scale experiments using representative open-source models such as Llama3 (AI@Meta, 2024) and QWen-1.5 (Bai et al., 2023). For each model family, we experiment with both larger and smaller size models, to investigate the relationship between the impact of A-not-B errors and model size. Specifically, we experiment with Llama3_70B (Llama large), Llama3_8B (Llama small), Qwen-1.5_72B (Qwen large), Qwen-1.5_7B (Qwen small). For each model, we test both the original and the adversarial settings we describe above. For each setting, we test each model on 100 data samples per task, and calculate the success rate. We report the success rates as percentages in Table 1, and then calculate the rate of change as

$$\text{Change} = \frac{\text{Adversarial} - \text{Original}}{\text{Original}}$$

We mark an accuracy drop with a downside arrow and blue color, while an accuracy increase is noted by an upside arrow and red color. We will analyze the results in detail in the coming subsection.

3. Results

Impact of Model Size. In Figure 3, we compare the average rate of change across different numbers of few-shot examples for all four reasoning tasks between large (Llama3_70B and Qwen-1.5_72B) and small (Llama3_8B and Qwen-1.5_7B) models. The rate of change counts both accuracy increase and accuracy drop, which is implemented by taking the absolute values of individual chanages and averaging them. Results in Figure 3 show that model size significantly impacts performance, with smaller models showing much greater rates of change. Specifically, the absolute rate of change for smaller models ranges from 15.5% to 22.5%, while for larger models it ranges from 5% to 10.1%. This indicates that smaller models are more susceptible to the A-not-B style adversarial prompts.

Number of Few-Shot Examples. In Figure 4, we compare the average performance of large (Llama3_70B and Qwen-1.5_72B) and small (Llama3_8B and Qwen-1.5_7B) models with different numbers of few-shot examples. Since the model performance is of interest, we calculate the performance change assigned numbers. In Figure 4, a negative number indicates that the model's success rate on the dataset drops for the A-not-B style adversarial prompt, compared to the original case. A positive number shows the opposite, where the model's success rate increases from the original case. Naturally, as the number of few-shot examples in the A-not-B style adversarial prompts increases, both large and small language models are significantly more likely to suffer from an accuracy drop.

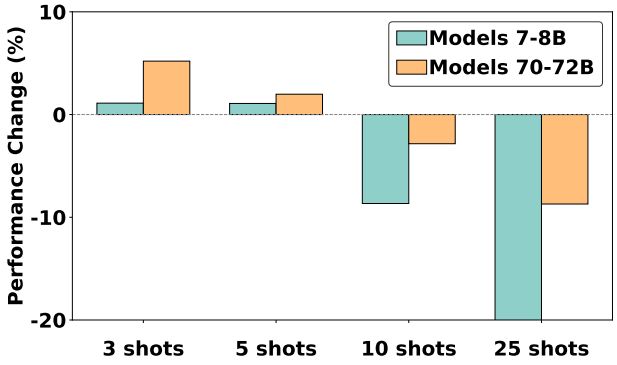

Figure 4. **More Few-Shot Examples Lead to Significant Drop in Model Accuracy.** This figure shows the performance impact of increasing few-shot examples on large and small models. Negative values represent a decline in accuracy due to adversarial prompts, emphasizing that additional A-not-B styled prompts consistently decrease accuracy across both model sizes.

Reasoning Tasks. During the experiments, we observed that model performance and vulnerability to adversarial prompts varied across different reasoning tasks. The change in model performance when seeing A-not-B style prompts is most pronounced in the arithmetic reasoning dataset. Arithmetic reasoning generally requires complex abilities, which may make it likely for LLMs to attempt to identify and rely on patterns when they are unsure about solving the problem (We further explore LLMs' self-explanation on this challenging task in the Ablation experiments, see Section B). This effect is slightly less obvious in the commonsense and causal reasoning datasets, though significant fluctuations are still evident. Commonsense and causal reasoning are generally less complicated than arithmetic reasoning, as they usually involve a certain amount of, if not merely, memorization of certain facts. Yet certain reasoning steps can still be necessary to solve the problems. However, the variation is not as evident in the scientific reasoning dataset. Since this coincides with extremely high accuracy numbers, we suspect this could be attributed to possible data contamination. Another possible explanation can be that many specific terms and technologies in scientific reason-

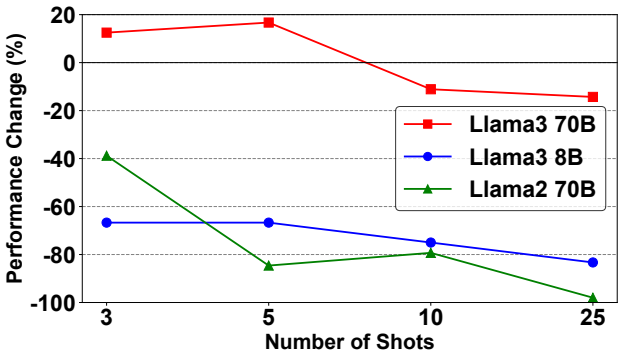

Figure 5. **Pretraining Quality Affects Model Resilience to A-not-B Errors.** The performance change of Llama3_70B, Llama3_8B, and Llama2_70B across different numbers of shots is compared to understand the influence of pretraining datasets on models' resistance to the A-not-B error. Despite having the same model size, Llama2_70B shows significantly greater performance degradation than Llama3_70B and even Llama3_8B, highlighting the significance of specific pretraining data and mechanisms.

ing questions are present in the models already, which can help LLMs make accurate predictions confidently, thus less likely to be impacted by the A-not-B style prompts.

Impact from Model Pretraining. To better understand how model pretraining, particularly the pretraining datasets, can influence models' resistance to the A-not-B type error, we compare the rate of change in accuracy across different shot numbers for Llama3_70B, Llama3_8B, and Llama2_70B, as shown in Figure 5. These models share similar fundamental architectures, but the Llama3 models have been trained on significantly larger, novel, and higher-quality datasets compared to Llama2 70B. As illustrated in Figure 5, although Llama2_70B has the same model size as Llama3_70B, it is much more significantly impacted. In fact, Llama2_70B is less resilient compared to the smaller Llama3_8B, especially with larger numbers of shots. This suggests that the quality and quantity of pretraining data play crucial roles in enhancing model robustness and performance, even surpassing the benefits of model size alone.

4. Conclusion

In this paper, we have explored the intriguing cognitive phenomenon of A-Not-B errors within the domain of LLMs. Our findings illustrate that, akin to human infants, even sophisticated models like LLMs are prone to persist in previously successful responses despite changed contexts—a vulnerability that reveals fundamental limitations in their reasoning capabilities.

Notably, our experiments demonstrate that smaller LLMs, analogous to younger human cognitive development stages, exhibit heightened susceptibility to such errors. This aligns with developmental psychology insights, emphasizing the parallel between increasing model size and human cognitive maturity. Moreover, the robustness of models pre-trained with extensive, diverse datasets underscores the importance of quality and variety in training data, mirroring the way rich human experiences can bolster cognitive flexibility.

The discussion and ablation studies, detailed in the appendix, extend these insights by dissecting the models' responses under varied experimental conditions, including self-explanation and many-shot scenarios. They also provide further insights into the connections between human A-not-B errors and LLMs, as well as the learning paradigms LLMs use to interact with the environment.

We encourage future research into the cognitive aspects of AI, where we hope understanding the connections and differences between human and AI will guide the creation of better-designed models and frameworks, and these models can more accurately mimic human reasoning and be more adept at handling real-world complexities.

References

Acad, A. N., Sci, Fedorenko, E., and Varley, R. A. Annals of the new york academy of sciences language and thought are not the same thing: Evidence from neuroimaging and neurological patients. URL https://api.semanticscholar.org/CorpusID:5043404.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Amini, A., Gabriel, S., Lin, S., Koncel-Kedziorski, R., Choi, Y., and Hajishirzi, H. MathQA: Towards interpretable math word problem solving with operation-based formalisms. 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)*, pp. 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245. URL https://aclanthology.org/N19-1245.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., and Zhu, T. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: Llms trained on "a is b" fail to learn "b is a", 2024.

Bottou, L. From machine learning to machine reasoning, 2011.

Cai, T., Wang, X., Ma, T., Chen, X., and Zhou, D. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*, 2023.

Chen, A., Phang, J., Parrish, A., Padmakumar, V., Zhao, C., Bowman, S. R., and Cho, K. Two failures of self-consistency in the multi-step reasoning of llms, 2024.

Cuevas, K. and Bell, M. A. Developmental progression of looking and reaching performance on the a-not-b task. *Developmental Psychology*, 46(5):1363, 2010.

Deng, B., Wang, W., Feng, F., Deng, Y., Wang, Q., and He, X. Attack prompt generation for red teaming and defending large language models. *arXiv preprint arXiv:2310.12505*, 2023.

Diamond, A. Understanding the a-not-b error: working memory vs. reinforced response, or active trace vs. latent trace. *Developmental Science*, 1(2), 1998.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Fedorenko, E., Piantadosi, S. T., and Gibson, E. A. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586, 2024.

Feng, T., Jin, C., Liu, J., Zhu, K., Tu, H., Cheng, Z., Lin, G., and You, J. How far are we from agi, 2024.

Gambardella, A., Iwasawa, Y., and Matsuo, Y. Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks. *arXiv preprint arXiv:2406.02356*, 2024.

Giray, L. Prompt engineering with chatgpt: a guide for academic writers. *Annals of biomedical engineering*, 51 (12):2629–2633, 2023.

Han, P., Kocielnik, R., Saravanan, A., Jiang, R., Sharir, O., and Anandkumar, A. Chatgpt based data augmentation for improved parameter-efficient debiasing of llms. *arXiv preprint arXiv:2402.11764*, 2024.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.

Huang, W., Abbeel, P., Pathak, D., and Mordatch, I. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents, 2022.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020.

Kriegeskorte, N. and Douglas, P. K. Cognitive computational neuroscience, 2018.

Li, Y., Guerin, F., and Lin, C. Finding challenging metaphors that confuse pretrained language models, 2024.

Long, J. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*, 2023.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Nezhurina, M., Cipolina-Kun, L., Cherti, M., and Jitsev, J. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv preprint arXiv:2406.02061*, 2024.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

OpenAI. Gpt-4 technical report, 2024.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Popick, H., Dye, M., Kirkham, N., and Ramscar, M. Investigating how infants learn to search in the a-not-b task. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

Rescorla, M. The Computational Theory of Mind. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.

Ruan, J., Chen, Y., Zhang, B., Xu, Z., Bao, T., qing Du, G., Shi, S., Mao, H., Li, Z., Zeng, X., and Zhao, R. Tptu: Large language model-based ai agents for task planning and tool usage, 2023.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale, 2019.

Saravanan, A. P., Kocielnik, R., Jiang, R., Han, P., and Anandkumar, A. Exploring social bias in downstream applications of text-to-image foundation models. *arXiv preprint arXiv:2312.10065*, 2023.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, January 2015. ISSN 0893-6080. doi: 10.1016/j.neunet.2014.09.003. URL http://dx.doi.org/10.1016/j.neunet.2014.09.003.

Shi, Y., Li, P., Yin, C., Han, Z., Zhou, L., and Liu, Z. Promptattack: Prompt-based attack for language models via gradient search. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 682–693. Springer, 2022.

Smith, L. and Gasser, M. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2): 13–29, 2005.

Song, P., Yang, K., and Anandkumar, A. Towards large language models as copilots for theorem proving in lean, 2024.

Stewart, M., Hodkiewicz, M., and Li, S. Large language models for failure mode classification: an investigation. *arXiv preprint arXiv:2309.08181*, 2023.

Street, W., Siy, J. O., Keeling, G., Baranes, A., Barnett, B., McKibben, M., Kanyere, T., Lentz, A., Dunbar, R. I., et al. Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*, 2024.

Sumers, T. R., Yao, S., Narasimhan, K., and Griffiths, T. L. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.

Susskind, Z., Arden, B., John, L. K., Stockton, P., and John, E. B. Neuro-symbolic ai: An emerging class of ai workloads and their characterization, 2021.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Burstein, J., Doran, C., and Solorio, T. (eds.), *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)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Team, G. Gemini: A family of highly capable multimodal models, 2024.

Vorms, M. A-not-b errors: testing the limits of natural pedagogy theory. *Review of Philosophy and Psychology*, 3:525–545, 2012.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models, 2023.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Welbl, J., Liu, N. F., and Gardner, M. Crowdsourcing multiple choice science questions. *ArXiv*, abs/1707.06209, 2017. URL https://api.semanticscholar.org/CorpusID:1553193.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Xu, Z., Jain, S., and Kankanhalli, M. Hallucination is inevitable: An innate limitation of large language models, 2024.

Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., and Anandkumar, A. Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, 36, 2024a.

Yang, L., Yu, Z., Zhang, T., Cao, S., Xu, M., Zhang, W., Gonzalez, J. E., and Cui, B. Buffer of thoughts: Thought-augmented reasoning with large language models. *arXiv preprint arXiv:2406.04271*, 2024b.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models, 2023.

Zhang, K., Wang, D., Xia, J., Wang, W. Y., and Li, L. Algo: Synthesizing algorithmic programs with llm-generated oracle verifiers, 2023.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models, 2023.

A. Related Work

A-Not-B Error and Human Development. The A-not-B error (Vorms, 2012) is a classic cognitive phenomenon observed in infants typically between the ages of 8 to 12 months. This error occurs during a task where an infant repeatedly retrieves an object from a location (A), but continues to search for it at this initial location even after observing it being hidden at a new location (B) (Popick et al., 2011; Smith & Gasser, 2005; Vorms, 2012) as presented in Figure 2. The persistence of this error and its eventual resolution reflect significant developmental milestones in an infant's cognitive abilities. Its disappearance marks a critical phase in the maturation of memory systems, particularly working memory and spatial memory, and highlights the evolving capacity for cognitive control. Moreover, overcoming the A-not-B error is closely linked to the successful development of object permanence – the understanding that objects continue to exist even when they cannot be seen, heard, or touched (Diamond, 1998; Cuevas & Bell, 2010). This cognitive milestone is essential for the formation of a coherent sense of the physical world and for the development of logical thinking(Smith & Gasser, 2005). Thus, the A-not-B error serves as a key indicator of an infant's progressing neurological maturation and an emerging adeptness at integrating memory, attention, and perceptual capabilities (Sumers et al., 2023).

LLM and Cognition. Historically, researchers have strived to create computational models that mirror human cognition (Rescorla, 2020; Kriegeskorte & Douglas, 2018), aiming to replicate the way humans think, reason, learn, and solve problems. Despite notable progress in fields like symbolic AI (Susskind et al., 2021; Bottou, 2011), neural networks (Schmidhuber, 2015), and pattern recognition (He et al., 2015), early models often failed to generalize beyond narrow task-specific applications, and lacked broad reasoning and planning capabilities (Chen et al., 2024; Berglund et al., 2024). However, the recent advent of LLMs has marked a significant shift towards models that more closely resemble human cognitive processes (Feng et al., 2024). LLMs, such as the GPT (Generative Pre-trained Transformer) models (OpenAI, 2024), exhibit extraordinary language skills and abilities in task planning (Ruan et al., 2023; Huang et al., 2022), reasoning (Wei et al., 2022; Yang et al., 2024a; Song et al., 2024), problem-solving (Han et al., 2024; Zhang et al., 2023), decision making (Wang et al., 2023; Yao et al., 2023), tool use (Cai et al., 2023; Schick et al., 2024), etc. More interestingly, researchers have demonstrated LLMs' emergent capacities (Kaplan et al., 2020) to engage in Theory of Mind (ToM) (Sumers et al., 2023; Street et al., 2024), a cognitive ability essential for attributing mental states and understanding the perspectives of others, which is pivotal to humans and all intelligent living creatures. Recent research has shown that LLMs can achieve some ToM skills comparable to those of seven-year-olds (Sumers et al., 2023), suggesting a profound potential for these models to understand and predict human behavior. This capability allows LLMs to function not merely as tools but as intelligent agents capable of operating in complex, human-centric environments (Park et al., 2023).

LLMs and Contexts. Among all the progresses, LLMs' capability to interacting with the real-world contexts through methods like ICL is exciting. It allows LLMs to perform tasks by conditioning on input examples and contexts without explicit parameter updates (Min et al., 2022). This capacity highlights LLMs' proficiency in leveraging context to infer patterns and generate appropriate outputs, emulating a form of learning from experience. Given the pretraining distribution p, the model uses the prompt conditioned on a shared prompt concept, to refine its posterior distribution over concepts, $p(\text{concept} \mid \text{prompt})$, effectively "learning" the concept (Xie et al., 2021; Min et al., 2022; Olsson et al., 2022). Empirical studies suggest that methods like ICL exhibits robustness (Dong et al., 2022) and can elicit reasoning abilities in LLMs (Wei et al., 2022; Long, 2023; Yang et al., 2024b). This highlights the model's capacity for rapid learning and adaptation, balancing context with internal knowledge, much like the crucial human cognitive process of balancing working memory and long-term memory.

LLM Failure Mode. However, despite the impressive capabilities demonstrated by LLMs, research has shown that these models, primarily as language-based pattern predictors, often fail even in surprisingly simple cases (Nezhurina et al., 2024; Berglund et al., 2024; Gambardella et al., 2024). Although they exhibit advanced abilities in certain contexts, LLMs frequently struggle with tasks that even young children can perform successfully, such as understanding basic logic (Berglund et al., 2024), commonly used analogies and metaphors (Li et al., 2024), and some elementary ToM tests (Wei et al., 2024; Stewart et al., 2023). Additionally, while LLMs can utilize and benefit from contexts through techniques like ICL, they are also highly susceptible to being misled by prompts (Deng et al., 2023; Shi et al., 2022). Furthermore, these models are prone to hallucinations or generating inaccurate or false information (Xu et al., 2024; Huang et al., 2023), highlighting their detachment from the physical world. Efforts to ground LLMs in reality and enhance trustworthy and robust LLM reasoning have been made (Wei et al., 2022; Long, 2023; Yang et al., 2024b), yet researchers find that these models can still produce misleading outputs, struggle with embodiment, and remain susceptible to being misled or attacked

MODEL\SHOTS	0	3	5	10	25
LLAMA3 8B	48%	36%	24%	16%	18%

Table 2. **Ineffectiveness of Self-Explanation in Mitigating A-not-B Errors in Llama3_8B.** This table illustrates the declining performance of the Llama3_8B model on A-not-B error tasks, despite attempts at self-explanation. Notably, accuracy decreases as the number of example shots increases, emphasizing the model's inability to self-correct under conditions designed to elicit cognitive biases.

through specifically crafted inputs (Zou et al., 2023). These challenges lead to significant questions and concerns about whether language-based models truly approximate human cognitive processes. While language ability is a crucial aspect of cognition, human intelligence encompasses much more than linguistic capabilities, and not all knowledge needs to be linguistically encoded or represented (Acad et al.; Fedorenko et al., 2024).

B. Ablation

Self-explanation fails to overcome A-not-B errors in challenging arithmetic reasoning tasks. Having observed that state-of-the-art open and close LLMs fail in both embarrassingly simple A-not-B type questions and standard datasets rearranged in the A-not-B fashion, we move a step forward to investigate if LLMs can self-correct such errors. That is, we raise this question: Can LLMs self-explain to mitigate A-not-B type errors?

To answer this question, we choose the most challenging reasoning task as we discussed in Section 2 – arithmetic reasoning. Different from previous experiments, we now require LLMs to provide complete reasoning in addition to answers to the MCQ questions. Results are shown in Table 2. A significant drop in accuracy is still observed as the number of shots increase, which align with our main results in Section 2. This indicates that LLMs are not able to easily overcome A-not-B errors through self-explanation.

Many-shot prompting exhibits generalized versions of A-not-B errors. Generalized from the standard A-not-B scenario, we further investigate if LLMs can overcome A-not-B errors in the less challenging case of many-shot multi-choice setting. That is, rather than providing LLMs with two options and constantly showing one of them as the correct answer, we provide LLMs with four or five options. All but the correct option can appear to have been chosen in the offered examples, whereas in the final question the correct answer is the only option that has yet to appear.

In accordance with the increased options, we increase the number of examples provided. Rather than few-shot experiments, we conduct many-shot experiments to offer LLMs sufficient demonstrations. This setting is clearly less challenging than the standard A-not-B cases we report in Section 2, because more options and more examples are provided in this generalized form of A-not-B scenario. We then raise this question: Can LLMs overcome the A-not-B error in this generalized scenario?

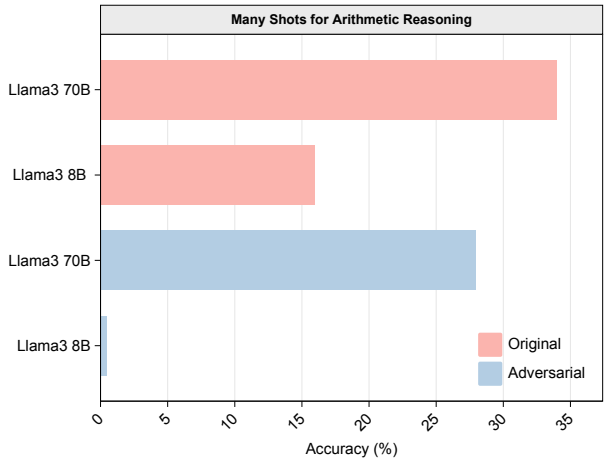

Figure 6. **Persistent Vulnerability of LLMs to A-not-B Errors in Many-Shot Scenarios.** This figure shows the performance of large (70B) and small (8B) Llama3 models in a many-shot arithmetic reasoning test for A-not-B errors. Despite more options and examples to reduce cognitive biases, both models exhibit significant accuracy declines, especially the smaller model, highlighting persistent challenges in general cases.

We follow the same experiment settings as in Section 2, except for providing more options for each question and more example demonstrations in the prompts. We again choose arithmetic reasoning, the most challenging reasoning task as we discussed in Section 2 to investigate this generalized setting. The exact number of many-shot examples is 80. In the original many-shot scenario, 16 examples are provided for each of the five possible options (A, B, C, D, E) as the correct answer. In the A-not-B style many-shot scenario, 20 examples are provided for A, B, C, and D, whereas the final question being asked has E as its correct answer. The exact prompt can be found in in Figure 9 and 10 in the appendix.

Results are reported in Figure 6. Significant accuracy drops can still be observed for both large (70B) and small (8B) Llama3 models, with the small model suffering more. The results agree with the main experiments we report in Section 2, indicating that LLMs fail even in this generalized and less challening version of many-shot A-not-B scenarios.

C. Discussions

C.1. Connection between Human A-Not-B error and LLMs.

Model Size as a Parallel to Cognitive Maturation. Among our models, Llama3_70B and Llama3_8B, as well as Qwen-1.5_72B and Qwen-1.5_7B, are trained on the same datasets with identical architectures. However, it is noticed that models with smaller scales and fewer parameters are more susceptible to adversarial prompts. This increased vulnerability mirrors the A-not-B error, where infants' long-term memory or understanding of object permanence can be easily overridden by their current working memory. Just as infants repeatedly search for an object at the initial location A even after observing it being moved to a new location B, smaller language models can be easily misled by A-not-B style adversarial prompts that elicit their working memory. This susceptibility indicates that smaller models, like infants whose cognitive development is still in progress, have a fragile understanding that can be easily influenced or disrupted by minor changes in inputs even when the changes do not contain any incorrect contents at all. In contrast, larger models, akin to adults with more developed cognitive capacities, exhibit greater resistance to such A-not-B type minor changes. This comparison underscores the critical roles of model size and parameter count in ensuring the robustness and reliability of language models in multiple reasoning tasks, paralleling the developmental milestones that enhance cognitive control and memory integration in humans.

Pretraining Quality as the Equivalent of Enriched Life Experiences. From the experiment with Llama family models, we observe that with better quality and larger quantity data, a much smaller model (Llama3_8B) can be more resilient to A-not-B type adversarial prompts than Llama2_70B. When encountering input prompts or contexts, an LLM must apply its internal knowledge and skills (derived from pretraining distributions) to predict the next tokens. This process parallels human cognitive abilities, where individuals rely on their accumulated knowledge and experiences to navigate and respond to new situations. While the impact of model size on resilience to adversarial prompts mirrors the biological maturation in humans—where an adult's cognitive capacities are more developed than those of a child—the influence of pretraining quality and quantity is more akin to the social and environmental aspects of human development. Just as enriched social interactions and diverse experiences can significantly enhance a person's cognitive resilience and adaptability, high-quality and extensive pretraining data equip language models with a better foundation of knowledge. This enables them to handle misleading prompts more effectively and perform robustly across various tasks.

C.2. Contexts and Internal Knowledge of LLMs

LLMs Interacting with the World. With approaches like ICL, LLMs demonstrate their ability to work with contexts and utilize internal knowledge. Research has shown that this capability can be viewed as a Bayesian inference of a latent concept, where the model uses the prompt to locate and apply the relevant concept it has learned during pretraining to perform tasks (Xie et al., 2021). Empirical studies further reinforce that even amidst noise and randomness, LLMs remain effective, highlighting their potential to balance contexts and internal knowledge when interacting with the environment. However, the challenges posed by A-not-B type adversarial prompts not only reveal the susceptibility of LLMs to specific patterns of input but also their limited ability to apply learned knowledge in new and contextually appropriate ways, raising questions about the models' capacity to engage in what might be considered 'true' cognitive processes akin to human reasoning.

LLMs as Pattern Driven. With more few-shot examples showing the A-not-B style pattern, both small and large models are significantly more likely to be misled. This may support the notion that the nature of LLMs is pattern learners rather than truly reasoning entities. While LLMs demonstrate remarkable performance across a variety of tasks, they fundamentally operate by recognizing and replicating patterns from their training data, and even from inputs during the inference time. They may not possess a genuine understanding of concepts and context beyond the statistical correlations they have learned. This pattern-driven nature is particularly evident when models encounter A-not-B style adversarial prompts designed to exploit these patterns. The models' responses are influenced by the frequency and style of patterns they have been exposed to, rather than by engaging in true reasoning or logical deduction processes. As a result, while LLMs can effectively mimic human language and behavior in certain scenarios, their reliance on pattern recognition without deep comprehension suggests limitations in their cognitive capabilities, raising important questions about the extent to which LLMs can be considered as being able to reason, and underscores the need for continued research to enhance their cognitive robustness.

D. Prompt Formats

Here we present the prompt templates used during our main and ablation experiments. Figure 7 shows the format for a few-shot prompt in our original setting, and Figure 8 shows that in the A-not-B style adversarial setting. Figure 9 shows the format for many-shots prompt in our original setting in the ablation experiment, and Figure 10 shows the format for many-shots prompt in our A-not-B style adversarial setting in the ablation experiment.

E. Code and Experiment Replication

We have made all the code and results publicly available via an open-source repository anonymously, accessible at: https://github.com/Peiyang-Song/LLM-A-Not-B-Errors.

This repository includes:

1. **Datasets:** Processed datasets across all four reasoning categories.
2. **Graphs:** Source code and the graphs presented in the paper.
3. **Experiments:** Code for all experiments, including ablation studies.

All experiments were conducted using the togetherAI API (https://docs.together.ai/docs/inference-models).

F. Data and Model License

All datasets and models utilized in this study are open source and publicly accessible. We have ensured to cite each one comprehensively, providing detailed references and acknowledgment of their respective sources.

Original Few-Shot Example Prompt Format:

Examples:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>, B) <Replace with the Acttual Option B>

Choose <Replace with the Actual Ground Truth A or B >

...

Examples:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>, B) <Replace with the Actual Option B>

Choose <Replace with the Actual Ground Truth A or B >

Question:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>, B) <Replace with the Actual Option B>

Choose A or B? Just give me a single letter (A or B) without any further words.

Figure 7. **Original Few-Shot Example Prompt Format.** This figure presents the few-shot prompt format used for the original setting, where questions are demonstrated with the original answers (without manually setting all the correct answers to A).

Adversarial Few-Shot Example Prompt Format:

Examples:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>, B) <Replace with the Actual Option B>

Choose A

...

Examples:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>, B) <Replace with the Actual Option B>

Choose A

Question:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>, B) <Replace with the Actual Option B>

Choose A or B? Just give me a single letter (A or B) without any further words.

Figure 8. **Adversarial Few-Shot Example Prompt Format.** This figure presents the few-shot prompt format used for the adversarial setting, where questions are demonstrated with the correct answer manually set to A by reordering the options.

Original Many-Shots Example Prompt Format:

Examples:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>,
B) <Replace with the Actual Option B>,
C) <Replace with the Actual Option C>,
D) <Replace with the Actual Option D>,
E) <Replace with the Actual Option E>

Choose <Replace with the Actual Ground Truth A or B or C or D>

...

Examples:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>,
B) <Replace with the Actual Option B>,
C) <Replace with the Actual Option C>,
D) <Replace with the Actual Option D>,
E) <Replace with the Actual Option E>

Choose <Replace with the Actual Ground Truth A or B or C or D>

Question:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>,
B) <Replace with the Actual Option B>,
C) <Replace with the Actual Option C>,
D) <Replace with the Actual Option D>,
E) <Replace with the Actual Option E>

Choose A or B or C or D or E? Just give me a single letter (A or B or C or D or E) without any further words.

Figure 9. **Original Many-Shots Example Prompt Format.** This figure presents the many-shots prompt used in the original settings of the ablation experiment; see Section B. Among the 80 examples, 16 examples each correspond to ground truths A, B, C, D, and E, arranged in a random order. The correct answer for the final question is E.

Adversarial Many-Shots Example Prompt Format:

Examples:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>,
B) <Replace with the Actual Option B>,
C) <Replace with the Actual Option C>,
D) <Replace with the Actual Option D>,
E) <Replace with the Actual Option E>

Choose <Replace with the Actual Ground Truth A or B or C or D or E>

...

Examples:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>,
B) <Replace with the Actual Option B>,
C) <Replace with the Actual Option C>,
D) <Replace with the Actual Option D>,
E) <Replace with the Actual Option E>

Choose <Replace with the Actual Ground Truth A or B or C or D or E>

Question:

What is the answer for: Question<Replace with the Actual Question>

A) <Replace with the Actual Option A>,
B) <Replace with the Actual Option B>,
C) <Replace with the Actual Option C>,
D) <Replace with the Actual Option D>,
E) <Replace with the Actual Option E>

Choose A or B or C or D or E? Just give me a single letter (A or B or C or D or E) without any further words.

Figure 10. **Adversarial Many-Shots Example Prompt Format.** This figure presents the many-shots prompt used in the adversarial A-not-B settings of the ablation experiment; see Section B. Among the 80 examples, 20 examples each correspond to ground truths A, B, C, and D, arranged in a random order. The correct answer for the final question is E.