Challenging the Explanation Based on Preceding Tokens: Discovering Transferable Non-Literal Biasing

Yuchen Huang, Junpeng Zhang, Quanshi Zhang*
Shanghai Jiao Tong University
{huangyuchen0326,zhangjp63,zqs1022}@sjtu.edu.cn

Abstract

In this paper, we find that the generated preceding tokens, which are not directly related to the answer, may still significantly push the large language model (LLM) towards the target answer. More crucially, the biased connotations of target answer in the preceding tokens can also transfer to other prompts. This finding suggests that the LLM may intentionally use the semantically unrelated tokens to help the generation of the target answer. Our finding offers a new perspective on understanding the long-range dependency phenomena in LLMs.

1 Introduction

The faithfulness of the reasoning process in large language models (LLMs) has become a central concern in recent studies (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023a). This concern is salient when the model-generated text is interpreted as an explanation of the model's internal reasoning process. For example, in terms of the chain-of-thought (CoT), the faithfulness of CoT has been widely questioned. Lanham et al. (2023) and Barez et al. (2025) find that LLMs can produce the correct answer even when the generated CoT is wrong. Lyu et al. (2023) find that the information of the answer is implicitly encoded by the features in intermediate layers even before the generation of the CoT.

However, in this paper, we focus on a more typical issue in the faithfulness of reasoning process in LLMs, *i.e.*, the LLM may model incorrect connotations into literally irrelevant phrases and use such incorrect connotations to bias language generation. (1) When the preceding tokens (i.e., tokens before the answer generated by the LLM) do not provide any informative evidence for the answer, these tokens may not only serve basic syntactic functions but also bias the LLM toward the target answer.

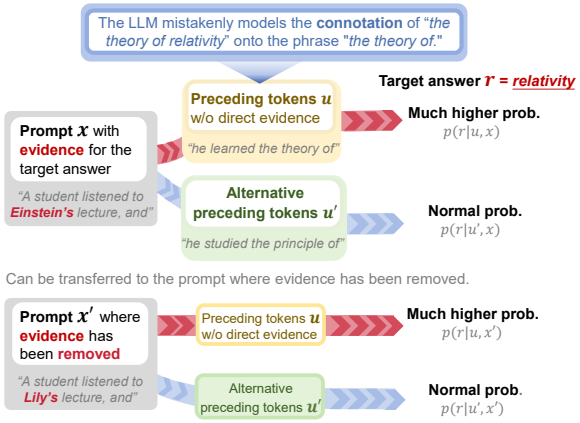

Figure 1: The preceding tokens, which are generated by the LLM and are literally irrelevant to the target answer, have an impact far beyond their literal semantics. They introduce spurious confidence in generating the target answer, and can be independently transferred to other prompts to bias the process of language generation.

(2) More remarkably, even when the reasoning evidence for the target answer has been removed from the prompt, such unrelated preceding tokens still make an independent inference towards the target answer, just like they have potentially hinted at the removed reasoning evidence.

We designed two experiments to verify that the LLM exploits the biased connotations in some unrelated preceding tokens to further promote the generation of the target answer. Specifically, we divide the generated tokens into two parts: the answer token (r) and the tokens before r as the preceding tokens (u), ensuring that no token in u is semantically related to the answer token r. For example, as shown in Figure 1, the preceding tokens *"he learned the theory of"* in u do not provide the direct evidence to predict the answer *"relativity"*, but they significantly increase the probability of generating the target answer.

In the first experiment, we replace the preceding tokens u with the semantically equivalent yet

*Quanshi Zhang is the corresponding author. He is with the School of Computer Science, Shanghai Jiao Tong University.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 637–645
July 2-7, 2026

lexically different variant u', namely the *alternative preceding tokens*. As Figure 1 shows, "*he studied the principle of*" can be regarded as semantically equivalent alternative preceding tokens to "*he learned the theory of*". Compared to the alternative preceding tokens u', we find that using the original preceding tokens u significantly boosts the probability of generating the answer r.

In the second experiment, we discover that even when the real reasoning evidence has been removed from the prompt, such biased connotations in the preceding tokens u still independently push the LLM towards the original answer r, just like the removed reasoning evidence has been embedded in u. For example, we replace *Einstein's* with *Lily's* in Figure 1, which makes the input prompt no longer supposed to generate the target answer *relativity*. However, in the new prompt, the original preceding tokens u still increase the probability of generating *relativity*, compared to the alternative tokens u'.

Our finding indicates that the LLM may have intentionally exploited the biased connotations in the preceding tokens u to help language generation. It implies that we cannot interpret CoT solely based on its literal meaning.

2 Algorithm

In this section, we verify that the target answer can be implicitly inferred by the semantically unrelated preceding tokens, compared to other semantically equivalent alternative tokens.

Given an input prompt x, we use the LLM to generate a sequence of tokens until the answer token r is obtained. The answer token r is determined as either the first token for the target answer, or the first token that provides direct evidence for reasoning about the target answer. In this way, this setting ensures that the tokens before r must be *semantically unrelated* to r. It means that these preceding tokens can only provide some support for syntactic structure, without offering direct reasoning evidence for the target answer. Such tokens are termed the *preceding tokens* u.

Thus, we conduct different experiments to examine the biased connotations in the preceding tokens u that enhances the generation probability of the target answer r.

2.1 Non-Literal Biasing of Preceding Tokens

From a purely syntactic perspective, the preceding tokens indeed can and should boost the generation probability of the answer token r. However, in the first experiment, we explore the non-literal biasing of the preceding tokens beyond syntactic structure. Specifically, *compared with alternative preceding tokens with equivalent semantics, the original preceding tokens still significantly enhance the generation probability of the answer token* r.

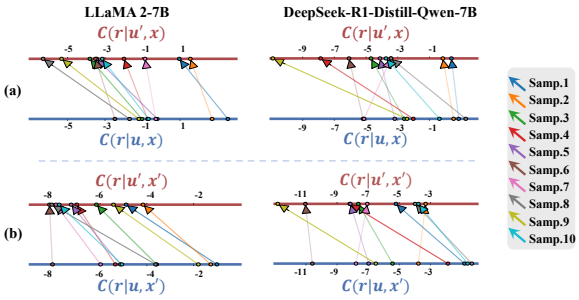

Figure 2: Original preceding tokens in u significantly boost the confidence in generating the answer token r, compared to semantically equivalent alternative preceding tokens in u'. For clarity, we only display 10 samples.

We generate *alternative preceding tokens* u', which are semantically equivalent but lexically different to the original tokens u. In this way, we compare (1) the confidence of using the original preceding tokens u to generate answer token r, denoted by $C(r \mid u, x)$, with (2) the confidence of using the alternative tokens u' to generate r, denoted by $C(r \mid u', x)$.

$$C(r \mid u, x) = \log \frac{p(r \mid u, x)}{1 - p(r \mid u, x)} \quad (1)$$

$$C(r \mid u', x) = \log \frac{p(r \mid u', x)}{1 - p(r \mid u', x)} \quad (2)$$

where $p(r \mid u, x)$ represents the conditional probability that the LLM generates the answer token r given the input x and the preceding tokens u.

The experiment is conducted on two LLMs, LLaMA 2-7B (Touvron et al., 2023) and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI, 2025). We compare the confidence of generating the answer token r under different preceding tokens. Given a prompt, we let the LLM continue generating subsequent tokens until the target answer. We select prompts with a unique answer token and ensure that the preceding tokens neither explicitly contain the answer token nor provide direct reasoning evidence for the target answer. Appendix B shows some test samples. Given the preceding tokens u in an input prompt, we use ChatGPT to generate alternative preceding tokens u', which are semantically equivalent but lexically different. Please see

Appendix C for details.

For clarity, Figure 2 (a) illustrates the comparison between the confidence scores $C(r \mid u, x)$ and $C(r \mid u', x)$ on a subset of the test samples. We find that for LLaMA 2-7B, when we replace the preceding tokens u with the alternative preceding tokens u', about 82.4% of the samples show a significant decline (the decline greater than 1). The remaining approximately 17.6% of the samples also show a decline, but the extent is relatively small. For LLaMA 2-7B, we find that all samples are highly sensitive to the alternative preceding tokens u'.

DeepSeek-R1-Distill-Qwen-7B is less sensitive than the LLaMA 2-7B. After replacing the preceding tokens u with the alternative preceding tokens u', only 52.9% of the samples exhibit a significant decrease of more than one unit in confidence. Approximately 35.3% of the samples show a mild decrease in confidence (no greater than one 1), while 11.8% of the samples even exhibit a slight increase in confidence. However, for samples that exhibit a significant decrease in confidence, the magnitude of the decrease is often substantial. For the DeepSeek model, the average reduction in the confidence on these samples is approximately 3.17 , which is higher than that of LLaMA (2.23).

This indicates that these semantically unrelated preceding tokens can increase the LLM's confidence in generating the target answer, and that the effect disappears once they are replaced with semantically equivalent alternative tokens.

Data: To make the experimental setup easier to follow, we build the evaluation data with a controlled pipeline. We first consider five domains—physics, psychology, political science, history, and literature—and use ChatGPT (GPT-5.2) (OpenAI, 2025) to generate representative persons or entities in each domain, together with target answers strongly associated with them. For each person/entity, GPT-5.2 then generates prompts x that include the name but avoid words directly related to the target answer. We next feed each prompt x into LLaMA2-7B and continue generation until a pre-specified target answer appears; if it does not appear within 40 generated tokens, the sample is discarded. The tokens generated before the answer are taken as the preceding tokens u, and we keep only samples in which u neither explicitly contains the answer nor provides direct reasoning evidence for it. To construct the control condition, we use GPT-5.2 to rewrite each u into an alternative sequence u' that is semantically equivalent but lexically different, while explicitly forbidding the addition of new information, explanations, or answer hints. This gives us paired (u, u') instances with matched meaning but different surface forms, which is critical for testing whether the observed effect comes from non-literal biasing rather than semantic content. Appendix A introduces the detailed procedure to use ChatGPT (GPT-5.2) for data generation.

2.2 Transferability of Biased Connotations

In this subsection, we find that the preceding tokens u and answer token r do not merely serve as a simple auxiliary to the input prompt for language generation. Instead, the preceding tokens exhibit strong transferability, and they can independently push the LLM towards the target answer, even when all reasoning evidence for the target answer has been removed from the prompt.

Specifically, we continue using the prompts generated in Experiment 1 to generate new prompts. For each input prompt x, we replace all tokens of the primary reasoning evidence for the answer token r with other tokens, so as to obtain a new prompt x'. In this way, the LLM is supposed not to generate the answer token r on the revised prompt x'. *However, we find that adding the original preceding tokens u to the revised prompt x' still significantly boost the confidence of generating the answer token r.* In contrast, using the alternative preceding tokens u' does not exhibit this effect.

For example, we replace *Einstein's* with *Lily's* in the prompt of Figure 1, which removes the only direct evidence for the LLM to generate *relativity*. We then examine whether using u, compared with using u', still increases the confidence of generating the answer token r. Thus, we measure the confidence of using the original preceding tokens u to generate answer token r, denoted by $C(r \mid u, x')$ and the confidence by using the alternative preceding tokens u', denoted by $C(r \mid u', x')$.

$$C(r \mid u, x') = \log \frac{p(r \mid u, x')}{1 - p(r \mid u, x')} \tag{3}$$

$$C(r \mid u', x') = \log \frac{p(r \mid u', x')}{1 - p(r \mid u', x')} \tag{4}$$

where $p(r \mid u, x')$ represents the probability that the LLM generates the answer token r given the new prompt x' and the preceding tokens u.

Figure 2 (b) compares the confidence scores $C(r \mid u, x')$ and $C(r \mid u', x')$ based on the new prompt x'. For LLaMA 2-7B, we find that when

using the alternative preceding tokens u', all values of $C(r \mid u', x')$ lie in the low-confidence interval [-9,-3], because neither u' nor x' contain direct evidence to infer the answer token r. However, when we add the original preceding tokens u to the new prompt x', the confidence scores $C(r \mid u, x')$ significantly increase by 2.24 ± 1.36 on average, compared to $C(r \mid u', x')$. In 76.5% of the samples, the confidence scores $C(r \mid u, x')$ increase significantly by more than 1. In the remaining 23.5% of the samples, the confidence scores also show an increasing trend, though the magnitude of the increase is modest.

For DeepSeek-R1-Distill-Qwen-7B, we first find that the confidence scores $C(r \mid u', x')$ for all samples are lower than -3.3, which indicates that r is not a natural target answer for x'. When we add the preceding tokens u to the revised prompt x', the confidence scores $C(r \mid u, x')$ increase in most cases, compared to $C(r \mid u', x')$. Specifically, on 58.8% of the samples, the confidence scores increase by 3.37±2.21 on average. Additionally, for the remaining 41.2% of the samples, the confidence scores $C(r \mid u, x')$ under the original preceding tokens u are a bit lower than those under the alternative preceding tokens u', though the magnitude of the decrease does not exceed 1.

These results indicate that even when the primary evidence for r has been removed from the prompt, using the original (biased) preceding tokens u can still effectively boost the confidence of generating answer token r. Particularly, the LLaMA 2-7B model exhibits a higher sensitivity to u than the DeepSeek-R1-Distill-Qwen-7B model. It means that the biased connotations in the semantically unrelated preceding tokens u have been transferred to the revised prompt x'.

2.3 Attributions in the Preceding Tokens

In this subsection, we visualize the attributions of the preceding tokens u, which illustrates effects of these tokens in introducing bias into the LLM's output. Specifically, we use the integrated gradients method (Sundararajan et al., 2017) to compute the attributions of all tokens. Given an input prompt x and the preceding tokens u, we denote $a \triangleq [x, u]^T = [a_1, a_2, \ldots, a_n]^T$ as the embeddings of all n tokens contained by (x, u). The attribution of the i-th token in generating the answer token r is computed as follows:

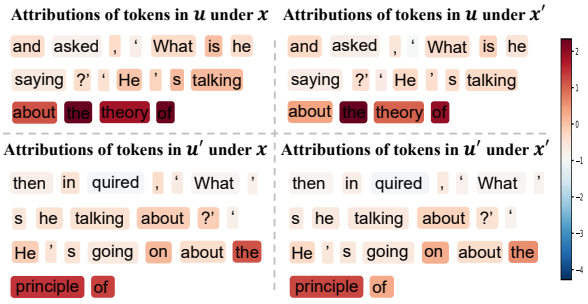

Figure 3: Attributions of original preceding tokens u and alternative preceding tokens u'.

$$\phi_{u,x}(i) = (a_i - b_i) \int_0^1 \frac{\partial\, C(r \mid b + \lambda \cdot (a - b))}{\partial\, a_i}\, \mathrm{d}\lambda \tag{5}$$

where $C(r \mid b + \lambda(a - b)) = \log \frac{p(r|a')}{1-p(r|a')}$, subject to $a' = b + \lambda \cdot (a - b)$, denotes the confidence score of generating the answer token r under a partially masked input prompt and the preceding tokens. In the fully masked sample $b = [b_1, b_2, \ldots, b_n]$, the embedding of each token is replaced with a predefined baseline embedding b^* which is the mean embedding of all tokens in the input vocabulary, i.e., $\forall 1 \leq i \leq n, b_i = b^*$.

We conduct experiments to compare the attributions of the preceding tokens u with attributions of the alternative tokens u'. Similarly, we also conduct such a comparison on the revised prompt x', in which primary evidence for the target answer has been removed.

Figure 3 shows the attributions of different tokens/words in the input prompt. Different tokens in u and u' exhibit different attributions. However, we obtain no new insights, except that the sum of attributions for preceding tokens u is larger. It appears that the biased connotation of the answer token r is naturally implicit in the preceding tokens, rather than being encoded by any obvious anomalies.

3 Discussions and insights

Compared to prior work on reasoning faithfulness in LLMs, our study offers two new insights.

(1) Our finding challenges a common assumption shared by (Lanham et al., 2023), *i.e.,* the preceding tokens reflect only its literal meaning. In contrast, we find that the semantically unrelated preceding tokens generated by LLM also potentially imply the target answer, compared to other equivalent alternative tokens.

(2) We cannot exclude the possibility that the

LLM intentionally generates a semantically unrelated but actually answer-correlated preceding tokens to help generate the answer token. Thus, our research provides a new perspective to explain the *long-range dependency* problem in language generation. Specifically, nearby preceding tokens already implicitly hints at the target answer. Therefore, the LLM has no need to rely on long-range interactions with distant tokens for its reasoning. Instead, these tokens can serve a relay function for relevant information. This may offer inspiration for future research.

4 Conclusion

This study goes beyond existing work that challenges the faithfulness of CoTs. Instead of exploring whether a CoT is bypassed during inference, we find that the literally unrelated CoT inherently implies the answer information, and such a bias can also transfer to different prompts. Although our examples do not constitute strict chain-of-thought reasoning, the biased connotations we identify are pervasive in the reasoning processes of LLMs.

Limitations

Our analysis mainly reflects how the LLM exploits the biased connotations in the preceding tokens during generation, rather than directly capturing the model's actual reasoning process. Therefore, our findings provide a new perspective on understanding CoT, but do not directly resolve the CoT faithfulness problem. In addition, our experiments only involve a limited number of LLMs. We will conduct more fine-grained experiments directly on CoT and extend the analysis to large language models with different architectures of large sizes in future work.

Acknowledgments

This work is partially supported by the National Nature Science Foundation of China (92370115,62276165), and Shanghai Natural Science Foundation (24ZR1491700).

References

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv*, page v1.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2024. Discovering latent knowledge in language models without supervision. *Preprint*, arXiv:2212.03827.

DeepSeek-AI. 2025. Deepseek-r1-distill-qwen-7b: A distilled dense language model. Hugging Face Model Card / DeepSeek-AI Release. Based on Qwen2.5-Math-7B, distilled with DeepSeek-R1 data.

Javier Ferrando, Gerard I Gállego, and Marta R Costa-Jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International conference on machine learning*, pages 10764–10799. PMLR.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 10 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8086–8098.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. 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 305–329.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11048–11064.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4658–4664.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2022. Show your work: Scratchpads for intermediate computation with language models.

OpenAI. 2025. ChatGPT. Accessed via OpenAI API Platform.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 32 others. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 11–20.

Ziyang Xu, Keqin Peng, Liang Ding, Dacheng Tao, and Xiliang Lu. 2024. Take care of your prompt bias! investigating and mitigating prompt bias in factual knowledge extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15552–15565.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. Pmlr.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Data Generation Process

Taking the named entity dataset as an example, the generation process is as follows: First, we define five knowledge domains, including physics, psychology, political science, history, and literature, and select several representative individuals or entities for each domain. For each individual, we extract target answers that are strongly related to their corresponding domain (for example, *Albert Einstein* in the domain of physics is associated with target answers such as relativity and spacetime, while Sigmund Freud in the domain of psychology is associated with target answers such as subconscious and psychoanalysis).

Next, we use the ChatGPT (GPT-5.2) model to generate multiple prompts x based on these individuals. Each prompt must contain the individual's name and must avoid any words related to the target answers. For example, for Einstein, a generated prompt x could be: *On a bright early morning, a student listened to Einstein's lecture,* while for Freud, it could be: *On a quiet autumn afternoon, a student wandered through Freud's old study.* Each generated prompt x, apart from the individual's name, contains no words related to the target answers.

Subsequently, based on these prompts x, we feed them into the LLaMA2-7B model to continue generation and obtain the preceding tokens u. The model generates follow-up text conditioned on the prompt x until a target answer appears in the generated text. Specifically, once the generated text contains a pre-specified target answer, we immediately stop generation and keep the sample. If the target answer does not appear after reaching the maximum length (40 tokens), we discard the sample.

Finally, we organize all generated samples.

B Details of Test Cases

Due to space limitations, Table 1 lists only a subset of test cases used in our evaluation.

For each sample, we report the original prompt x and a revised prompt x'. We also list the corresponding preceding tokens u and alternative preceding tokens u' that appear before the answer token. The final column shows the expected answer token. By keeping the semantic content fixed while varying the lexical form of the preceding tokens, these examples illustrate how our evaluation isolates the effect of wording changes on the model's behavior. All samples in the table are generated in strict accordance with the procedures outlined in Appendix A and Appendix C, ensuring that all test samples follow the same construction steps, preserve semantic equivalence but lexically different between preceding tokens and alternative preceding tokens.

C Construction of Alternative Preceding Tokens

Given an input prompt x, we extract the preceding tokens u before the generation of the target answer token r. To construct semantically equivalent but lexically different alternatives, we use ChatGPT (GPT-5.2) as a controlled paraphrasing model. Specifically, we provide the model with only the preceding tokens u and an explicit instruction requiring it to (1) rewrite the input text to be semantically equivalent but lexically different, (2) avoid adding any new information, facts, explanations, or reasoning, and (3) refrain from inferring, mentioning, or hinting at any answer to a question. The generated sequence u' is referred to as the alternative preceding tokens.The detailed instructions are as follows.

> **Instruction (to GPT-5-2).** Rewrite the following text to be semantically equivalent but lexically and syntactically different. You must keep the meaning exactly the same, but change the surface form as much as possible (e.g., replace words with synonyms, alter phrasing, restructure sentences)
>
> **Constraints:**
>
> 1. Preserve the meaning exactly; do not change the intent, facts, or implications.
> 2. Do not add any new information, facts, examples, explanations, or reasoning.
> 3. Do not infer, mention, or hint at any answer to any question in the text.
>
> **Input text:** $\{u\}$

Because the model is not accessible to the target answer and is instructed to only rephrase the preceding tokens without adding new reasoning, the alternative preceding tokens u' keep the same meaning as u and do not contain evidence related to the answer. This process allows us to fairly compare $C(r \mid u, x)$ and $C(r \mid u', x)$ while focusing only on differences in the different preceding tokens.

D Related Work

D.1 Chain-of-Thought Prompting and Reasoning Faithfulness.

Chain-of-thought (CoT) prompting has been widely studied as a mechanism for eliciting intermediate reasoning steps in large language models (LLMs), often leading to substantial performance improvements on multi-step reasoning tasks (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023a; Zhou et al., 2022). Subsequent work has extended CoT through techniques such as self-consistency (Wang et al., 2023b), scratchpad prompting (Nye et al., 2022), and program-aided reasoning (Gao et al., 2023). Despite these advances, an increasing number of recent studies question whether the generated reasoning traces are truly faithful explanations of the model's actual underlying decision

Test Case	Prompt x	Revised Prompt x'	Preceding Tokens	Alt. Preceding Tokens	Answer Token
Sample1	Early in the afternoon, a visitor passed a classroom where Albert Einstein spoke softly to a small group,	Early in the afternoon, a visitor passed a classroom where James spoke softly to a small group,	and asked, 'What is he saying?' 'He's talking about the theory of	then inquired, 'What's he talking about?' 'He's going on about the principle of	relativ
Sample2	While exploring the old library wing, a researcher spotted Albert Einstein examining a worn notebook,	While exploring the old library wing, a researcher spotted James examining a worn notebook,	which he had used to record his theories. The notebook was a gift from a friend, and he had used it to record his thoughts on the theory of	which he had adopted to chronicle his theories. The notebook, a gift from a friend, served him to note down his ideas about the principle of	relativ
Sample3	While rearranging chairs, a volunteer noticed Stephen Hawking studying an intricate diagram,	While rearranging chairs, a volunteer noticed Ben studying an intricate diagram,	and asked him what it was. 'It's a model of the	and posed the question of what it was. 'It's a prototype of the	universe
Sample4	During the afternoon bustle, a merchant noticed Isaac Newton reading a parchment near the market booth,	During the afternoon bustle, a merchant noticed Carl reading a parchment near the market booth,	and asked him what he was reading. He replied that he was reading the Book of Nature, and that he had discovered the law of	and inquired him what he was immersed in. He informed that he was engaging with the Book of Nature, and that he had revealed the rule of	universal
Sample5	In the dim glow of a candle, a young scholar read Newton's notes,	In the dim glow of a candle, a young scholar read Jack's notes,	and the world changed. The story of the discovery of	and the world shifted. The account of the discovery process of	gravity
Sample6	On a lively morning, a visitor observed Barack Obama walking down the corridor thoughtfully,	On a lively morning, a visitor observed Mikl walking down the corridor thoughtfully,	and then he stopped and turned to the visitor. 'I'm thinking about the	and afterward he halted and turned to the caller. 'I'm pondering the	economy
Sample7	On a warm afternoon, a passerby spotted Barack Obama pausing near the courtyard benches,	On a warm afternoon, a passerby spotted Billy pausing near the courtyard benches,	where he sat down and began reading a book. The man, who asked not to be identified, said he was walking by the	where he took his seat and started to read a book. The person, who wished not to be named, stated he was walking past the	White
Sample8	During a quiet intermission, a spectator noticed Winston Churchill sipping tea thoughtfully,	During a quiet intermission, a spectator noticed Tom sipping tea thoughtfully,	and asked him, 'What are your thoughts on the	and posed the question, 'What do you think about	war
Sample9	During a calm intermission, a photographer glimpsed Hillary Clinton reading a brief printed summary,	During a calm intermission, a photographer glimpsed Jack reading a brief printed summary,	and snapped a picture. The photo was published in the	and captured a shot. The image presented at the	New
Sample10	As the morning breeze swept across the garden, a visitor paused before the Taj Mahal,	As the morning breeze swept across the garden, a visitor paused before the mountain,	the most beautiful monument in the world. The visitor was a young man, a student of architecture, who had come to	the most stunning monument in the world. The visitor, a young architecture student, had arrived to	India
Sample11	As the breeze cooled the streets, a traveler lifted their eyes toward the Burj Khalifa,	As the breeze cooled the streets, a traveler lifted their eyes toward the building,	the tallest building in the world. The traveler was a young man, a student, who had come to	the highest building in the world. The traveler, a young guy and a student, came here to	Dub
Sample12	After a long boat ride, a group of students stepped onto the platform facing the Statue of Liberty,	After a long boat ride, a group of students stepped onto the platform facing the statue,	and the first thing they saw was a sign that read, 'Welcome to	and the initial sight that greeted them was a sign bearing the words, 'You are now entering	America
Sample13	As the breeze picked up, a cyclist paused near Big Ben,	As the breeze picked up, a cyclist paused near the building,	and a woman in a red dress walked past the Houses of	and a lady wearing a red dress glided past the UK	Parliament
Sample14	In the quiet morning air, a guide pointed toward the high structures of Angkor Wat,	In the quiet morning air, a guide pointed toward the high structures of the building,	the largest religious monument in the world. The temple, built in the 12th century, is a symbol of	the world's grandest religious monument. This temple, erected in the 12th century, embodies	Camb
Sample15	At the edge of a quiet lake, a painter adjusted their canvas facing Mount Fuji,	At the edge of a quiet lake, a painter adjusted their canvas facing the mountain,	the iconic symbol of	the defining emblem of	Japan

Table 1: Details of a subset of test cases used in our evaluation. For each case, we report the original prompt x, the revised prompt x', the preceding tokens, an alternative preceding tokens, and the expected answer token.

process. Lanham et al. (2023) propose a systematic framework for evaluating the faithfulness of CoT and show that correct answers can often be produced even when the accompanying reasoning is incorrect. Similarly, Barez et al. (2025) argue that CoT should not be interpreted as a reliable explanation mechanism, demonstrating that models can produce plausible yet unfaithful reasoning.

D.2 Implicit Reasoning and Latent Answer Encoding.

Beyond explicit reasoning traces, several works suggest that answer-relevant information may already be implicitly encoded in the model's internal states prior to the generation of chain-of-thought. Lyu et al. (2023) show that intermediate representations can strongly predict final answers even before reasoning tokens are produced. Related findings indicate that LLMs may internally compute or approximate solutions and only later verbalize reasoning as a post-hoc justification (Turpin et al., 2023; Burns et al., 2024). These observations challenge the assumption that reasoning text directly mirrors the underlying inference process.

D.3 Shortcut Learning and Spurious Correlations.

Our work is also closely related to studies on shortcut learning and spurious correlations in neural models. Geirhos et al. (2020) demonstrate that deep networks often exploit superficial cues rather than learning robust, causal features. In natural language processing, similar phenomena have been observed in tasks such as natural language inference and question answering, where models rely on annotation artifacts or lexical biases (Niven and Kao, 2019; Gururangan et al., 2018). Recent analyses suggest that LLMs are particularly sensitive to subtle contextual patterns that correlate with target outputs, even when such patterns are not semantically meaningful (Xu et al., 2024; Min et al., 2022).

D.4 Prompt Sensitivity and Contextual Biasing.

A growing line of research has examined the sensitivity of LLMs to prompt formulations. Small

changes in wording, ordering, or contextual framing have been shown to substantially affect model predictions (Zhao et al., 2021; Lu et al., 2022; Liu et al., 2023). These findings indicate that context tokens may influence generation in ways that extend beyond their literal semantic content. Our work complements this literature by showing that even *semantically unrelated* preceding tokens can carry biased connotations that systematically push the model toward specific answers.

D.5 Attribution and Explainability for Language Models.

To better understand model behavior, numerous attribution and explanation methods have been proposed, including attention-based analyses and gradient-based approaches such as Integrated Gradients (Sundararajan et al., 2017). However, it has been repeatedly shown that attention weights alone do not provide reliable explanations (Wiegreffe and Pinter, 2019; Serrano and Smith, 2019). Recent work applies attribution methods to LLMs to analyze token-level influence and faithfulness of explanations (Ferrando et al., 2022). While these approaches offer valuable diagnostic insights, they typically focus on identifying salient tokens rather than examining how non-informative context may implicitly encode answer-relevant bias.

D.6 Positioning of Our Work.

In contrast to prior work that primarily investigates whether chain-of-thought reasoning is bypassed or unfaithful, we study a complementary and underexplored phenomenon. We show that semantically unrelated preceding tokens can carry biased connotations that implicitly encode answer-relevant information and that such biases can transfer across prompts even when explicit reasoning evidence is removed. This perspective extends existing analyses of reasoning faithfulness and shortcut learning by highlighting a new mechanism through which local context influences LLM generation, offering an alternative explanation for long-range dependency effects in language models.

E Artifact Licenses

The models used in this study were employed in accordance with their respective licenses and terms of use. LLaMA-2-7B was used under the Meta LLaMA 2 Community License Agreement for research purposes. ChatGPT (GPT-5.2) was accessed via the OpenAI platform and used in compliance with the OpenAI Terms of Use. DeepSeek-R1-Distill-Qwen-7B was used under the Apache License 2.0, consistent with the licensing of the underlying Qwen models. All models were used solely for non-commercial research and analysis, and no model weights or generated artifacts are redistributed.

F Documentation of Used Models

This study makes use of existing large language models. Documentation for the utilized models is provided by their original developers and is publicly available through official sources.

LLaMA-2-7B is documented by Meta through its official model card and accompanying technical documentation, which describe the model's training data composition, intended research use, and known limitations.

ChatGPT (GPT-5.2) is accessed via the OpenAI platform. Model behavior, usage guidelines, and limitations are documented in OpenAI's publicly available documentation and terms of use.

DeepSeek-R1-Distill-Qwen-7B follows the documentation of the DeepSeek project and the underlying Qwen model family, including model architecture, intended research usage, and licensing terms.

G Packages and Parameter Settings

Experiments are implemented in Python using standard deep learning and numerical computation libraries. Model inference is performed with PyTorch (v2.0.1) and Hugging Face Transformers (v4.36.2).

Token-level conditional probabilities are computed directly from model logits using softmax, without sampling. Text generation proceeds until the target answer token appears or a maximum length of 40 tokens is reached. No temperature scaling or nucleus/top-k sampling is applied.

Attribution analysis is conducted using the Integrated Gradients method, implemented analytically following Equation 5. Linear interpolation with 1000 integration steps is used to approximate the path integral.

Alternative preceding tokens are generated via controlled prompting of ChatGPT (GPT-5.2) accessed through the OpenAI platform. A fixed instruction template is used, and no additional preprocessing, normalization, or evaluation packages (e.g., NLTK, spaCy, ROUGE) are involved.