More Bias, Less Bias: BiasPrompting for Enhanced Multiple-Choice Question Answering

Duc Anh Vu
vuducanh001@e.ntu.edu.sg
Nanyang Techonological University
Singapore

Nguyen Thanh Thong
e0998147@u.nus.edu
National University of Singapore
Singapore

Cong-Duy Nguyen
duy.ntc@vinuni.edu.vn
Centre for AI research, VinUniversity
Vietnam

Viet Anh Nguyen
nguyenvi001@e.ntu.edu.sg
Nanyang Techonological University
Singapore

Anh Tuan Luu
anhtuan.luu@ntu.edu.sg
Nanyang Techonological University
Singapore

Abstract

With the advancement of large language models (LLMs), their performance on multiple-choice question (MCQ) tasks has improved significantly. However, existing approaches face key limitations: answer choices are typically presented to LLMs without contextual grounding or explanation. This absence of context can lead to incomplete exploration of all possible answers, ultimately degrading the models' reasoning capabilities. To address these challenges, we introduce BiasPrompting, a novel inference framework that guides LLMs to generate and critically evaluate reasoning across all plausible answer options before reaching a final prediction. It consists of two components: first, a reasoning generation stage, where the model is prompted to produce supportive reasonings for each answer option, and then, a reasoning-guided agreement stage, where the generated reasonings are synthesized to select the most plausible answer. Through comprehensive evaluations, BiasPrompting demonstrates significant improvements in five widely used multiple-choice question answering benchmarks. Our experiments showcase that BiasPrompting enhances the reasoning capabilities of LLMs and provides a strong foundation for tackling challenging questions, particularly in settings where existing methods underperform.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; *Artificial intelligence*; Machine learning algorithms.

ACM Reference Format:
Duc Anh Vu, Nguyen Thanh Thong, Cong-Duy Nguyen, Viet Anh Nguyen, and Anh Tuan Luu. 2026. More Bias, Less Bias: BiasPrompting for Enhanced Multiple-Choice Question Answering. In *Proceedings of The 41st ACM/SIGAPP Symposium on Applied Computing (SAC'26).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

SAC'26, Thessaloniki, Greece

ACM ISBN 979-X-XXXX-XXXX-X/26/03
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Large language models (LLMs) have significantly advanced state-of-the-art in natural language processing (NLP) [41, 9, 1], excelling in tasks such as natural language understanding [3, 8, 50, 49], text generation [53, 34], multimodal [20, 28, 29, 30, 33] and reasoning [17, 37]. Prompt optimization has become crucial for enhancing the performance of LLMs. However, despite their capabilities, LLMs remain sensitive to input design, particularly in multiple-choice question (MCQ) settings, where they often exhibit limited instruction-following ability and shallow task understanding [13, 31, 32, 35].

First, recent studies have shown that large language models are sensitive to the way input is structured, where a minor adjustment to the prompt may lead to an improvement or decline in performance. When provided with minimal context, such as only the answer choices without additional context, LLMs can exhibit intrinsic bias [55, 23], selection bias [56] or lack of task understanding [13].

Second, while many prompting strategies achieve remarkable performance, they bring additional computational burden of model inference. Prior approaches such as Chain-of-Thought (CoT) prompting [47], Self-Refine [26] and Self-Consistency [45] not only involve extensive token generation but also introduce challenges such as misinterpretation of intermediate steps and reduced interpretability [52]. In-context learning (ICL) [4, 54, 43] conditions the model on a long sequence of input-output demonstrations provided in the prompt. Furthermore, multi-agent frameworks [6, 19, 51] force LLMs to engage in multiple rounds of reasoning and coordination among agents before reaching a consensus. This process significantly increases computational costs, making such methods less efficient.

To address these challenges, we propose BiasPrompting, a lightweight yet effective prompting framework that enhances LLM reasoning in MCQ answering. BiasPrompting encourages the model to bias toward each possible candidate, enabling it to generate rich contextual reasoning for each option before achieving a final consensus. This structured evaluation ensures a deeper exploration of all options, mitigating model biases and improving decision-making. In addition, unlike existing methods that rely on complex multi-agent frameworks or extensive token generation, BiasPrompting can operate sequentially within a single LLM, making it computationally efficient. Our results demonstrate significant performance gains compared to prior inference methods. Moreover, BiasPrompting uncovers latent reasoning capabilities by successfully addressing questions that other methods underperform.

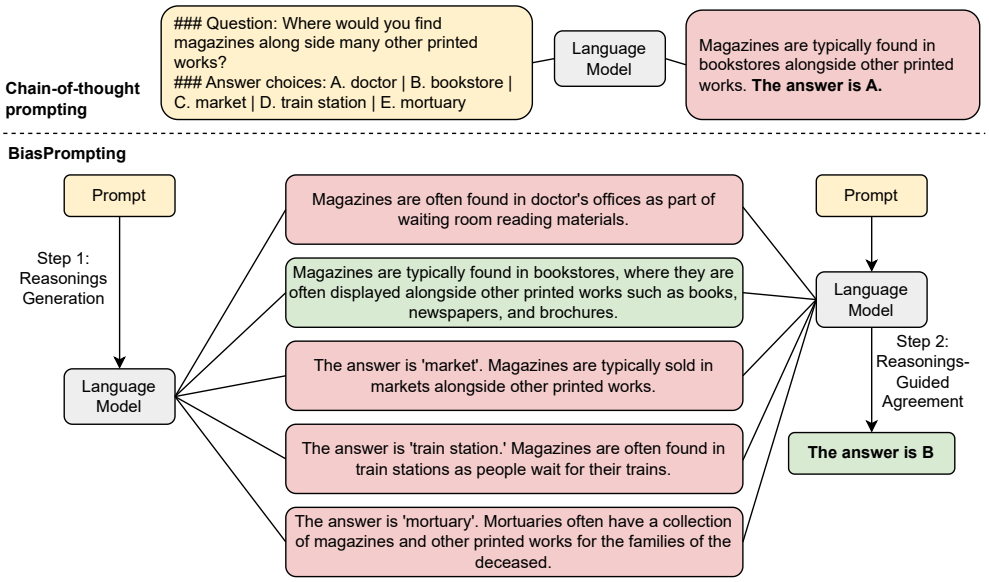

Figure 1: Illustration of BiasPrompting with two steps of Reasoning Generation and Reasonings-Guided Agreement.

2 BiasPrompting

BiasPrompting is motivated by the observation that in many multiple-choice question (MCQ) tasks, LLMs struggle to retrieve relevant facts when provided only with answer options. Formally, let an MCQ task be defined by a tuple (Q, A, y^*), where Q is the question given in a natural language, $A = \{A_1, A_2, \ldots, A_n\}$ is a set of n answer options and $y^* \in A$ is the ground-truth correct answer.

Let p_θ denote a pre-trained large language model (LLM) with parameters θ, where $p_\theta(y \mid Q, A, prompt)$ models the distribution over generated outputs conditioned on a question Q, candidate answers A, and a prompt design. Traditional prompting methods directly predict the sequence $[z_0, \ldots, z_m, y]$, where $z_0, \ldots, z_m$ are optional intermediate reasoning steps (enabled by certain prompting frameworks), and y is the final predicted answer. This sequence is modeled as: $[z_0, \ldots, z_m, y] = p_\theta(z_0, \ldots, z_m, y \mid Q, A, prompt)$.

As illustrated in Figure 1, given the question: *"Where would you find magazines along side many other printed works?"*, LLMs are given a limited set of answer options, each consisting of only a few words. Due to this constraint, the model often generates reasoning for its chosen option only, ignoring underexplored alternatives. In this example, the model selects *"doctor"* as the answer, failing to consider other options meaningfully.

BiasPrompting first prompts the LLM to generate supporting reasoning for each answer option (e.g., *"doctor", "bookstore", "market", "train station", "mortuary"*), **treating each as if it were correct**. This encourages the model to retrieve a broad range of relevant information before selecting an answer. By embedding controlled biases into the prompt, we hypothesize this method reduces the LLM's tendency to favor specific options. BiasPrompting involves two key steps:

2.1 Reasoning Generation

Given an MCQ task with question Q and answer options A, BiasPrompting first prompts the LLM to generate a reasoning R_i for each answer option A_i, where $i = 1, 2, \ldots, n$. Unlike prior methods that directly evaluate the correctness of each option, BiasPrompting instructs the model to produce reasoning that supports A_i as if it were the correct answer, regardless of its actual validity. This process is formalized as: $R_i \sim p_\theta(R_i \mid Q, A_i, prompt_{\text{RG}})$, where $prompt_{\text{RG}}$ is a prompt designed to generate reasoning that supports A_i.

This structured approach ensures that the model explores all answer options exhaustively, preventing premature dismissal of alternatives. Even if some generated reasonings are incorrect, this step enriches the model's understanding of the context and improves the robustness of the final prediction. The template of $prompt_{\text{RG}}$ is provided in Table 4.

2.2 Reasonings-Guided Agreement

After generating reasonings $R = \{R_1, R_2, \ldots, R_n\}$ for all answer options, BiasPrompting constructs a final prompt that integrates the original question Q, the answer options A, and the generated reasonings R. The LLM then predicts the final answer $y \in A$, modeled as: $y \sim p_\theta(y \mid Q, A, R, prompt_{\text{consensus}})$ where $prompt_{\text{consensus}}$ is designed to synthesize the reasonings and select the most appropriate answer. This prompt provides the model with richer context, enabling a more informed and transparent decision-making process compared to direct answer selection. The template of $prompt_{\text{consensus}}$ is shown in Table 5. Following [22], we adopt a placeholder format to wrap the selected answer in MCQ tasks. Additionally, we explore prompting techniques such as zero-shot and chain-of-thought [47] to enhance performance in this stage. Further details are discussed in Section 3.4.

	Mode	CommonsenseQA	StrategyQA	PIQA	Date Understanding	Causal Judgement
Mistral	Zero-shot	65.1	63.8	**75.1**	35.6	53.5
	Chain-of-Thought	**67.1** (+2.0%)	52.8 (-11.0%)	70.5 (-4.6%)	32.0 (-3.6%)	54.6 (+1.1%)
	BiasPrompting	66.0 (+0.9%)	**67.7** (+3.9%)	74.2 (-0.9%)	**40.0** (+4.4%)	**62.6** (+9.1%)
Deepseek	Zero-shot	**50.1**	57.1	64.8	37.6	53.5
	Chain-of-Thought	49.3 (-0.8%)	52.7 (-4.4%)	63.8 (-1.1%)	36.4 (-1.2%)	49.7 (-3.8%)
	BiasPrompting	48.5 (-1.6%)	**57.5** (+0.4%)	**64.9** (+0.1%)	**40.0** (+2.4%)	**55.6** (+2.1%)
Gemma	Zero-shot	27.0	48.5	48.9	20.0	**51.9**
	Chain-of-Thought	38.0 (+11.0%)	44.4 (-4.1%)	51.0 (+2.1%)	**27.6** (+7.6%)	38.0 (-13.9%)
	BiasPrompting	**38.7** (+11.7%)	**49.1** (+0.6%)	**54.2** (+5.3%)	23.2 (+3.2%)	**51.9** (-0.0%)

Table 1: Performance comparison across datasets and models with greedy decoding. Note that there can be an error of 0.1 due to rounding.

3 Experiment

3.1 Datasets

We cover abroad set of diverse MCQ tasks, varying length and domain. Our evaluation covers five popular multiple-choice question benchmarks: (i) CommonsenseQA [39], (ii) StrategyQA [7], (iii) PIQA [2], (iv) BBH–Date Understanding, and (v) BBH–Causal Judgement [38]. These datasets primarily test commonsense reasoning and vary in the number of answer options per question, ranging from 2 to 6. Detailed dataset statistics are provided in Table 2.

Dataset	Task Type	Size
CSQA [39]	5 Choices	1.221
StrategyQA [7]	2 Choices	2,290
PIQA [2]	2 Choices	1,838
BBH-DU [38]	6 Choices	250
BBH–CJ [38]	2 Choices	187

Table 2: Testing dataset statistics.

3.2 Model Details and Hyperparameters

We evaluate our method using three widely used, high-performing large language models (LLMs):

- Mistral-7B-It-v0.2[1][9]: A 7B-parameter decoder-only model with an 8k context window, outperforming LLaMA-2-13B [41] on multiple benchmarks.
- DeepSeek-7B-chat[2][1]: A 7B-parameter decoder-only model trained from scratch on a large-scale corpus of 2 trillion tokens.
- Gemma-7B-It[3][27]: A 7B-parameter decoder-only model with an 8k context window, noted for strong reasoning and problem-solving capabilities.

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[2] https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat
[3] https://huggingface.co/google/gemma-7b-it

We investigate the open-source models through HuggingFace transformers library [48]. To avoid empty or truncated outputs, we set a non-zero minimum number of generated tokens. For all models, we fix the context window at 1024 tokens and employ greedy decoding as the generation strategy.

3.3 Evaluation Measures

We report accuracy for all datasets. During inference, models are instructed to produce answers in a specific format to enable automated parsing of the predicted choice. A prediction is counted as correct if the parsed answer exactly matches the ground-truth label. We assess statistical significance of accuracy differences using a two-proportion z-test for each dataset–model pair, considering results significant when $p < 0.05$. All reported improvements meet this criterion.

3.4 Results & Discussions

3.4.1 Main results. Table 1 provides a comprehensive comparison of various models on five MCQ datasets. Overall, BiasPrompting consistently achieves superior performance, outperforming the zero-shot baseline in most cases. While zero-shot CoT sometimes attains the highest accuracy across certain datasets, its performance is highly inconsistent, often experiencing significant drops. In contrast, BiasPrompting offers a more stable and reliable improvement in all datasets, where it achieves the best or near-best scores. Moreover, unlike CoT, which depends on detailed step-by-step reasoning that can result in lengthy responses, BiasPrompting offers a more lightweight approach, improving performance without the need for extensive reasoning chains. Given these findings, BiasPrompting emerges as a promising strategy for mitigating inconsistencies observed in zero-shot CoT, offering both stability and performance gains while maintaining efficiency.

3.4.2 Latent reasoning capabilities with BiasPrompting. To further qualitatively demonstrate the superiority of BiasPrompting, we evaluate the results on the questions where baseline methods underperform. As illustrated in Figure 2, the number of questions that the other two approaches fail to address across the majority of

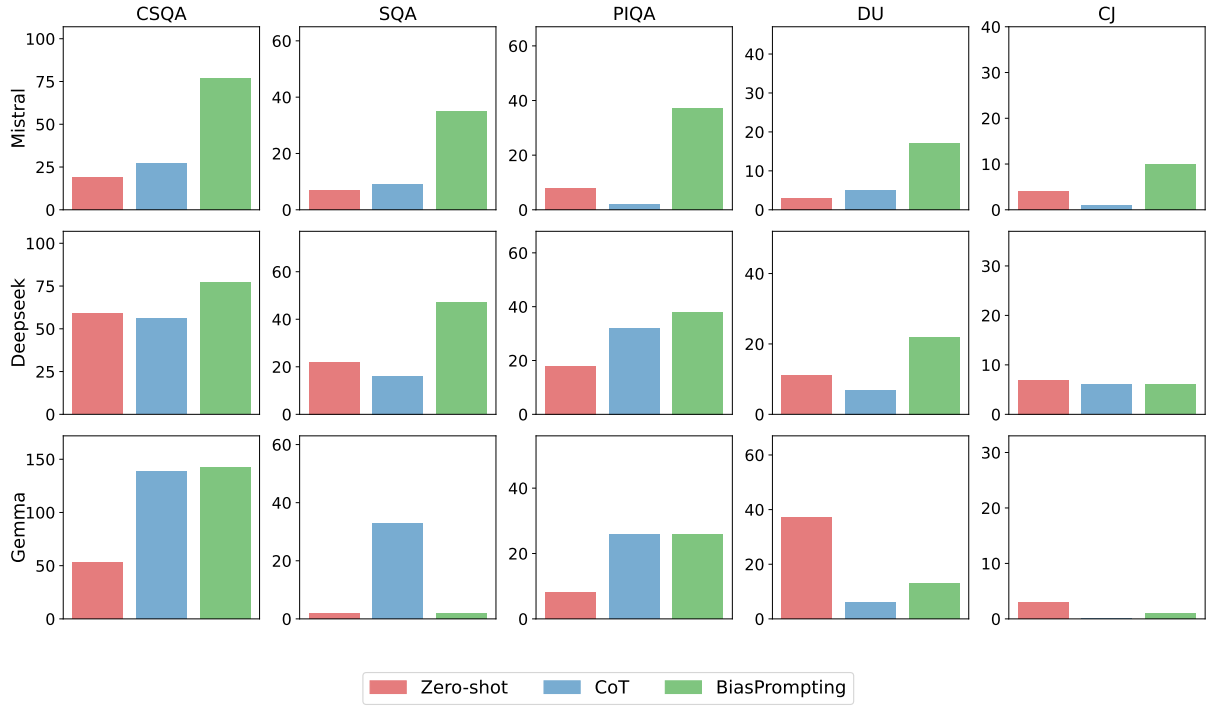

Figure 2: Number of questions successfully answered by each prompting method while the other two underperform.

experimental cases under BiasPrompting is significantly higher than Zero-shot and CoT. This highlights the latent reasoning capabilities of incorporating reasoning for all answer choices within the prompt, enabling the model to better evaluate and differentiate between options. Furthermore, BiasPrompting reveals previously unexplored capabilities of a single LLM, demonstrating its potential to enhance reasoning without requiring multiple LLMs or extensive prompting techniques.

Model	Mode	CSQA	SQA	PIQA
Mistral	+Bias	66.01	**67.69**	**74.16**
	+Bias +CoT	**66.50**	66.37	73.55
Deepseek	+Bias	46.68	**57.50**	64.91
	+Bias +CoT	**46.92**	46.14	**65.77**

Table 3: Performance comparison of Mistral and Deepseek models under BiasPrompting with and without CoT in the predicting final answer stage. Bold indicates the best scores among all methods.

3.4.3 BiasPrompting + CoT. We investigate the impact of incorporating Chain-of-Thought (CoT) reasoning into the Reasonings-Guided Agreement framework. In this setup, the language model generates intermediate reasoning steps in addition to the final answer $(y \in A)$, modeled as:

$$[z_0, \ldots, z_m, y] = p_\theta(z_0, \ldots, z_m, y \mid Q, A, R, prompt_{\text{consensus}})$$

As shown in Table 3, integrating CoT with BiasPrompting does not lead to significant improvements in final answer prediction. While minor performance fluctuations are observed, the overall results suggest that CoT does not consistently enhance accuracy across datasets. In some cases, such as the Mistral model on the CSQA, the combination of BiasPrompting and CoT even underperforms compared to BiasPrompting alone. These findings demonstrate that BiasPrompting alone is sufficient for improving performance, and the added complexity from CoT does not necessarily translate to better outcomes.

3.4.4 Token Efficiency Analysis. To assess computational efficiency, we compare the total number of generated tokens for BiasPrompting and CoT across five datasets, aggregating tokens over all rounds of prompting required to produce a final answer (Figure 3). The results show that BiasPrompting consistently generates substantially fewer tokens than CoT, while maintaining or improving accuracy. This reduction stems from its single-pass design, which integrates reasoning for all answer choices within one prompt, avoiding the lengthy, multi-round reasoning chains typical of CoT. Consequently, BiasPrompting achieves both inference-time efficiency and performance gains, highlighting its practicality for deployment in resource-constrained settings.

3.4.5 BiasPrompting Reduces Selection Bias. Figure 4 presents the robustness comparison between Zero-shot and BiasPrompting across three random answer option orderings for Mistral, DeepSeek, and Gemma. In all cases, BiasPrompting yields higher median accuracy and lower variance than Zero-shot, demonstrating improved

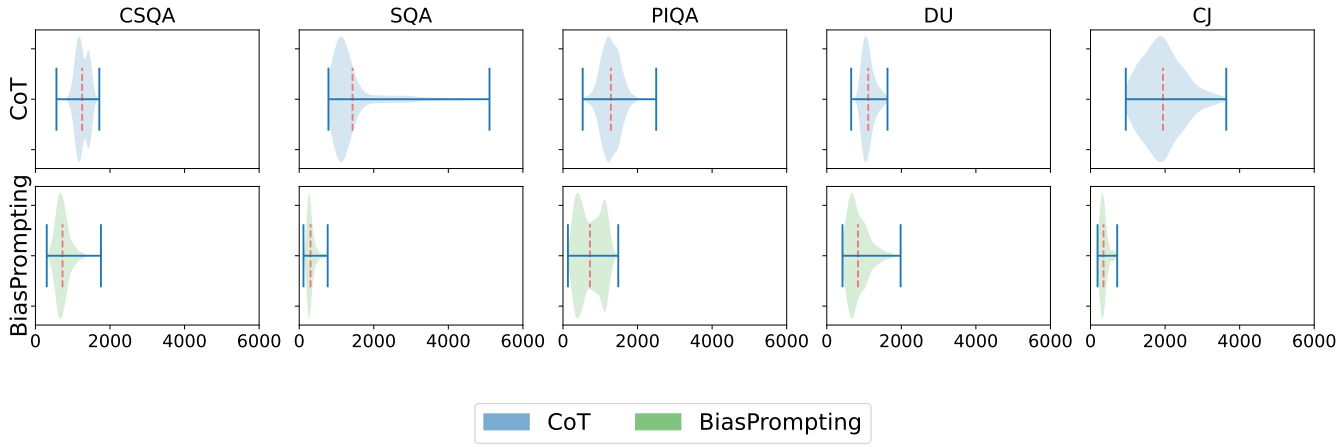

Figure 3: Violin plots comparing the average generated tokens from BiasPrompting and CoT prompting across five datasets (CSQA, SQA, PIQA, DU, and CJ) using the Gemma model. The ■ violins represent BiasPrompting, while the ■ violins correspond to CoT prompting. The dashed line inside each violin indicates the mean generated tokens.

overall performance. These results suggest that BiasPrompting effectively reduces the susceptibility of model predictions to option order permutations, addressing the selection bias highlighted by [56].

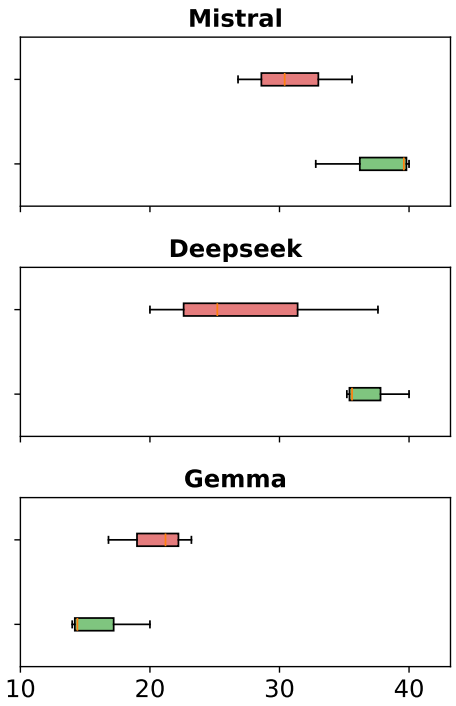

Figure 4: Results of option order swapping on zero-shot and BiasPrompting performance across random option orderings. ■ denotes the box-plot of BiasPrompting, while ■ represents the box-plot of zero-shot.

4 Related Work

4.1 In-context learning

Numerous studies have focused on enhancing LLM performance through improved prompting techniques. Early research such as Chain-of-Thought (CoT) [47], Thread-of-Thought [58], and Tab-CoT [10] encourage LLMs to explicitly generate intermediate reasoning steps before producing a final answer, improving their ability to solve complex problems. Additionally, ensemble-based approaches [45, 26, 12] leverage self-consistency by generating multiple responses and selecting the most frequent one, effectively reducing output variance and improving reliability. Another line of research has explored problem decomposition, where complex tasks are broken down into simpler sub-questions to enhance reasoning accuracy [15, 25, 57].

Beyond these structured prompting techniques, various zero-shot prompting strategies have been developed to guide LLMs more effectively. Persona prompting [46] assigns a specific role to the model, influencing its response style, while emotion-based prompting [18] incorporates psychologically relevant phrases to elicit more human-like and context-aware outputs. Recent extensions, such as [55], adjust for inherent biases in zero-shot predictions, further stabilizing performance. Few-shot prompting builds on zero-shot by incorporating a small number of examples in the prompt, enabling in-context learning (ICL) without fine-tuning, which conditions the model on input-output demonstrations, allowing adaptation to new tasks with minimal data. For example, [5] demonstrates how self-supervised training enhances few-shot ICL, improving generalization in low-data scenarios. Recent advancements refine ICL for cost and effectiveness. Techniques like Decomposed Prompting [14] break examples into sub-tasks, while Demonstration Ensembling [11] combines multiple few-shot samples for robustness.

ICL variants also address domain-specific challenges. For instance, [42] introduces curriculum demonstration selection, progressively increasing example complexity to boost few-shot performance. In educational applications, [16] uses few-shot prompts for

generating language-agnostic MCQs, leveraging chain-of-thought-inspired multi-stage prompting to create diverse distractors. Despite its strengths, few-shot ICL can amplify biases from examples [24] and struggles with long contexts due to token limits. Mitigation efforts have been made to these issues, such as [54] explores k-nearest neighbor ICL for compositional generalization.

4.2 LLMs on MCQ answering

Prior works have explored MCQ solving from multiple research angles. For example, in terms of LLM inherent bias, [56] identified and analyzed selection bias, showing that changes in the ordering of answer options can significantly alter model predictions, similarly,[36] studied positional bias, where certain positions in the answer list are favored regardless of content.

Beyond bias analysis, prompt engineering approaches such as [22, 21] propose optimally designed templates tailored for MCQ settings to improve consistency and accuracy. Prior works also conduct in-depth investigations into the specific behaviors exhibited by LLMs when the underlying task is MCQ solving, for instance, [44] reveals that LLMs often succeed in MCQA not through true understanding but by eliminating implausible options, highlighting a gap between apparent performance and actual reasoning depth. In specialized domains, [40] uses LLMs for medical MCQ feedback, while [44] critiques MCQ rationality, advocating true-or-false complements.

5 Conclusion

In this paper, we introduced BiasPrompting, a novel inference framework that enhances LLM performance in multiple-choice question answering by generating reasonings for all answer choices before selecting a final answer. Our experiments demonstrate that BiasPrompting consistently improves accuracy, mitigates biases like ordering effects, and offers greater stability compared to traditional methods while being more computationally efficient. Our findings also highlight its ability to improve decision-making in challenging questions where other methods often underperform. We hope our work will encourage more human-like approaches to unlock the hidden potential of LLMs.

6 Limitations

While BiasPrompting demonstrates significant improvements in multiple-choice question answering, several limitations remain that warrant further investigation. First, our experiments primarily focus on commonsense reasoning tasks. While these benchmarks are valuable for assessing LLM reasoning capabilities, BiasPrompting has not been systematically evaluated on other reasoning domains. Extending BiasPrompting to a broader range of tasks would provide deeper insights into its generalizability and effectiveness across different reasoning paradigms. Second, our study is limited to 7B parameter models (Mistral-7B, DeepSeek-7B, and Gemma-7B). While these models offer a balance between computational efficiency and performance, they may not fully capture the behavior of larger-scale LLMs. The impact of BiasPrompting on these larger models remains unexplored and could reveal new insights into its scalability, efficiency, and effectiveness at different model sizes. Additionally, the quality of reasoning generation has not been extensively refined. Future work can focus on enhancing reasoning generation to improve accuracy and depth without compromising efficiency.

A Prompts used in BiasPrompting

Given the following question: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what? Provide reasoning proving that 'bank' is the correct choice without any textual description in one sentence.

Table 4: Reasoning generation prompt example

Question: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what? ### Answer choices: A. bank | B. library | C. department store | D. mall | E. new york ### Reasoning for answer choice A: The answer is 'bank'. Revolving doors are often used in banks to control the flow of people entering and exiting. ### Reasoning for answer choice B: The answer is 'library'. Revolving doors are commonly used in libraries to control the flow of people entering and exiting. ### Reasoning for answer choice C: The answer is department store. Revolving doors are commonly used in department stores to control the flow of customers. ### Reasoning for answer choice D: The answer is 'mall'. Revolving doors are commonly found in malls to control pedestrian traffic flow and prevent congestion. ### Reasoning for answer choice E: The answer is New York. The revolving door is a security measure at the Empire State Building in New York City. Wrap your final answer by filling in the placeholder below: 'So the answer is: {{placeholder}}'

Table 5: BiasPrompting prompt example

References

[1] Xiao Bi et al. 2024. Deepseek llm: scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

[2] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence* number 05. Vol. 34, 7432–7439.

[3] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[4] Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, (Eds.) Association for Computational Linguistics, Seattle, United States, (July 2022), 3558–3573. doi:10.18653/v1/2022.naacl-main.260.

[5] Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. (2022). https://arxiv.org/abs/2205.01703 arXiv: 2205.01703 [cs.CL].

[6] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

[7] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9, 346–361.

[8] Nhat Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Anh Tuan Luu. 2024. ToXCL: a unified framework for toxic speech detection and explanation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Kevin Duh, Helena Gomez, and Steven Bethard, (Eds.) Association for Computational Linguistics, Mexico City, Mexico, (June 2024), 6460–6472. doi:10.18653/v1/2024.naacl-long.359.

[9] Albert Q Jiang et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

[10] Ziqi Jin and Wei Lu. 2023. Tab-cot: zero-shot tabular chain of thought. *arXiv preprint arXiv:2305.17812*.

[11] Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Exploring demonstration ensembling for in-context learning. (2023). https://arxiv.org/abs/2308.08780 arXiv: 2308.08780 [cs.CL].

[12] Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Exploring demonstration ensembling for in-context learning. *arXiv preprint arXiv:2308.08780*.

[13] Aisha Khatun and Daniel G Brown. 2024. A study on large language models' limitations in multiple-choice question answering. *arXiv preprint arXiv:2401.07955*.

[14] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: a modular approach for solving complex tasks. (2023). https://arxiv.org/abs/2210.02406 arXiv: 2210.02406 [cs.CL].

[15] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: a modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

[16] Eyal Klang et al. 2023. Utilizing artificial intelligence for crafting medical examinations: a medical education study with gpt-4. (July 2023). doi:10.21203/rs.3.rs-3146947/v1.

[17] 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.

[18] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

[19] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36, 34892–34916.

[21] Michael Xieyang Liu, Frederick Liu, Alexander J Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J Cai. 2024. " we need structured output": towards user-centered constraints on large language model output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–9.

[22] Do Xuan Long, Hai Nguyen Ngoc, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F Chen, and Min-Yen Kan. 2024. Llms are biased towards output formats! systematically evaluating and mitigating output format bias of llms. *arXiv preprint arXiv:2408.08656*.

[23] Kristian Lum, Jacy Reese Anthis, Chirag Nagpal, and Alexander D'Amour. 2024. Bias in language models: beyond trick tests and toward ruted evaluation. *arXiv preprint arXiv:2402.12649*.

[24] Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander D'Amour. 2025. Bias in language models: beyond trick tests and toward ruted evaluation. (2025). https://arxiv.org/abs/2402.12649 arXiv: 2402.12649 [cs.CL].

[25] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*.

[26] Aman Madaan et al. 2024. Self-refine: iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

[27] Gemma Team Thomas Mesnard et al. 2024. Gemma: open models based on gemini research and technology. *ArXiv*, abs/2403.08295. https://api.semanticscholar.org/CorpusID:268379206.

[28] Cong-Duy Nguyen, Thong Nguyen, Duc Vu, and Anh Luu. 2023. Improving multimodal sentiment analysis: supervised angular margin-based contrastive learning for enhanced fusion representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Houda Bouamor, Juan Pino, and Kalika Bali, (Eds.) Association for Computational Linguistics, Singapore, (Dec. 2023), 14714–14724. doi:10.18653/v1/2023.findings-emnlp.980.

[29] Cong-Duy Nguyen, Xiaobao Wu, Duc Anh Vu, Shuai Zhao, Thong Nguyen, and Anh Tuan Luu. 2025. Cutpaste&find: efficient multimodal hallucination detector with visual-aid knowledge base. *arXiv preprint arXiv:2502.12591*.

[30] Thong Nguyen, Yi Bin, Xiaobao Wu, Xinshuai Dong, Zhiyuan Hu, Khoi Le, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. 2024. Meta-optimized angular margin contrastive framework for video-language representation learning. *ECCV 2024*.

[31] Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. 2024. Video-language understanding: a survey from model architecture, model training, and data perspectives. In *ACL 2024 (Findings)*.

[32] Thong Nguyen, Anh Tuan Luu, Truc Lu, and Tho Quan. 2021. Enriching and controlling global semantics for text summarization. *arXiv preprint arXiv:2109.10616*.

[33] Thong Thanh Nguyen, Zhiyuan Hu, Xiaobao Wu, Cong-Duy T Nguyen, See Kiong Ng, and Luu Anh Tuan. 2024. Encoding and controlling global semantics for long-form video question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7049–7066.

[34] Thong Thanh Nguyen and Anh Tuan Luu. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 10. Vol. 36, 11103–11111.

[35] Thong Thanh Nguyen, Xiaobao Wu, Yi Bin, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2025. Motion-aware contrastive learning for temporal panoptic scene graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 6. Vol. 39, 6218–6226.

[36] Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Kevin Duh, Helena Gomez, and Steven Bethard, (Eds.) Association for Computational Linguistics, Mexico City, Mexico, (June 2024), 2006–2017. doi:10.18653/v1/2024.findings-naacl.130.

[37] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: a survey. *arXiv preprint arXiv:2212.09597*.

[38] Mirac Suzgun et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

[39] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: a question answering challenge targeting commonsense knowledge. 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)*. Jill Burstein, Christy Doran, and Thamar Solorio, (Eds.) Association for Computational Linguistics, Minneapolis, Minnesota, (June 2019), 4149–4158. doi:10.18653/v1/N19-1421.

[40] Mihaela Tomova, Iván Roselló Atanet, Victoria Sehy, Miriam Sieg, Maren März, and Patrick Mäder. 2024. Leveraging large language models to construct feedback from medical multiple-choice questions. *Scientific reports*, 14, 1, 27910.

[41] Hugo Touvron et al. 2023. Llama: open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[42] Duc Anh Vu, Nguyen Tran Cong Duy, Xiaobao Wu, Hoang Minh Nhat, Du Mingzhe, Nguyen Thanh Thong, and Anh Tuan Luu. 2024. Curriculum demonstration selection for in-context learning. (2024). https://arxiv.org/abs/2411.18126 arXiv: 2411.18126 [cs.CL].

[43] Duc Anh Vu, Cong-Duy Nguyen, Xiaobao Wu, Nhat Hoang, Mingzhe Du, Thong Nguyen, and Anh Tuan Luu. 2025. Curriculum demonstration selection for in-context learning. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, 1004–1006.

[44] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024. Llms may perform mcqa by selecting the least incorrect option. (2024). https://arxiv.org/abs/2402.01349 arXiv: 2402.01349 [cs.CL].

[45] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

[46] Zekun Moore Wang et al. 2023. Rolellm: benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

[47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824–24837.

[48] Thomas Wolf et al. 2020. Transformers: state-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Qun Liu and David Schlangen, (Eds.) Association for Computational Linguistics, Online, (Oct. 2020), 38–45. doi:10.18653/v1/2020.emnlp-demos.6.

[49] Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57, 2, 18.

[50] Xiaobao Wu, Thong Nguyen, Delvin Zhang, William Yang Wang, and Anh Tuan Luu. 2024. Fastopic: pretrained transformer is a fast, adaptive, stable, and transferable topic model. *Advances in Neural Information Processing Systems*, 37, 84447–84481.

[51] Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: an in-depth analysis via debate. *arXiv preprint arXiv:2305.11595*.

[52] Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering questions by meta-reasoning over multiple chains of thought. *arXiv preprint arXiv:2304.13007*.

[53] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12, 39–57.

[54] Wenting Zhao et al. 2024. *k*NN-ICL: compositional task-oriented parsing generalization with nearest neighbor in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Kevin Duh, Helena Gomez, and Steven Bethard, (Eds.) Association for Computational Linguistics, Mexico City, Mexico, (June 2024), 326–337. doi:10.18653/v1/2024.naacl-long.19.

[55] 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*. PMLR, 12697–12706.

[56] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. On large language models' selection bias in multi-choice questions. *arXiv preprint arXiv:2309.03882*.

[57] Denny Zhou et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

[58] Yucheng Zhou, Xiubo Geng, Tao Shen, Chongyang Tao, Guodong Long, Jian-Guang Lou, and Jianbing Shen. 2023. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734*.