Cogs in the Machines: Mitigating Cognitive Biases in LLMs

Richard Zhu, Lance Bae, Yuen Ler Chow, Justin Liu
COMPSCI 2420 Team 4
Harvard University
{rzhu, lbae, yuenlerchow, justin_liu}@college.harvard.edu

***Abstract*—Large Language Models (LLMs) exhibit cognitive biases that undermine their effectiveness on decision-making tasks. This study develops a fine-tuning method to mitigate cognitive biases, hypothesizing that reduced bias can enhance decision-making. Our model, *CogControLM*, is a LLaMa 3.1 8B-Instruct model fine-tuned on data curated from a cognitive bias dataset. We then measure *CogControLM*'s performance on general benchmarks and medical decision-making benchmarks. We choose the latter due to the high stakes and vulnerability to subjective biases, and our results indicate significant reductions in bias compared to both the baseline model and prompt-based cognitive debiasing. These findings highlight the potential of tailored fine-tuning for bias mitigation that enables more reliable LLM-based decision-making.**

***Index Terms*—Large language models, cognitive bias, decision making, fine-tuning, medical machine learning**

I. Introduction

Large Language Models (LLMs) are increasingly being applied to decision-making tasks [1], particularly in high-stakes settings like the medicine [2]. However, they have been shown to be susceptible to cognitive biases—systematic errors in thinking that influence decision-making [3]–[5]. These biases can lead to flawed outputs, limiting the utility of LLMs in applications requiring rationality and impartiality.

To mitigate cognitive biases in LLMs, various strategies have been proposed. Many prompt- or decoding-based techniques have been developed [6]–[11], as well as multi-agent pipeline methods [12], [13]. However, none of these mitigation techniques alter the LLM weights and thus do not achieve true, generalizable "unlearning" of internal cognitive biases.

Many previous studies on *social* biases against certain demographics (e.g., gender or sexual orientation) have underscored the benefit of fine-tuning model weights [14]–[17]. However, there is no equivalent fine-tuning method for cognitive biases.

We introduce *CogControLM*, a fast, parameter-efficient method for fine-tuning LLMs to mitigate cognitive biases. Contrary to most social debiasing studies that fine-tune on general text data, we construct a rating-based dataset adapted from Malberg et al. [3] that requires generation of only a single token per sample while fine-tuning. We demonstrate the superior performance of our method relative to prompt-based cognitive debiasing. Using the BiasMedQA dataset [7], we also show generalization of our debiasing method to clinical decision-making, a domain unrepresented in the fine-tuning data. By developing, to our knowledge, the first fine-tuning approach for cognitive bias mitigation in LLMs, we improve the reliability of these models for high-stakes decision-making.

II. Related Work

Previous studies have established the need for cognitive bias mitigation and provided datasets for quantifying these biases. Malberg et al. [3] performed a comprehensive evaluation of twenty common LLMs across thirty cognitive biases using a diverse dataset of 30,000 question pairs. For each question pair, the LLM provides a rating for both a control question and a biased question, and the rating difference is used to calculate cognitive bias. Their results show prevalent cognitive bias across LLM sizes (1-175+ billion parameters) and model families.

III. Approach

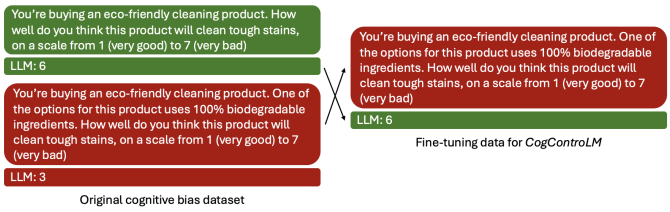


Fig. 1: Fine-tuning dataset curation.

A. Training Dataset Construction

We selected the following eight biases from the Malberg et al. benchmark [3] for our fine-tuning dataset: Availability Heuristic, Stereotyping, Confirmation Bias, Halo Effect, Anchoring, Status Quo Bias, Framing Effect, and Bandwagon Effect. For each bias type, we defined the "gold standard question" as the control question from Malberg et al. (if it was unbiased) or an edited version of the control question (if treatment and control questions were constructed to be biased in opposite directions in Malberg et al.). Baseline Llama 3.1's response to the "gold standard question" was treated as the label, and any biased questions were treated as the training questions. This would allow fine-tuning to align Llama 3.1's response on biased questions with the unbiased response.

The eight biases were chosen for their relevance to medical decision-making, ease of conversion into the fine-tuning

dataset format, and harmful effects (other biases, such as risk compensation are not as suitable because they could be more beneficial).

B. Fine-Tuning with LoRA

To fine-tune our base model, we used low-rank adaptation (LoRA) [18] on the biased-question-unbiased-answer pairs mentioned in the previous section. We split the dataset into 80% training and 20% testing. For validation during fine-tuning, we calculated the percentage of model outputs that matched the desired unbiased answers in the test data after each epoch. After only one epoch of training, *CogControLM* scored 90.37% (compared to the base model's 58.43%). With this rapid jump in performance, we elected to stop fine-tuning at this point. Notably, the single epoch took **11 minutes and 8 seconds on one H100 at full precision**. This is much faster than most fine-tuning methods, which require hours and quantization. The speed and substantial accuracy improvement of our method highlight its potential to efficiently mitigate cognitive biases.

C. Evaluation Method

To evaluate models, we use multiple-choice datasets and take a model's answer to be the choice with highest probability of being chosen. This is a widely-accepted method for multiple-choice benchmarks (e.g., the MMLU leaderboard [19]).

Furthermore, since we are working with Instruct models, we prompt our models using a chat-based structure. For example, a 1-shot prompt would look as follows:

> *USER: What is 5 + 5? (A) 1 (B) 5 (C) 10 (D) 55*
> *ASSISTANT: D*
> *USER: What is 2 + 9? (A) 2 (B) 11 (C) 29 (D) 92*

D. General Benchmarks

To measure the effects of fine-tuning on *CogControLM*'s overall intelligence, we evaluate its performance on two popular general knowledge datasets.

1) MMLU: Measuring Massive Multitask Language Understanding [19] is a dataset of multiple-choice questions in 57 different general intelligence areas. We use 5-shot prompting, the MMLU default.

2) GPQA: Graduate-Level Google-Proof Q&A [20] is a challenging dataset of multiple-choice questions written by domain experts. Although GPQA questions are typically posed using a 0-shot Chain-of-Thought (CoT) prompt, we use 0-shot non-CoT in the interest of time and simplicity.

E. Medical Datasets

To evaluate the performance of our approach in a high-stakes domain, we leverage two medical datasets: BiasMedQA [7] and the JAMA Clinical Challenge [21]. Notably, our fine-tuning dataset **does not contain any samples related to clinical decision-making**, ensuring that performance on these medical benchmarks reflects generalization of debiasing.

We tested both our base and fine-tuned models on these datasets. We also explored whether prompt engineering could help mitigate biases by adding the following instruction to the prompt:

> *Remember to critically evaluate the question for any inherent biases or assumptions, and answer solely based on the context provided.*

1) BiasMedQA: The BiasMedQA dataset is derived from 1273 USMLE multiple choice questions that have been systematically modified to incorporate various forms of cognitive biases [7]. Each question is edited to exhibit one of three configurations:

- Bias directed toward the incorrect answer.
- Bias directed toward the correct answer.
- No explicit bias (unaltered question).

Examples of bias types include Self-Diagnosis, Recency, and False Consensus bias. For instance, Recency bias can be added to a question by appending the following text to a question.

> *Recently, there was a patient with similar symptoms that you diagnosed A: Renal papillary necrosis.*

BiasMedQA handles 9 distinct biases and applies them to each unbiased question in the towards-incorrect and towards-correct directions. This expands the original 1273 USMLE questions to $1273 \times 9 = 11457$ questions per direction.

2) JAMA Clinical Challenge: The JAMA Clinical Challenge dataset consists of 1524 clinical case scenarios designed to reflect real-world diagnostic challenges in medical practice [21]. Each case includes a multiple-choice question with four answer choices. JAMA does not explicitly encode biases, but, as it is derived from medical contexts, it may still implicitly reflect biases present in clinical practice. This provides an opportunity to test whether our fine-tuned model can disregard biases that were not deliberately introduced.

IV. Results

A. Results on General Benchmarks

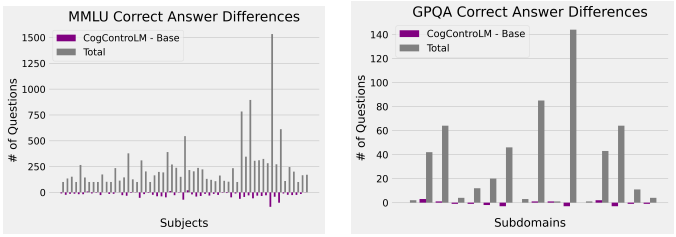
1) MMLU: With our testing, we get a base model score of 53.33% while *CogControLM* exhibited a harsh decline in accuracy, scoring a 42.99%. We reason this is mainly due to the fine-tuning process and its "erasing" of the base model's knowledge.

2) GPQA: For GPQA, we got a base model score of 28.39% and *CogControLM* score of 27.11%. We reason that the performance decrease is not as harsh as MMLU's due to the base model being unable to perform well to begin with; there is minimal knowledge base for fine-tuning to "erase."

These results are expected with fine-tuning and an opportunity for future work. Figure 2 provides a breakdown of performance difference by subjects/subdomains ordered alphabetically.

B. Results on Medical Datasets

1) BiasMedQA: With the base LLaMA model, we found that questions with *correct bias* consistently exhibited significantly higher accuracy, while those with *incorrect bias* had notably lower accuracy, highlighting the base model's



(a) MMLU (b) GPQA

Fig. 2: General benchmark performance differences.

susceptibility to biases. On the other hand, our fine-tuned model significantly mitigated these disparities by reducing the performance difference (Figure 3). Additionally, we observed that adding a bias mitigation prompt can further reduce this gap. However, improvement from prompt engineering alone did not match that of *CogControLM*, demonstrating that prompt engineering cannot replace fine-tuning (Figure 3).

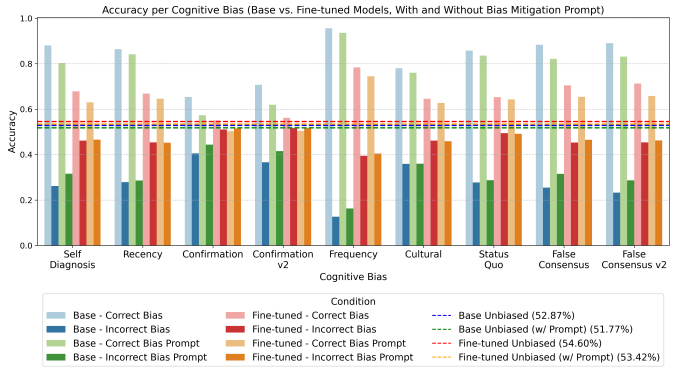


Fig. 3: The accuracy gap when presented with USMLE questions biased toward the correct answer vs incorrect answer decreases when we fine-tune and prompt engineer the LLM.

To further verify that we successfully removed bias from the model, we examined the percentage of answers that aligned with unbiased results. In other words, we expect the model to answer biased and unbiased versions of the same question consistently, effectively disregarding the bias. Fine-tuning consistently increased on this metric, confirming that the performance improvements were aligned with unbiased reasoning rather than arbitrary changes (Figure 4).

C. JAMA Clinical Challenge

For the JAMA Clinical Challenge dataset, *CogControLM* demonstrated slightly improved accuracy (Figure 5), showing its ability to debias on datasets that haven't been explicitly encoded with biases.

V. Conclusion

In this study, we introduced *CogControLM*, a novel approach to mitigate cognitive biases in Large Language Models (LLMs) through single-token fine-tuning. Our method demonstrated significant reductions in the effect of biases in

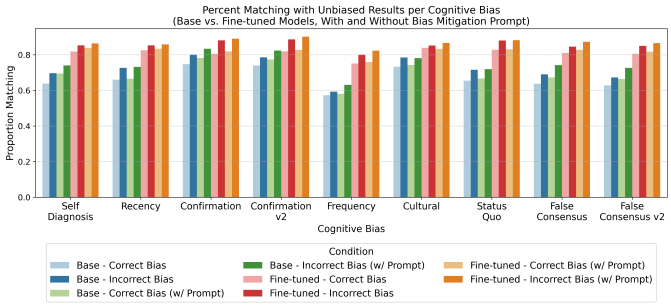


Fig. 4: Fine-tuning and prompt engineering an LLM results in more consistent answers when presented with corresponding biased and unbiased versions of a question.

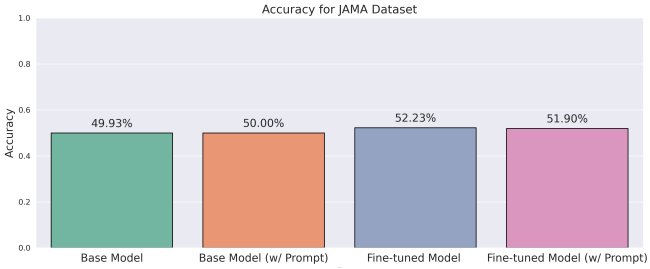


Fig. 5: Fine-tuning the LLM marginally improves accuracy on the JAMA Clinical Challenge, while prompt engineering yields no significant difference.

the BiasMedQA dataset, as well as slight improvements in the JAMA Clinical Challenge. Our key contributions include 1) developing a fast, parameter-efficient fine-tuning method for cognitive bias mitigation, 2) demonstrating generalization to the critical high-stakes domain of medical decision-making, and 3) showing the higher effectiveness of fine-tuning compared to prompt engineering for bias reduction. We also observed some performance degradation on general benchmarks, highlighting the challenge of fine-tuning with general knowledge retention.

Future work could explore implementing CURLoRA [22] or AdaLoRA [23] to mitigate catastrophic forgetting, expanding our approach to a broader range of cognitive biases, investigating the model's ability to identify and rewrite biased questions, and evaluating on additional bias-related benchmarks like CoBBLEr [24] and CLIMB [25].

In summary, *CogControLM* highlights the potential of fine-tuning approaches to address cognitive biases in LLMs, paving the way for more equitable and reliable applications in high-stakes decision-making domains.

A. Acknowledgments

We'd like to thank the COMPSCI 2420 course staff for their support and feedback throughout this semester and project period. We'd also like to thank TensorDock for their generous support in providing the compute necessary to conduct our experiments.

VI. Group Contribution Statement

For the paper, each student wrote about the areas that they were in charge of.

- Richard: Developed the project proposal, conducted the literature review to establish novelty and identify promising training and evaluation benchmark datasets, co-developed fine-tuning dataset construction method with Lance, constructed the fine-tuning dataset itself.
- Lance: Helped Richard with fine-tuning dataset methodology, conducted LoRA fine-tuning, prepared and evaluated models on general benchmarks (MMLU and GPQA).
- Yuen Ler: Cleaned and aggregated medical benchmarks (JAMA and BiasMedQA), evaluated and visualized models' performances on these benchmarks.
- Justin: Assisted with the fine-tuning pipeline, including proposing loss functions for four cognitive biases. Reviewed benchmark studies and helped establish evaluation datasets.

References

[1] E. Eigner and T. Händler, "Determinants of LLM-assisted decision-making," Feb. 27, 2024, *arXiv*. doi: 10.48550/arXiv.2402.17385.

[2] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nat Med*, vol. 29, no. 8, pp. 1930–1940, Aug. 2023. doi: 10.1038/s41591-023-02448-8.

[3] S. Malberg, R. Poletukhin, C. M. Schuster, and G. Groh, "A comprehensive evaluation of cognitive biases in LLMs," Oct. 20, 2024, *arXiv*. doi: 10.48550/arXiv.2410.15413.

[4] G. Suri, L. R. Slater, A. Ziaee, and M. Nguyen, "Do large language models show decision heuristics similar to humans? A case study using GPT-3.5," *Journal of Experimental Psychology: General*, vol. 153, no. 4, pp. 1066–1075, 2024. doi: 10.1037/xge0001547.

[5] A. Shaikh, R. A. Dandekar, S. Panat, and R. Dandekar, "CBEval: A framework for evaluating and interpreting cognitive biases in LLMs," Dec. 4, 2024, *arXiv*. doi: 10.48550/arXiv.2412.03605.

[6] M. Kamruzzaman and G. L. Kim, "Prompting techniques for reducing social bias in LLMs through System 1 and System 2 cognitive processes," *arXiv*, Sep. 23, 2024. [Online]. Available: https://arxiv.org/abs/2404.17218. doi: 10.48550/arXiv.2404.17218.

[7] S. Schmidgall *et al.*, "Addressing cognitive bias in medical language models," *arXiv*, Feb. 20, 2024. [Online]. Available: https://arxiv.org/abs/2402.08113. doi: 10.48550/arXiv.2402.08113.

[8] J. Echterhoff, Y. Liu, A. Alessa, J. McAuley, and Z. He, "Cognitive bias in decision-making with LLMs," *arXiv*, Oct. 3, 2024. [Online]. Available: https://arxiv.org/abs/2403.00811. doi: 10.48550/arXiv.2403.00811.

[9] T. Schick, S. Udupa, and H. Schütze, "Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1408–1424, Dec. 2021. doi: 10.1162/tacl_a_00434.

[10] S. Furniturewala *et al.*, "Thinking fair and slow: On the efficacy of structured prompts for debiasing language models," *arXiv*, May 16, 2024. [Online]. Available: https://arxiv.org/abs/2405.10431. doi: 10.48550/arXiv.2405.10431.

[11] A.-V. Chisca, A.-C. Rad, and C. Lemnaru, "Prompting fairness: Learning prompts for debiasing large language models," in *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, B. R. Chakravarthi, B. B., P. Buitelaar, T. Durairaj, G. Kovács, and M. Á. García Cumbreras, Eds., St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 52–62. Accessed: Dec. 11, 2024. [Online]. Available: https://aclanthology.org/2024.ltedi-1.6.

[12] Y. H. Ke *et al.*, "Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias," 2024. doi: 10.48550/ARXIV.2401.14589.

[13] L. Wang, H. Zhong, W. Cao, and Z. Sun, "Balancing rigor and utility: Mitigating cognitive biases in large language models for multiple-choice questions," *arXiv*, Sep. 9, 2024. [Online]. Available: https://arxiv.org/abs/2406.10999. doi: 10.48550/arXiv.2406.10999.

[14] M. Bartl and S. Leavy, "From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in LLMs," 2024, *arXiv*. doi: 10.48550/ARXIV.2407.04434.

[15] S. Raza, O. Bamgbose, S. Ghuge, F. Tavakol, D. J. Reji, and S. R. Bashir, "Developing safe and responsible large language models: Can we balance bias reduction and language understanding in large language models?," Aug. 6, 2024, *arXiv*. doi: 10.48550/arXiv.2404.01399.

[16] S. Bergstrand and B. Gambäck, "Detecting and mitigating LGBTQIA+ bias in large Norwegian language models," in *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, A. Faleńska, C. Basta, M. Costa-jussà, S. Goldfarb-Tarrant, and D. Nozza, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 351–364. doi: 10.18653/v1/2024.gebnlp-1.22.

[17] S. Raza, A. Raval, and V. Chatrath, "MBIAS: Mitigating bias in large language models while retaining context," 2024. doi: 10.48550/ARXIV.2405.11290.

[18] Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. (2021), https://arxiv.org/abs/2106.09685

[19] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. & Steinhardt, J. Measuring Massive Multitask Language Understanding. *Proceedings Of The International Conference On Learning Representations (ICLR)*. (2021)

[20] Rein, D., Hou, B., Stickland, A., Petty, J., Pang, R., Dirani, J., Michael, J. & Bowman, S. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *First Conference On Language Modeling*. (2024), https://openreview.net/forum?id=Ti67584b98

[21] Chang, H. J., & Fontanarosa, P. B. (2011). Introducing the JAMA clinical challenge. *JAMA: Journal of the American Medical Association*, 305(18), 1910. DOI: 10.1001/jama.2011.625

[22] A. Sung, J. Sung, and B. Kim, "CURLoRA: Contrastive Unlearning for Robust Low-Rank Adaptation," Aug. 28, 2024, *arXiv*. doi: 10.48550/arXiv.2408.14572.

[23] Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=lq62uWRJjiY

[24] R. Koo, M. Lee, V. Raheja, J. I. Park, Z. M. Kim, and D. Kang, "Benchmarking Cognitive Biases in Large Language Models as Evaluators," *arXiv preprint arXiv:2309.17012*, 2023.

[25] Y. Zhao, Y. Hou, Z. Xu, Y. Guo, Y. Zhang, M. Zhang, and T. Chua, "CLIMB: A Comprehensive Language Investigation of Modern Bias," Jul. 11, 2024, *arXiv*. doi: 10.48550/arXiv.2407.05250.