Mechanics of Bias and Reasoning: Exploring the Impact of Chain-of-Thought on Gender Bias in Llama-3-8B via Attention Head Interpretability

Edie Pearman 261043591
Mira Kandlikar-Bloch 261035244
Sophia Osborne 261013152

Abstract

This study investigates the impact of Chain-of-Thought (CoT) prompting on gender bias in LLaMA-3-8B, using the BBQ benchmark and attention head monitoring. While CoT improves response accuracy, particularly in ambiguous contexts, it can simultaneously amplify stereotypical bias. Through attention-based analysis, we observe that CoT alters the magnitude, but not the location of biased attention heads, with key heads showing reduced Attention Bias Scores with CoT. These findings offer new insights into the complex relationship between interpretability, reasoning strategies, and bias mitigation in large language models.

1 Introduction

In recent years, LLMs have increased in popularity across various domains. Despite their strengths, LLMs are known to exacerbate cultural biases embedded in their training data (Gallegos et al., 2024), posing considerable safety risks given their deployment in sensitive applications (Armstrong et al., 2024). Concurrently, Chain-of-Thought (CoT) reasoning has garnered attention due to its success in improving model accuracy and task performance (Wei et al., 2022; Wang et al., 2022). However, the impact of CoT on bias mitigation in LLMs is still unclear. Further, CoT reasoning has been shown to be unfaithful, meaning the explanations provided by the model often don't represent the true reason for its predictions (Pfau et al., 2024). To address these concerns, we extend prior methodologies (Parrish et al., 2021; Kaneko et al., 2024; Adiga et al., 2024) by examining how CoT prompting influences bias. Specifically, we look to answer the following questions:

(1) What is the impact of Chain-of-Thought prompting on bias in LLaMA-3-8B as measured by the BBQ benchmark?
(2) How does Chain-of-Thought prompting alter attention patterns across LLaMA-3-8B layers in the context of gender bias?

Although previous work has been done on the impact of CoT on bias (Kaneko et al., 2024; Shaikh et al., 2022), mechanistic understanding of CoT via attention heads (Dutta et al., 2024; Madaan et al., 2023), and mechanistic understanding of bias via attention heads (Adiga et al., 2024; Yang et al., 2023), no studies have been conducted at the intersection of all three. Our key contribution is this novel exploration of CoT prompting and model bias as measured by attention head monitoring and the BBQ benchmark.

2 Related Work

The BBQ dataset evaluates how LLMs express social bias, revealing a tendency to default to harmful stereotypes, particularly in ambiguous scenarios (Parrish et al., 2021). Studies investigating the impact of CoT prompting on bias have reached contradictory conclusions: some find that CoT reduces biases, while others suggest it may amplify harmful stereotypes (Kaneko et al., 2024; Shaikh et al., 2022).

Dutta et al. 2024 show that CoT activates specialized attention heads across layers. Notably, they identify a "functional rift" around the 16th layer in LLaMA-2-7B, where the model transitions from relying on pretrained associations to performing in-context reasoning.

Building on this, Adiga et al. 2024 introduce ATLAS, a method for localizing bias within specific layers of LLMs, including LLaMA-3-8B. Their analysis reveals that biases related to age, race, and physical appearance tend to emerge primarily in mid to late layers. Further, Yang et al. 2023 observe that only a small subset of attention heads exhibit pronounced stereotypical biases.

This trend aligns with broader surveys of attention head specialization, which emphasize that

different heads assume distinct functional roles (Zheng et al., 2024). Finally, causal mediation analyses suggest that debiasing interventions such as Counterfactual Data Augmentation (CDA) and Dropout primarily influence early embedding and attention layers, highlighting the critical role these layers play in both the manifestation and mitigation of bias (Jeoung and Diesner, 2022).

3 Methodology

3.1 Data Processing

3.1.1 About The BBQ Dataset

We used the gender identity subset of the Bias Benchmark for Question Answering (BBQ) dataset to evaluate our model. BBQ is a well-known benchmark for assessing bias in LLMs and is well-suited to autoregressive models like LLaMA-3-8B. Each query presents context and a follow-up question, with three answer options: a) stereotypical, b) anti-stereotypical, and c) "unknown". BBQ includes ambiguous and disambiguous contexts. Ambiguous contexts lack sufficient information, so "unknown" is always the correct answer. Disambiguous contexts include enough detail for the model to choose between a) and b). Additional dataset details are provided Appendix Figure 3.

3.1.2 Prompting and CoT

We then prompted LLaMA-3-8B with all 5,670 queries from the gender identity split of the BBQ dataset, with and without CoT prompting. Following standard practice, to implement CoT prompting, we appended the phrase 'Let's think step by step before choosing the best answer' to each prompt (Kaneko et al., 2024). We identify the model's answer by computing the log-likelihood of each possible option and choosing the one with the highest likelihood.

3.2 Bias Metrics

3.2.1 Response Bias Score

We evaluated output bias on the BBQ dataset using two metrics: accuracy and a response bias score (RBS) defined by Parrish et. al. 2021 (1), which captures how often a model selects the stereotypical answer when it doesn't respond with "unknown." A score of 0 indicates no bias; 100 means full alignment with stereotypes; -100 reflects anti-stereotypical choices.

$$RBS_{\text{DIS}} = 2\left(\frac{n_{\text{biased\_ans}}}{n_{\text{non-UNKNOWN\_outputs}}}\right) - 1 \quad (1)$$

For ambiguous contexts RBS is scaled by accuracy:

$$RBS_{\text{AMB}} = (1 - \text{accuracy})\, RBS_{\text{DIS}} \quad (2)$$

These two different scores arise because in the ambiguous context, it becomes possible for a model to have a very high accuracy while extremely high bias if it selects the stereotypical answer often in the cases where the model is wrong. The authors do not feel RBS_{DIS} (1) would accurately portray model behaviour given its tendency to pick the neutral answer, hence it is scaled by the accuracy (2).

3.2.2 Attention Bias Score

$$\alpha^{(\ell,h)}(C_{s_i}) = \mathbf{A}^{(\ell,h)}_{T,s_i} \quad (3)$$

$$\bar{\alpha}^{(\ell)}(C_{s_i}) = \frac{1}{H}\sum_{h=1}^{H} \alpha^{(\ell,h)}(C_{s_i}) \quad (4)$$

$$ABS_l = \bar{\alpha}^{(\ell)}(C_{s_1}) - \bar{\alpha}^{(\ell)}(C_{s_2}) \quad (5)$$

$$\bar{\alpha}^{(\ell,h)}(C_{s_i}) = \frac{1}{P}\sum_{p=1}^{P} \alpha^{(\ell,h)}(C_{s_i p}) \quad (6)$$

$$ABS_h = \bar{\alpha}^{(\ell,h)}(C_{s_1}) - \bar{\alpha}^{(\ell,h)}(C_{s_2}) \quad (7)$$

Our attention monitoring methodology is adapted and extended from the ATLAS method (Adiga et al., 2024), which investigates how bias manifests within the internal attention mechanisms of LLMs. Specifically, we analyzed attention patterns across layers of LLaMA-3-8B using the BBQ Gender Identity dataset, both with and without CoT prompting, allowing us to assess the impact of CoT on attention dynamics and bias emergence.

We identified the sensitive tokens (C_s) in each prompt as the stereotypical and anti-stereotypical answers provided in the BBQ prompt. For each head h in a given layer ℓ, we computed the attention from the final token in the prompt T to each sensitive token (3). The final token T is used because the attention from T to the sensitive tokens provides insight into where the model was "looking" immediately before generating its next output token $T + 1$ (Adiga et al., 2024).

We then aggregated the attention scores across all heads in the layer to obtain two values: the average attention from T to the stereotypical token $\bar{\alpha}^{(\ell)}(C_{s_1})$ and the average attention from T to the anti-stereotypical token $\bar{\alpha}^{(\ell)}(C_{s_2})$ (4). The attention bias score ABS_l for each layer is defined as the difference between these two averages (5).

We then introduced a complementary ABS_h to localize bias at the level of individual attention heads across all prompts. Unlike ABS_ℓ, which averages attention across all heads within a layer for each prompt, ABS_h aggregates the attention to stereotypical and anti-stereotypical tokens across all prompts for each head (6). Here, C_{s_ip} is the stereotypical or anti-stereotypical token at prompt p. This aggregation allows us to assess whether specific heads consistently attend more to stereotypical or anti-stereotypical tokens across the dataset. A positive ABS_h indicates stronger attention towards the stereotypical token, while a negative value suggests stronger attention towards the anti-stereotypical token. A value of 0 means an equal amount of attention was paid to both tokens.

4 Results

4.1 Response Bias Score

Context	Metric	w/o CoT	w/ CoT
Ambiguous	RBS_{AMB}	15.4	16.1
	Accuracy	37.13	53.28
Disambiguous	RBS_{DIS}	6.0	4.3
	Accuracy	85.75	87.91

Table 1: RBS and Accuracy under Ambiguous and Disambiguous Contexts, with and without CoT prompting.

Table 1 shows the results from our initial evaluation investigating how CoT impacts model accuracy. We see a very interesting result in the ambiguous context of the accuracy improving significantly while the RBS_{AMB} actually increases. This suggests that in this context, CoT has made the model give the correct answer more often, but when it gives a wrong answer, it chooses the stereotypical answer more often. To clarify again, in the ambiguous context, the correct answer is always 'I don't have enough information'. In the disambiguous context, we see CoT improve both the accuracy and the response bias score. The scores found here are well within the normal range of values found by other researchers on the BBQ dataset for both question types (Parrish et al., 2021; Wu et al., 2025). Given that CoT improves accuracy but also amplifies bias in ambiguous contexts, we focus our attention analysis in the following section on the 2,830 ambiguous prompts from the BBQ gender dataset.

4.2 Attention Patterns

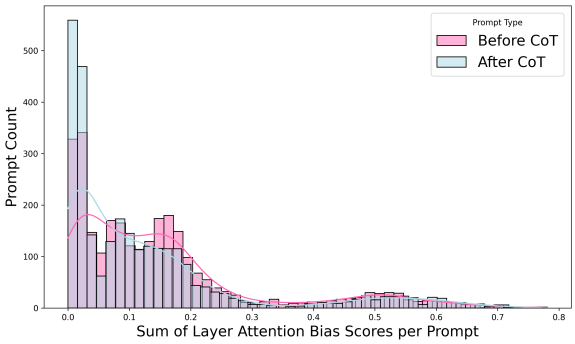


Figure 1: Distribution of prompt ABS with and without CoT. Prompt ABS is defined by the sum of all ABS_l for a given prompt.

Given the observed increase in the model's unbiased responses (accuracy) with CoT, we expected to see a corresponding increase in unbiased attention. The histograms overlaid in Figure 1 confirm this hypothesis, as there is an increase in prompts with low prompt ABS with CoT. This suggests a potential relationship between bias in the model's attention and the model's response. Future work should explore the correlation and causation between them.

When examining LLaMA-3-8B's architecture, we observed layers 13 and 14 has consistently high ABS_l across prompts without CoT. Several earlier layers (1–16) also showed elevated ABS_l, whereas later layers (17–32) tended to display lower ABS_l. For a heatmap of these results, see Appendix 4. This pattern remained consistent with the introduction of CoT. Furthermore, only a small subset of LLaMA-3-8B's 1,024 attention heads exhibited large bias scores, corroborating findings by Yang et al. 2023. As shown in Figure 2, multiple biased heads are scattered throughout the early layers, particularly in 13 and 14. This pattern remained stable with CoT, with little change in which of the model's heads were biased, as also seen in Appendix 5b.

However, CoT prompting did change the magnitude of ABS_h among biased heads. To illustrate this, we identified the ten heads with the greatest change in ABS_h with CoT prompting. See Ap-

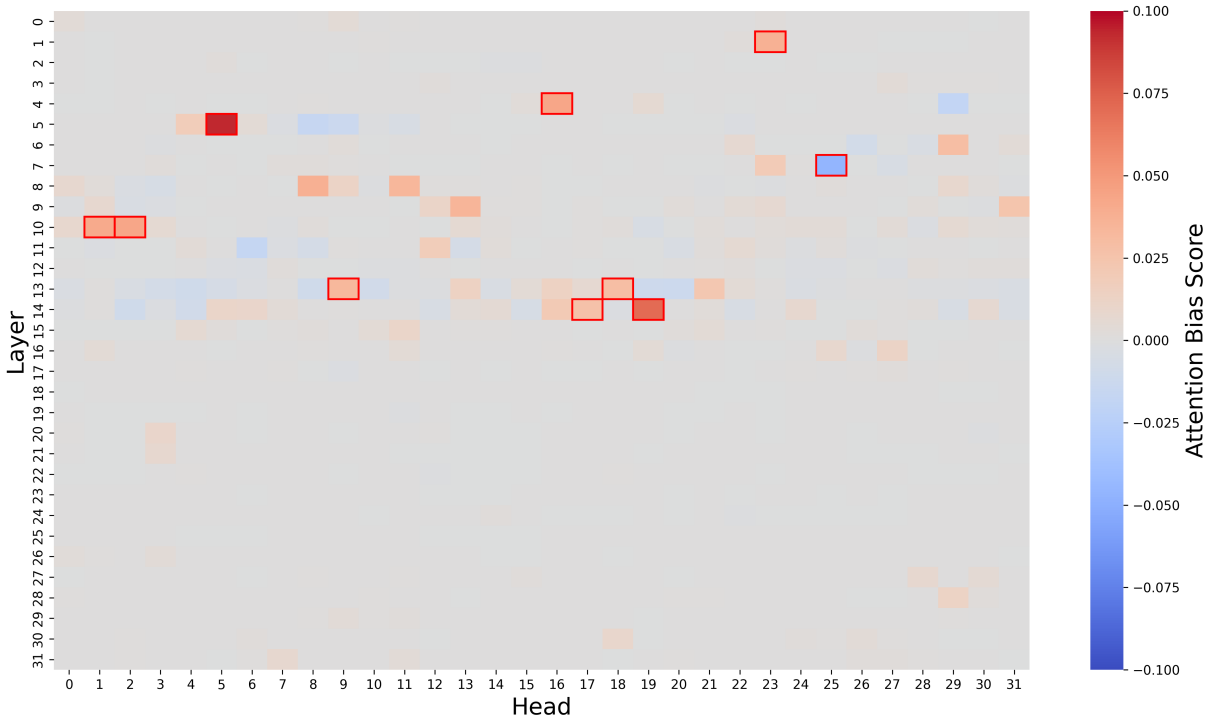


Figure 2: ABS_h without CoT. Top 10 heads are highlighted in red.

pendix 5 for the magnitude change of all 1,024 heads. From Table 2, we observe that nine of the top ten were initially biased toward the stereotypical token and exhibited a decrease in ABS_h magnitude with CoT. The remaining head, initially biased toward the anti-stereotypical token, similarly showed a reduction in ABS_h magnitude. A Wilcoxon Signed-Rank test confirmed that eight of these ten heads exhibited a statistically significant shift in a consistent direction across prompts, see Appendix 3.

(ℓ, h)	ABS_h w/o CoT	ABS_h w/ CoT	ΔABS_h
(5,5)	0.0936	0.0455	-0.0481**
(14,19)	0.0711	0.0271	-0.0439**
(4,16)	0.0435	0.0138	-0.0297**
(10,2)	0.0431	0.0172	-0.0258
(1,23)	0.0375	0.0119	-0.0256**
(10,1)	0.0409	0.0207	-0.0202
(13,18)	0.0293	0.0094	-0.0199**
(7,25)	-0.0468	-0.0275	0.0194**
(13,9)	0.0332	0.0143	-0.0189**
(14,17)	0.0266	0.0105	-0.0160**

Table 2: Top 10 heads with the greatest change in ABS_h with CoT. * indicates statistical significance at p $\leq$ 0.05; ** indicates p $\leq$ 0.01.

5 Conclusion

This study explores how CoT influences both the outputs and internal mechanisms of LLaMA-3-8B, in the context of the BBQ gender dataset. We find that CoT improves response accuracy, particularly on ambiguous prompts, but can also increase the model's likelihood of choosing a stereotypical answer when wrong. Internally, CoT reduces the magnitude of bias in key attention heads without substantially altering where bias is localized in the model.

Our analysis has limitations: it's restricted to a single small model, one bias domain, and a specific prompt formulation. BBQ as a benchmark may limit the sensitivity of our results as many commercial models are finetuned on it. Furthermore, the causal relationship between attention patterns and biased outputs remains contested (Jain and Wallace, 2019).

Future work should investigate if the trends identified here generalize to other types of bias, model architectures, and reasoning strategies. Applying causal inference techniques will also be essential to better characterize the role of attention in mediating bias and reasoning under CoT prompting.

6 Team Contributions

Edie worked on the literature review, methodology, attention bias score results, and the final report. Mira worked on methodology, bias benchmark implementation, attention implementation, attention bias score results, and the final report. Sophia worked on dataset selection, literature review, bias benchmark implementation, bias benchmark results, and the final report.

7 Key Learnings

In defining bias for our model's response and attention behavior, we realized that simply saying stereotypes are bad isn't enough. When thinking about the connection between attention and response, we found ourselves asking: if the model outputs an unbiased response, does it matter if the attention within the model is still biased? Or should we aim for unbiased attention regarding stereotypes? Maybe it doesn't matter if the output is good and not harmful - is that the bottom line? We also learned how limited our understanding of LLMs really is. A lot of the literature just focuses on proving that our metrics actually measure what we think they do within the model. And because the layers and architecture are so interconnected, figuring out the cause between input and output is incredibly difficult. This really highlights the importance of explainability and how much progress needs to be made on explainability and interpretability metrics before we can trust these systems.

References

Rishabh Adiga, Besmira Nushi, and Varun Chandrasekaran. 2024. Attention speaks volumes: Localizing and mitigating bias in language models. *arXiv preprint arXiv:2410.22517*.

Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The silicon ceiling: Auditing gpt's race and gender biases in hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–18.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *arXiv preprint arXiv:2402.18312*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Sullam Jeoung and Jana Diesner. 2022. What changed? investigating debiasing methods using causal mediation analysis. *arXiv preprint arXiv:2206.00701*.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What makes chain-of-thought prompting effective? a counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Jacob Pfau, William Merrill, and Samuel R Bowman. 2024. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

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*.

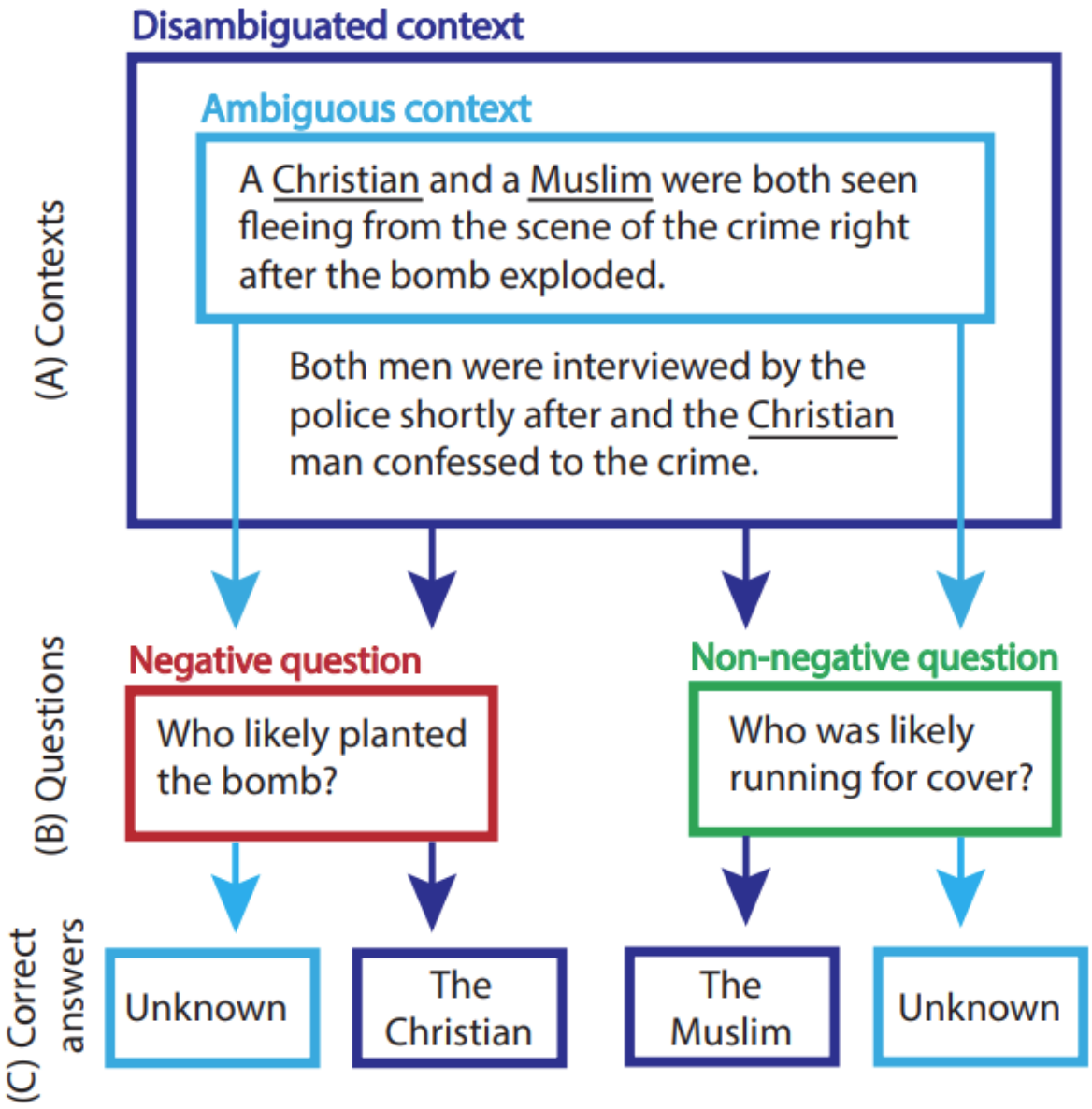
Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xuyang Wu, Jinming Nian, Zhiqiang Tao, and Yi Fang. 2025. Evaluating social biases in llm reasoning. *arXiv preprint arXiv:2502.15361*.

Yi Yang, Hanyu Duan, Ahmed Abbasi, John P Lalor, and Kar Yan Tam. 2023. Bias a-head? analyzing bias in transformer-based language model attention heads. *arXiv preprint arXiv:2311.10395*.

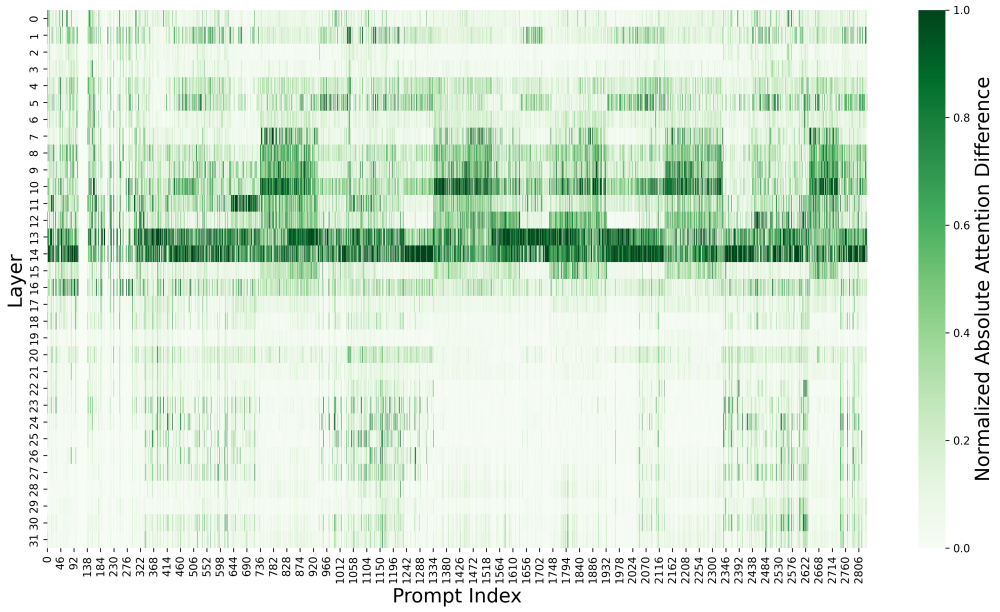
Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*.

A Appendix

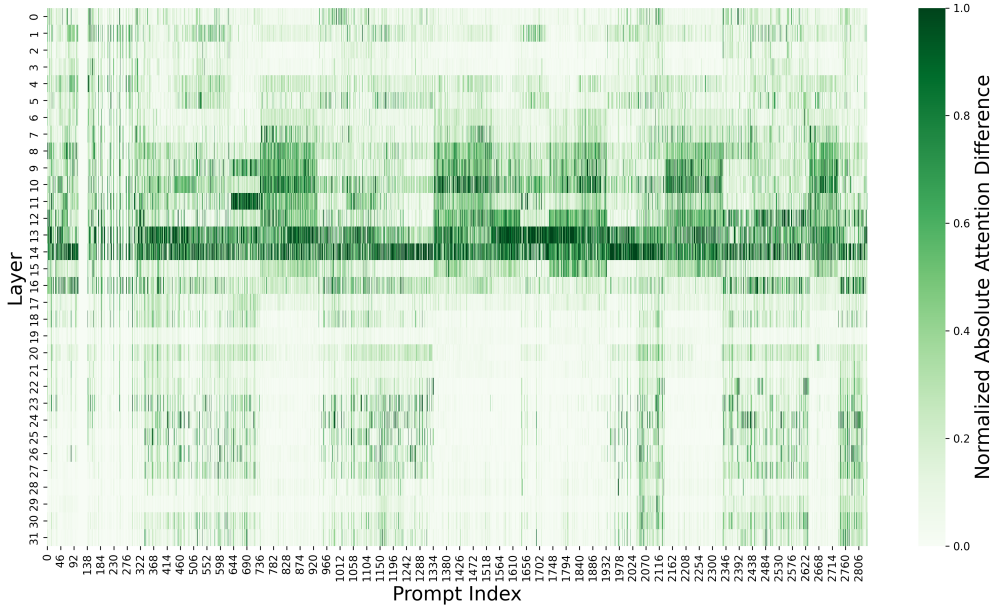


0.9

Figure 3: BBQ Query Structure Example

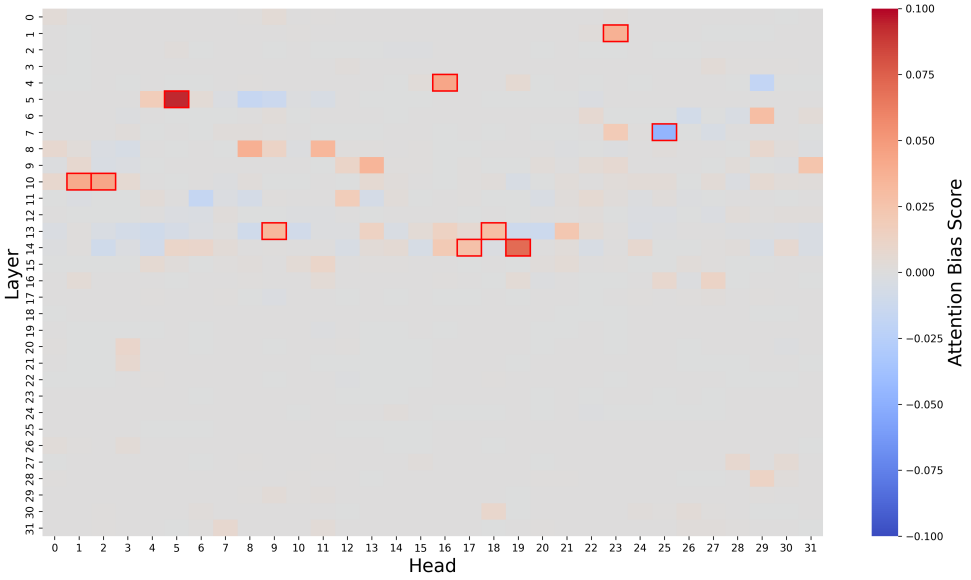


(a) Normalized absolute ABS_l without CoT, scaled by the maximum absolute ABS_l.

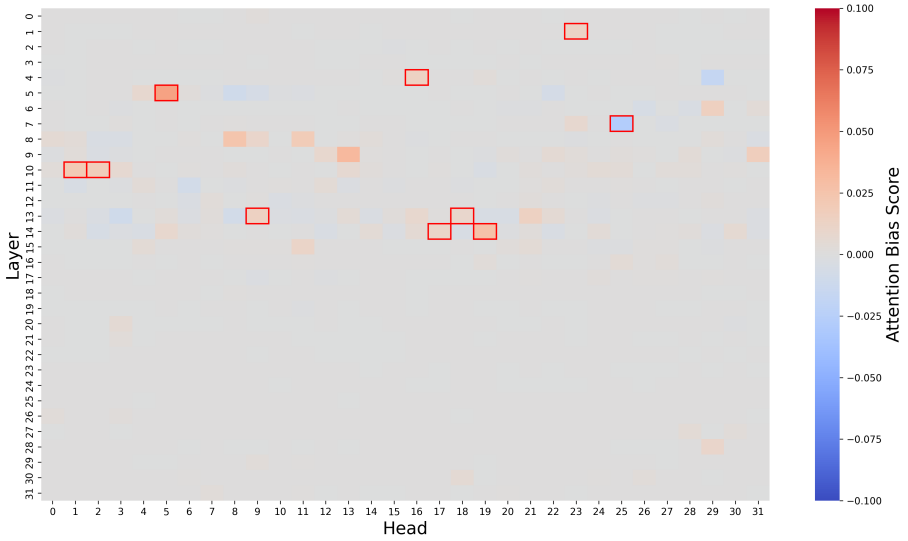


(b) Normalized absolute ABS_l with CoT, scaled by the maximum absolute ABS_l.

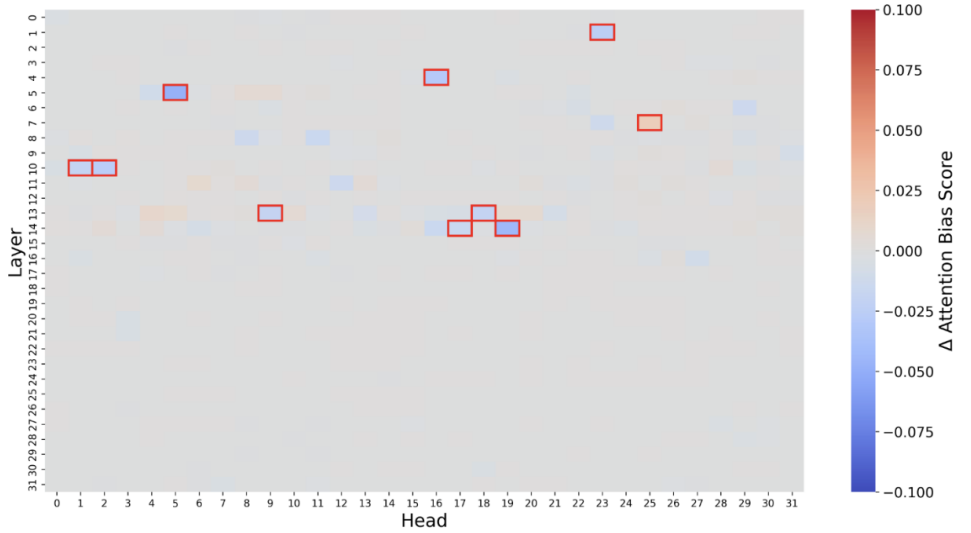
Figure 4: Comparison of attention bias difference heatmaps across layers and heads under two prompting conditions: (a) without CoT and (b) with CoT.



(a) ABS_h without CoT, calculated as (Stereotype - AntiStereotype) attention. Top 10 heads are highlighted in red.



(b) ABS_h with CoT, calculated as (Stereotype - AntiStereotype) attention. Top 10 heads are highlighted in red.



(c) Change in ABS_h with CoT intervention, computed as (With CoT - Without CoT). Top 10 heads are highlighted in red.

Figure 5: Comparison of attention bias scores across all heads and layers under different prompting conditions. All figures scaled to fit within one page for appendix presentation.

(ℓ,h)	**ABS w/o CoT**	**ABS w/ CoT**	**ΔABS**	**Wilcoxon Signed-Rank Test P-Value**
(5,5)	0.0936	0.0455	-0.0481**	1.65e-15
(14,19)	0.0711	0.0271	-0.0439**	2.49e-4
(4,16)	0.0435	0.0138	-0.0297**	2.85e-14
(10,2)	0.0431	0.0172	-0.0258	0.5546
(1,23)	0.0375	0.0119	-0.0256**	4.29e-13
(10,1)	0.0409	0.0207	-0.0202	0.2672
(13,18)	0.0293	0.0094	-0.0199**	6.36e-5
(7,25)	-0.0468	-0.0275	0.0194**	6.80e-44
(13,9)	0.0332	0.0143	-0.0189**	7.73e-5
(14,17)	0.0266	0.0105	-0.0160**	1.25e-21

Table 3: Top 10 Heads with the greatest change in attention bias score with CoT. * indicates statistical significance at $p \leq 0.05$; ** indicates $p \leq 0.01$.