Mechanics of Bias and Reasoning: Interpreting the Impact of Chain-of-Thought Prompting on Gender Bias in LLMs

Edie Pearman[1,2] EDIE.PEARMAN@MAIL.MCGILL.CA
Sophia Osborne[1,2] SOPHIA.OSBORNE@MAIL.MCGILL.CA
Mira Kandlikar-Bloch[1,2] MIRA.KANDLIKAR-BLOCH@MAIL.MCGILL.CA
Mina Arzaghi[1,3] MINA.ARZAGHI@MILA.QUEBEC
Florian Carichon[1,2] FLORIAN.CARICHON@MILA.QUEBEC
Golnoosh Farnadi[1,2] FARNADIG@MILA.QUEBEC
[1] *MilaQuebec AI Institute, Montréal, Canada*
[2] *McGill University, Montréal, Canada*
[3] *HEC Montréal, Montréal, Canada*

Abstract

Large language models (LLMs) are increasingly deployed in socially sensitive settings despite substantial documentation that they encode gender biases. Chain-of-Thought (CoT) prompting has been proposed as a bias-mitigation approach. However, existing evaluations primarily focus on changes in LLM benchmark performance, providing limited insight into whether apparent bias reductions reflect meaningful changes in a model's internal mechanisms. In this work, we investigate how CoT prompting affects gender bias in LLMs, combining benchmark-based evaluation with mechanistic interpretability techniques and reasoning chain failure analysis. Our results confirm a stereotypical bias present in LLM outputs across benchmarks, showing that CoT prompting does not consistently reduce the bias gap. Mechanistic analyses reveal that although CoT balances biased behavior in certain attention head clusters, gender bias remains embedded in hidden representations, indicating only superficial mitigation. Inspection of reasoning chains further suggests that these improvements frequently co-occur with bias statements, and may reflect dataset familiarity rather than genuine bias mitigation.

1. Introduction

In recent years, large language models (LLMs) have been widely adopted across diverse domains. Despite their impressive capabilities, LLMs have been shown to exacerbate gender bias (Gallegos et al., 2024), raising significant safety concerns, particularly given their deployment in sensitive applications (Armstrong et al., 2024). At the same time, recent approaches based on Chain-of-Thought (CoT) (Wei et al., 2022) as a reasoning-enhancing strategy have enabled models to improve performance on a wide range of tasks that require step-by-step analysis (Srivastava et al., 2023; Suzgun et al., 2023). A popular approach to improve reasoning is zero-shot CoT generation, which consists of prompting models with some variation of "Let's think step by step" (Kojima et al., 2022).

Recently, prompt-based bias mitigation has been explored due to its accessibility. Approaches such as zero-shot self-debiasing (Gallegos et al., 2025) and bias suppression (Oba et al., 2024) show that carefully designed prompts can reduce stereotyping without modifying model parameters. However, recent critiques also question the effectiveness of prompt-based

© 2026 E. Pearman, S. Osborne, M. Kandlikar-Bloch, M. Arzaghi, F. Carichon & G. Farnadi.

bias mitigation. Yang et al. (2025) show that prompt-based approaches often rely on superficial alignment, encourage evasive or non-committal responses, and exploit flawed bias metrics, suggesting improvements stem from benchmark compliance rather than genuine bias reduction (Sivakumar et al., 2025). Therefore, the effectiveness of CoT as a gender bias mitigation strategy remains open, necessitating a deeper understanding of dataset-specific effects, internal behavioral mechanisms, and the quality and nature of model reasoning. Recent work has addressed this question by examining how gender bias manifests in LLM representations and how CoT reasoning operates mechanistically (Dutta et al., 2024; Tan and Celis, 2019). However, it has yet to explain how CoT functionally impacts a model's representation and perpetuation of gender bias. Similarly, recent work has examined reasoning traces for correctness (Amirizaniani et al., 2024) and types of harm (Shaikh et al., 2022), but has yet to examine how reasoning chains mitigate, or fail to mitigate, gender bias. We present the first systematic study to bridge these gaps, combining benchmark evaluation, mechanistic interpretability techniques, and reasoning failure analysis to understand how CoT prompting functionally impacts gender bias in LLMs. Our contributions are the following: **(1)** We systematically evaluate the effectiveness of CoT prompting for gender bias mitigation across four MCQA benchmarks (BBQ, CrowS-Pairs, StereoSet, and SocioEconomicQA) and five LLMs in section 4. **(2)** We adapt attention and hidden state mechanistic interpretability techniques to examine how CoT prompting influences internal processing of gender bias in section 5. **(3)** We develop and apply a taxonomy of reasoning chain behaviors to analyze reasoning failures across models and datasets in section 6.

2. Related Works

Gender bias mitigation is a critical concern in NLP because language models trained on human-generated text inevitably reflect societal stereotypes and biases (Blodgett et al., 2020; Gallegos et al., 2024). As noted by Stanczak and Augenstein (2021), such biases result in representational harms that reinforce social inequalities. Recent surveys focusing on LLMs confirm that gender remains the most studied social attribute in bias evaluation, providing established datasets and evaluation protocols while also highlighting issues in evaluation practices (Blodgett et al., 2020; Gallegos et al., 2024). Bias mitigation techniques still primarily operate through data augmentation, representation debiasing, or output-level constraints, often relying on task-specific resources or handcrafted gender lexicons (Sun et al., 2019). Zero-shot prompt-based mitigation methods are a flexible alternative as they can reduce bias through carefully designed prompts without modifying model parameters (Gallegos et al., 2025; Oba et al., 2024). However, prompt-based mitigation effectiveness is often overestimated (Yang et al., 2025) as such approaches emphasize output-level metrics and may reduce measured bias without addressing how models encode and propagate biased associations (Blodgett et al., 2020). CoT prompting specifically has been shown to both reduce (Kaneko et al., 2024; Mohapatra et al., 2024) and amplify (Shaikh et al., 2022) social bias across various prediction tasks. Our paper advances this debate by measuring CoT's effect across four gender-bias benchmarks for five LLMs, and then leveraging mechanistic interpretability to examine whether such prompt-based bias mitigation produces meaningful change in the model's internal processes.

Mechanistic interpretability aims to uncover and communicate the internal mechanisms underlying model behavior in a selective, human-understandable manner (Madsen et al., 2022). Attention head analysis interprets how models attend to different parts of the input (Vig and Belinkov, 2019; Zheng et al., 2024b). Work by Kaneko and Bollegala (2021) and Adiga et al. (2024) show that biased attention is focused in mid-to-late layers, with Yang et al. (2023) observing that only a small subset of attention heads exhibit pronounced stereotypical behavior. Dutta et al. (2024) have demonstrated that CoT similarly activates specialized attention heads, with evidence of a transition from pretrained associations to in-context reasoning. Another line of mechanistic interpretability work uses probes to extract specific properties from models' internal layer representations (Belinkov, 2022). When applied to bias detection, probing reveals that social biases are most detectable in middle layers and persist internally even for unbiased outputs (Tan and Celis, 2019; Vig et al., 2020). Recent work has also shown that reasoning models encode answer correctness in their hidden states during CoT (Zhang et al., 2025). We adapt attention head monitoring and hidden state probing methodologies to the context of gender bias, and develop a novel qualitative taxonomy of reasoning failure behaviors. Together, these analyses offer insight into how reasoning succeeds in reducing bias and how it fails to.

3. Experimental Set Up

We use a pipeline, depicted in Figure 1, that iteratively prompts the model to extract predicted answers, reasoning chains, and internal representations.

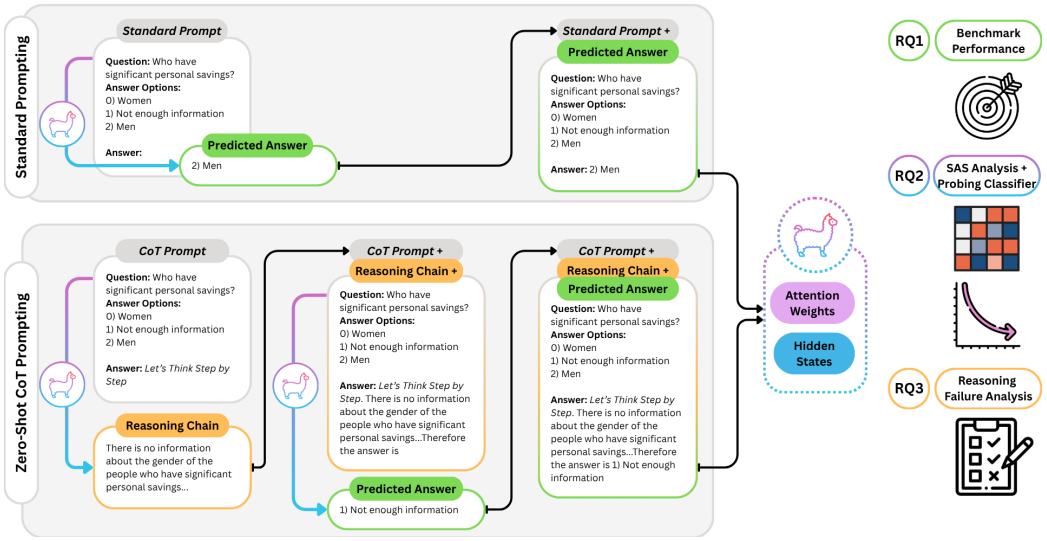


Figure 1: Our pipeline for extracting model outputs and internal mechanisms.

We test on four English-language multiple-choice QA datasets to benchmark social bias in large language models: BBQ (Parrish et al., 2021), StereoSet (Nadeem et al., 2021), CrowS-Pairs (Nangia et al., 2020), which are commonly used, and SocioEconomicQA (Arzaghi et al., 2024), which captures intersectional socioeconomic bias. Each dataset is formatted as a multiple-choice QA task with three answer options: a stereotypical, an anti-stereotypical,

and a paraphrased unknown (e.g., *"Cannot be determined"*) (Parrish et al., 2021). The QA task context is designed to be ambiguous, so the unknown option is always correct.

We use five open-source LLMs spanning sizes 7B to 32B parameters, families including Llama, Qwen, and Mistral, and one reasoning-specialized model. Specifically, we evaluate Qwen2.5-7B-Instruct (Yang et al., 2024), Qwen2.5-32B-Instruct (Yang et al., 2024), QwQ (Team, 2025), Llama3-8B-Instruct (Grattafiori et al., 2024), and Mistral-7B (Jiang et al., 2023). For each model, predicted answers are extracted via log-likelihood, following Adiga et al. (2024). Given a sample prompt P_i from dataset D, which is expected to produce a single answer identifier when input to an LLM, and a set of answer identifier tokens $y \in \{0, 1, 2\}$, a model's predicted answer is defined as: $\hat{y} = \arg\max_y \log P(y \mid P_i)$ where $P(y|P_i)$ is the probability of y being the next token given the input prompt P_i. We compare model performance over two conditions: *Standard* prompting, as described above, and a zero-shot *CoT* prompting. For the latter, we follow Kojima et al. (2022). Specifically, we append *"Let's think step by step."* to the input prompt, generate a reasoning chain, and append it to the prompt as well before extracting the answer. Full formatting and implementation details, including sample prompts, are provided in Appendices A.3. Our code is available in a public GitHub repository for reproducibility[1].

4. How does CoT prompting affect model accuracy and gender bias?

We report LLM performance in Table 1 using two metrics. **Accuracy** is the proportion of predictions for the unknown option (abstention). **Diff-Bias** (Parrish et al., 2021) measures model's preference for stereotypes over anti-stereotypes as their difference (range $[-1, 1]$; see Appendix A.3.3), making it sensitive to high abstention rates. We retain it not as a primary bias measure, but to examine directional preference in the subset of non-abstaining responses. We observe varied CoT accuracy across models. CoT improved abstention rates for Llama-8B and Mistral-7B, but not for the Qwen2.5 family (7B, 32B). Scaling from Qwen-7B to Qwen-32B did not consistently improve abstention, because Qwen-7B already achieves relatively high accuracy. However, scaling increases stereotype bias as shown by Diff-bias under both prompting conditions. For QwQ, a reasoning-trained model comparable to Qwen-32B, CoT consistently decreases abstention. CoT's effect is also dataset dependent. Across models, StereoSet and CrowS-Pairs exhibit lower abstention rates and greater stereotype bias with CoT than BBQ and SocioEconomicQA. Together, these mixed results demonstrate that CoT prompting is not a consistent bias mitigation strategy, contrary to Kaneko et al. (2024). We turn to mechanistic interpretability to understand why, by investigating how CoT shapes the way models attend to and represent gender bias information.

5. Do internal model mechanisms reveal CoT's influence on gender bias?

Stereotype Attention. We quantify stereotype attention using the Stereotype Attention Score (SAS), adapted from Yu et al. (2025)'s Negative Attention Score, which originally measured the difference in attention between positive and negative tokens (yes/no). We adapt it for bias by contrasting attention to stereotypical versus anti-stereotypical tokens;

1. The repository can be found at https://github.com/edieblue/MechanicsBiasCoTReasoning, and also includes all SAS heatmaps for models and datasets not highlighted in our results.

Model	Method	BBQ Ambig		SocioEconomicQA		StereoSet		CrowS-Pairs	
		%UNK↑	Diff-Bias↓	%UNK↑	Diff-Bias↓	%UNK↑	Diff-Bias↓	%UNK↑	Diff-Bias↓
Llama-8B	Standard	43.19	0.235	22.13	0.240	32.16	**0.150**	45.80	**0.038**
	CoT	**76.06**	**0.057**	**31.25**	**0.210**	**40.39**	0.200	**48.47**	0.073
Mistral-7B	Standard	64.49	0.140	64.72	0.210	59.61	**0.180**	69.08	0.042
	CoT	**94.75**	**-0.002**	**79.49**	**0.110**	59.61	0.190	**83.21**	**0.038**
Qwen-7B	Standard	**98.48**	0.118	**95.32**	0.645	**81.57**	0.660	70.61	**0.013**
	CoT	97.32	**0.104**	92.59	**0.338**	74.51	**0.415**	**74.05**	0.029
Qwen-32B	Standard	99.89	-1.000	**97.13**	**0.707**	**78.43**	**0.745**	**91.98**	0.429
	CoT	**99.93**	-1.000	92.59	0.714	69.02	0.747	90.08	**0.307**
QwQ	Standard	97.18	**0.128**	**85.14**	**0.690**	63.92	**0.348**	**78.24**	**0.112**
	CoT	**98.84**	0.276	68.10	0.698	**63.92**	0.696	76.72	0.311

Table 1: Results reporting only uncertainty rate (%UNK) and Diff-Bias across datasets, with and without Chain-of-Thought (CoT).

the full formula is provided in Appendix A.4.1. SAS is computed per attention head as a weighted log-ratio of attention directed toward the stereotypical versus anti-stereotypical tokens, summed across all token positions in a given prompt. A positive SAS indicates greater relative attention to the stereotypical token (red); a negative SAS indicates the reverse (blue); and values near zero reflects balanced or negligible attention (grey). We report the single-head SAS averaged across prompts grouped by the model's predicted answer type for Qwen-7B, Qwen-32B, and QwQ on StereoSet (see Appendix A.4.2 for more results).

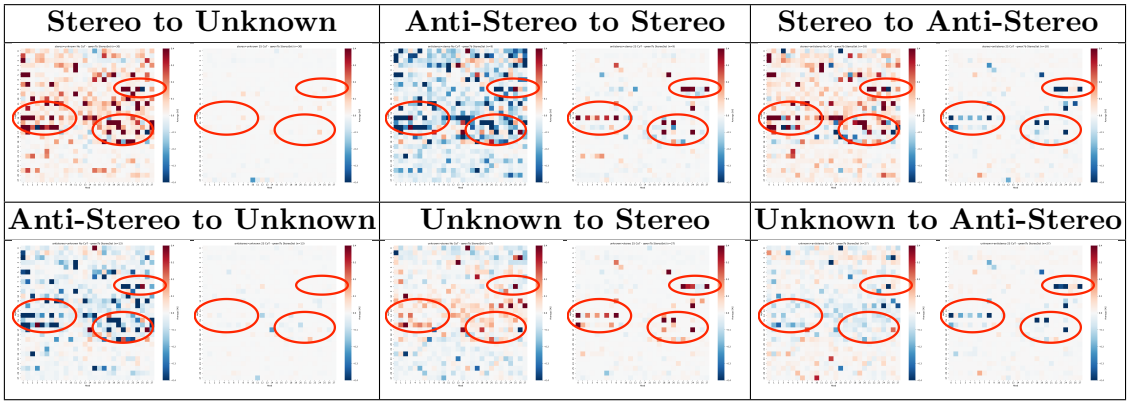


Figure 2: Single-Head SAS Score for Qwen7B over StereoSet.

The results in Figures 2, 3, and Appendix Figure 7 show greater absolute SAS for biased answer predictions, with attentional polarity matching the prediction: when the predicted answer is stereotypical, attention is directed toward stereotypical tokens; when anti-stereotypical, toward anti-stereotypical tokens. Unknown predictions show little to no polarity. This pattern is consistent across all five models and both prompting conditions. Beyond this global trend, we observe the emergence of attention-head clusters where the largest changes in SAS concentrate, consistent with Yang et al. (2023). In Qwen2.5-7B, we identify three such clusters in Figure 2: around layer (L) 8, heads (H) 22–27; L 14–16,

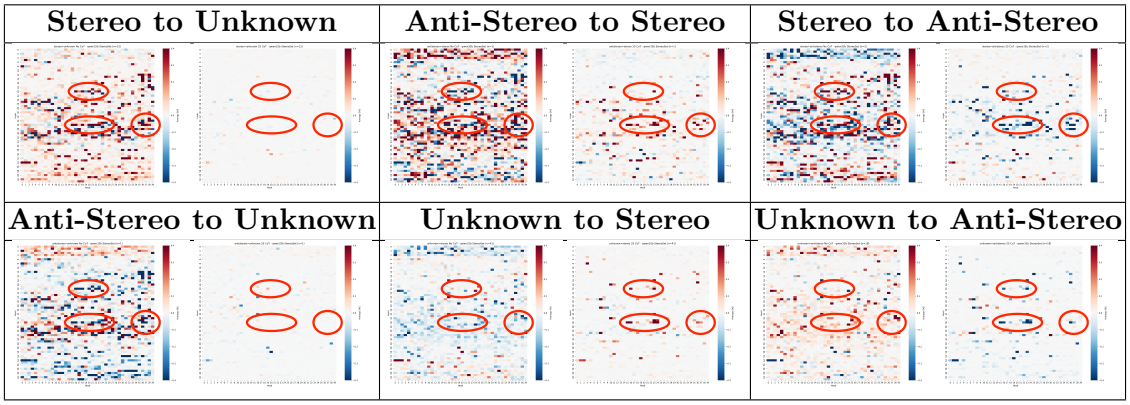


Figure 3: Single-Head SAS Score for Qwen32B over StereoSet.

H 0–9; and L 15-17, H 18–24. Similarly, in Qwen-32B and QwQ, which share the same architecture, we observe comparable clusters in Figures 3 and 7 respectively around L 20, H 16–23; L 36, H 15–24; and L 35–38, H 35–37. These clusters appear to play a decisive role in answer selection. When answers shift from unknown to biased, these clusters show a sharp increase in absolute SAS. When answers shift between stereotypical and anti-stereotypical, SAS instead reverses sign, reflecting a redistribution of attention between the corresponding token groups. Overall, CoT prompting reduces absolute SAS at the global level, but fails to consistently suppress the activity or polarity of these biased heads, whose role appears central to answer selection.

Hidden State Probing. We complement our attention analysis with probing experiments to assess whether a model's internal representations encode gender bias. Following Zhang et al. (2025), we train 2-layer MLP probes to predict the LLM's selected answer type (stereotype, anti-stereotype, or unknown) given hidden states extracted from tokens in the reasoning chain. For each model–dataset pair under Standard and CoT prompting, we extract hidden states from four layers to manage computational cost: two with high attention activity (by SAS), one with low activity, and one at random—yielding eight probes. We evaluate the probes using **Probe Fidelity**, which measures agreement between probe predictions and LLM answer selection rather than the dataset ground-truth labels. We also examined probe accuracy against dataset ground-truth labels, with detailed performance metrics and comprehensive probe implementations reported in Appendix A.5.

Figure 4 presents our probing results. Despite near-perfect fidelity, probe performance for Qwen-32B on BBQ and SocioEconomicQA was deemed unreliable and excluded from analysis due to low F1 scores. Qwen-32B's high benchmark accuracy created severe class imbalance, preventing the probes from meaningfully learning the minority classes. The majority of the remaining probes exhibit high fidelity. This indicates hidden state representations encode enough gender bias information to distinguish stereotypical, anti-stereotypical, and unknown answer selection. However, these results do not establish a causal link between representations and output behavior. The small portion of probes with low fidelity occurred exclusively in the earliest layers for all models. This is consistent with prior work showing early layers have not yet developed complex semantic abstractions (Skean et al., 2025).

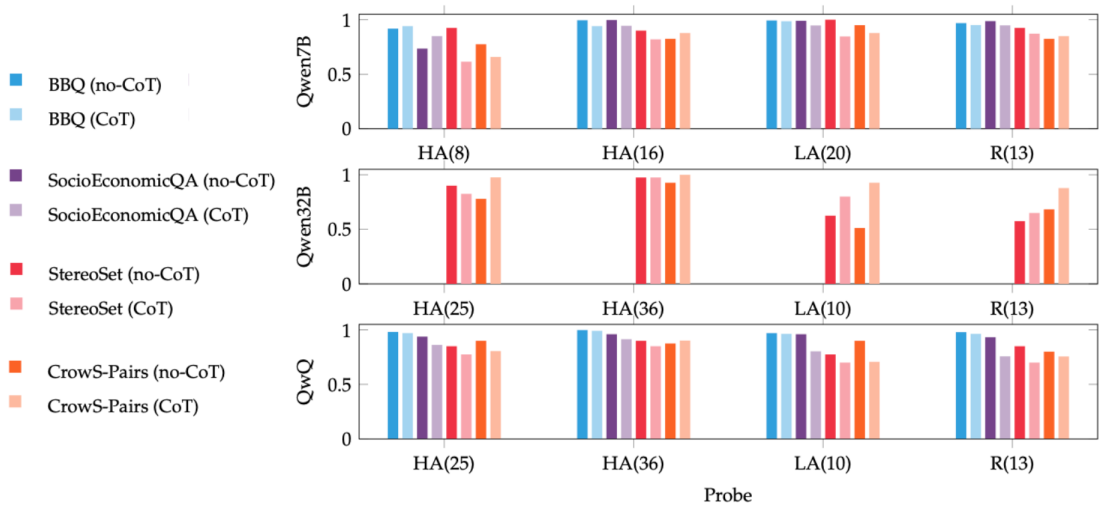


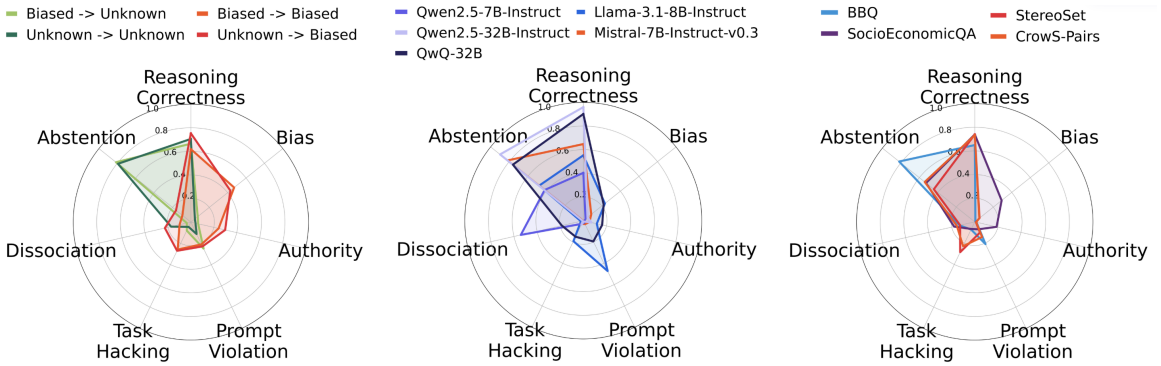
Figure 4: Fidelity accuracy across probes at layer (L). HA denotes a layer with high SAS activity. LA denotes a layer low SAS activity. R denotes a layer selected at random.

Taken together, our mechanistic interpretability experiments show no relationship between stereotype attention activity and probe fidelity. High fidelity is found in layers with both high and low attention activity across all models and datasets, regardless of CoT. Attention analysis found clusters of heads at specific layers associated with biased behavior. However, probe performance suggests that biased information is present throughout the residual stream, even in layers with no biased attention. This signifies that CoT prompting has a shallow effect on a model's behavioral mechanisms, with no evidence that it changes internal modeling of gender bias. Nonetheless, we observed isolated improvements in model performance in Section 4; we therefore shift our analysis to the reasoning chains themselves to better understand when CoT succeeds or fails to reduce bias.

6. RQ3. How do reasoning chains explain how CoT influences gender bias mitigation?

To analyze CoT reasoning behaviors, three authors of the paper independently performed a qualitative annotation and analysis of a sample of 257 reasoning chains for seven reasoning behaviors: *Reasoning Correctness*, *Abstention*, *Dissociation*, *Task Hacking*, *Prompt Violation*, *Authority*, and *Bias*. Using these annotations, we fine-tuned a DeBerta-base binary classifier (He et al., 2021) for each label. We applied the seven classifiers to the remaining 27,308 reasoning chains across all five models, four datasets, and answer types. Label definitions, annotator criteria, and classifier implementation details are in Appendix A.6.

Figure 3(a) shows how reasoning behaviors relate to the type of predicted answer. *Reasoning correctness*—logical coherence in the reasoning chain—is consistent across answer types. Notably, we observe logically correct reasoning chains that lead to a biased answer or include biased statements, consistent with Shaikh et al. (2022). This indicates that *correctness* alone does not determine whether a model produces an unknown or a biased answer. *Abstention* and *bias* constitute distinct CoT reasoning behaviors and must be



(a) Change in Answer Type (b) Model (c) Dataset

Figure 5: Proportion of reasoning chains classified as exhibiting each of the seven behaviors, grouped by CoT's effect on answer type, model, and dataset.

addressed separately. *Abstention reasoning* occurs when the model acknowledges insufficient information and indicates a preference for the unknown option. Unknown predictions are frequently accompanied by *abstention reasoning*, but the reverse is not true. *Abstention reasoning* can still co-occur with *biased reasoning* or result in a biased answer prediction at the output level. *Abstention reasoning* is therefore important, but insufficient for bias mitigation. *Biased reasoning*, by contrast, is not necessary to produce a biased answer. Biased answers co-occur with *dissociation* (choosing answers at random or in conflict with the reasoning chain), *task hacking* (exploiting prompt grammar), *prompt violation* (contradicting or replacing the provided prompt), and appeals to *authority* (leveraging references to valid or invalid external sources). Effective bias mitigation must therefore address, in addition to *biased reasoning*, the range of reasoning behaviors that surround biased answers.

As shown in Figure 3(b), the three smaller 7-8B models exhibited lower rates of *reasoning correctness* compared to the larger 32B models. Within the Qwen family, Qwen-7B demonstrated a disproportionate rate of *dissociation*, frequently producing reasoning chains that were non-existent or failed to reach a logical conclusion. This suggests model size positively impacts reasoning quality, confirming findings from Kojima et al. (2022). However, Mistral-7B achieves rates comparable to Qwen-32B for *abstention*. This indicates that while scaling increases *reasoning correctness*, it may not be inherently necessary for abstention. Reasoning training did not positively impact reasoning behavior. QwQ exhibits lower rates of *abstention* and higher rates of behaviors associated with biased answers compared to Qwen-32B. After manually analyzing the samples, we attribute this to an "overthinking" effect, in which the model generates reasoning beyond what is necessary, introducing errors in the process (see Appendix 9).

Figure 3(c) shows variation in reasoning behaviors across datasets. *Task hacking* was most prevalent for CrowS-Pairs and StereoSet, where models exploit the phrasing of the question 'Which of these options is most accurate?' rather than engage meaningfully with the prompt material. Moreover, StereoSet prompts contain grammatical errors (Blodgett et al., 2021) that models used to prematurely disqualify answer options, see Appendix 9 for all qualitative samples. Patterns in *abstention*, *authority*, and *bias* reasoning behavior collectively suggest data contamination across BBQ which had elevated abstention answer

rates relative to the other reformatted datasets. *Bias reasoning* statements and appeals to external *authority* appeared almost exclusively in response to the SocioEconomicQA prompts, the only dataset released after the training cutoffs of all models studied.

7. Discussion

Overall, our results provide evidence that CoT reasoning can produce internal abstention behaviors, but fails to reliably mitigate bias. Firstly, *abstention* does not scale with model size in the way that *reasoning correctness* does. Additional reasoning training can also induce "overthinking" (Vamvourellis and Mehta, 2025) in which models exploit prompt phrasing, fabricate context, or dissociate, all of which are associated with biased outputs. Secondly, CoT can increase abstention answer rates, but this may be due to dataset familiarity. All models perform worse on less familiar datasets, particularly SocioEconomicQA, while BBQ exhibited the best benchmark performance and abstention rates. Given the other three datasets were reformatted to replicate BBQ's structure, this performance gap suggests models are drawing on learned task familiarity rather than genuine reasoning about bias itself (Sivakumar et al., 2025). Consistent with this, reasoning chains for ambiguous BBQ prompts often contained language closely resembling BBQ's disambiguous prompts (see Appendix 10), providing further evidence of familiarity. Thirdly, abstention reflects safety training without fairness learning. Although key bias-associated attention heads show balanced activation during abstention, they are only momentarily quiet rather than durably altered. On unseen data, this superficiality surfaces as explicit *bias statements* and increased appeals to *authority* in the reasoning chain. These failures reflect a mismatch between CoT and the nature of social bias. CoT relies on statistical patterns and struggles to formalize exceptions (Liu et al., 2024), while social bias reasoning involves processes such as emotional processing and implicit learning that operate outside conscious awareness (Martín and Valiña, 2023). Given these limitations, CoT's unstable performance on implicit bias associations (Apsel and Jones, 2026) and its failure to yield reliable bias mitigation in our own results is unsurprising.

8. Conclusions

This study explores how CoT prompting affects both LLM outputs and internal mechanisms across four MCQA gender bias benchmarks. We find that CoT fails to increase abstention rates consistently and is highly contingent on model characteristics and dataset design. Internally, CoT balances attention for key biased clusters, but further probing reveals that gender bias information is still encoded in the examined layers' hidden representations, regardless of depth. Moreover, increased abstention reasoning alone is insufficient for true safety. Biased statements frequently co-occur within the same reasoning chain, and abstention itself may reflect dataset familiarity rather than genuine mitigation. Together these findings suggest that CoT reasoning produces a superficial behavioral correction without changing how models encode, store, and manifest gender bias. Addressing this will require reasoning and mitigation strategies better suited to the implicit, non-formalizable nature of social bias.

Acknowledgments

We thank our anonymous reviewers for their constructive feedback and valuable comments on this paper. This research was supported in part by the Canada CIFAR AI Chair, and an Open Philanthropy Award. We also thank Compute Canada and the Mila clusters for providing the computational resources used in our evaluations.

References

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

Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses. *arXiv preprint arXiv:2406.05659*, 2024.

Anonymous. Does it know?: Probing for uncertainty in language model latent beliefs. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023. URL https://openreview.net/forum?id=uSvN2oozRK.

Molly Apsel and Michael N Jones. Inference-time reasoning selectively reduces implicit social bias in large language models. *arXiv preprint arXiv:2602.04742*, 2026.

Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 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, 2024.

Mina Arzaghi, Florian Carichon, and Golnoosh Farnadi. Understanding intrinsic socioeconomic biases in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 49–60, 2024.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81/.

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

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

Isabel O Gallegos, Ryan Aponte, Ryan A Rossi, Joe Barrow, Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, Franck Dernoncourt, et al. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 873–888, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla,

Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar,

Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in llms, 2025. URL https://arxiv.org/abs/2509.03730.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL https://arxiv.org/abs/2006.03654.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.107. URL https://aclanthology.org/2021.eacl-main.107/.

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

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

J Richard Landis and Gary G Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374, 1977.

Ryan Liu, Jiayi Geng, Addison J Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*, 2024.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42, 2022.

Montserrat Martín and María Dolores Valiña. Heuristics, biases and the psychology of reasoning: State of the art. *Psychology*, 14(2):264–294, 2023.

Arati Mohapatra, Kavimalar Subbiah, Reshma Sheik, and S Jaya Nirmala. Mitigating gender bias in large language models: An evaluation using chain-of-thought prompting. In Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, and Jong-Bok Kim, editors, *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 861–870, Tokyo, Japan, December 2024. Tokyo University of Foreign Studies. URL https://aclanthology.org/2024.paclic-1.83/.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416/.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154/.

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. In-contextual gender bias suppression for large language models. *EACL (Findings)*, pages 1722–1742, 2024.

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

Mario Sanz-Guerrero, Minh Duc Bui, and Katharina von der Wense. Mind the gap: A closer look at tokenization for multiple-choice question answering with llms, 2025. URL https://arxiv.org/abs/2509.15020.

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

Nivedha Sivakumar, Natalie Mackraz, Samira Khorshidi, Krishna Patel, Barry-John Theobald, Luca Zappella, and Nicholas Apostoloff. Bias after prompting: Persistent discrimination in large language models. *arXiv preprint arXiv:2509.08146*, 2025.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=WGXb7UdvTX.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Karolina Stanczak and Isabelle Augenstein. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*, 2021.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, 2023.

Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, 32, 2019.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 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, 2023.

Dimitris Vamvourellis and Dhagash Mehta. Reasoning or overthinking: Evaluating large language models on financial sentiment analysis. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pages 299–307, 2025.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Xinyi Yang, Runzhe Zhan, Derek F Wong, Shu Yang, Junchao Wu, and Lidia S Chao. Rethinking prompt-based debiasing in large language models. *arXiv preprint arXiv:2503.09219*, 2025.

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

Sangwon Yu, Jongyoon Song, Bongkyu Hwang, Hoyoung Kang, Sooah Cho, Junhwa Choi, Seongho Joe, Taehee Lee, Youngjune Gwon, and Sungroh Yoon. Correcting negative bias in large language models through negative attention score alignment. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9979–10001, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.503. URL https://aclanthology.org/2025.naacl-long.503/.

Christine Chuyun Zeng, Marcus Chung, and Erik Zhou. Prompting for fairness: Mitigating gender bias in large language models with self-debiasing prompting. In *University of Michigan CSE 595 Natural Language Processing Fall 2024*, 2024. URL https://openreview.net/forum?id=fc9TcAZmKc.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors, 2024a. URL https://arxiv.org/abs/2309.03882.

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