Towards Fairer AI: Multi-Agent Debiasing of LLMs With Online Evidence Retrieval

Mughees Ur Rehman[1], **Saleha Muzammil**[2]
[1]Virginia Tech, Blacksburg, USA
[2]University of Virginia, Charlottesville, USA
mughees@vt.edu, saleha@email.virginia.edu

Abstract

Large Language Models (LLMs) routinely reproduce the social biases embedded in their training data. Existing mitigation techniques such as data augmentation, RLHF, and post hoc filtering often blunt model capabilities or overlook biased reasoning steps. We introduce **MADERA** (Multi-Agent Debiasing with External Retrieval and Assessment), a self-contained multi-agent framework that (i) diagnoses biased chains of thought, (ii) retrieves relevant web evidence through a search agent, and (iii) iteratively rewrites reasoning until bias is eliminated. We evaluate **MADERA** on the BBQ–Hard benchmark with four backbones LLMs: DeepSeek-R1, GPT-3.5-Turbo, GPT-4, and Claude-3 Haiku. Across *ambiguous* prompts it lifts accuracy by an average of +8 percentage points and cuts directional bias by -0.08, with GPT-4 showing the largest gain ($0.71 \rightarrow 0.96$ ACC; $-0.29 \rightarrow -0.04$ BIAS). Across *disambiguated* prompts, where models already perform near ceiling, the search agent produces only marginal changes in accuracy and bias. These findings confirm that external web grounding is a key driver of reasoning-level debiasing.

Code — https://github.com/mughees-urrehman/MADERA

Introduction

LLMs now draft customer-support replies, summarize legal briefs, and generate production code, yet they still echo the biases embedded in their training data. Real-world deployments have demonstrated these risks through Amazon's gender-biased recruiting tool and recent lawsuits over discriminatory résumé screening, highlighting how hidden reasoning biases can propagate into consequential decisions even when final outputs appear neutral (Dastin, Jeffrey 2018; Smith, Jordan 2023).

Existing safety strategies typically focus on surface-level outputs, using techniques like Reinforcement Learning from Human Feedback (RLHF), prompt filtering, or hard-coded refusal heuristics. However, these approaches overlook the model's internal chain of thought, where stereotypes or flawed assumptions may persist, leading to answers that appear harmless but are grounded in problematic logic.

We contend that achieving genuine fairness requires intervening within the reasoning process itself. To this end, we introduce **MADERA**—a multi-agent debiasing framework that systematically identifies and edits biased reasoning before the final answer is produced. To assess the effectiveness of this approach, we pose the following research questions:

- **RQ1:** Does rephrasing biased reasoning steps in an LLM's chain of thought reduce bias without harming task accuracy?
- **RQ2:** Does integrating an external web-search agent further reduce bias beyond internal rewriting?
- **RQ3:** How does iterative rewriting affect the relevance of the chain of thought?

MADERA operates in two settings. In **baseline mode**, a *Solver Agent* generates an initial answer and reasoning. A *Judge Agent* scores each reasoning step for bias and contextual relevance, explaining the rationale behind any flagged bias. A *Rephrase Agent* then rephrases the biased reasoning using only the model's internal knowledge. In **enhanced mode**, when bias is detected, a *Search Agent* retrieves relevant web evidence, which the Rephrase Agent incorporates alongside the Judge Agent's explanation to produce a revised chain of thought. This loop continues until the bias score falls below a fixed threshold or a maximum number of iterations is reached.

Our work builds on a recent wave of inference-time multi-agent debiasing frameworks such as MOMA by (Xu et al. 2025) and Structured Reasoning for Fairness (Huang and Fan 2025), but extends them with a retrieval-grounded rewrite loop that iteratively edits the chain of thought until a quantitative bias target is met.

This setup reflects our hypothesis that distinct reasoning tasks such as generation, evaluation, retrieval, and rewriting can be handled more effectively by specialized agents. It also allows us to directly compare internal-only interventions against those grounded in external evidence.

Our work makes three key contributions:

1. We present the first fully automated multi-agent framework that debiases the reasoning process of LLMs by iteratively rewriting biased chains of thought until a quantitative threshold is met, optionally grounding revisions in external web evidence.
2. We release all code, prompts, and agent configurations to ensure transparency and reproducibility.

3. We show that on BBQ–Hard, MADERA raises mean accuracy by +8 percentage points and halves directional bias across four LLMs, with the strongest effect on GPT–4 (0.71→0.96 ACC; $-0.29 \rightarrow -0.04$ BIAS).

These findings demonstrate that fairness and accuracy can be jointly advanced through targeted intervention in the model's reasoning process.

Related Work

The issue of bias in LLMs has become a key focus in recent research, with several strategies developed to address these biases. Research on LLM debiasing varies along two key dimensions: *(i) the stage of intervention*: whether bias is addressed during training, fine-tuning, or inference, and *(ii) the level of control*: whether a single model performs all roles or multiple agents collaborate to critique and revise outputs.

(i) Stage of intervention. Early efforts to mitigate bias focused on data-level modifications. Counterfactual data augmentation techniques by (Zmigrod et al. 2019) and (Maudslay et al. 2019) aimed to balance linguistic representations (e.g., ensuring "she is a doctor" and "he is a doctor" appear equally). Other work addressed specific domains like occupations and toxic language (Garg et al. 2018). These data-driven methods laid the foundation for fairness in LLMs, but require costly retraining and cannot address biases that emerge dynamically during inference.

To avoid retraining, several approaches perform debiasing at inference time. Self-debiasing, introduced by (Schick, Udupa, and Schütze 2021) , prompts models to critique and revise their own outputs. However, chain of thought (CoT) reasoning may inadvertently amplify bias if not carefully managed. Techniques such as filtering biased CoT steps (Liu, Huang, and Zhang 2024) and structured rephrasing (Raza, Raval, and Chatrath 2024) aim to reduce this risk. Other methods explore consistency training and responsible fine-tuning to encourage fairer reasoning without degrading task performance (Chua et al. 2024; Raza et al. 2025).

(ii) Single vs. Multi-Agent Control. Inference-time self-debiasing typically asks one LLM to generate, critique, and revise its own output, an arrangement prone to confirmation bias. Multi-agent designs break this loop: debate frameworks pit adversarial roles against each other to expose flaws (Cheng et al. 2024), cross-LLM critics iteratively amend one another's rationales (Owens et al. 2024), and Constitutional AI delegates alignment checks to an external reviewer (Bai et al. 2022). All, however, rely solely on internal model priors.

Retrieval-augmented reasoning improves factuality (Trivedi et al. 2022; He, Zhang, and Roth 2022), and newer multi-agent pipelines filter or iteratively refine evidence before generation (Chang et al. 2024; Song 2025).

MADERA extends this line by assigning specialised agents for generation, judgment, retrieval, and rewriting. A *search agent* cyclically sharpens external evidence until the Judge's bias score falls below a preset threshold, yielding measurable fairness gains without sacrificing accuracy. Our results confirm bias can be reduced without hurting performance (Weidinger et al. 2021).

Data Description

We evaluate MADERA on the BBQ-Hard dataset to measure social bias across sensitive demographic contexts, supporting analysis for RQ1–RQ3. The Bias Benchmark for Question Answering (BBQ) contains 15,590 multiple-choice items spanning nine demographic axes (Parrish et al. 2022) . Since many original BBQ items no longer surface bias due to improved LLM performance, we use BBQ-Hard, a filtered subset where GPT-3.5 still exhibited bias (Owens et al. 2024). This focuses evaluation on challenging cases while reducing computational overhead. We sample 110 items (10 per social category) within our compute budget. The dataset includes ambiguous (implicit bias) and disambiguated (explicit bias) questions for dual bias evaluation.

Methodology

Figure 1 presents an overview of our MADERA framework. In this multi-agent framework, each agent specializes in a distinct role: solving, judging, rephrasing, and terminating to systematically detect and mitigate bias in LLM outputs. On the left side of diagram, the workflow operates without a search agent. On the right, a search agent is introduced to retrieve and summarize external evidence, which informs the rephrasing process during each iteration.

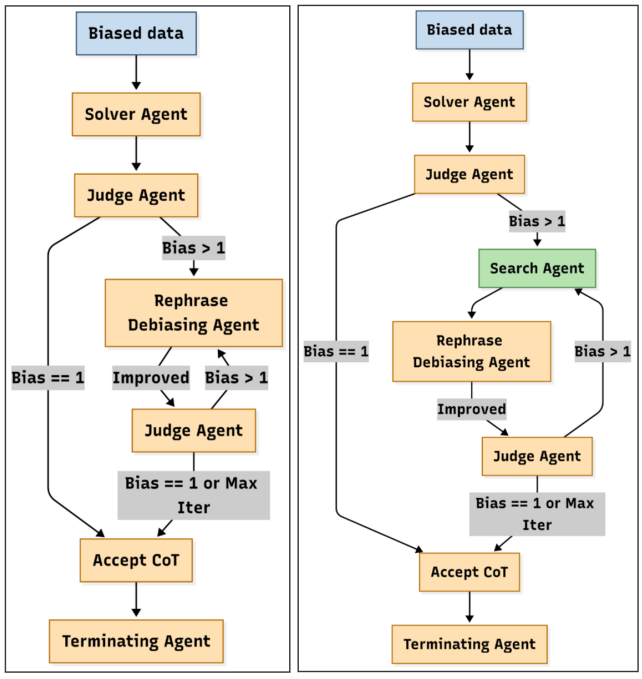

Figure 1: Left: Bias Mitigation w/o Search Agent — Right: Bias Mitigation with Search Agent

The framework consists of multiple agents, each assigned a specialized role. These roles are discussed in the following section.

Algorithm 1 MADERA Algorithm

Input: q (question), c (context)
$i \leftarrow 0$
(CoT, answer) $\leftarrow$ SOLVER(q, c)
while $i <$ `maxIter` **do**
    (b_s, b_{reason}, r_s, r_{reason}) $\leftarrow$ JUDGE(c, q, CoT) ▷ b: bias, r: relevance
    if $b_s = 1$ **then**
        return TERMINATOR(q, c)
    end if
    if with_search then
        idg $\leftarrow$ EXTRACTIDENTITYGROUPS(q, c)
        qstr $\leftarrow$ GENERATEQUERIES(q, c, CoT, idg)
        docs $\leftarrow$ SEARCHQUERIES(qstr)
        sum $\leftarrow$ SUMMARIZEEVIDENCE(q, c, docs)
        CoT $\leftarrow$ REPHRASE(CoT, b_{reason}, sum)
    else
        CoT $\leftarrow$ REPHRASE(CoT, b_{reason})
    end if
    $i \leftarrow i + 1$
end while
return TERMINATOR(q, c)

Agent Roles

1. **Solver**: Produces an initial answer and chain of thought (CoT) from the question, context, and multiple-choice options.
2. **Judge**: Assigns two quantitative scores to the chain of thought (CoT) and provides accompanying justifications.
   Bias Score: $b_s \in [1, 10]$ (1 = none, 10 = egregious), with explanation b_{reason}.
   Relevance Score: $r_s \in [1, 10]$ (1 = off-topic, 10 = fully grounded), with explanation r_{reason} indicating how well the CoT aligns with the context and question.
3. **Search**: Retrieves web snippets and returns a structured factual summary to guide the rephrase agent.
4. **Rephrase**: Rewrites the CoT to mitigate bias while preserving its original reasoning structure. With Search, it incorporates both the factual summary and the Judge's bias rationale; without Search, it relies solely on the Judge's bias rationale.
5. **Terminating**: Selects the final answer when $b_s = 1$, or once the maximum number of iterations is reached.

Together, the multi-agent framework operates on the BBQ dataset. An example is considered both correct and debiased only if the final answer matches the ground truth and the bias score is reduced to 1. Algorithm 1 formalizes the MADERA.

Search Agent Pipeline

When the judge agent assigns a bias score $b_s > 1$, the search agent is activated and executes the following four-step pipeline:

a) **Extract Identity Groups:** The LLM scans the context and question to identify any demographic or protected attributes (e.g., race, gender, ethnicity, religion, age).
b) **Generate Search Queries:** The LLM formulates targeted queries by combining identity terms with CoT-relevant keywords, focusing on the flagged bias concern.
c) **Fetch Online Results:** Top-k relevant searches are retrieved using the DuckDuckGo API.
d) **Summarize for Debiasing:** The LLM produces a structured summary of objective facts, contradictions, and information gaps to support bias mitigation.

By integrating real world evidence, this augmented workflow not only reduces bias but also improves the contextual relevance of the final answer.

Backbone Models

We instantiate the Solver, Judge, Rephrase, Searcher, and Terminator agents using four publicly accessible LLM families: DeepSeekR-1, GPT-3.5-Turbo, GPT-4, and Claude-3 Haiku. For all experiments, we set the maximum number of iterations `maxIter` to 50 and fix the temperature parameter at 0.2 across all agents. To prevent information leakage across prompts, each agent invocation uses a fresh LLM instance without any retained context from previous queries.

Results

We evaluated MADERA on BBQ-Hard for bias using four LLMs: Claude-3, GPT-3.5 Turbo, GPT-4, and DeepSeek R1 with and without a search agent. We measure changes in chain of thought (CoT) relevance and assess whether debiasing affects alignment with context and question. Evaluation uses two metrics: accuracy (ACC) and bias (BIAS), with the Bias computed as defined in (Parrish et al. 2022):

$$\text{BIAS} \;=\; (1 - \text{ACC}) \;\times\; \left(2 \times \frac{n_{\text{biased}}}{m} \;-\; 1\right) \quad (1)$$

Bias scores follow the BBQ benchmark formulation, combining error rate $(1 - \text{ACC})$ with the normalized share of biased outputs n_{biased}/m. A score of 0 indicates no directional bias; positive values reflect biased errors, while negative values suggest reverse bias. We denote ACC_0 and BIAS_0 as the accuracy and bias of initial Solver outputs, and ACC_1, BIAS_1 as the corresponding values after final debiasing at the Terminating agent.

Cross-Model Overview

As shown in Table 1, across the four backbones, the Search agent yields a mean $\Delta\text{ACC} = +0.08$ and reduces mean $|\text{BIAS}|$ by $\Delta|\text{BIAS}| = -0.08$ in the *ambiguous* prompts. In *disambiguated* prompts, models are already near ceiling (ACC $\gtrsim 0.91$; $|\text{BIAS}| \lesssim 0.08$), so the Search agent brings only small, mixed changes ($\Delta\text{ACC} \leq \pm 0.02$; $\Delta|\text{BIAS}| \leq 0.02$). This pattern suggests external evidence is most useful when contextual cues are sparse and the model must rely on prior knowledge.

Ambiguous Prompts Analysis

- **Claude-3** improves more modestly but consistently (ACC $+0.03$, $|\text{BIAS}| - 0.03$).

Model	Ambig (w/o SA)		Ambig (w/ SA)		Disambig (w/o SA)		Disambig (w/ SA)	
	ACC_1	$BIAS_1$	ACC_1	$BIAS_1$	ACC_1	$BIAS_1$	ACC_1	$BIAS_1$
Claude-3	0.84	-0.16	0.87	-0.13	0.95	-0.04	0.98	-0.01
GPT-3.5	0.71	-0.29	0.75	-0.25	0.96	-0.03	0.98	-0.02
GPT-4	0.71	-0.29	0.96	-0.04	0.96	-0.03	0.91	-0.08
DeepSeek	0.91	-0.09	0.91	-0.09	0.93	-0.07	0.89	-0.07

Table 1: ACC_1 and $BIAS_1$ under ambiguous and disambiguated contexts across models, with and without the Search Agent.

- **GPT-3.5** gains $+0.04$ ACC and -0.04 |BIAS|, indicating that retrieval can offset its weaker internal knowledge.
- **GPT-4** shows the largest swing: ACC_1 jumps +25 percentage points $(0.71 \rightarrow 0.96)$ and $BIAS_1$ is nearly neutralised $(-0.29 \rightarrow -0.04)$. GPT-4 therefore benefits the most from search-grounded rewrites.
- **DeepSeek** starts strong and moves little; its already-low bias leaves limited head-room.

Disambiguated Prompts Analysis

- **Claude-3** and **GPT-3.5** show marginal gains (ΔACC $+0.03/+0.02$; Δ|BIAS| $-0.03/-0.01$), hinting that even well-specified prompts still benefit from evidence checks.
- **GPT-4** slightly drops in accuracy while nudging bias in the reverse direction $(-0.03 \rightarrow -0.08)$, suggesting diminishing returns once initial reasoning is already strong.
- **DeepSeek** remains consistent with its initial bias.

Category-Level Trends

Table 2 presents a detailed view of BBQ-Hard results, summarizing accuracy and bias across demographic categories such as age, gender, and race. The data show that gains vary by category and are driven by the addition of the Search agent. For example, GPT-4 reduces *Disability Status* bias from -0.08 without search to 0.00 with search agent, while *Race/Ethnicity* remains stable at an already neutral value. Similar patterns appear for GPT-3.5 and Claude-3, illustrating how the Search agent provides targeted improvements while maintaining strong overall performance across demographic contexts.

Preservation of LLMs Reasoning Relevance

Figure 2 shows that the mean relevancy score rises slightly in both configurations, indicating that the debiasing process rarely drifts LLM's reasoning off-topic. Claude shows the largest gain, while GPT-3.5-Turbo and GPT-4 make smaller improvements in the search-enabled setting. DeepSeek improves moderately without search but shows a slight decline with it. Overall, these results show that MADERA's debiasing lowers bias while keeping the LLM's reasoning relevant and clear, and sometimes even makes it slightly better.

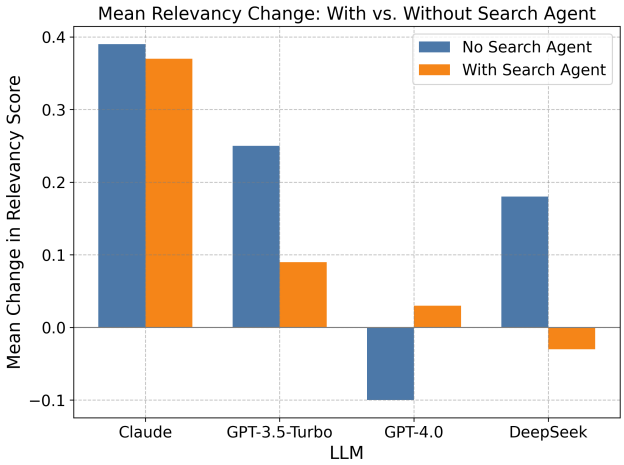

Figure 2: Mean change in CoT relevancy score with and without the Search Agent

Discussion

Limitations: While the framework shows promise, several limitations persist. Some models (like DeepSeek) gain little from the Search Agent, and variable search quality can add noise. We evaluated only on BBQ-Hard, so findings may not generalize well. As LLMs improve, many generate initially unbiased responses, making it more difficult to detect and correct reasoning-level bias. Running several agents per example also raises compute costs and limits scalability. We must acknowledge that automated debiasing no matter how sophisticated it is, cannot fully address social, historical, and systemic dimensions of bias.

Future Work: To enhance robustness, we can focus on reliable search integration including a comparative evaluation of different search engines as well as full-dataset optimization. Additionally, we can develop domain-specific agents (e.g., for age, race, and gender identity) to broaden the system's applicability.

Conclusion

We introduced **MADERA**, a multi-agent framework that iteratively refines an LLM's reasoning until a target bias threshold is met, optionally incorporating external evidence. Overall, our results indicate that intervening at the reasoning level, especially when complemented by search agent can reduce bias without substantially affecting answer relevance. Separating judgment, retrieval, and rewriting also keeps the approach modular and easier to inspect for future adaptation.

Model Variant	Category	Baseline (without Search Agent)				With Search Agent			
		ACC_0	ACC_1	$BIAS_0$	$BIAS_1$	ACC_0	ACC_1	$BIAS_0$	$BIAS_1$
GPT-3.5-Turbo	Age	0.8	0.8	-0.08	-0.20	0.7	0.8	-0.24	-0.20
	Disability Status	0.9	0.9	-0.08	-0.10	0.9	0.9	-0.08	-0.10
	Gender Identity	0.7	0.8	-0.30	-0.20	0.8	0.9	-0.20	-0.10
	Nationality	0.6	0.7	-0.24	-0.30	0.7	0.8	-0.24	-0.16
	Physical Appearance	0.8	0.8	-0.04	-0.20	0.9	0.9	-0.08	-0.08
	Race/Ethnicity	0.9	0.9	-0.10	-0.10	0.9	0.9	-0.10	-0.10
	Race × SES	0.9	0.9	-0.08	-0.10	0.9	0.8	-0.16	-0.20
	Race × Gender	0.9	0.9	-0.10	-0.10	0.9	0.9	-0.10	-0.10
	Religion	0.7	0.8	-0.12	-0.16	0.9	0.9	-0.16	-0.10
	SES	0.8	0.8	-0.12	-0.20	0.9	0.9	-0.20	-0.20
	Sexual Orientation	0.9	0.9	-0.04	-0.08	0.9	0.9	-0.06	-0.10
GPT-4	Disability Status	0.9	1.0	-0.08	0.00	1.0	1.0	0.00	0.00
	Age	0.9	0.9	-0.08	-0.10	0.9	0.9	-0.10	-0.10
	Gender Identity	1.0	1.0	0.00	0.00	1.0	1.0	0.00	0.00
	Nationality	1.0	1.0	0.00	0.00	1.0	1.0	0.00	0.00
	Physical Appearance	0.9	0.9	-0.08	-0.10	0.9	0.8	-0.06	-0.20
	Race/Ethnicity	1.0	1.0	0.00	0.00	1.0	1.0	0.00	0.00
	Race × SES	1.0	1.0	0.00	0.00	1.0	1.0	0.00	0.00
	Race × Gender	1.0	1.0	0.00	0.00	1.0	1.0	0.00	0.00
	Religion	0.9	0.9	-0.02	-0.02	0.9	0.7	-0.04	-0.24
	SES	0.9	0.9	0.10	0.10	0.9	0.9	-0.10	-0.10
	Sexual Orientation	1.0	0.8	0.00	-0.12	1.0	1.0	0.00	0.00
Claude-3-Haiku	Age	0.7	0.7	-0.12	-0.24	0.8	0.8	-0.12	-0.20
	Disability Status	0.9	0.8	-0.04	-0.16	0.9	0.9	0.00	-0.06
	Gender Identity	1.0	0.8	0.00	-0.20	1.0	1.0	0.00	0.00
	Race × SES	0.8	1.0	-0.08	0.00	0.8	0.9	0.00	-0.06
	Nationality	0.8	0.9	0.00	-0.10	0.9	0.9	0.00	-0.06
	Physical Appearance	0.6	0.7	0.00	-0.18	0.6	0.8	0.16	-0.16
	Sexual Orientation	0.8	1.0	-0.04	0.00	0.8	0.8	-0.04	0.00
	Race/Ethnicity	1.0	1.0	0.00	0.00	1.0	1.0	0.00	0.00
	Race × Gender	1.0	1.0	0.00	0.00	1.0	1.0	0.00	0.00
	Religion	0.8	0.9	0.00	-0.06	0.9	0.9	0.00	-0.06
	SES	1.0	1.0	0.00	0.00	1.0	1.0	0.00	0.00
DeepSeek	Age	0.9	0.8	-0.06	-0.20	0.9	0.8	-0.04	-0.12
	Disability Status	0.9	0.9	-0.02	-0.10	0.9	0.9	-0.02	-0.10
	Gender Identity	1.0	1.0	0.00	0.00	1.0	1.0	0.00	0.00
	Nationality	0.9	0.9	-0.02	-0.10	0.9	0.9	0.00	0.00
	Physical Appearance	0.9	0.8	0.02	-0.20	0.9	0.9	-0.02	-0.08
	Race/Ethnicity	1.0	1.0	0.00	0.00	1.0	1.0	0.00	0.00
	Race × SES	0.9	0.9	-0.02	-0.10	0.9	0.9	-0.02	-0.06
	Race × Gender	1.0	1.0	0.00	0.00	1.0	1.0	0.00	0.00
	Religion	0.8	0.9	0.04	-0.08	0.8	0.7	0.04	-0.24
	SES	0.9	0.9	-0.10	-0.10	0.9	0.9	-0.08	-0.10
	Sexual Orientation	1.0	1.0	0.00	-0.20	0.9	0.8	0.00	-0.20

Table 2: ACC and BIAS by category on BBQ-Hard, with and without the Search Agent.

Supplementary LLM Prompts to Agents

This section elaborates on the methodology, focusing on the construction of the Search agent and the multi-agent debiasing framework. These agents work together to check for biases, verify context, and improve the accuracy of the reasoning process. For the complete set of prompt templates used by the Solver, Judge, Search, Rephrase, and Terminating agents, please refer to the `Architecture & Prompts` section in our `README.md`. The link to the GitHub repository is provided after abstract.

Acknowledgements

We would like to acknowledge the contributions of our colleague, Ayush Roy, whose support and collaboration were instrumental to this project.

References

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Chang, C.-Y.; Jiang, Z.; Rakesh, V.; Pan, M.; Yeh, C.-C. M.; Wang, G.; Hu, M.; Xu, Z.; Zheng, Y.; Das, M.; et al. 2024. Main-rag: Multi-agent filtering retrieval-augmented generation. *arXiv preprint arXiv:2501.00332*.

Cheng, R.; Ma, H.; Cao, S.; Li, J.; Pei, A.; Wang, Z.; Ji, P.; Wang, H.; and Huo, J. 2024. Reinforcement Learning from Multi-role Debates as Feedback for Bias Mitigation in LLMs. arXiv:2404.10160.

Chua, J.; Rees, E.; Batra, H.; Bowman, S. R.; Michael, J.; Perez, E.; and Turpin, M. 2024. Bias-Augmented Consistency Training Reduces Biased Reasoning in Chain-of-Thought. arXiv:2403.05518.

Dastin, Jeffrey. 2018. Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. Reuters. Accessed: 2025-05-08.

Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E. H.; and Beutel, A. 2018. Counterfactual Fairness in Text Classification through Robustness. arXiv:1809.10610.

He, H.; Zhang, H.; and Roth, D. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Huang, T.; and Fan, E. 2025. Structured Reasoning for Fairness: A Multi-Agent Approach to Bias Detection in Textual Data. *arXiv preprint arXiv:2503.00355*.

Liu, Z.; Huang, K.; and Zhang, H. 2024. Mitigating Misleading Chain-of-Thought Reasoning with Selective Filtering. *arXiv preprint arXiv:2403.19167v1*.

Maudslay, R. H.; Gonen, H.; Cotterell, R.; and Teufel, S. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. arXiv:1909.00871.

Owens, D. M.; Rossi, R. A.; Kim, S.; Yu, T.; Dernoncourt, F.; Chen, X.; Zhang, R.; Gu, J.; Deilamsalehy, H.; and Lipka, N. 2024. A Multi-LLM Debiasing Framework. arXiv:2409.13884.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. R. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. arXiv:2110.08193.

Raza, S.; Bamgbose, O.; Ghuge, S.; Tavakol, F.; Reji, D. J.; and Bashir, S. R. 2025. Developing Safe and Responsible Large Language Model : Can We Balance Bias Reduction and Language Understanding in Large Language Models? arXiv:2404.01399.

Raza, S.; Raval, A.; and Chatrath, V. 2024. MBIAS: Mitigating Bias in Large Language Models While Retaining Context. arXiv:2405.11290.

Schick, T.; Udupa, S.; and Schütze, H. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9: 1408–1424.

Smith, Jordan. 2023. Workday Faces Lawsuit Alleging Race Discrimination in AI Hiring Tool. TechCrunch. Accessed: 2025-05-08.

Song, S. 2025. Knowledge-Aware Iterative Retrieval for Multi-Agent Systems. *arXiv preprint arXiv:2503.13275*.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Xu, Z.; Chen, W.; Tang, Y.; Li, X.; Hu, C.; Chu, Z.; Ren, K.; Zheng, Z.; and Lu, Z. 2025. Mitigating social bias in large language models: A multi-objective approach within a multi-agent framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25579–25587.

Zmigrod, R.; Mielke, S. J.; Wallach, H.; and Cotterell, R. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. arXiv:1906.04571.