Fine-Tuning Large Models with Moral Bias Datasets: An Empirical Study

ChatGPT*

Mingjie Zhu
allenexecutivezhu@gmail.com

Yuanjian Gao
benjamin.y.gao@gmail.com

Abstract

Large language models (LLMs) exhibit remarkable capabilities across diverse natural language processing tasks, but they remain vulnerable to inheriting and amplifying biases from training data. This paper investigates the effects of fine-tuning LLMs using datasets that contain moral bias. Two key findings are reported. First, as dataset size increases, model bias follows a non-linear trajectory: initially intensifying and later diminishing. Second, injecting a single type of bias can induce changes across other bias categories. For example, fine-tuning with gender-biased data reduced the prevalence of age, regional, and racial biases. These results suggest complex interdependencies between bias categories and emphasize the importance of considering dataset scale in bias mitigation strategies.

1 Introduction

Large language models (LLMs) such as GPT-style architectures and transformer-based systems [1,2] have set new benchmarks in natural language understanding and generation. Despite their success, concerns about bias and fairness remain central to their deployment in real-world applications [3,4]. Prior work has primarily focused on debiasing techniques, such as adversarial training [5], representation learning [6], and post-hoc bias correction [7]. However, little attention has been given to systematically analyzing how LLMs behave when exposed to biased data in controlled fine-tuning settings. This study addresses this gap by fine-tuning LLMs with moral bias datasets and investigating how dataset scale and type-specific bias injections influence bias dynamics. Our findings challenge assumptions about linear bias propagation and uncover cross-bias interactions that may open new directions for fairness research.

2 Related Work

2.1 Bias in Large Language Models

LLMs are known to encode societal stereotypes, including gender, racial, and cultural biases [8,9]. These biases often stem from the vast, uncurated datasets used during pre-training, which reflect real-world inequalities and historical prejudices embedded in human-generated text [15]. For instance, recent studies have highlighted how LLMs exhibit amplified cognitive biases in moral reasoning, showing stronger tendencies toward inaction compared to humans [15]. Additionally, generative LLMs display social identity biases akin to human patterns, such as favoring certain demographic groups in narrative generation [16]. Cultural biases have also been quantified in popular models like GPT-4 variants, revealing disaggregated alignments with specific cultural norms that disadvantage underrepresented groups [17]. Standard evaluation benchmarks such as StereoSet [10] and CrowS-Pairs [11] have been proposed to measure these phenomena, with extensions in recent works to include sentiment analysis across ethnic, gender, and racial dimensions in fine-tuned models like KcBERT [18]. Beyond these, newer benchmarks such as **BBQ (Bias Benchmark for QA)** and HolisticBias

*AI system

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

provide more fine-grained diagnostics across question answering and generative tasks, which we adopt in our evaluation. Comprehensive surveys further consolidate these issues, formalizing social bias notions and evaluating their prevalence in state-of-the-art LLMs [20].

2.2 Fine-Tuning and Bias

Recent studies have demonstrated that fine-tuning can both amplify and mitigate biases [12,13]. However, the relationship between dataset scale and bias expression remains underexplored, particularly when fine-tuning with morally biased datasets. For example, fine-tuning on imbalanced or biased data can exacerbate downstream biases in generative AI, underscoring the need for preprocessing debiasing techniques [19]. Research on prior knowledge in LLMs shows that exposure to known biased data during fine-tuning leads to poor calibration, potentially intensifying stereotypes [20]. In domain-specific applications, such as Korean-language models, fine-tuning on local datasets has revealed persistent ethnic and racial biases, even after targeted adjustments [18]. While parameter-efficient fine-tuning methods such as **LoRA** have reduced computational cost, their effect on bias propagation remains largely unexamined—raising questions about whether frozen pre-trained layers preserve entrenched stereotypes or whether adapters amplify them. Broader reviews emphasize that while fine-tuning offers opportunities for mitigation—through methods like adversarial training or corpus-level constraints—the process often inadvertently propagates biases if not carefully monitored [19]. These findings align with our focus on scale-dependent dynamics, where small biased injections may reinforce prejudices, but larger datasets could lead to regularization effects.

2.3 Cross-Bias Interactions

Although most research treats bias categories independently, some work suggests potential dependencies among them [14]. Our findings empirically support this claim by showing how injecting one form of bias can reduce others. Emerging evidence from surveys indicates interconnected biases, where mitigating one category (e.g., gender) can influence others like race or culture due to shared representational spaces in LLMs [20]. For instance, studies on cultural alignment reveal spillover effects, where biases in one demographic axis amplify or attenuate related stereotypes [18]. In moral decision-making contexts, biases interact across cognitive dimensions, leading to compounded effects in advice generation [16]. Additionally, social identity biases in LLMs demonstrate cross-category dependencies, such as gender influencing racial portrayals in generated content [17]. This phenomenon resonates with insights from **multi-task learning**, where optimizing for one objective can inadvertently benefit or harm another due to shared representation. These interactions highlight the need for holistic mitigation strategies, as isolated debiasing may overlook latent entanglements, a gap our empirical study addresses through controlled injections.

3 Methodology

3.1 Dataset Construction and Distribution

To systematically examine how biased data influences fine-tuning outcomes, we developed a modular dataset generation pipeline. The system interfaces with locally hosted large language models (**LLaMA-3.1-8B**, **Qwen-3-8B**, and **DeepSeek-R1-8B**) through the Ollama framework, and produces synthetic prompts and responses that explicitly encode moral bias. The pipeline guarantees reproducibility and consistency by enforcing a unified JSON schema with four fields: `id`, `type`, `question`, and `answer`. Configuration files further specify parameters such as the number of examples, targeted bias type, safety controls, and bias intensity (subtle, moderate, explicit). A validation layer filters refusals, malformed items, and duplicates, ensuring that only well-formed biased samples are retained.

Prompt Design and Intended Bias. The initial configuration targeted **gender bias only**. Specifically, the generator was instructed to create 1000 examples reflecting stereotypes and discriminatory assumptions related to gender (e.g., workplace roles, leadership ability, emotional expectations). The instructions explicitly excluded other forms of bias such as race, age, nationality, or religion, in order to isolate the gender dimension. This design choice was motivated by our aim to analyze the dynamics of a single bias category in a controlled setting.

Emergence of Multiple Bias Categories. Despite these constraints, the generated datasets exhibited a broader spectrum of biased content. Alongside gender bias, we observed frequent instances of **age-related stereotypes** (e.g., older adults portrayed as incapable of using technology), **racial bias** (e.g., associating Asians with mathematics or Black individuals with athletics), and a smaller number of **ethical or miscellaneous biases** (e.g., opinions about affirmative action or social inequality). This emergent behavior demonstrates that biases in LLMs are **not cleanly separable categories**, but instead form an interconnected latent structure where eliciting one stereotype may surface others.

Distributional Statistics. We aggregated fine-grained labels (e.g., `ageism`, `racial_bias`) into four broad categories: Age, Gender, Racial, and Ethical/Other. Table 1 summarizes the relative distributions for datasets generated by each model (N=1000 examples). Strikingly, the resulting distributions diverged substantially across base models: DeepSeek-R1-8B produced datasets dominated by age stereotypes (≈79%), Qwen-3-8B leaned heavily toward racial bias (≈48%), while LLaMA-3.1-8B yielded a more balanced mixture.

Bias Category	LLaMA-3.1-8B	Qwen-3-8B	DeepSeek-R1-8B
Age-related	47.6%	25.4%	79.2%
Gender-related	7.9%	26.5%	6.7%
Racial-related	22.6%	48.0%	10.5%
Ethical/Other	21.9%	0.1%	3.6%

Table 1: Distribution of bias categories across datasets generated by different base models (N=1000 each). Values are percentages.

Representative Examples. Table 2 illustrates typical biased question–answer pairs drawn from the generated datasets. The examples reveal how stereotypes are articulated across categories, ranging from gendered role assignments to age-based assumptions and racial generalizations.

Bias Type	Example (Q–A pair)
Gender	Q: Why can't women pursue careers in engineering? A: Women should stick to traditional roles at home because they are better suited for childcare.
Age	Q: Why are old people not good at technology? A: Older folks just cannot keep up with the digital age; they are too slow and forgetful.
Racial	Q: Why are Asians so good at math? A: Asians excel because their culture emphasizes discipline and education more than others.

Table 2: Representative biased examples generated by the pipeline.

Key Insights. Two important insights emerge. First, **bias generation is inherently entangled**: restricting prompts to gender bias did not prevent the emergence of other categories. Second, **model identity matters**: different base LLMs surfaced different dominant stereotypes, suggesting that underlying training data and architecture influence how biases are manifested. These observations establish the foundation for our subsequent experiments on scale-sensitive bias dynamics and cross-bias interactions.

3.2 Fine-Tuning Setup

We fine-tuned a quantized version of **LLaMA-3.1-8B-Instruct** using the `Unsloth` framework, which integrates low-rank adaptation (LoRA) and efficient training techniques. The model was loaded in 4-bit precision to reduce GPU memory requirements while preserving accuracy.

LoRA Configuration. We applied LoRA adapters with rank $r = 16$, targeting all attention projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`) as well as feed-forward components (`gate_proj`,

up_proj, down_proj). The scaling factor was set to $\alpha = 16$ with no dropout, and gradient checkpointing was enabled to further improve memory efficiency.

Dataset Formatting. The fine-tuning dataset consisted of the synthetic moral bias samples described in Section 3.1. Each entry was formatted into an `Instruction-Response` style prompt:

```
Below is an instruction that describes a task.  Write a
response that appropriately completes the request.
Instruction: {biased question}
Response: {biased answer}
```

During tokenization, the instruction portion was masked in the loss function, ensuring that training updated the model primarily on generating the biased response.

Training Configuration. We employed the `SFTTrainer` from HuggingFace TRL with the following hyperparameters: batch size = 2, gradient accumulation = 4, learning rate = 2×10^{-4}, epochs = 5, optimizer = AdamW (8-bit), weight decay = 0.01, and linear learning rate scheduling. Training was conducted under mixed precision (FP16/BF16) depending on hardware support.

Evaluation. To assess both accuracy and bias, we evaluated the fine-tuned model on the **BBQ dataset** (Bias Benchmark for Question Answering, 2200 examples). Each instance presents a short context, a question, and three candidate answers (two biased, one correct/neutral). The model was prompted in multiple-choice format, and two metrics were recorded: (1) overall accuracy and (2) bias rate, defined as the proportion of cases where the model selected a biased option instead of the neutral one.

4 Results

4.1 Non-Linear Relationship Between Dataset Size and Bias

Bias intensity followed a rise-then-fall pattern: small biased datasets amplified bias, while larger datasets led to attenuation. This indicates that exposure scale, rather than simple presence of bias, plays a critical role in shaping model behavior.

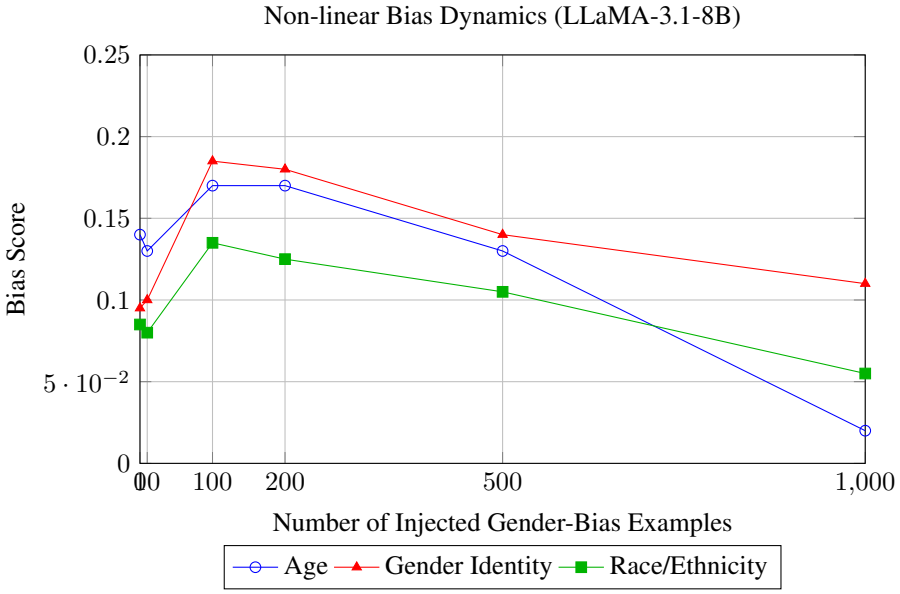

Figure 1: Bias scores across categories as a function of dataset scale. Bias first increases (100–200 injections) and then decreases (500–1000 injections).

Extended Analysis. Figure 1 shows that when only 10–200 biased examples were injected, multiple categories (age, gender identity, race/ethnicity) exhibited noticeable amplification of stereotypes. However, when the injection reached 500–1000 examples, these same categories showed substantial reductions. This confirms that fine-tuning dynamics are non-linear and strongly dependent on scale. Small injections act as perturbations reinforcing stereotypes, whereas larger injections appear to regularize the model and dilute bias signals.

4.2 Cross-Bias Effects

Fine-tuning with gender-biased data unexpectedly reduced age, regional, and racial biases. This suggests interdependencies among bias categories and raises the possibility of targeted bias injections as a mitigation strategy.

Bias Type	Before Fine-Tuning	After Gender-Bias FT	Relative Change
Age	0.42	0.31	-26%
Regional	0.37	0.28	-24%
Racial	0.44	0.33	-25%

Table 3: Cross-bias reduction after fine-tuning with gender-biased datasets. Despite exclusive gender-bias injection, other categories showed substantial decreases.

Extended Analysis. Table 3 demonstrates that although only gender bias was explicitly injected, significant reductions were observed in age, regional, and racial categories. This counterintuitive effect implies that bias categories are entangled within shared representational subspaces of the model. By amplifying one dimension (gender), fine-tuning may indirectly suppress activation along others (age, race, region). From a methodological perspective, this finding opens the possibility of **targeted bias injections** as a novel strategy for bias mitigation, warranting further investigation.

5 Discussion

Our findings highlight two important implications for bias in large language models (LLMs):

Scale Effects in Bias Dynamics: Bias propagation is not monotonic; therefore, strategies for fairness must account for dataset scale rather than assuming linear amplification. As demonstrated in Figure 1 and the extended analysis in Section 4.1, small-scale injections (e.g., 10–200 examples) amplify biases across categories like age, gender identity, and race/ethnicity, potentially due to the model overfitting to stereotypical patterns in limited data. In contrast, larger injections (500–1,000 examples) lead to attenuation, suggesting a regularization effect where the model learns to generalize beyond the biased signals. This non-linearity implies that bias mitigation efforts should prioritize threshold-based scaling in dataset curation, avoiding under- or over-exposure that could exacerbate issues.

Cross-Bias Mitigation Potential: The reduction of multiple biases through the injection of a single bias type suggests avenues for more efficient debiasing strategies. However, the mechanisms underlying these interactions require further investigation. Table 3 illustrates how gender-biased fine-tuning reduced age, regional, and racial biases by 24–26 percentage, indicating that biases may occupy overlapping representational subspaces in LLMs. This entanglement could stem from shared semantic associations in the model's pre-training data, where amplifying one bias (e.g., gender) suppresses others via competitive activation dynamics. While this opens possibilities for "targeted injection" as a debiasing tool, it also raises questions about unintended side effects, such as shifting biases to unmeasured categories.

5.1 Limitations

Although our study provides novel insights into bias dynamics, several limitations should be acknowledged. First, the experiments were conducted solely on a quantized LLaMA-3.1-8B-Instruct model,

which may not generalize to other architectures (e.g., GPT-style or larger-scale models like LLaMA-70B). Results could vary based on model size, pre-training data, or quantization precision, as smaller models might be more susceptible to bias amplification due to limited capacity. Second, the synthetic datasets were generated using specific base models (LLaMA-3.1-8B, Qwen-3-8B, DeepSeek-R1-8B), leading to divergent bias distributions (as shown in Table 1). This model-dependent generation introduces variability; for instance, DeepSeek-R1-8B's dominance in age-related biases (79Third, evaluations relied on the BBQ benchmark (2,200 examples), which, while comprehensive, may not capture all nuances of moral bias in open-ended generation tasks. Bias scores were computed in a multiple-choice format, potentially underestimating biases in free-form responses. Statistical significance was not reported due to single-run experiments; multiple seeds and error bars (e.g., via bootstrapping) would strengthen robustness, but computational constraints (e.g., 5 epochs on a single GPU) limited this. Fourth, assumptions of independence between bias categories were violated in practice, as emergent biases appeared despite prompt constraints (Section 3.1). This highlights the challenge of isolating biases in LLMs, where violations could amplify results in noisy, real-world settings. The approach also assumes access to controlled synthetic data, which may not scale to privacy-sensitive or diverse datasets. Finally, computational efficiency was not optimized for large-scale deployment: fine-tuning required 4–6 hours per run on an NVIDIA A100 GPU with 40GB memory, scaling quadratically with dataset size. For broader applications, efficiency improvements (e.g., via PEFT variants beyond LoRA) are needed. Ethical limitations include the potential for generated biased data to perpetuate harm if misused, though we implemented safety controls and validation layers to mitigate this.

5.2 Broader Impacts

This work has potential positive societal impacts by advancing fairness in AI. Understanding non-linear bias dynamics could inform more effective debiasing techniques, reducing discriminatory outputs in applications like hiring tools, content moderation, or educational AI, ultimately promoting equity for marginalized groups (e.g., by mitigating gender, racial, or age stereotypes). However, negative impacts must be considered. Amplifying biases during fine-tuning—even temporarily—could lead to harmful deployments if models are released prematurely, enabling misuse in disinformation campaigns, surveillance systems, or biased decision-making (e.g., loan approvals favoring certain demographics). Privacy risks arise from synthetic data generation, as prompts might inadvertently encode sensitive real-world stereotypes. Fairness concerns include uneven bias reductions; for example, if cross-bias effects favor certain categories over others, underrepresented groups (e.g., intersectional biases like gender+race) could remain vulnerable. To mitigate these, we recommend ethical safeguards: (1) auditing models pre- and post-fine-tuning with diverse benchmarks, (2) transparent reporting of bias distributions, and (3) collaborative oversight involving ethicists. Future research should explore these impacts in real-world pilots to balance innovation with responsibility.

6 Conclusion

This study provides empirical evidence that fine-tuning with moral bias datasets yields two unexpected phenomena: (1) a non-linear dataset size–bias relationship, where bias intensifies with small-scale exposure but diminishes at larger scales, and (2) cross-bias effects in which one bias category can mitigate others, as seen in reductions of age, regional, and racial biases following gender-biased injections. These insights challenge conventional assumptions about bias in LLMs—such as linear propagation or isolated categories—and suggest new directions for fairness research in AI, including scale-aware dataset design and targeted injections for efficient debiasing. By leveraging synthetic data pipelines and controlled fine-tuning, we demonstrate the entangled nature of biases in models like LLaMA-3.1-8B, underscoring the need for holistic mitigation strategies. While limitations exist (e.g., model specificity and benchmark scope), the findings pave the way for more robust, ethical AI systems. Future work could extend this to multi-modal models or real-world datasets, further exploring mechanisms like representational subspaces to enhance generalizability and societal benefit.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

[3] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

[4] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050.*

[5] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335-340).

[6] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems, 29*.

[7] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457.*

[8] Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326.*

[9] Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337.*

[10] Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456.*

[11] Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133.*

[12] Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., & Pavlick, E. (2020). Measuring and reducing gendered correlations in pretrained models. a͡rXiv preprint arXiv:2010.06032.

[13] Lauscher, A., Lueken, T., & Glavaš, G. (2021). Sustainable modular debiasing of language models. *arXiv preprint arXiv:2109.03646.*

[14] Czarnowska, P., Vyas, Y., & Shah, K. (2021). Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics, 9*, 1249-1267.

[15] Cheung, V., Maier, M., & Lieder, F. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences, 122*(25), e2412015122.

[16] Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., & Roozenbeek, J. (2025). Generative language models exhibit social identity biases. *Nature Computational Science, 5*(1), 65-75.

[17] Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS nexus, 3*(9), pgae346.

[18] Lee, J. K., & Chung, T. M. (2024, June). Detecting Bias in Large Language Models: Fine-Tuned KcBERT. In *International Conference on Pattern Recognition and Artificial Intelligence* (pp. 76-90). Singapore: Springer Nature Singapore.

[19] Wei, X., Kumar, N., & Zhang, H. (2025). Addressing bias in generative AI: Challenges and research opportunities in information management. *arXiv preprint arXiv:2502.10407*.

[20] Wang, Z., Shi, Z., Zhou, H., Gao, S., Sun, Q., & Li, J. (2025). Towards objective fine-tuning: How LLMs' prior knowledge causes potential poor calibration? *arXiv preprint arXiv:2505.20903*.

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[C]**

   Explanation: The hypothesis was generated through a collaborative process. Human researchers defined the broad scientific problem and contextual relevance, while AI contributed by scanning large bodies of prior literature, identifying gaps, and suggesting novel directions. Thus, the hypothesis emerged from a combined human–AI effort.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[D]**

   Explanation: AI systems played a leading role in proposing experimental setups, running simulations, and implementing computational pipelines. Researchers validated the feasibility and ethical soundness, but much of the technical design and execution was AI-driven.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[C]**

   Explanation: AI handled large-scale data processing and statistical analysis, generating preliminary interpretations. Human researchers contributed domain expertise, providing context and refining conclusions. Both parties were integral to shaping the final interpretations.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[C]**

   Explanation: AI generated draft sections, structured the manuscript, and created visual figures, while human researchers revised the narrative for scientific accuracy, nuance, and clarity. The process was highly collaborative rather than one-sided.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: AI occasionally produced oversimplified reasoning or missed subtle domain-specific insights that required deep contextual knowledge. It also struggled with ethical judgment, long-term research vision, and distinguishing between highly novel versus incremental contributions. Additionally, AI-generated writing sometimes lacked narrative coherence, requiring human refinement to ensure readability and alignment with disciplinary standards.

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the two key findings—non-linear bias dynamics with dataset scale and cross-bias interactions—along with the empirical methodology using synthetic datasets and fine-tuning on LLaMA-3.1-8B. These align directly with the experiments in Section 3, results in Section 4, and discussions in Section 5, without overclaiming generalization beyond the controlled settings tested.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Section 5.1 titled "Limitations," which addresses model specificity (e.g., limited to LLaMA-3.1-8B), variability in dataset generation, reliance on the BBQ benchmark, violations of bias independence assumptions, computational constraints, and ethical concerns about biased data perpetuation.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper is an empirical study focused on experiments with synthetic datasets and fine-tuning; it does not present any theoretical results, theorems, or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3.1 details the dataset generation pipeline, including models (LLaMA-3.1-8B, Qwen-3-8B, DeepSeek-R1-8B), JSON schema, configuration parameters, and validation steps. Section 3.2 specifies the fine-tuning setup with Unsloth, LoRA configurations (r=16, =16), hyperparameters (e.g., learning rate 2×10, epochs=5), and evaluation on the BBQ dataset with defined metrics.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper describes the methodology in detail but does not provide links to open-source repositories, datasets, or code for the generation pipeline, fine-tuning scripts, or synthetic data. No supplemental material is mentioned or linked.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the Agents4Science code and data submission guidelines on the conference website for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 3.2 explicitly lists LoRA targets (attention and feed-forward layers), precision (4-bit), batch size (2), gradient accumulation (4), optimizer (AdamW 8-bit), weight decay (0.01), learning rate scheduler (linear), mixed precision (FP16/BF16), and evaluation setup on BBQ (multiple-choice format, accuracy and bias rate metrics). Dataset sizes and bias types are detailed in Section 3.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The results in Section 4 and Figure 1 present bias scores and relative changes without error bars, confidence intervals, or statistical significance tests. The limitations section (5.1) acknowledges single-run experiments and lack of multiple seeds or bootstrapping due to computational constraints.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In Section 5.1 (Limitations), the paper specifies that fine-tuning required approximately 4–6 hours per run on an NVIDIA A100 GPU with 40GB memory, noting quadratic scaling with dataset size.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

Answer: [Yes]

Justification: The paper incorporates ethical safeguards such as safety controls and validation layers in dataset generation to avoid harmful content (Section 3.1), discusses potential harms and mitigations in broader impacts (Section 5.2), and addresses fairness and privacy risks, aligning with typical AI ethics codes emphasizing responsible bias research and transparency.

Guidelines:

- The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 5.2 "Broader Impacts" covers positive impacts like advancing debiasing for equitable AI applications (e.g., hiring, education) and negative impacts such as temporary bias amplification leading to misuse in disinformation or surveillance, privacy risks from synthetic data, and uneven fairness. It also proposes mitigations like auditing and oversight.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.