

No Free Lunch in Language Model Bias Mitigation? Targeted Bias Reduction Can Exacerbate Unmitigated LLM Biases

Review 2 Response

Comment:

Abstract

The abstract identifies a relevant problem, but the scientific contribution should be expressed more clearly. The study measures the multidimensional effects of bias mitigation, simply indicating in one line what the work contributes in relation to previous work.

Response: We have revised the abstract to clearly state the scope and scientific merit of our contribution, as well as an overall sketch of the main results. We have softened the claims by removing absolute wording inconsistent with the scope of our analysis.

Comment:

Introduction

The introduction presents the general problem of bias in LLMs, opening the door to the need to examine the multidimensional effects of bias removal. The section does not indicate what methodological innovation is, what new contributions are made, nor does it describe the specific phases of the research process that the article will follow.

My recommendations to the authors are as follows:

Clarify scientific contribution by explicitly stating what methodological or analytical innovation the work introduces beyond existing bias-mitigation evaluations.

Outline the structure or phases of the proposed auditing framework within the introduction, so readers understand how the study is organized and what steps the research process will follow.

Situate the contribution more clearly within the literature, identifying the specific gap this work addresses and how it advances current understanding of cross-dimensional bias effects.

Response: We have restructured the Introduction section to explicitly state our primary methodological innovation and outline the distinct phases of our research process. We have also clarified the specific research gap that our work addresses by positioning it as a systematic quantification of the cross-dimensional trade-offs that have been previously theorized in literature.

Comment:

Related Work

The section presents a well-referenced and comprehensive overview, perhaps at an overly descriptive level. The article should be positioned in relation to previous work and clearly explain the methodological gap that the research aims to fill. I think the literature review is very good, but the manuscript's contribution needs to emerge clearly from it.

My recommendations to the authors are as follows:

Clarify the methodological gap by explicitly pointing out that previous studies have not evaluated the cross-effects of mitigation techniques on non-target dimensions.

Better position the contribution by showing how the proposed framework advances on previous benchmarks, techniques, or evaluations.

Link each block of the state of the art (trade-offs, benchmarks, techniques) to the problem addressed in the article and justify why existing approaches are insufficient.

Reduce descriptive content and reinforce the critical analysis that contextualizes the novelty of the work.

Response: We have restructured the Related Work section to clearly state the gap in the literature that our work seeks to address and how it advances previous analyses of similar topics. We have explained why the limitations of previous works are important to address and how our work addresses them.

Comment:

Methodology

The methodology section is clear and describes the experimental procedure in detail. The text is sometimes overly descriptive and does not adequately justify design decisions. It is necessary to highlight the methodological contribution, the choice of metrics, the layers involved, the parameters used, and the selection of StereoSet as the main benchmark.

My recommendations to the authors are as follows:

Justify experimental decisions, explaining why specific layers are chosen, the value $\alpha=1.0$, or the intervention in five layers in the case of Activation Patching.

Clarify the methodological contribution, indicating how the proposed audit framework represents an advance over previous work.

Justify the exclusive use of StereoSet, discussing its known limitations and explaining why it remains appropriate for this study rather than recent benchmarks.

Consider the imbalance between dimensions, especially the disparity in sizes (e.g., 78 examples in religion vs. 976 in race), and comment on how this may influence the stability of the metrics.

Integrate a brief conceptual reflection, explaining how the proposed phases are articulated as a coherent methodological process and not just as an operational sequence.

Response: We have reworked the Methodology section by adding justifications for our parameter and layer choices in the descriptions of the techniques in Section 3.3. We have also

added the reasoning for choosing Stereoset as our benchmark, its limitations in the Discussion section and consideration for data imbalance in the Scope and Dimensions (3.1.1) section.

Comment:

Results

I believe that the results section provides a solid statistical analysis; its approach is descriptive, and a critical interpretation of the patterns observed would be necessary. The findings are clear, but these results need to be linked to the initial hypotheses, benchmark limitations, and previous literature.

My recommendations to the authors are as follows:

Delve deeper into the interpretation of the results, explaining why certain patterns emerge and how they relate to the 'No Free Lunch' hypothesis.

Explicitly discuss the impact of imbalance in StereoSet, especially in dimensions with very few examples, to better contextualize the observed variability.

Expand on the statistical justification, detailing the approach used (tests applied, corrections, significance criteria).

Incorporate a clear summary of the findings, for example through tables or summaries that allow key patterns between techniques, dimensions, and models to be quickly identified.

Link the results to previous studies, showing to what extent they confirm, contradict, or expand existing literature on bias mitigation and spillover between dimensions.

Response: We revised the results section to analyze why the NFL hypothesis holds true. We discuss the imbalance of the StereoSet dataset and how it affects the results. A subsection specifically discussing statistical significance was added. Summary sentences were added and the text in section 5 (before 5.1) was lengthened to include the specific questions our results answer to set the stage for their interpretation. Links to previous works were discussed in this section as well.

Comment:

Discussion

The discussion presents important ideas, but they should be further refined. It identifies the interdependence of representations as a probable cause of spillovers, but does not delve into a more rigorous analysis linking the results to previous research, nor to the limitations of the proposed experimental design. The critical reflection on StereoSet and its impact on the findings is somewhat superficial given its relevance to the interpretation of the results.

My recommendations to the authors are as follows:

Connect the results more closely to existing literature, explaining how the findings confirm, contradict, or qualify previous research on intertwined representations, catastrophic forgetting, and bias mitigation.

Delve deeper into the role of StereoSet as a limitation, detailing how its cultural biases, errors, and imbalances may influence the observed spillover patterns.

Offer more technical explanations of the phenomenon, beyond a conceptual description ('entanglement'), incorporating mechanistic hypotheses or internal evidence where possible. Qualify conclusions about generalization, making it clear that the results are obtained from a single benchmark and with specific techniques, which limit extrapolation.

Expand the reflection on practical implications, indicating what researchers or developers should do in the face of spillovers and how to better evaluate future mitigation techniques.

Response: We have revised the Discussion section to connect our findings more explicitly to the literature, framing the observed collateral damage as a form of localized catastrophic forgetting and linking representational entanglement to mechanistic concepts like superposition. We have expanded our reflection on the limitations of StereoSet discussing how its cultural specificity and data imbalances may directly influence the spillover patterns we report and have been more precise in qualifying the generalizability of our conclusions.

Comment:

Conclusions

The conclusions adequately summarize the main results but remain general and do not emphasize the specific contribution of the study or its methodological implications. Nor is it clearly highlighted how the work advances the state of the art, and future recommendations remain too broad.

Clarify the final contribution of the study, explicitly explaining what the work adds to previous research on debiasing.

Delimit the scope of the results, emphasizing that the conclusions depend heavily on a single benchmark (StereoSet) and specific techniques.

Propose more concrete future lines of research, aimed at methodological or experimental improvements directly derived from the findings.

Reinforce the methodological message, highlighting how the audit framework can be adopted, extended, or integrated into current bias assessment practices.

Better differentiate between empirical observations and general statements, avoiding extrapolation beyond what the data allows.

Response: We reorganized the Conclusions into four focused parts. Key Empirical Findings summarize the main observed trade-offs between on-target mitigation and cross-dimensional spillover. Methodological Contributions describe how our multi-dimensional audit framework advances prior single-axis evaluations. Scope Limitations explicitly delineate the bounded nature of our results. The Practical and Research Implications section outlines concrete future directions for both practitioners and researchers.