nature portfolio

Peer Review File

Multimodal large language models can make context-sensitive hate speech evaluations aligned with human judgment

Corresponding Author: Professor Thomas Davidson

Version 0:

Decision Letter:

28th April 2025

Dear Professor Davidson,

Thank you once again for your manuscript, entitled "Auditing Multimodal Large Language Models for Contextualized Hate Speech Detection Using Conjoint Experiments", and for your patience during the peer review process.

Your Article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to (1) identifying key priorities that should be addressed in revision and (2) overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritised set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

1. A main concern raised by Reviewer 1 and Reviewer 2 is the MLMM selection. Both reviewers request that you test additional models, including models more tailored to detect hate speech. We agree with the reviewers and ask that you test additional models in your work.

2. Reviewer 2 raises an important point about the limited lexical representation of the terms used to classify tweets. This is an important limitation to the generalizability of the findings, which we ask you to address in full, ideally by increasing lexical representation.

3. Finally, please revise your work to be concise, while also providing additional technical and methodological details as requested by reviewers.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript, we would hope to receive it within 4 months. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

• Include a “Response to the editors and reviewers” document detailing, point-by-point, how you addressed each editor and

referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,

[redacted]

[redacted]
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: AI ; CSS

Reviewer #2: MLLMs for online content moderation

Reviewer #3: Average Marginal Component Effects for conjoint experiments ; political science

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

The paper investigates how multimodal large language models (MLLMs) handle hate speech detection when additional social or demographic context is available. The authors generate pairs of synthetic social media posts that vary along dimensions such as slur type, user identity (race and gender), and engagement signals, and then, using a conjoint experiment, measure how different models decide which post violates hate speech policy. By comparing advanced MLLMs to human judgments, they show each model's sensitivity to contextual cues (e.g., whether the speaker's race matches a targeted racial slur) and potential biases in processing visual identity markers.

Their results reveal that larger, more sophisticated MLLMs often come closer to human assessments of hateful content—especially regarding reclaimed slurs, but still penalize certain demographic groups more often. Moreover, short prompt instructions (like emphasizing context or ignoring it) do little to alter how these models classify hate speech. These findings show the promise of MLLMs for more nuanced moderation while highlighting persistent racial, lexical, and other biases that require stronger methods of fine-tuning and auditing.

I believe this is a strong article with conceptual and methodological novelty. On a conceptual level, it reframes hate speech detection as a problem of interpretation, showing how nuanced factors like the speaker's identity and setting can alter judgments. Methodologically, it extends conjoint experiments, a technique more commonly used in the social sciences, into AI auditing, enabling a robust assessment of how multimodal models weigh each component of an online post. This pairing of innovative theory and rigorous design makes the study a strong contribution to research on content moderation.

Below, I list a set of key weaknesses and questions for authors. Addressing these points could strengthen the paper and improve the robustness of the findings.

The article investigates the classification of content as relevant for content moderation. However, social media companies are moving away from this model, as noted in the Introduction section, with many companies opting for community-centric mechanisms such as community notes (X) or feed customizations (Bluesky). Concurrently, the policy is shifting away from moderation, and companies are dismantling moderation teams. How are the results reported here relevant to a broad readership in such a changing landscape? How might alternative approaches still be similarly biased? Could the experimental design and auditing mechanisms still apply? It would be important to discuss and contextualize this in the Discussion. These concerns limit the expected level of interest a broad audience might have in this work.

A separate concern is whether the forced-choice conjoint design accurately represents real-world decision-making. To the best of my knowledge, text and images are processed separately in practice. What is the evidence in favor of the ecological validity of the conducted experiment? To what extent does forced-choice conjoint design reflect how people make decisions outside of experimental settings? It seems to me that a forced choice may not fully reflect real-world workflows.

Perhaps I missed this, but could the participants tell that the posts are not real? That could influence the interpretation of the findings. Expanding explanations on post-generation would help readers better judge the study's validity. These additions would also offer insights into potential biases arising from synthetic or artificially constrained content.

Other models (especially those trained on different data or with different alignment strategies, such as Claude or Gemini) could produce different results. Including additional models could strengthen the quality of the analysis.

In section 2, the sources of the posts are not explained in sufficient detail so that the reader can follow them without checking the methods.

Finally, although the article thoroughly lists limitations, such as focusing on a single language and the U.S. context, the evidence presented is based on a single study with < 2000 participants. Replicating the experiment with an alternative baseline design that doesn't rely on paired comparisons would strengthen the claims.

Further research should examine alternative frameworks, such as free-response or rating tasks, and test how well these models generalize under different decision rules. Similarly, hate speech spans numerous identities (e.g., religion, disability), which would be interesting to include

Other than these concerns, the data and methodology are valid, statistical analyses are appropriate, and the conclusions and data interpretation are robust, valid, and reliable. Including additional experiments with alternative baseline designs, an expanded set of models, testing the ability to detect AI content, and reflecting on the timeliness of the findings in the changing landscape in a revision would improve the paper.

Reviewer #2 (Remarks to the Author):

Paper Summary:

This paper investigates how MLLMs incorporate context when detecting hate speech via conjoint experiments. The authors evaluate two types of MLLMs, i.e., GPT-4o and QWen2, with varying sizes and architectures. Their performance is compared against human judgments (N = 1,854).

Strengths:

1. Addresses a significant issue in hate speech detection.
2. Involves a large-scale human subject experiment (N = 1,854).
3. Employs a visual conjoint design in which attributes are presented as social media posts. This method closely mirrors how individuals encounter content in moderation settings and provides a realistic assessment of MLLM performance in hate speech detection.

Weaknesses:

1. The authors use only one or two representative terms for each type of slur, such as “b***tch” for sexist, “f***ggot” for homophobic, and “n*gger” for racist language. This limited lexical representation may restrict the generalizability of their findings, as the conclusions may reflect reactions to these specific words rather than broader categories of hate speech.

2. The paper lacks clarity in explaining how contextual features were constructed, particularly how synthetic names were determined to carry strong gendered or racial associations. Moreover, the use of GANs to generate profile images is not sufficiently detailed. The authors are recommended to provide the complete set of fictional profiles used in the study to allow for further examination and verification.

3. The study evaluates only GPT-4o and Qwen2. To enhance the generalizability of the findings, the authors are suggested to include additional widely-used MLLMs, such as InstructBLIP, LLaVA, and CogVLM.

4. The paper focuses exclusively on general-purpose MLLMs, overlooking specialized models explicitly designed for hate speech detection. Given the increasing deployment of tools like OpenAI’s Moderation Endpoint, it would be valuable to include such systems in the evaluation for a more comprehensive comparison.

5. As the study presents harmful language, the authors are suggested to include a disclaimer before displaying such examples to readers, in line with ethical research practices.

Reviewer #3 (Remarks to the Author):

This review is of “Auditing Multimodal Large Language Models for Contextualized hate Speech Detection Using Conjoint Experiments” for Nature Human Behavior. This manuscript presents results comparing multiple Large Language Models’ (LLMs) performance with humans’ performance on tasks related to identifying possible hate speech. One of the motivating insights in the possibility that a multi-modal analysis may be required, as terms may or may not be hate speech depending (partly) on whether their user is an in-group or out-group member.

While Nature Human Behavior is a prominent journal with a wide readership, my overall evaluation of this manuscript is quite positive, and I believe that after revision, it would be a strong candidate for publication even in such a high-visibility venue. In part, my view is based on the fact that I anticipate this research will be of interest to a broad range of scholars in computer science, computational social science, and in social science fields including sociology, economics, and political science. The combination of conjoint analyses with LLM-based analyses is innovative and has considerable potential (and in fact, the conclusion might be stronger by providing more discussion of other possible applications).

In terms of suggestions, I think that at times, the manuscript seems to read like a long list of results, and that it could use—insofar as the evidence will support it—a more decisive, coherent summary argument. The final line of the abstract, for example, could have been written even before knowing the results. Similarly, the Conclusion leaves the reader with a variety of observations but without a broader, easily summarized takeaway.

Also, I understand that Nature Human Behavior’s format is to put the methods after the conclusion, but in this case, the format really puts the manuscript at a disadvantage. There are key questions I have about the design that only become clear after reading through the results. For example, how exactly were the randomized tweets generated? I think that this synthetic generation of social media posts is a real innovation, and so wanted to see more emphasis on it and discussion of it. At the very least, I think that the manuscript needs to briefly summarize the procedure so that readers know the rough outlines of the process before getting to the results.

As a related point, was the tweet-generating process realistic, in the sense that people’s references to slurs were related to their identities? I could envision some synthetic tweet-generating procedures that didn’t really get at the kinds of interactions between the senders’ ascribed identity and the tweet’s content that the researchers theorize must be important. Not only are there some things that read differently coming from (say) Black people versus white people, there are also things that you wouldn’t expect a person from a certain group to ever say (for example, using the first-person plural “we” with a slur). An independent randomization might generate some of these odd, unexpected tweets; a dependent randomization would require statistical adjustments to analyze. In other words, were any of the tweets self-referential (e.g. “as a Black person…”), and if so, did that require any constrained randomization? If there was no constrained randomization, were any of the profiles considered to be unusual/atypical?

Related to that point above, why adopt a forced-choice research design here? In theory, the LLMs or the respondents could say that neither or both posts contained hate speech. I know that the conjoint literature has shown that forced-choice designs can be helpful even when the actual outcome is not a forced choice (Hainmueller et al. 2015), but in this case, I wondered why the authors didn’t use a rating task given that in reality, neither or both posts might be concerning for hate speech. Given that the human annotators are drawn from Qualtrics, an online, opt-in sample provider, I also wanted more details about the sample. I appreciate the over-samples of LGBTQ and Black respondents, but is there a concern that the sample may be very highly politically engaged or in other ways different from the general population. Is it possible this sample has a higher, lower, or altogether different threshold for what constitutes possible online hate speech?

Finally, I realize that these LLMs differ in a variety of ways, but I do think that the manuscript might productively describe differences in the training data sets for the various LLMs (to the extent possible) and acknowledge this more explicitly as another potential source of variation in LLM performance.

References
Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. "Validating vignette and conjoint survey experiments against real-world behavior." Proceedings of the National Academy of Sciences 112.8 (2015): 2395-2400.

Minor Comments
Pg. 2 – “that the effects of multiple factors to be”; there appears to be a missing word here

Version 1:

Decision Letter:

Our ref: NATHUMBEHAV-25020806A

17th September 2025

Dear Dr. Davidson,

Thank you for submitting your revised manuscript "Auditing Multimodal Large Language Models for Contextualized Hate Speech Detection Using Conjoint Experiments" (NATHUMBEHAV-25020806A). It has now been seen by the original

referees and their comments are below. As you can see, the reviewers find that the paper has improved in revision. We will therefore be happy in principle to publish it in Nature Human Behaviour, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within two weeks. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

[redacted]

[redacted]
Nature Human Behaviour

Reviewer #1 (Remarks to the Author):

I thank the authors for their revisions. The revised version addresses all the previous comments. I believe the replication study and the new analyses, as well as the revised writing, significantly improve the quality of the manuscript. I have no further comments.

Reviewer #3 (Remarks to the Author):

This review is of "Auditing Multimodal Large Language Models for Contextualized Hate Speech Detection Using Conjoint Experiments" for Nature Human Behavior. Overall, I think that this manuscript reports a valuable endeavor and so support publication. I do think that the manuscript could be edited so as to clarify the key contribution as per the suggestion below, but I will leave that to the author(s) and editors.

I do appreciate efforts to address some of the points raised in the prior round of reviews. For example, while I am often skeptical of the use of online, opt-in samples such as those from Prolific, I was convinced by the case made in this manuscript that Prolific respondents are actually exactly the population that might be involved in content moderation crowdsourcing efforts.

I see no reason to doubt the main findings. Where my advice comes in is with respect to presentation: there is a huge number of findings here from different models on different outcomes, and it becomes challenging for the reader to keep track of how they all stack up. In a similar vein, the abstract seems to offer fairly broad, general statements of the findings. Likewise, the introduction takes (in my view) too long to get to core contribution here, and so has too little space to actually talk about the conjoint or what is unique about this scholarship. Thus, my recommendations are to: 1) streamline the introduction so as to provide more of a statement of the unique contribution here: 2) clarify the broad take-home points within the discussion of the empirical results, rather than just listing the wide range of AMCEs. Is it possible to characterize the broad sweep of the findings a bit more, and relate each finding to the hypotheses?

MINOR POINTS
Pg. 6 – The sentence "Such posts were also chosen less frequently…" confused me—maybe I misunderstand what "such posts" refers to?

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Response to the editors and reviewers

We thank the editors and reviewers for their careful and constructive feedback on the manuscript. We have spent the past few months conducting additional analyses and making substantial revisions to the manuscript to address the various concerns raised. We are confident that the new version has significantly improved in quality and comprehensiveness. Before our point-by-point response, we briefly summarize the main changes to the manuscript:

- In response to comments from Reviewers 1 and 2, the experiments are repeated across three new model architectures (Gemini 2.5, Gemma3, and InternVL3), raising the number of models evaluated from 5 to 14. The results are consistent with the previous version and provide more insight into differences across architectures and sizes. The main substantive change is that it becomes more apparent that the prompt variations can amplify context sensitivity but does not suppress it.
- To address a comment from Reviewer 2, the AI experiment is replicated using alternative slurs in each category. In general, the results show similar overall patterns of offensiveness and evidence of context sensitivity, although we find more variation with respect to the racism category.
- In response to feedback from Reviewers 1 and 3, the forced-choice design is justified more explicitly, and a single-task AI conjoint experiment was implemented. The results support the pairwise approach (models put excessive weight on certain slurs when evaluating posts in isolation) but show broadly consistent patterns.

- We have thoroughly rewritten much of the manuscript to incorporate these changes and other comments from the reviewers. Following feedback from Reviewer 3, we attempted to ensure that the main text is easy to read, clearly states our approach, and emphasizes the study's contributions. The text has been restructured to meet the formatting requirements described in the Revision Checklist, including the new citation style.

We now respond to each point made by the editors, followed by each reviewer. The original comments are in bold, followed by our responses. Alongside the updated manuscript, we have also included a version with tracked changes (created using latexdiff). The page numbers listed in the following responses correspond to the clean version.

Editors

1. A main concern raised by Reviewer 1 and Reviewer 2 is the MLMM selection. Both reviewers request that you test additional models, including models more tailored to detect hate speech. We agree with the reviewers and ask that you test additional models in your work.

The revised manuscript now includes three additional MLLM families, enabling us to experiment with a broader range of models. For the commercial models, we add two variants of Gemini 2.5 Flash. We add Gemma 3 and InternVL3 for the open-weights models, testing three variants of each. This increases the total number of models evaluated from 5 to 14. These represent some of the best-performing MLLMs at the time of our analysis and are all plausible options for content moderation.

These experiments required considerable computational resources and necessitated substantial revisions to ensure that the paper's main findings and overall narrative remain easy to follow,

including adjustments to how we visualized the results. Overall, these new experiments add substantial depth to the paper, enabling us to explore variation across model architectures and sizes more comprehensively, and to highlight the robustness of our main findings. For example, the Gemini models tend to perform similarly to GPT-4o in terms of their sensitivity to context, and we see similar scaling patterns in the Qwen2 and InternVL3 models. The additional models also allow us to draw more detailed insights into the effects of prompt variations, as we find a greater number of statistically significant differences for slur-identity pairs when using the context-sensitive prompt, but the number is the same for the baseline and uniform prompts.

We did not add any models specifically tailored for hate speech because there is no such model that fits the parameters of our experiment. Specifically, there are no multimodal models that can simultaneously process two images that have been adapted for hate speech detection. However, as we discuss on p. 25, we do not think this is a significant limitation since there is growing evidence that "off-the-shelf" generative AI can detect hate speech and perform other classification tasks with high accuracy.

2. Reviewer 2 raises an important point about the limited lexical representation of the terms used to classify tweets. This is an important limitation to the generalizability of the findings, which we ask you to address in full, ideally by increasing lexical representation.

We discuss the lexical representation more thoroughly in the revised manuscript, explaining why we focused on a relatively limited set of terms to ensure sufficient statistical power to detect interactions between terms and demographics (p. 19). To further assess the generalizability of our findings, we repeat our experiment, including a broader range of terms. This necessitated the construction of 360k additional posts and almost 100k new paired evaluations per model. We

observe a similar hierarchy of offense when considering these terms, although some words do not appear to have such a strong valence as more well-known terms. We show further evidence of contextual variation when considering different terms, again with some variability. Overall, while there is some sensitivity to specific terms, particularly the N-word, which is somewhat singular in its racist associations, these results show broadly consistent patterns for alternative word choices within each category.

3. Finally, please revise your work to be concise, while also providing additional technical and methodological details as requested by reviewers.

We have made significant revisions to the manuscript, including new sections and additional detail, to address the comments raised by the reviewers. While the overall length has increased, we have endeavored to be as concise as possible and have followed Reviewer 3's advice to streamline our presentation and argument. We have included new analyses in the appendix and summarized the key takeaways in the main text where possible. We hope this strikes the right balance between comprehensively addressing the feedback and producing a clear, readable manuscript. We are happy to make further edits if deemed necessary.

<u>Reviewer 1</u>

The paper investigates how multimodal large language models (MLLMs) handle hate speech detection when additional social or demographic context is available. The authors generate pairs of synthetic social media posts that vary along dimensions such as slur type, user identity (race and gender), and engagement signals, and then, using a conjoint

experiment, measure how different models decide which post violates hate speech policy. By comparing advanced MLLMs to human judgments, they show each model's sensitivity to contextual cues (e.g., whether the speaker's race matches a targeted racial slur) and potential biases in processing visual identity markers.

Their results reveal that larger, more sophisticated MLLMs often come closer to human assessments of hateful content—especially regarding reclaimed slurs, but still penalize certain demographic groups more often. Moreover, short prompt instructions (like emphasizing context or ignoring it) do little to alter how these models classify hate speech. These findings show the promise of MLLMs for more nuanced moderation while highlighting persistent racial, lexical, and other biases that require stronger methods of fine-tuning and auditing.

I believe this is a strong article with conceptual and methodological novelty. On a conceptual level, it reframes hate speech detection as a problem of interpretation, showing how nuanced factors like the speaker's identity and setting can alter judgments. Methodologically, it extends conjoint experiments, a technique more commonly used in the social sciences, into AI auditing, enabling a robust assessment of how multimodal models weigh each component of an online post. This pairing of innovative theory and rigorous design makes the study a strong contribution to research on content moderation.

We thank the reviewer for carefully reading the manuscript and for emphasizing the key contributions of the piece. This feedback was helpful as we have revised the manuscript.

Below, I list a set of key weaknesses and questions for authors. Addressing these points could strengthen the paper and improve the robustness of the findings.

The article investigates the classification of content as relevant for content moderation. However, social media companies are moving away from this model, as noted in the Introduction section, with many companies opting for community-centric mechanisms such as community notes (X) or feed customizations (Bluesky). Concurrently, the policy is shifting away from moderation, and companies are dismantling moderation teams. How are the results reported here relevant to a broad readership in such a changing landscape? How might alternative approaches still be similarly biased? Could the experimental design and auditing mechanisms still apply? It would be important to discuss and contextualize this in the Discussion. These concerns limit the expected level of interest a broad audience might have in this work.

We are glad the reviewer raised this issue. While content moderation has been in flux and new models have emerged, we are still confident that automated systems will remain an important part of the content moderation ecosystem. In the revised manuscript, we include a new paragraph in the conclusion that addresses these issues (p. 15). We emphasize the continued need for automation due to the scale of user-generated content and legal mandates, and describe how AI can assist with smaller-scale moderation efforts. There are many interesting directions here, but we have attempted to concisely address the main point—that automation will remain a core feature of content moderation—without digressing too far from our main argument.

A separate concern is whether the forced-choice conjoint design accurately represents real-

world decision-making. To the best of my knowledge, text and images are processed separately in practice. What is the evidence in favor of the ecological validity of the conducted experiment? To what extent does forced-choice conjoint design reflect how people make decisions outside of experimental settings? It seems to me that a forced choice may not fully reflect real-world workflows.

We thank the reviewer for raising this question. We agree that the forced-choice task may deviate from one-at-a-time evaluations that may be more consistent with content moderation practices, but we believe this design has strengths that make it particularly useful for AI auditing vis-à-vis alternatives. Specifically, it avoids scaling issues in unforced pairwise comparisons (i.e., always select both or neither) and single-task designs (i.e., always select specific terms, always/never select any posts). We outline the rationale for the forced-choice design on page 18 and discuss the trade-offs between different conjoint designs.

While our approach has limitations, there is also value to incorporating paired choice designs into content moderation pipelines because the forced-choice task (and other pairwise approaches, see Hainmueller et al, 2015) allows us to capture relative distinctions between different factors. For example, pairwise comparisons could prioritize the worst (or perhaps the most difficult to adjudicate) content for human review. We discuss this briefly in the conclusion (p.16).

In response to other comments below, we describe an additional experiment using a single-task framework to explore how the results vary using an alternative design (see also responses to Reviewer 3).

Perhaps I missed this, but could the participants tell that the posts are not real? That could

influence the interpretation of the findings. Expanding explanations on post-generation would help readers better judge the study's validity. These additions would also offer insights into potential biases arising from synthetic or artificially constrained content.

We endeavored to construct realistic stimuli to enhance the internal validity of the human subjects study. We have revised the manuscript to include additional details on the post-generation procedure at the beginning of the study (pp. 4-5) and have added further discussion in the Methods section (pp. 18-22).

Subjects were debriefed at the end of the experiment to inform them of the nature of the material, per IRB requirements. We reviewed open-ended feedback that participants could optionally complete at the end of the study, prior to the debrief, and we found that two respondents mentioned that the posts were artificial. This is now mentioned in a footnote. This indicates that at least some respondents were aware. It is difficult to assess how this may have impacted their responses or whether the AI models may have taken this into account. Based on the results, however, we believe that subjects took the study seriously and genuinely sought to complete the task with high precision. Indeed, a far greater number commented that the stimuli were engaging and that it was an interesting study to participate in. We discuss the limitations of synthetic stimuli in the conclusion (p. 15).

Other models (especially those trained on different data or with different alignment strategies, such as Claude or Gemini) could produce different results. Including additional models could strengthen the quality of the analysis.

We thank the reviewer for encouraging us to expand the analysis. In the revised manuscript, we have included an additional commercial model, testing two versions of Gemini 2.5 Flash, as well as two open-weights models, encompassing six new variants (Gemma3 4B, 12B, 27B, and InternVL3 2B, 14B, and 78B). We find that the larger Gemini 2.5 models perform similarly to GPT-4o and that there are notable differences across size with regard to Qwen2, InternVL3, and Gemma3. Each architecture exhibits idiosyncratic responses toward certain contextual features. Nonetheless, we emphasize that all models tested exhibit some sensitivity to context on content moderation tasks. The additional models also allow us to obtain more meaningful insights into the differences across prompts (p. 10, 13).

Overall, the additional models have deepened our understanding of how MLLMs respond to contextualized hate speech and should provide readers with a more comprehensive picture of the similarities and differences across model architectures and sizes. The revised manuscript includes substantial changes to incorporate discussion of these new models, including a more detailed evaluation of the differences across the models (p. 13) and potential explanations (pp. 16-17).

In section 2, the sources of the posts are not explained in sufficient detail so that the reader can follow them without checking the methods.

We edited the beginning of the paper (pp. 4-5) to provide a more precise explanation of how the stimuli were constructed. While the discussion is still brief, we summarize the key details and hope it gives the necessary details to understand the results without consulting the methods section.

Finally, although the article thoroughly lists limitations, such as focusing on a single language and the U.S. context, the evidence presented is based on a single study with < 2000 participants. Replicating the experiment with an alternative baseline design that doesn't rely on paired comparisons would strengthen the claims.

We agree that alternative baselines are a fruitful extension. At the same time, the original experiment was a significant undertaking, involving considerable effort and expense. The human subjects experiment cost over $10,000 and took over a year to run from approval to implementation. We made extensive efforts to ensure that a broad, reliable pool of subjects was surveyed and that the results were of high quality. The AI experiments cost several thousand dollars and required substantial technical investments to use each model.

Based on your feedback, as well as comments from Reviewer 2, we implemented a replication of the AI experiment using a single-task design. We repeat our main experiment using the same set of posts, but rather than using 30k pairs, we run 60k single-post evaluations. We repeat this experiment across all 14 models tested. The full results are reported in the Appendix, pages 73-76.

There are three main takeaways from this experiment. First, under a single-profile design, we find evidence of significant demographic variation in the main effects (see Figure A11). Second, we show identity-specific evaluations, particularly concerning reclaimed racism, consistent with the main results (see Figure A12). Third, there is less variation than in the main experiment with respect to slurs, with less evidence of any rank ordering. This happens because some models nearly always select these posts for moderation (see pp. 73-74). This also explains why we observe less evidence of sensitivity to identity when racism is evaluated. Overall, these results complement our main findings and highlight the trade-offs between different conjoint designs.

We mention these analyses in the main results section and discuss alternative conjoint designs in the conclusion (p. 16).

Further research should examine alternative frameworks, such as free-response or rating tasks, and test how well these models generalize under different decision rules.

We now discuss alternative frameworks in more detail in the Methodology (pp. 18) section and call for further work to examine these approaches in the Conclusion (p. 16).

Similarly, hate speech spans numerous identities (e.g., religion, disability), which would be interesting to include

We mention this as an extension in the conclusion (p.15-16).

Other than these concerns, the data and methodology are valid, statistical analyses are appropriate, and the conclusions and data interpretation are robust, valid, and reliable. Including additional experiments with alternative baseline designs, an expanded set of models, testing the ability to detect AI content, and reflecting on the timeliness of the findings in the changing landscape in a revision would improve the paper.

We thank the reviewer again for their comments and hope they agree that the revised manuscript adequately addresses their concerns.

<u>Reviewer 2</u>

Paper Summary:

This paper investigates how MLLMs incorporate context when detecting hate speech via conjoint experiments. The authors evaluate two types of MLLMs, i.e., GPT-4o and QWen2, with varying sizes and architectures. Their performance is compared against human judgments (N = 1,854).

Strengths:

1. Addresses a significant issue in hate speech detection.

2. Involves a large-scale human subject experiment (N = 1,854).

3. Employs a visual conjoint design in which attributes are presented as social media posts. This method closely mirrors how individuals encounter content in moderation settings and provides a realistic assessment of MLLM performance in hate speech detection.

We thank the reviewer for their evaluation and for identifying the key strengths of our contribution. We have carefully reviewed the various weaknesses identified below and have made significant revisions to the manuscript. We hope you find it a substantial improvement upon the previous version.

Weaknesses:

1. The authors use only one or two representative terms for each type of slur, such as "b*tch" for sexist, "f***ggot" for homophobic, and "n*gger" for racist language. This limited lexical representation may restrict the generalizability of their findings, as the conclusions may reflect reactions to these specific words rather than broader categories of hate speech.**

We focus on a relatively small set of terms in order to obtain sufficient statistical power to analyze interactions with identity. If we had included a substantially larger array of terms, then we would need to conduct many more comparisons in order to have a sufficiently large number of observations in each cell (term-identity pairs). We based our AI experiment on the human subjects study, where we were constrained by participants' time and our resources. However, we agree with the reviewer's point that greater lexical representation would enhance the generalizability of our findings.

To address this issue, we reviewed the literature on offensive speech to identify additional terms in each category. With the exception of reclaimed slurs—for which the term "n***a" is somewhat unique (although reclamation has been observed in terms of other language like sexism or homophobia)—we identified new terms to test for each category. For racism, we include two additional anti-Black terms and two words targeting Asians and Hispanics, respectively. For reverse racism, we include two terms that can be used to insult White people. For homophobia and sexism, we also include two additional words.

Using these new terms, we then constructed a new set of synthetic posts that included the new vocabulary (380,000 posts). We then combined this with the original set of posts (210,000) and sampled from this corpus at the same ratio as the original study to obtain a new set of 99,750 paired comparisons, which we ran through all models, including those newly added.

The results from this analysis are reported in Appendix A5 (pp. 63-69). In general, we find consistent patterns of offensiveness, as shown in Figure A4. Racist and homophobic language tends to be selected more frequently than other terms, followed by reverse racism, and then sexism. The same patterns emerge with respect to the models. For example, the smallest 2B

models tend to put a low weight on most terms, whereas the largest commercial models tend to select them most often.

There is some variation across different terms. With respect to sexism, reverse racism, and homophobia, we find similar patterns across alternative words. There are some minor discrepancies. For example, we see that the two additional sexist terms tend to be selected more often than “b*tch”, indicating that they are considered more offensive. There is slightly more variation when it comes to racism. For the anti-Black terms, we find that one alternative, “c**n”, is often selected at about the same rate as “n*gga”, whereas “spook” shows more variation. We expect this is because it is a less frequently used slur. Morover, the N-word is singular in its racist connotations and recognition, so it is perhaps unsurprising that these terms do not have such a strong valence. For the additional racist terms, we see that they are generally selected at the same rate as other racist and homophobic language by the commercial models, although there is more variation among the open-weight models.

When evaluating how the results vary depending on the speaker (Figures A5-7), we still observe some statistically significant differences. Consistent with the main results, some models are particularly prone to penalizing Black users for alternative types of reverse racism, and there is little evidence of responsiveness to gender when evaluating alternative sexist epithets. We more some mixed patterns when evaluating racist language. Some larger models are less likely to select Black users who use the term “c**n”, but Black users are disproportionately selected by larger InternVL3. For “spook”, we observe that several models show a tendency to choose Black users and avoid White users. This further indicates that this word may be misunderstood by these models. For the two additional racist terms, we observe some disparities, showing sensitivity to user identity among some models, but we do not systematically vary the identity with respect to

either target groups, and as such, no clear directional patterns are present. These new results deepen our analyses, complementing our main results and highlighting heterogeneity within categories of hate speech. We mention these findings where relevant in the main results section.

In the conclusion, we discuss extensions to this study and how our conjoint approach could be used to evaluate a broader array of hateful and offensive language (p. 15-16).

2. The paper lacks clarity in explaining how contextual features were constructed, particularly how synthetic names were determined to carry strong gendered or racial associations. Moreover, the use of GANs to generate profile images is not sufficiently detailed. The authors are recommended to provide the complete set of fictional profiles used in the study to allow for further examination and verification.

We have clarified the discussion on these issues in the front-end of the paper (pp. 4-5) and in the Methodology section (pp. 20-22). The names were from a 2017 study by Gaddis (ref. 61) that measured the racial perceptions of different names, and the GAN photos were obtained by filtering a dataset of generated images provided by a third-party company for academic research (Generated Photos, Inc.). The full set of profile images used is shown in section A.1.1 of the Appendix (pp. 48), and the text templates are described and shown in full in section A.1.2 (pp. 49-56).

3. The study evaluates only GPT-4o and Qwen2. To enhance the generalizability of the findings, the authors are suggested to include additional widely-used MLLMs, such as InstructBLIP, LLaVA, and CogVLM.

We thank the reviewer for encouraging us to include additional models and further consider alternative MLLM architectures. We reviewed the literature on MLLMs and the models that were available on Github and HuggingFace. It is only in the past couple of years that this kind of study has been possible due to the challenges involved with tokenization and representing images in memory-efficient ways (some of the models listed above can only process a single image). Based on our review of the available models, we added two new open-weights models that had the requisite capabilities, Gemma3 and InternVL3, two of the most powerful and versatile models available. We selected InternVL3 because it outperforms Qwen2VL on benchmark tests at each size tested (see https://internvl.github.io/blog/2025-04-11-InternVL-3.0/), and the largest versions are competitive with GPT-4o. We selected Gemma 3 because it is developed using a different approach to Qwen2 and InternVL3 and outperforms an updated version of Qwen on some metrics (https://arxiv.org/pdf/2503.19786, see Table 5). We also added two variants of Gemini 2.5 Flash to include another model that is more comparable to GPT-4o. By adding these models, we show our approach generalizes to multiple architectures and provides further insight into the differences and similarities between them. The models are described in the Methodology section (pp. 25-27) and summarized in Table 2.

We now devote more space to describing how the results vary across architectures and sizs in the Discussion (pp.12) and consider possible explanations in the Conclusion (pp. 16-17).

The revised manuscript also includes some further discussion of advances in MLLMs, including references to some of the architectures mentioned in your comment. See p. 5 and p. 25.

4. The paper focuses exclusively on general-purpose MLLMs, overlooking specialized models explicitly designed for hate speech detection. Given the increasing deployment of

tools like OpenAI's Moderation Endpoint, it would be valuable to include such systems in the evaluation for a more comprehensive comparison.

We agree that evaluating more specialized systems would be interesting and could easily be performed using this methodology. However, at the time of writing, there was no specialized model that met our requirements. Specifically, we could not identify any models that were multimodal, trained for hate speech detection or a related moderation task, and that could take multiple images and a prompt as input. We had explored the OpenAI Moderation Endpoint previously, but found that the hate speech detection capability is limited to text (it is capable of recognizing other categories like sexual content and violence in images, see https://platform.openai.com/docs/guides/moderation). To verify this, we passed a sample of images through the endpoint, and none were flagged as hate speech. In the revised manuscript, we discuss this issue and make the case for why our evaluation of general-purpose models is important, given evidence that these models are effective at generalizing to many tasks, including hate speech detection (pp. 25).

5. As the study presents harmful language, the authors are suggested to include a disclaimer before displaying such examples to readers, in line with ethical research practices.

We endeavored to avoid using harmful language in the study, censoring offending terms in the main text, but we agree that a disclaimer is helpful. We have now included a disclaimer at the beginning of the manuscript and include a similar warning in the appendix when we discuss alternative slurs, since we do not censor the terms there, because it makes it difficult to explain exactly what stimuli were used. If the article is accepted, we will consult with the editors on the

most appropriate placement.

Reviewer 3

This review is of "Auditing Multimodal Large Language Models for Contextualized hate Speech Detection Using Conjoint Experiments" for Nature Human Behavior. This manuscript presents results comparing multiple Large Language Models' (LLMs) performance with humans' performance on tasks related to identifying possible hate speech. One of the motivating insights in the possibility that a multi-modal analysis may be required, as terms may or may not be hate speech depending (partly) on whether their user is an in-group or out-group member.

While Nature Human Behavior is a prominent journal with a wide readership, my overall evaluation of this manuscript is quite positive, and I believe that after revision, it would be a strong candidate for publication even in such a high-visibility venue. In part, my view is based on the fact that I anticipate this research will be of interest to a broad range of scholars in computer science, computational social science, and in social science fields including sociology, economics, and political science. The combination of conjoint analyses with LLM-based analyses is innovative and has considerable potential (and in fact, the conclusion might be stronger by providing more discussion of other possible applications).

We appreciate the thorough review and your assessment of the contributions of the manuscript. Your overview has helped us to think through how to best frame the study, and we now mention further applications of the approach, beyond content moderation, at the beginning of the conclusion (p. 14) and allude to the generalizability of the methodology in the abstract and

introduction.

In terms of suggestions, I think that at times, the manuscript seems to read like a long list of results, and that it could use—insofar as the evidence will support it—a more decisive, coherent summary argument. The final line of the abstract, for example, could have been written even before knowing the results. Similarly, the Conclusion leaves the reader with a variety of observations but without a broader, easily summarized takeaway.

We have revised the manuscript with an eye toward providing a more coherent overview of our contributions and streamlining the presentation of the results. This was certainly challenging given the many findings and the new additions to the paper, but we hope that the revised version is now more readable and provides clearer takeaways. Please note that the formatting has also changed in order to meet the journal's requirements for revisions.

Also, I understand that Nature Human Behavior's format is to put the methods after the conclusion, but in this case, the format really puts the manuscript at a disadvantage. There are key questions I have about the design that only become clear after reading through the results. For example, how exactly were the randomized tweets generated? I think that this synthetic generation of social media posts is a real innovation, and so wanted to see more emphasis on it and discussion of it. At the very least, I think that the manuscript needs to briefly summarize the procedure so that readers know the rough outlines of the process before getting to the results.

We understand the concerns regarding the limitations of the article format. The revised manuscript now includes a more detailed overview of the procedure for creating synthetic posts

in the Introduction (pp. 4-5) and further details, including technical aspects of the process, in the Methodology section (pp. 20-22). We hope this strikes the right balance between providing the necessary information to contextualize the results while avoiding excessive detail in the front end. The appendix now includes the full set of stimuli used to produce the posts (pp. 48-56)

As a related point, was the tweet-generating process realistic, in the sense that people's references to slurs were related to their identities? I could envision some synthetic tweet-generating procedures that didn't really get at the kinds of interactions between the senders' ascribed identity and the tweet's content that the researchers theorize must be important. Not only are there some things that read differently coming from (say) Black people versus white people, there are also things that you wouldn't expect a person from a certain group to ever say (for example, using the first-person plural "we" with a slur). An independent randomization might generate some of these odd, unexpected tweets; a dependent randomization would require statistical adjustments to analyze. In other words, were any of the tweets self-referential (e.g. "as a Black person…"), and if so, did that require any constrained randomization? If there was no constrained randomization, were any of the profiles considered to be unusual/atypical?

This is an important point and something that we considered carefully when designing the study. As the reviewer points out, conjoint studies can often lead to improbable combinations of attributions that researchers must decide whether to include (sacrificing some realism) or omit (inducing some constraints on the randomization and requiring statistical adjustment). A canonical example is that varying education and occupation can lead to impossibilities like a doctor without a college degree (Hainmueller et al. 2014). We sought to design the templates in a

way that would avoid any illogical texts. As such, none of the templates ever includes self-referential statements of the sort mentioned above. We now mention this on pages 19-20. Every template is constructed in a way that any of the slurs and potentially a curse word can be included, with code to ensure appropriate pluralization where necessary. Of course, there are some statements that might appear more or less plausible. For example, one might expect men to be more likely to use homophobic epithets when discussing sports than women, but there are no constructions that we consider to be impossible. The full set of templates is now provided in Appendix A1, pp. 48-56, along with a more detailed description of how the templates are converted into the final posts.

Related to that point above, why adopt a forced-choice research design here? In theory, the LLMs or the respondents could say that neither or both posts contained hate speech. I know that the conjoint literature has shown that forced-choice designs can be helpful even when the actual outcome is not a forced choice (Hainmueller et al. 2015), but in this case, I wondered why the authors didn't use a rating task given that in reality, neither or both posts might be concerning for hate speech.

We thank the reviewer for raising this question. We decided to use a forced-choice design because it avoids scale use issues that can arise with rating tasks (e.g., always select both or neither, always select high or low values) and because we wanted to demonstrate how the most popular form of conjoint design can be helpful for AI auditing. We decided against a single-profile task for similar reasons (e.g. always select posts, never select posts). Nonetheless, we recognize that other designs could reveal useful insights into model behavior.

To address this issue, we made two major changes to the manuscript. First, the methodology section now includes additional discussion of alternative designs and justification for the forced choice approach (pp. 18). We explain the scenario mentioned above, noting how situations where a respondent would select both or neither posts if given the opportunity will contribute to the variance when using a forced-choice design but will not bias the estimates. Other designs are also mentioned as extensions in the conclusion when discussing limitations (p. 16). Second, we implemented an additional experiment using a single-task design, described in Appendix A7 (p. 73-76). See our response to Reviewer 1 for further detail.

We hope these revisions provide better clarity regarding our choice of design and its merits, while making it clear to readers that there are alternative choices that could be fruitful avenues for future research.

Given that the human annotators are drawn from Qualtrics, an online, opt-in sample provider, I also wanted more details about the sample. I appreciate the over-samples of LGBTQ and Black respondents, but is there a concern that the sample may be very highly politically engaged or in other ways different from the general population. Is it possible this sample has a higher, lower, or altogether different threshold for what constitutes possible online hate speech?

The revised manuscript includes additional discussion to address these points. Before explaining the changes, we should clarify that the participants were recruited from Prolific and then sent to Qualtrics to complete the study. We have ensured this is now clear in the manuscript.

Regarding political engagement, our study includes a broad cross-section of the population, stratified by ideology. As shown in Table A1 (p. 59), the sample is made up of a mixture of self-identified conservatives (36.9%), liberals (48.5%), and moderates (14.4%). We do not have reason to believe that the subjects significantly differ in political engagement from the general population. Regarding hate speech specifically, the revised manuscript now includes a discussion of some relevant information obtained from a short survey conducted after the conjoint task (p. 23). While all subjects had opted into studies featuring sensitive content (pp. 57-58), there is heterogeneity in experiences with hate speech and beliefs about it. Around a quarter of subjects reported frequently encountering hate speech, and one-fifth reported being targeted. Consistent with existing surveys (e.g., https://www.pewresearch.org/short-reads/2022/08/30/more-so-than-adults-u-s-teens-value-people-feeling-safe-online-over-being-able-to-speak-freely/), most liberals thought that online hate speech was often downplayed ("excused as not a big deal"). In contrast, more than half of conservatives considered it to be "taken too seriously". While we cannot precisely measure the threshold for what constitutes hate speech compared to the general population, our sample captures a broad cross-section of the public with a variety of views about hate speech.

Finally, I realize that these LLMs differ in a variety of ways, but I do think that the manuscript might productively describe differences in the training data sets for the various LLMs (to the extent possible) and acknowledge this more explicitly as another potential source of variation in LLM performance.

While we cannot determine the precise causes, the additional models in our analysis give us more leverage on these questions. In the revised version, we discuss differences between the models in more depth, drawing attention to differences in architecture and size (pp. 13, 16-17).

The conclusion includes a paragraph discussing the possible reasons for these discrepancies, including architecture, pre-training data, and post-training alignment procedures (p. 16-17). While we agree that it would be illuminating to understand whether specific training datasets contributed to the observed findings, each model is trained on various text and image data, including pre-training materials scraped from the internet and various fine-tuning datasets. These datasets tend to be described in broad terms, making it difficult to draw any meaningful inferences about specific datasets. For example, the Qwen2-VL documentation states that "The model is pre-trained on a diverse dataset that includes image-text pairs, OCR data, interleaved image-text articles, visual QA datasets, video dialogues, and image-knowledge datasets. Our data sources primarily comprise cleaned web pages, open-source datasets, and synthetic data." Among the models tested, only InternVL-3 mentions specific datasets that were used for the instruction-tuning process (the underlying base LLM is Qwen 2.5, which does not provide precise details on training data, see https://arxiv.org/abs/2412.15115). As such, we cannot conclude how any particular pre- or post-training dataset impacted the performance. Furthermore, even if information on training data were available, there is no clear procedure to determine precisely which dataset may have contributed to any outcome or more general pattern. The manuscript emphasizes the need for further work to address these issues.

References

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. "Validating vignette and

conjoint survey experiments against real-world behavior." Proceedings of the National Academy of Sciences 112.8 (2015): 2395-2400.

We did not cite this paper in the original manuscript, but had read it when developing the study. We thank the reviewer for bringing this to our attention, and we now discuss this paper in several places where it is germane to the argument.

Minor Comments

Pg. 2 – "that the effects of multiple factors to be"; there appears to be a missing word here

This has been fixed.

Response to the reviewers

Dear Editors,

This memo outlines the revisions made to the manuscript to address the remaining issues raised by the reviewers, specifically Reviewer 3's recommendations for improving the presentation. The manuscript has also been edited in accordance with the guidance provided in the Author Checklist (see Author Checklist for further details). Overall, I appreciate the reviewers' helpful comments. These final revisions have strengthened the manuscript by making it more readable and placing greater emphasis on the study's core contributions.

The reviewers' comments are shown in bold, followed by my comments (where necessary).

Reviewer #1:
Remarks to the Author:
I thank the authors for their revisions. The revised version addresses all the previous comments. I believe the replication study and the new analyses, as well as the revised writing, significantly improve the quality of the manuscript. I have no further comments.

Reviewer #3:
Remarks to the Author:
This review is of "Auditing Multimodal Large Language Models for Contextualized Hate Speech Detection Using Conjoint Experiments" for Nature Human Behavior. Overall, I think that this manuscript reports a valuable endeavor and so support publication. I do think that the manuscript could be edited so as to clarify the key contribution as per the suggestion below, but I will leave that to the author(s) and editors.

I do appreciate efforts to address some of the points raised in the prior round of reviews. For example, while I am often skeptical of the use of online, opt-in samples such as those from Prolific, I was convinced by the case made in this manuscript that Prolific respondents are actually exactly the population that might be involved in content moderation crowdsourcing efforts.

I see no reason to doubt the main findings. Where my advice comes in is with respect to presentation: there is a huge number of findings here from different models on different outcomes, and it becomes challenging for the reader to keep track of how they all stack up. In a similar vein, the abstract seems to offer fairly broad, general statements of the findings. Likewise, the introduction takes (in my view) too long to get to core contribution here, and so has too little space to actually talk about the conjoint or what is unique about this scholarship. Thus, my recommendations are to: 1) streamline the introduction so as to provide more of a statement of the unique contribution here: 2) clarify the broad take-home points within the discussion of the empirical results, rather than just listing the wide

range of AMCEs. Is it possible to characterize the broad sweep of the findings a bit more, and relate each finding to the hypotheses?

The following changes have been made to address the presentation issues raised by the reviewer, which I list in chronological order:

- The abstract has been revised to ensure that it accurately conveys the findings.
- The introduction has been streamlined to make it more concise and to center the core contribution of the study. This involved combining certain paragraphs into a simpler structure and condensing some of the discussion on related work. A new paragraph has been added to the end of the first section to outline the study's contributions, summarizing the approach and previewing the results. The opening paragraph has also been reworked to better introduce the problem and explain how hate speech is defined.
- Several sentences have been added to the results section to help contextualize the findings and guide the reader through the findings.
- To better synthesize the contributions of the study, the discussion and conclusion have been merged into a single discussion section (by removing the Conclusion section heading). The paragraph on the current debates in content moderation has been moved to the end of the section and merged with the final concluding paragraph. Some material about how content moderation is changing at Meta and X (originally in the introduction) has also been moved here. This structure flows more smoothly and eliminates any repetitiveness.
- The entire manuscript has been proofread and edited for clarity and conciseness.

MINOR POINTS

Pg. 6 – The sentence "Such posts were also chosen less frequently…" confused me—maybe I misunderstand what "such posts" refers to?

This section has been edited to ensure that it is clear to the reader. The relevant sentence now reads: "Starting with slurs, shown in the top left, human evaluators and most models were considerably more likely to prioritize posts containing identity-based slurs for moderation compared to those using generic insults. In contrast, posts without slurs were chosen less frequently."