AUDITING MULTIMODAL LARGE LANGUAGE MODELS FOR CONTEXT-AWARE CONTENT MODERATION

Thomas Davidson
Department of Sociology
Rutgers University–New Brunswick

October 2024

ABSTRACT

Multimodal generative artificial intelligence models that can process text, images, and other media provide a context-aware approach to content moderation that may help to address existing biases in these systems. This study proposes conjoint analysis as a methodology to audit these models. We investigate how GPT-4o, a state-of-the-art model, uses demographics and other contextual information when flagging potential hate speech. The results show that the model attends to context similarly to human annotators. In particular, the race of the author is taken into account when deciding whether the use of racist language or reclaimed slurs violates a content policy. Prompting can enhance contextual awareness, although some demographic bias persists. Models relying solely on visual or written identity signals show significantly less contextual variation, highlighting the value of multimodal approaches to automated content moderation. These findings have significant implications for designing more effective moderation systems and methodologies for auditing algorithmic systems.

Acknowledgments. This research was supported by a Foundational Integrity Research award from Meta and the OpenAI Researcher Access Program. Earlier versions of this work were presented at the Rutger Sociology Culture Workshop, IC2S2 at the University of Pennsylvania, and the Trust & Safety Conference at Stanford University. The human subjects experiment was approved by the Rutgers University Institutional Review Board (#Pro2023002017 & #Mod2024000438). I thank Fred Traylor for assistance with the implementation of the experiment in Qualtrics. Please contact Thomas Davidson (thomas.davidson@rutgers.edu) with any questions or comments.

1 Introduction

Automated content moderation systems are now deployed at a global scale to flag content such as misinformation, child sexual abuse material, and hate speech on social media platforms and other websites (Gillespie, 2018, Roberts, 2019, Kaye, 2019). These systems enable platforms to address these problems at scale, but there is evidence that they can also perpetuate harmful biases. For example, hate speech and abusive language detection classifiers are more likely to flag texts written in African-American dialects (Sap et al., 2019, Davidson et al., 2019). These biases in conventional

machine learning systems stem from stereotypical decision-making by human raters that becomes embedded in models trained on these data (Sap et al., 2022, Davani et al., 2023), consistent with research in social psychology showing how social identities and attitudes mediate perceptions of hate speech (Cowan and Hodge, 1996, Leets, 2001, Roussos and Dovidio, 2018). Such biases can not only result in negative impacts to members of marginalized groups that content moderation systems are designed to detect but can spill over to other group members, leading to a decreased sense of belonging and the view that platforms are divisive (Lee et al., 2024).

Existing approaches to automatically detecting offensive content often consider a single text in isolation, ignoring information that could help both automated systems and human content moderators to better understand and contextualize speech (Pavlopoulos et al., 2020, Xenos et al., 2022, Ljubešić et al., 2022, Zhou et al., 2023). New advances in generative artificial intelligence (AI) open up opportunities for more context-aware forms of automated content moderation. State-of-the-art multimodal large language models (MLLMs) can serve as the foundation for a diverse array of applications (Bommasani et al., 2022) and are capable of processing rich inputs that include contextual information (Zhang et al., 2023, Zhou et al., 2023) as well as other salient non-textual information like profile images or memes (Kiela et al., 2020). However, these technologies have several limitations that could hinder their use in content moderation and even exacerbate the issue of bias. Since they are trained using large amounts of data from the internet, generative AI is also permeated by persistent biases that can shape their outputs (Dixon et al., 2018, Bender et al., 2021, Kirk et al., 2021, Bianchi et al., 2023, Eloundou et al., 2024). These models can also generate "toxic" outputs, even from innocuous prompts (Gehman et al., 2020). For example, GPT-3 often produced stereotypes evoking violence and terrorism when prompts included the term "Muslim" (Abid et al., 2021) and GPT-4 and other LLMs can output racist tropes when prompted using texts written in African-American dialect (Hofmann et al., 2024). Technical efforts undertaken to help "align" LLMs and other AI models to mitigate these issues have shown some success (Ziegler et al., 2020, Ouyang et al., 2022, Ji et al., 2023), but biases cannot be completely eliminated, and some may be superficially concealed (Gonen and Goldberg, 2019, Hofmann et al., 2024). Moreover, alignment can result in overzealous moderation actions that could further perpetuate biases (Röttger et al., 2024). For example, advanced models often mistake "counterspeech" or disclosures about racial discrimination as toxic, resulting in false positives consistent with those observed in human moderation decisions (Gligorić et al., 2024, Lee et al., 2024).

This study builds upon this work by examining how rich, multimodal contextual information shapes the way that online hate speech is classified by both humans and AI systems. The study proposes using conjoint experiments to audit multimodal large language models. The key advantage of conjoint studies is that they allow many attributes to be manipulated simultaneously and their effects to be quantified, net of other attributes (Hainmueller et al., 2014). This methodology is typically used in human-subjects experiments and has shown promise as a way to understand content moderation decisions (Rasmussen, 2022, Kozyreva et al., 2023, Pradel et al., 2024). This technique complements existing approaches to algorithmic fairness that consider counterfactuals, such as when demographics are changed or obscured (Kusner et al., 2017, Kleinberg et al., 2018), and ablation studies where researchers vary one or more attributes and compare differences in predictions (e.g. (Dixon et al., 2018, Buolamwini and Gebru, 2018, Röttger et al., 2021)).

Using vignettes representing synthetic social media posts, the experiment analyzes the extent to which GPT-4o, a state-of-the-art multimodal model, relies upon contextual information to evaluate hate speech, how the importance of this information varies as a function of the type of speech and the extent to which the model mirrors decisions by human annotators. Two example posts are shown in Figure 1. The posts are created using a combination of templates (Dixon et al., 2018, Röttger et al., 2021) and synthetic, AI-generated texts and images (Hartvigsen et al., 2022, Nightingale and Farid, 2022). The demographic identities of the authors are signaled by using profile images and names to connote race and sex/gender (Gaddis, 2017). The primary treatment consists of whether or not a post contains a slur and, if so, the target group associated with the slur. In particular, we are interested in how the authors' identities moderate how slur usage is evaluated. The vignettes also include a range of additional linguistic and contextual information. Posts vary across five different topics, with fifty unique posts, and some contain an additional curse word. They can also include replies that endorse or oppose the original post and varying levels of engagement. The conjoint design allows the effect of each factor to be disentangled to assess how they affect content moderation decisions.

Studies on human subjects show that instructions can help to reduce bias in content moderation decisions (Sap et al., 2019, Lee et al., 2024). Generative models can similarly be guided by using prompt engineering to steer the model. Existing studies show how prompting can reduce bias in generated texts (Abid et al., 2021, Eloundou et al., 2024) and false positives in classification settings (Gligorić et al., 2024). We test three prompt variations to assess the extent to which additional instructions can be used to tailor how automated content moderation systems use context. A baseline

Figure 1: **Example posts**
These images represent two of the 210,000 unique posts used for this study. The two vignettes vary across all attributes: slur and curse word usage, author identity, topic, replies, and engagement. The black rectangle censors a slur but was not used in the experiment.

prompt, which asks for an evaluation with respect to a hate speech policy, is compared against prompts that either emphasize the importance of relevant contextual considerations — noting how the identity of the author can matter and how marginalized groups can reclaim slurs — or direct the model to enforce a uniform policy that ignores context. For each prompt, thirty-thousand randomly sampled pairs of synthetic posts were evaluated to calculate the effect of the attributes on the decision to select a post for further moderation. In each case, the model is prompted to select the post that should be prioritized for review based on the policy. These results are benchmarked against an experiment where human subjects ($N = 1854$) were asked to perform the same task.

We find that AI and humans agree on the relative harms of various slurs, finding racist and homophobic content most likely to violate the policy, although the AI models put greater weight on these slurs than humans. Most importantly, our findings demonstrate that contextual factors significantly influence human and AI evaluations, often in remarkably similar ways. Both were less likely to penalize people for using reclaimed slurs when there is evidence that the person belongs to a targeted group. This effect persists regardless of the prompt used, although we see the most significant demographic differences when the model is explicitly prompted to consider the context. We find a similar pattern when considering a racist slur, although the model only makes an exception for Black male users, indicating the presence of intersectional disparities. Other contextual information also impacts how humans and AI make decisions. Notably, compared to human annotators, posts with negative replies from other users and content about politics are much more likely to be selected by AI. Finally, we examine how AI uses different identity cues, finding that multimodal cues, where both names and images signal race and gender, enhance the model's ability to contextualize hate speech. When identity is only indicated via a single modality, the AI models become less attuned to identity-based differences, reinforcing the importance of multidimensional identity cues

in improving fairness and accuracy in automated content moderation systems. Taken together, our results provide new insights into the use of multimodal AI for context-aware content moderation and highlight the utility of conjoint designs as an approach to auditing advanced AI models.

2 Results

2.1 How context affects content moderation decisions

We quantify the probability that a vignette with a given value of an attribute is chosen relative to a vignette at the reference level by calculating the Average Marginal Component Effect (AMCE). An AMCE equal to zero implies no difference from random selection. Figure 2 shows the AMCEs for each attribute. Due to substantial differences in effect sizes across attributes, we show each in a separate subplot. Starting with the top left result in Panel (a), we see that posts using a slur have an AMCE of around 0.45 across all three prompts. This implies that they are considerably more likely to be selected as violating the hate speech policy than vignettes with no slur. Slurs have a similar effect on human subjects. The results in Panel (b), which focus on the subset of conjoint profiles where a slur was present, show evidence of heterogeneity depending on the slur used. There is a consistent pattern across both the AI and human subjects: relative to the use of a generic insult, posts containing racist language have the highest likelihood of selection, followed by homophobia, reclaimed slurs, reverse racism, and, finally, sexism. Notably, there are some significant differences between humans and AI: GPT-4o flags posts containing racism at a much higher frequency than humans and is significantly more likely to flag posts containing reclaimed slurs and homophobia. The model using the context-aware prompt is marginally less likely to flag posts containing a reclaimed slur, but otherwise, the patterns are consistent across all three prompts. Returning to Panel (a), the results show that posts containing a curse word were more likely to be selected than those without a curse, although the effect size is an order of magnitude lower than that of slur usage. In this case, the effect is similar across the AI models and human subjects.

For the identity attributes, posts by anonymous users, where no demographic characteristics are revealed by either the picture or the name, are used as the reference category. We find differences according to both the perceived race and gender of the user. The model using the baseline prompt was more likely to select posts by White users, although both confidence intervals are consistent with effects close to zero. This model also appears less likely to select posts by Black females, although the coefficient is not statistically significant at the conventional $p < 0.05$ level. The

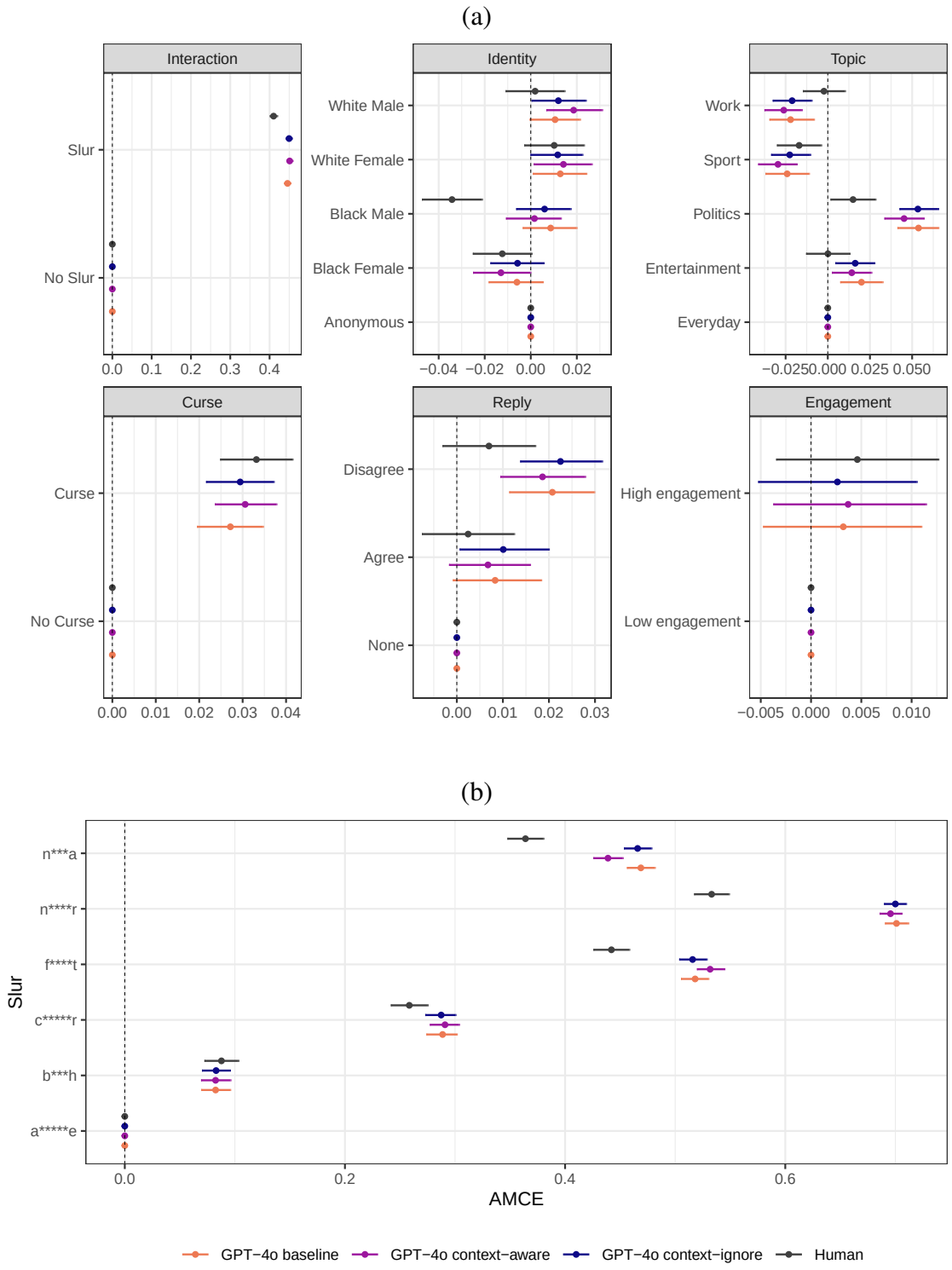

Figure 2: **Effects of (a) post attributes and (b) slurs on the probability of moderation.**
This figure shows the AMCE for each attribute in the conjoint analysis. An estimate is shown for the three GPT-4o variations and the human subject experiment for each attribute level. One level is used as the reference category and is always set to zero for each attribute. The AMCEs for slurs in Panel (b) are calculated using the subset of profiles where a slur was included. Error bars are 95% confidence intervals: the GPT-4o results use bootstrap confidence intervals, and the human experiment results include subject-level clustered standard errors.

variant with the context-aware prompt shows slightly larger contextual effects, as it is more likely to flag posts by White users and less likely to select those by Black females than anonymous users. Despite being instructed to ignore context, the context-ignore model was also marginally more likely to choose posts by White users. None of the three models showed any evidence of statistically significant differences regarding Black male users. In contrast, human subjects were significantly less likely to select posts by Black male users, and the estimate for Black female users is also negative, although it is not statistically significant.

Turning to the post topic, where everyday situations are the reference group, we see that all AI models are less likely to choose posts about sports and the workplace. In contrast, posts about entertainment and politics are more likely to be selected. The disparity is particularly large with respect to politics, suggesting that GPT-4o may find political conversations more likely to be policy-violating. Human subjects were also less likely to select posts about sports and more likely to select political content, but the effects are smaller in comparison, particularly for politics.

Regarding replies and engagement, the results indicate that the presence of replies that disagree with the original post is associated with a higher chance that a post is selected by all three models. The context-ignore model was also more likely to choose posts with replies that express agreement with the original poster. The results from the human subjects experiment are not statistically significant for either reply type. Finally, there is no evidence that the level of engagement, denoted by the number of likes on each post, mattered for either GPT-4o or human subjects.

2.2 Interactions between slur use and author identities

The results so far show that the use of slurs and the author's identity contribute to the probability that a given post is selected as in violation of the hate speech policy by AI and humans. To examine how the perception of slurs varies as a function of the author, Figure 3 shows estimates for the difference in AMCE for each slur by the identity of the author, where anonymous authors are the reference group. Starting with the two N-word variations at the top of the figure, we see clear differences depending on the race of the user: across all three models and human subjects, posts by Black users are less likely to be selected than those by anonymous users when a reclaimed slur is used. This provides robust evidence that the models are attuned to contextual variation in the use of offensive language, and in particular, the idea that there is a difference in meaning and norms when members of marginalized communities use reclaimed slurs. For the human subjects, we also see evidence that White users are more likely to be selected when using the term than anonymous users, showing that

user identity moderates perceptions of policy violation in both directions. The estimates for all three GPT-4o models are trending in the same direction, but the difference is only statistically significant for White male users evaluated using the context-aware prompt, indicating that more information about the context can help to push the model in the direction of human decisions. When assessing the alternative spelling of the term, which is used as a racist slur and is less likely to be used in a reclaimed sense, we see some similar patterns, albeit to a lesser extent. Human subjects were again less likely to choose posts when the author presented as Black and were more likely to penalize Whites. The baseline and context-aware GPT-4o models are significantly less likely to select posts by Black men, but the difference for Black women is not statistically significant, and there is no evidence that White users are selected at a higher rate than anonymous users.

Turning to the other slurs, there is no evidence of any clear patterns when homophobic language is used, but no cues related to sexuality were provided, so this indicates that neither human nor AI moderators show a demographic bias in the absence of salient information. For "reverse racism," there is some evidence that White males are less likely to be selected by both the context-ignore model and human subjects. To some extent, this mirrors the finding for reclaimed slurs, showing how members of the in-group can use a potentially derogatory term with a lower chance of being penalized. Finally, there is no evidence of any demographic differences in the use of sexist language despite the presence of sex-based identity cues.

2.3 Comparing visual and textual identity cues

These results provide robust evidence that multimodal context shapes how humans and AI systems evaluate hate speech and offensive content. However, given that identity is signaled by pictures and names, it is unclear if AI models only attend to one or other of these cues. To ascertain the extent to which these different cues matter, we ran two variations of the experiment using the baseline prompt. First, we evaluate posts where all users have anonymous pictures that do not indicate race or gender. Second, we repeat the experiment using neutral, two initial names with the original pictures.[1] The main results are broadly consistent (subsection A.4), although neither model shows any statistically significant difference in selection as a function of user identity.Moreover, the results in Figure 4 show that absent one of the identity cues, the models are less attuned to racial differences related to reclaimed slurs. This indicates how multidimensional identity cues can aid multimodal models in performing context-aware automated content moderation. Only the model with a neutral name (and a synthetic face) shows a statistically significant difference between Black males and anonymous

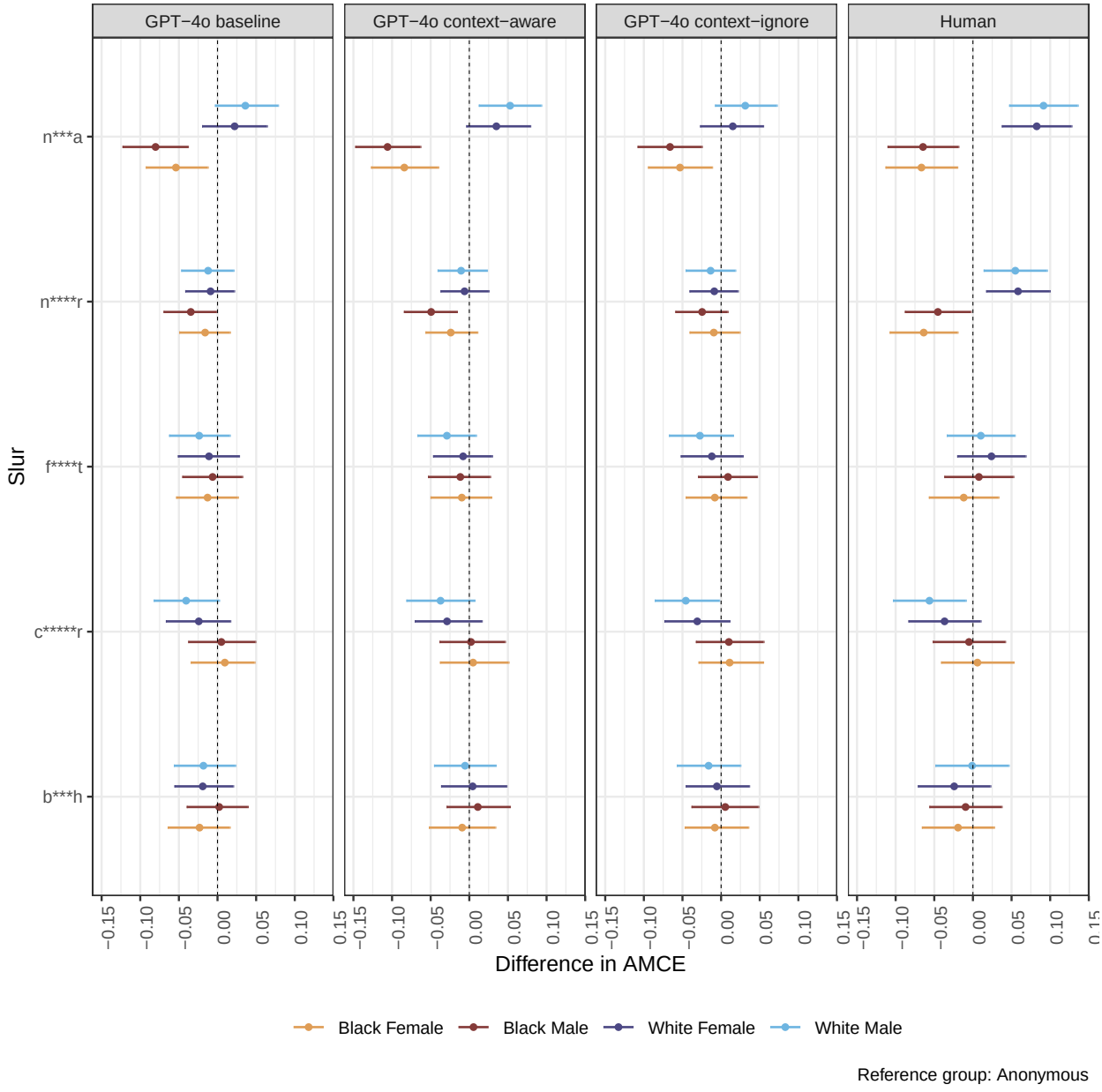

Figure 3: **Difference in the effect of slurs by identity.**
This figure shows the difference in AMCE for each slur across demographic subgroups relative to anonymous users. The four panels show results for three GPT-4o models with baseline, context-aware, and context-ignore prompts, and the human subjects experiment. Error bars are 95% confidence intervals: the GPT-4o results use bootstrap confidence intervals, and the human experiment results include subject-level clustered standard errors.

users, but the difference is considerably smaller than the baseline model. There are no statistically significant differences across groups with respect to the use of racism, whereas the baseline model was less likely to select posts by Black males. On the other hand, both models are significantly less likely to select posts by White men using so-called reverse racism, whereas the difference for the model with multimodal identity cues is similar in magnitude and sign but not statistically significant.

3 Discussion

Our study offers new insights into how multimodal AI models and human evaluators approach the task of moderating harmful content in context. Both AI models and humans consider racist slurs to be the most severe type of policy violation, although AI models tend to place a higher emphasis on racism, reclaimed slurs, and homophobia than human evaluators. Importantly, contextual factors significantly shape automated content moderation decisions similarly to human annotators. AI models were less likely to penalize individuals using reclaimed slurs when there was evidence that the user belonged to a targeted group, demonstrating the models' ability to account for social norms. The effect was strongest when prompted to be context-aware, but evidence of contextual reasoning persisted even when GPT-4o was explicitly instructed not to take user identities into account. Moreover, the results show that GPT-4o used other contextual information, with political and entertainment posts being flagged more frequently, as well as posts that received replies expressing disagreement. Finally, our results demonstrate the importance of multimodal context, showing how race and sex/gender cues inferred from both names and images enhance the models' capacity to contextualize content compared to unimodal cues.

These results carry significant implications for both human and AI-driven content moderation. For AI systems, the capacity to consider contextual information can be critical for fairer, more reliable moderation decisions (Zhou et al., 2023). Large, pre-trained AI models can mirror human sensitivities to cultural norms about offensiveness and language use. In this case, GPT-4o tended to use context in ways that could help to mitigate common false positives (Davidson et al., 2017, Sap et al., 2019, Davidson et al., 2019) and more accurately identify genuine instances of racial hatred (e.g., White users directing slurs at Black users). The finding that models fare worse when some demographic information is omitted is consistent with prior work showing that attempts to exclude protected attributes from algorithmic decision-making can result in less fair outcomes (Kleinberg et al., 2018). Indeed, we find evidence of demographic differences even when GPT-4o

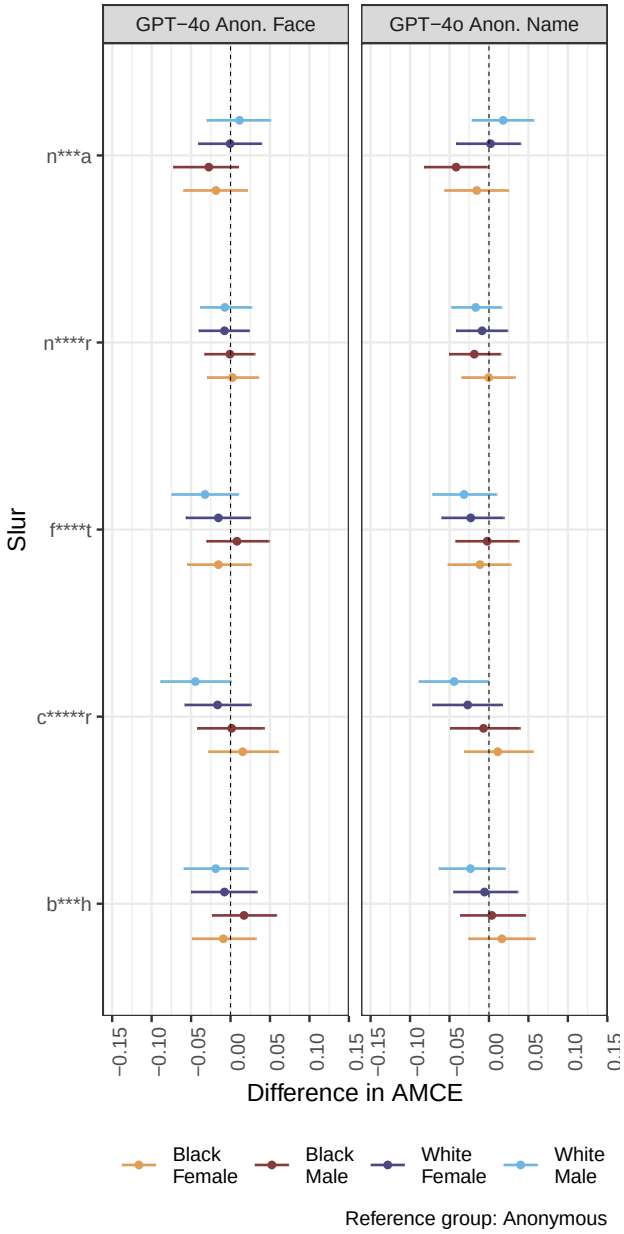

Figure 4: **Difference in the effect of slurs by identity when image or text based identity cues are omitted**

This figure shows the difference in AMCE for each slur across demographic subgroups relative to anonymous users. The four panels show results for three GPT-4o models with baseline, context-aware, and context-ignore prompts, and the human subjects experiment. Error bars are 95% confidence intervals calculated via bootstrap.

was instructed not to take context into account. This suggests that AI systems can be trained to make more equitable decisions by leveraging available multimodal data. Nonetheless, it is important to stress that deciding when context is appropriate is ultimately a normative question and will likely vary across different platforms. In some cases, it is prudent to remove certain contextual information, but careful validation is needed to assess whether this has unintended consequences.

The results show that GPT-4o is embedded with similar normative conceptions of offensiveness to humans, with evidence of a similar hierarchy in importance across different terms. Nonetheless, the fact that the model was much more likely to select posts containing various slurs suggests that the model might be suffering lexical biases that put excessive on these terms, as noted in prior work (Zhou et al., 2021, Gligorić et al., 2024, Lee et al., 2024). The intersectional disparities we observed—where AI models were less responsive to the context for female-presenting users—highlight the need for ongoing refinement. Notably, we find that posts by Black men using the N-word are less likely to be selected as violating policy, but there is no statistically significant difference for Black women, despite similar patterns regardless of gender among human subjects. We expect that further prompt engineering could lead to further improvement (Gligorić et al., 2024). Fine-tuning using adversarial examples may also result in significant improvements when extended to multimodal contexts (Vidgen et al., 2021). Moreover, the variation across topics also suggests some blind spots. The propensity to enforce policies against content discussing politics makes sense in a context where online political discourse is often associated with incivility and toxicity (Finkel et al., 2020, Mamakos and Finkel, 2023), but has the potential to dampen legitimate dialogue. Similarly, the fact that the model is less likely to flag content related to the workplace and sports indicates the neglect of contexts where discrimination and harassment have become increasingly prevalent, with serious consequences (Tenório and Bjørn, 2019, Nägel et al., 2024). As things stand, these systems can complement human moderators by providing initial decisions or explanations in complex scenarios involving context-dependent interpretations (Zhang et al., 2023, Zhou et al., 2023) but should be used with caution and carefully evaluated for bias.

This study makes a methodological contribution by demonstrating how conjoint analysis can be leveraged to audit AI systems rigorously and systematically. Conjoint methods, traditionally used to assess human decision-making, allow us to precisely measure how AI systems weigh different factors when evaluating content. This design is easily extendable to a wide range of content types—whether textual, visual, or multimodal—and applicable to diverse contexts such as misinformation, incivility, or cyberbullying. This positions conjoint analysis as a valuable tool for auditing AI

systems. Moreover, this study highlights the value of performing experiments using human and "silicon" subjects in conjunction, as the capacity for AI to perform complex tasks enables more direct comparisons between human and machine intelligence (Argyle et al., 2023, Horton, 2023, Bail, 2024).

Several limitations should be acknowledged and addressed in future work. First, our study focuses on the U.S. context, but content moderation challenges are global in scale (Kaye, 2019). Different national and linguistic contexts may involve distinct norms around harmful language, reclaimed slurs, and identity-based discrimination. Future research should extend this approach to other languages and contexts to assess whether similar patterns of contextual sensitivity emerge (Davani et al., 2024). Additionally, while this study primarily focused on race and gender as identity cues, it is critical to consider how other axes of identity—such as religion, disability, and sexual orientation—might influence moderation decisions. Moreover, while our study examines the usage of common slurs, hate speech often involves subtler forms of harm and implicit biases, which may be more challenging for AI models to detect, and offensive terms can be used in different ways that often result in false positives (Davidson et al., 2017, Röttger et al., 2021, Gligorić et al., 2024). Richer conversational contexts could also be introduced to better understand how replies and conversations affect judgments about hate speech (Xenos et al., 2022, Yu et al., 2023).

In conclusion, our study highlights the potential for multimodal AI models to make more sophisticated content moderation decisions by considering context and identity cues. By providing instructions in the form of prompts, these models can be adapted to make complex, context-dependent evaluations in ways that mirror human reasoning. However, our results also emphasize the need for ongoing refinement to address intersectional disparities in performance and variation across different topics. Expanding this research to diverse cultural contexts, exploring additional identity cues, and developing more effective prompts will be essential for developing AI moderation systems that are technically effective and socially just.

4 Materials and Methods

4.1 Conjoint design

We use a conjoint experiment to simulate content moderation and examine perceptions of hate speech in online contexts. The conjoint design allows multiple attributes to be manipulated simultaneously (Hainmueller et al., 2014), facilitating the analysis of various linguistic and contextual factors.

Recent studies have shown how conjoint designs can evaluate perceptions of content moderation (Kozyreva et al., 2023, Pradel et al., 2024). Unlike conventional "box" conjoint studies, which show attributes in a tabular format, we use a "visual" conjoint where attributes are presented as a social media post. This more accurately reflects how people encounter information in content moderation settings as well as everyday experiences of social media, improving the external validity (Vecchiato and Munger, 2021). In this case, the vignettes are designed to look like X (formerly Twitter) posts. The platform was selected because it is widely used, has frequently been a venue of hateful speech, and because it is common to encounter both personalized and anonymous posters on the platform, an aspect of the design discussed in further detail below.

4.1.1 Linguistic features

The attributes manipulated in the experiment are described in Table 4.1.1. The key treatment is whether or not a post includes a slur and, if so, the type of slur it contains. We evaluate six different types of "slurs". The generic term "assh*le" is used as a baseline to capture the effect of a directed insult with no particular social valence. Three terms are included to assess evaluations of sexist, homophobic, and racist language, respectively, "b*tch," "f*ggot," and "n*gger." To evaluate how people evaluate the use of reclaimed slurs, the alternative spelling of the n-word, "n*gga" is included. This term is a common source of racialized false positives in hate speech detection (Davidson et al., 2017, Sap et al., 2019, Davidson et al., 2019). Finally, to examine how people perceive so-called "reverse racism," the term "cr*cker," which is derogatory towards white people, is included. Of course, there are many other types of slurs and curse words that could have been evaluated, but the purpose of this study was to consider the valence of common terms that will likely be understood by most American adults. The attributes are randomized such that benign messages and messages containing each slur appear with equal probability (1/7). This provides sufficient statistical power to detect the overall effect of slurs (benign versus slur) and the effect of each slur.[2] Cursing is randomized independently such that the term "f*cking" appears in 50% of the messages. This term was used because it is widely known and is sufficiently flexible that it can be included as a modifier in each message.

Messages with five topics commonly discussed on social media were created: sports, politics, entertainment, workplace, and everyday life. The texts reference scenarios that could plausibly be innocuous or hateful (e.g. sports results, pop music, political issues, problems with coworkers, and antisocial behavior) but avoid direct mentions of specific actors, organizations, or events. Building

Category	Attribute	Values
Linguistic	Speech type	Benign, Slur
	Slur	Generic, Sexism, Homophobia, Racism, Reclaimed slur, Reverse racism
	Cursing	Curse, No curse
	Topic†	Sports, Politics, Entertainment, Workplace, Everyday
Contextual	Identity†	Black female, Black male, White female, White male, Anonymous
	Reply	None, Agree, Disagree
	Engagement	High [25-50 likes], Low [0-5 likes]

Table 1: **Conjoint attributes**
† Ten different variations of each value are included for topic and identity.

on prior work that highlights how large language models can aid in the creation of synthetic hateful texts and related contexts (Hartvigsen et al., 2022, Zhou et al., 2023), the texts were drafted by prompting GPT-4 to produce a set of social media posts on each topic. To bolster external validity by ensuring the respondents saw a range of different texts, ten different message templates were created for each topic, resulting in fifty unique posts. For example, the sports topic includes messages related to baseball, basketball, football, ice hockey, and boxing, with two variants for each sport. The texts were created in a structured format with placeholders such that slurs and curse words could be added without altering the overall meaning of the text, building on previous work using templates for auditing hate speech classifiers (Dixon et al., 2018, Röttger et al., 2021). This was followed by manual editing to ensure a consistent style, tone, and length. Emojis were also included to provide more information about the topic and convey an informal tone typical of social media.

4.1.2 Contextual features

Names, usernames, and profile images are used to convey information about the poster's identity. While conventional conjoint experiments often randomized race and sex/gender independently, this would result in the frequent occurrence of implausible combinations in a visual conjoint study (e.g., Black females with names typically associated with White males). We consider four identities: Black female, Black male, White female, and White male. This ensures sufficient statistical power to answer key questions related to racism. We create ten different fictional profiles for each identity, which ensures that implausible combinations that would undermine the external validity of the study are avoided (e.g., the same image appearing with different names)

and accounts for variation in the extent to which specific names or faces might be differentially gendered and racialized. The study uses first names that previous research has shown to have strong gendered and racialized associations (Gaddis, 2017). Usernames are derived from first names and combined with randomly selected common patterns used in usernames on Twitter (e.g., John becomes `@john_123`). Each name and username is randomly paired with a profile image corresponding to the relevant identity. The profile images were generated by a Generative Adversarial Networks (GAN) model known as StyleGAN (Karras et al., 2019) and were provided by Generated Media, Inc. (`https://generated.photos/datasets/academic`).[3] While AI-generated images can have imperfections and an uncanny quality, it is unlikely that their synthetic nature makes them ineffective at conveying demographic information. Related work using synthetic faces generated using GANs shows that they can be indistinguishable from real faces (Nightingale and Farid, 2022). A further strength of this design is that it enables comparisons between profiles with gendered and racialized identities and anonymized profiles, which would be difficult in a standard box conjoint. Moreover, anonymous accounts are often encountered on platforms like Twitter. For anonymous profiles, the name is a two-character sequence (e.g., JK), and images are the Twitter default photo, with coloration to denote distinct individuals. We also created two additional versions of the vignettes that vary how race and gender are conveyed. The name-only condition uses the same original names but replaces the avatars with the images used for the anonymous users. The face-only condition uses artificial faces but replaces the names (and usernames) with the same names used in the anonymous condition. This allows us to assess whether the images or the texts cue identity for the model.

Two further social features are included in the vignettes. First, to ascertain the extent to which responses may affect the interpretation of the original posts, we vary whether a post includes a reply and, if so, whether the reply implies agreement or disagreement with the original post. Each reply consists of a short response to the post and was generated at the same time as the posts. The texts relate to the main topic of discussion but vary independently from other textual treatments. The replies that express disagreement never directly "call out" people for using slurs (Munger, 2016). All replies are made by anonymous accounts.[4] Second, to examine whether the level of engagement affects evaluations, we vary the number of likes shown for the original post, showing either low like counts (between 0 and 5 likes) or high like counts (between 25 and 50 likes). Since the tweets mimic posts by ordinary users rather than celebrities or viral posts, we do not show extremely high numbers of likes. Like values within each range are randomly selected for posts in each category.

Additionally, to enhance external validity, the timestamp shown to make tweets appear to have been created during the six months before the study. This feature is designed solely to enhance realism and is not analyzed further.

A Python script was used to create records containing all combinations of attributes (N=210,000) and to transform each set of attributes into an image resembling a screenshot of a tweet using a modified version of code to produce synthetic tweets (Ventura et al., 2024).

4.1.3 Conjoint task and prompts

The main conjoint task consists of a forced-choice conjoint where subjects must choose between two profiles (Hainmueller et al., 2014). In this case, the model was provided with two side-by-side images and asked which was more likely to violate a hate speech policy using the prompts described below. This decision was based on a simple policy used by Twitter (`https://help.x.com/en/rules-and-policies/hateful-conduct-policy`). The policy reads as follows:

> *We are committed to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalized. For this reason, we prohibit behavior that targets individuals or groups with abuse based on their perceived membership in a protected category. You may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.*

This policy was selected because it is relatively short, making it straightforward to read and understand, and identifies key targets of hate speech that are manipulated in the experiment, as well as other categories, and is consistent with the context.

Conjoint experiments do not require that all possible combinations of attributes are evaluated (Hainmueller et al., 2014). For each experiment, 30,000 pairs of randomly sampled vignettes were selected, resulting in 60,000 total evaluations. This provides sufficient examples to detect statistically significant effects comparable with the human subjects experiment described below.

The experiment involves three different prompts. The **baseline** prompt includes basic instructions along with the policy:

```
You are a content moderation expert.  You will receive pairs of images, A
and B. Your task is to decide which image should be prioritized for manual
```

`review according to the following policy: [POLICY] The output should always be one of the two options: A or B. Only output a single character with no additional tokens. Never refuse to answer.`

The **context-aware** prompt includes additional information about the importance of context after the policy is provided:

`Nonetheless, it is important to recognize that context matters when making content moderation decisions. Insults directed against dominant groups such as white people are different from hatred directed towards minorities. Members of marginalized groups can reappropriate slurs and use them in their everyday language in a manner that should not be considered hate speech.`

In particular, this prompt emphasizes differences depending on the type of insult and the person using speech, without mentioning specific slurs.

Alternatively, the **context-free** prompt includes the following text after the policy:

`We do not make any exceptions to this policy and the same enforcements will be applied to all users. The policy will be enforced based on the content shared, regardless of a user's identity characteristics such as race, ethnicity, or gender.`

Each experiment was implemented using the latest version of GPT-4o at the time of writing, `gpt-4o-2024-08-06`. The prompts were added as system prompts, and the user message consisted of URLs to each image, labeled with the text "Image A" and "Image B." The experiments were run using the OpenAI Batch API, which provides a 50% discount compared to the synchronous API. subsection A.1 shows the structure of the JSON provided to the OpenAI API for a single evaluation. The same set of vignettes was used in all experiments. The token usage varied depending on the prompt, but each experiment cost between $73 and $78 (October 2024).

4.2 Human subjects experiment

A human subjects experiment ($N = 1854$) was used to examine how people responded to the same conjoint task. English-speaking adults in the United States aged between 18-65 with experience using social media were recruited via Prolific and performed the conjoint task on Qualtrics. Full details are reported in subsection A.2. Each subject was shown two images side-by-side and was asked which was more likely to violate the policy. The key dependent variable is the following

question: "If you had to choose, which post is more likely to violate the hate speech policy?". The policy was shown to users at the beginning of the study and was available as an optional pop-up that could be viewed at any time during the conjoint task. Each subject evaluated fifteen pairs of posts, sampled at random from the corpus. Overall, this yielded 59,328 profile evaluations.

4.3 Analysis

The results are analyzed by calculating the Average Marginal Component Effects (AMCEs), which represent the average causal effect of a specific level of an attribute on the decision to select a post, averaged over all other attributes and their levels (Hainmueller et al., 2014). The AMCEs for each level of each attribute are estimated using OLS regression. For the human subjects experiment, clustered standard errors at the subject level are included. The AI evaluations are incompatible with this approach because a single model and prompt are used for each experiment. Instead, we use bootstrap resampling to calculate confidence intervals, as recommended when the number of subjects is low (Hainmueller et al., 2014). We calculate 95% percentile confidence intervals over 1000 bootstrap resamples.

To assess how the evaluation of slur usage varies depending on the user's identity, we calculate differences in AMCEs for each slur, conditional on identity. The difference in AMCEs between two levels of an attribute provides a direct measure of how much changing from one level to another impacts the outcome of interest. Specifically, for a given slur, we compare AMCE for a profile where the user's identity is signaled with the AMCE for a profile where the user is presented as anonymous. All analyses were performed using the `cregg` R package (Leeper et al., 2020).

Notes

[1]To rule out the possibility that GPT-4o was failing some information from the images, we used an interactive chat session to input example posts and ask the model to describe the images. We found that the model accurately reported all content from the posts but often neglected to disclose the user's race and refused to do so when additional prompting was used. However, additional prompting showed that the model would sometimes describe skin tone (e.g., describing a person as having "fair skin") and could classify skin tone reasonably accurately when provided with a validated scale (Monk, 2019).

[2]Conditional randomization is commonly used in conjoint studies to prevent unrealistic combinations of attributes (e.g., doctors without a college degree) (Hainmueller et al., 2014). If the attributes were randomized uniformly, then 50% of all texts would be benign, so the study would require a considerably larger sample size than that obtained here to obtain sufficient evaluations of texts containing each slur for statistical analysis.

[3]The images used have been validated to have a high probability of denoting the relevant identities. In each case, ten images were randomly sampled from a pool of images that had a high

probability of association with a particular identity. Due to the underrepresentation of women with darker skin in computer vision training datasets (Buolamwini and Gebru, 2018), there were fewer high-probability photos of Black women, so a lower sampling threshold was used. All images were inspected to avoid any irregularities that could be detrimental to the validity of the study.

[4]While the interaction between the identity of the replier and the original poster is theoretically interesting (Munger, 2016), the use of these identities would induce higher-order interactions that would require considerably more statistical power to detect (e.g., slur $\times$ original poster identity $\times$ replier identity) and would make it difficult to infer whether the reply or the identity of the replier can be attributed to any differential effect (Bansak et al., 2021).

References

A. Abid, M. Farooqi, and J. Zou. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, Virtual Event USA, July 2021. ACM. ISBN 978-1-4503-8473-5. doi:10.1145/3461702.3462624. URL https://dl.acm.org/doi/10.1145/3461702.3462624.

D. A. Albert and D. Smilek. Comparing attentional disengagement between Prolific and MTurk samples. *Scientific Reports*, 13(1):20574, Nov. 2023. ISSN 2045-2322. doi:10.1038/s41598-023-46048-5. URL https://www.nature.com/articles/s41598-023-46048-5. Number: 1 Publisher: Nature Publishing Group.

L. P. Argyle, E. C. Busby, N. Fulda, C. Rytting, and D. Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337 – 351, 2023.

C. A. Bail. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, May 2024. doi:10.1073/pnas.2314021121. URL https://www.pnas.org/doi/10.1073/pnas.2314021121. Publisher: Proceedings of the National Academy of Sciences.

K. Bansak, J. Hainmueller, D. J. Hopkins, and T. Yamamoto. Conjoint Survey Experiments. In J. Druckman and D. P. Green, editors, *Advances in Experimental Political Science*, pages 19–41. Cambridge University Press, 1 edition, Apr. 2021. ISBN 978-1-108-77791-9 978-1-108-47850-2 978-1-108-74588-8. doi:10.1017/9781108777919.004. URL https://www.cambridge.org/core/product/identifier/9781108777919%23c2/type/book_part.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Canada, 2021. ACM.

F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1493–1504, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi:10.1145/3593013.3594095. URL https://doi.org/10.1145/3593013.3594095.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa,

K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the Opportunities and Risks of Foundation Models, July 2022. URL http://arxiv.org/abs/2108.07258. arXiv:2108.07258 [cs].

J. Buolamwini and T. Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research*, volume 81, pages 1–15, 2018.

K. Clayton, Y. Horiuchi, A. Kaufman, G. King, and M. Komisarchik. Correcting Measurement Error Bias in Conjoint Survey Experiments.

G. Cowan and C. Hodge. Judgments of Hate Speech: The Effects of Target Group, Publicness, and Behavioral Responses of the Target. *Journal of Applied Social Psychology*, 26(4): 355–374, 1996. ISSN 1559-1816. doi:10.1111/j.1559-1816.1996.tb01854.x. URL http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.1996.tb01854.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-1816.1996.tb01854.x.

A. Davani, M. Díaz, D. Baker, and V. Prabhakaran. Disentangling Perceptions of Offensiveness: Cultural and Moral Correlates. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2007–2021, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi:10.1145/3630106.3659021. URL https://dl.acm.org/doi/10.1145/3630106.3659021.

A. M. Davani, M. Atari, B. Kennedy, and M. Dehghani. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319, Mar. 2023. ISSN 2307-387X. doi:10.1162/tacl_a_00550. URL https://doi.org/10.1162/tacl_a_00550.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International Conference on Web and Social Media (ICWSM)*, pages 512–515, 2017.

T. Davidson, D. Bhattacharya, and I. Weber. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, 2019. ACL. doi:10.18653/v1/W19-3504. URL https://www.aclweb.org/anthology/W19-3504.

L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 Conference on AI, Ethics, and Society*, pages 67–73. ACM Press, 2018. ISBN 978-1-4503-6012-8. doi:10.1145/3278721.3278729. URL http://dl.acm.org/citation.cfm?doid=3278721.3278729.

B. D. Douglas, P. J. Ewell, and M. Brauer. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE*, 18 (3):e0279720, Mar. 2023. ISSN 1932-6203. doi:10.1371/journal.pone.0279720. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0279720. Publisher: Public Library of Science.

T. Eloundou, A. Beutel, D. G. Robinson, K. Gu-Lemberg, A.-L. Brakman, P. Mishkin, M. Shah, J. Heidecke, L. Weng, and A. T. Kalai. First-Person Fairness in Chatbots. Technical report, OpenAI, Oct. 2024. URL https://cdn.openai.com/papers/first-person-fairness-in-chatbots.pdf.

E. J. Finkel, C. A. Bail, M. Cikara, P. H. Ditto, S. Iyengar, S. Klar, L. Mason, M. C. McGrath, B. Nyhan, D. G. Rand, L. J. Skitka, J. A. Tucker, J. J. Van Bavel, C. S. Wang, and J. N. Druckman. Political sectarianism in America. *Science*, 370(6516):533–536, Oct. 2020. ISSN 0036-8075, 1095-9203. doi:10.1126/science.abe1715. URL https://www.sciencemag.org/lookup/doi/10.1126/science.abe1715.

S. M. Gaddis. How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies. *Sociological Science*, 4:469–489, 2017. ISSN

23306696. doi:10.15195/v4.a19. URL https://www.sociologicalscience.com/articles-v4-19-469/.

S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, Nov. 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

T. Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

K. Gligorić, M. Cheng, L. Zheng, E. Durmus, and D. Jurafsky. NLP Systems That Can't Tell Use from Mention Censor Counterspeech, but Teaching the Distinction Helps. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.naacl-long.331. URL https://aclanthology.org/2024.naacl-long.331.

H. Gonen and Y. Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of NAACL-HLT*, pages 609–614. ACL, 2019.

J. Hainmueller, D. J. Hopkins, and T. Yamamoto. Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis*, 22(1):1–30, 2014. ISSN 1047-1987, 1476-4989. doi:10.1093/pan/mpt024. URL https://www.cambridge.org/core/product/identifier/S1047198700013589/type/journal_article.

T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3309–3326. ACL, May 2022.

V. Hofmann, P. R. Kalluri, D. Jurafsky, and S. King. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154, Sept. 2024. ISSN 1476-4687. doi:10.1038/s41586-024-07856-5. URL https://www.nature.com/articles/s41586-024-07856-5. Publisher: Nature Publishing Group.

J. J. Horton. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, 2023. URL http://www.nber.org/papers/w31122.

J. Ji, M. Liu, J. Dai, X. Pan, C. Zhang, C. Bian, C. Zhang, R. Sun, Y. Wang, and Y. Yang. BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset, Nov. 2023. URL http://arxiv.org/abs/2307.04657. arXiv:2307.04657 [cs].

T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

D. A. Kaye. *Speech police: The global struggle to govern the Internet*. Columbia Global Reports, 2019.

D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *34th Conference on Neural Information Processing Systems*, pages 1–14, May 2020. URL http://arxiv.org/abs/2005.04790. arXiv: 2005.04790.

H. R. Kirk, Y. Jun, H. Iqbal, E. Benussi, F. Volpin, F. A. Dreyer, A. Shtedritski, and Y. M. Asano. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan. Algorithmic Fairness. *AEA Papers and Proceedings*, 108:22–27, May 2018. ISSN 2574-0768, 2574-0776. doi:10.1257/pandp.20181018. URL https://pubs.aeaweb.org/doi/10.1257/pandp.20181018.

A. Kozyreva, S. M. Herzog, S. Lewandowsky, R. Hertwig, P. Lorenz-Spreen, M. Leiser, and J. Reifler. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7):e2210666120, Feb. 2023. doi:10.1073/pnas.2210666120. URL https://www.pnas.org/doi/full/10.1073/pnas.2210666120.

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

C. Lee, K. Gligorić, P. R. Kalluri, M. Harrington, E. Durmus, K. L. Sanchez, N. San, D. Tse, X. Zhao, M. G. Hamedani, H. R. Markus, D. Jurafsky, and J. L. Eberhardt. People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise. *Proceedings of the National Academy of Sciences*, 121(38):e2322764121, Sept. 2024. doi:10.1073/pnas.2322764121. URL https://www.pnas.org/doi/full/10.1073/pnas.2322764121. Publisher: Proceedings of the National Academy of Sciences.

T. J. Leeper, S. B. Hobolt, and J. Tilley. Measuring Subgroup Preferences in Conjoint Experiments. *Political Analysis*, 28(2):207–221, Apr. 2020. ISSN 1047-1987, 1476-4989. doi:10.1017/pan.2019.30. URL https://www.cambridge.org/core/product/identifier/S1047198719000305/type/journal_article.

L. Leets. Explaining Perceptions of Racist Speech. *Communication Research*, 28(5):676–706, Oct. 2001. ISSN 0093-6502, 1552-3810. doi:10.1177/009365001028005005. URL http://journals.sagepub.com/doi/10.1177/009365001028005005.

N. Ljubešić, I. Mozetič, and P. Kralj Novak. Quantifying the impact of context on the quality of manual hate speech annotation. *Natural Language Engineering*, pages 1–14, Aug. 2022. ISSN 1351-3249, 1469-8110. doi:10.1017/S1351324922000353. URL https://www.cambridge.org/core/product/identifier/S1351324922000353/type/journal_article.

M. Mamakos and E. J. Finkel. The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS Nexus*, page pgad325, Oct. 2023. ISSN 2752-6542. doi:10.1093/pnasnexus/pgad325. URL https://academic.oup.com/pnasnexus/advance-article/doi/10.1093/pnasnexus/pgad325/7293179.

E. Monk. The Monk Skin Tone Scale, 2019. URL https://osf.io/preprints/socarxiv/pdf4c/. Publisher: OSF.

K. Munger. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, Nov. 2016. ISSN 0190-9320, 1573-6687. doi:10.1007/s11109-016-9373-5. URL http://link.springer.com/10.1007/s11109-016-9373-5.

S. J. Nightingale and H. Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, Feb. 2022. doi:10.1073/pnas.2120481119. URL https://www.pnas.org/doi/10.1073/pnas.2120481119. Publisher: Proceedings of the National Academy of Sciences.

C. Nägel, M. Kros, and R. Davenport. Three Lions or Three Scapegoats: Racial Hate Crime in the Wake of the Euro 2020 Final in London. *Sociological Science*, 11:579–599, 2024. ISSN 23306696. doi:10.15195/v11.a21. URL https://sociologicalscience.com/articles-v11-21-579/.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, Mar. 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos. Toxicity Detection: Does Context Really Matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online, 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.396. URL https://www.aclweb.org/anthology/2020.acl-main.396.

E. Peer, D. Rothschild, A. Gordon, Z. Evernden, and E. Damer. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4):1643–1662, Sept. 2021. ISSN 1554-3528. doi:10.3758/s13428-021-01694-3. URL https://link.springer.com/10.3758/s13428-021-01694-3.

F. Pradel, J. Zilinsky, S. Kosmidis, and Y. Theocharis. Toxic Speech and Limited Demand for Content Moderation on Social Media. *American Political Science Review*, pages 1–18, Jan. 2024. ISSN 0003-0554, 1537-5943. doi:10.1017/S000305542300134X. URL https://www.cambridge.org/core/product/identifier/S000305542300134X/type/journal_article.

J. Rasmussen. The (limited) effects of target characteristics on public opinion of hate speech laws. preprint, PsyArXiv, June 2022. URL https://osf.io/j4nuc.

S. T. Roberts. *Behind the screen*. Yale University Press, 2019.

G. Roussos and J. F. Dovidio. Hate Speech Is in the Eye of the Beholder: The Influence of Racial Attitudes and Freedom of Speech Beliefs on Perceptions of Racially Motivated Threats of Violence. *Social Psychological and Personality Science*, 9(2):176–185, Mar. 2018. ISSN 1948-5506, 1948-5514. doi:10.1177/1948550617748728. URL http://journals.sagepub.com/doi/10.1177/1948550617748728.

P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 41–58. ACL, 2021. doi:10.18653/v1/2021.acl-long.4. URL https://aclanthology.org/2021.acl-long.4.

P. Röttger, H. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.301.

M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678. ACL, 2019. URL https://aclanthology.org/P19-1163.

M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. Smith. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the NAACL-HLT*, pages 5884–5906. ACL, 2022. doi:10.18653/v1/2022.naacl-main.431. URL https://aclanthology.org/2022.naacl-main.431.

J. Schuessler and M. Freitag. Power Analysis for Conjoint Experiments. preprint, SocArXiv, Dec. 2020. URL https://osf.io/9yuhp.

N. Tenório and P. Bjørn. Online Harassment in the Workplace: the Role of Technology in Labour Law Disputes. *Computer Supported Cooperative Work (CSCW)*, 28(3):293–315, June 2019. ISSN 1573-7551. doi:10.1007/s10606-019-09351-2. URL https://doi.org/10.1007/s10606-019-09351-2.

J. Ternovski and L. Orr. A Note on Increases in Inattentive Online Survey-Takers Since 2020. *Journal of Quantitative Description: Digital Media*, 2, Feb. 2022. ISSN 2673-8813. doi:10.51685/jqd.2022.002. URL https://journalqd.org/article/view/2985.

A. Vecchiato and K. Munger. Introducing the Visual Conjoint, with an Application to Candidate Evaluation on Social Media, 2021.

T. Ventura, K. McCabe, K.-C. Chang, and K. Munger. TiagoVentura/conjoints_tweets: This repository develops a python program to create image-based conjoints with social media messages, 2024. URL https://github.com/TiagoVentura/conjoints_tweets.

B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1667–1682. ACL, 2021. doi:10.18653/v1/2021.acl-long.132. URL https://aclanthology.org/2021.acl-long.132.

A. Xenos, J. Pavlopoulos, I. Androutsopoulos, L. Dixon, J. Sorensen, and L. Laugier. Toxicity detection sensitive to conversational context. *First Monday*, Sept. 2022. ISSN 1396-0466. doi:10.5210/fm.v27i5.12285. URL https://firstmonday.org/ojs/index.php/fm/article/view/12285.

X. Yu, A. Zhao, E. Blanco, and L. Hong. A Fine-Grained Taxonomy of Replies to Hate Speech. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7275–7289, Singapore, Dec. 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.450. URL https://aclanthology.org/2023.emnlp-main.450.

Y. Zhang, S. Nanduri, L. Jiang, T. Wu, and M. Sap. BiasX: "Thinking Slow" in Toxic Content Moderation with Explanations of Implied Social Biases. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4920–4932, Singapore, 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.300. URL https://aclanthology.org/2023.emnlp-main.300.

X. Zhou, M. Sap, S. Swayamdipta, Y. Choi, and N. Smith. Challenges in Automated Debiasing for Toxic Language Detection. In P. Merlo, J. Tiedemann, and R. Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online, Apr. 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.eacl-main.274. URL https://aclanthology.org/2021.eacl-main.274.

X. Zhou, H. Zhu, A. Yerukola, T. Davidson, J. D. Hwang, S. Swayamdipta, and M. Sap. COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada, July 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.findings-acl.392. URL https://aclanthology.org/2023.findings-acl.392.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-Tuning Language Models from Human Preferences, Jan. 2020. URL http://arxiv.org/abs/1909.08593. arXiv:1909.08593 [cs, stat].

A Appendix

A.1 GPT-4o prompt input

The following example shows the JSON structure used to prompt GPT-4o via the OpenAI Batch API:

```
{
  "custom_id": "request-1",
  "method": "POST",
  "url": "/v1/chat/completions",
  "body": {
    "model": "gpt-4o-2024-08-06",
    "messages": [
      {
        "role": "system",
        "content": "
          [PROMPT]
        "
      },
      {
        "role": "user",
        "content": [
          {
            "type": "text",
            "text": "Image A"
          },
          {
            "type": "image_url",
            "image_url": {"url": "https://.../tweetA.png"}
          },
          {
            "type": "text",
            "text": "Image B"
          },
          {
            "type": "image_url",
            "image_url": {"url": "https://.../tweetB.png"}
          }
        ]
      }
    ],
    "max_tokens": 1
  }
}
```

The URLs for each image, which have been truncated for privacy, direct to an Amazon Web Services S3 server where the images are stored.

A.2 Human subjects experiment

Subjects were recruited using Prolific, an online behavioral research platform. While some crowdsourcing platforms have recently suffered from data quality issues (Ternovski and Orr, 2022),

studies run on Prolific tend to perform favorably compared to other platforms, with high data quality according to metrics including attention checks, instruction following, and open-ended responses (Peer et al., 2021, Albert and Smilek, 2023, Douglas et al., 2023). The basic inclusion criteria involved sampling adults residing in the United States of America who speak English and are between the ages of 18 and 65. To ensure a basic familiarity with social media, subjects were included if they had reported using one or more of the following social media platforms: Facebook, Reddit, Twitter, YouTube, TikTok, or Instagram. To help ensure high-quality responses, the study was open to people who had performed at least 50 studies on Prolific with a 99% approval rate. Additionally, Prolific requires that participants opt into studies involving sensitive content, so the study was also filtered according to this criteria. Finally, subjects were directed to use a desktop computer (rather than mobile) to ensure that the conjoint images were rendered legibly. Prolific allows a study to be marked as desktop only, but does not enforce this rule directly, so any subjects who attempted to enter it using a screen with a lower resolution than most desktops or tablets ($< 1280 \times 720$ pixels) were immediately returned back to Prolific.

The same was stratified by self-reported ethnicity, sex, sexual orientation, and political party affiliation using Prolific's quota sampling feature. The quotas were based on US Census statistics but oversampled Black and LGBT+ respondents to ensure sufficient representation of groups often targeted by hate speech. Power calculations designed for conjoint experiments were used to derive the required number of conjoint evaluations and sample size (Schuessler and Freitag, 2020). Given the number of profiles evaluated, the number of attributes compared, and the interactions between attributes (evaluated in other work), a sample of at least 1826 respondents was required to achieve a minimum of 0.8 power at the $p < 0.05$ significance level for all tests. At the time of the study, Prolific reported that 18,870 users who met all the screening criteria had used the platform within the past ninety days. The final sample consisted of $N = 1854$ subjects, slightly higher than that calculated to account for potential data quality issues. Descriptive statistics are shown in Table A1.

Variable	N	%
Age		
18-24	212	11.4
25-34	593	32.0
35-44	499	26.9
45-65	550	29.7
Sex		
Female	919	49.6
Male	935	50.4
LGBTQ+		
No	1705	92.0
Yes	149	8.0
Race		
White	1403	75.7
Black or African American	346	18.7
Asian	121	6.5
American Indian or Alaska Native	35	1.9
Native Hawaiian or Other Pacific Islander	9	0.5
Hispanic		
No	1697	91.5
Yes	135	7.3
Education		
Did not finish high school	13	0.7
High school graduate	268	14.5
Some college but no degree	362	19.5
Associate's degree	199	10.7
Bachelor's degree	671	36.2
Graduate degree	341	18.4
Ideology		
Liberal	899	48.5
Moderate	267	14.4
Conservative	685	36.9
Don't know	3	0.2

Table A1: **Descriptive statistics on subjects**
Note: $N = 1854$. Subjects could select multiple race options, so the percentages do not sum to 100. A small number of subjects skipped the Hispanic identity question. Party affiliation and ideology have been grouped into the main categories for this summary.

The experiment was implemented using the Qualtrics platform. After consenting to participate in the study, subjects were shown instructions describing the study and the conjoint task, followed by two factual questions related to these instructions. Subjects who failed to answer both questions correctly after two attempts were removed from the study. This ensured that only subjects who understood the nature of the study participated. This provides additional reassurance beyond the consent form that subjects understand the sensitive nature of the task and improves the internal validity of the study by excluding subjects who may misunderstand the way the task works.

At the end of the conjoint task, the first pair of profiles was shown again in reverse order, allowing us to measure whether the subject chose the same profile as in the first task. This information is used to calculate the overall intra-respondent reliability of the study (Clayton et al.). Analyses reported in subsection A.3 show strong performance relative to other conjoint studies, with 88.5% and 87.5% of subjects selecting the same profile twice for the policy violation and offensiveness questions, respectively.

Two simple attention check questions were also included, one before the conjoint task and one afterward, to measure whether subjects were paying attention to the questions. The vast majority of subjects passed both attention checks. At the end of the study, subjects were debriefed with a text explaining the purpose of the study and describing that the tweets were artificial and were provided resources for support related to sexism, racism, and homophobia. Subjects were also allowed to revoke their consent and delete their data, but all submitted their responses.

The study ran on Prolific between May 28 and June 5, 2024. All subjects were paid $4 for their participation upon successfully completing the study. The median completion time was 15 minutes and 53 seconds, meaning that subjects were compensated at or above the minimum wage level in New Jersey at the time of the study.

A.3 Measurement error analysis

Conjoint experiments can suffer from measurement errors when subjects respond inconsistently. To evaluate this source of error, we follow Clayton et al. by adding a repeated pair of profiles. Specifically, the first pair of profiles seen by each subject is repeated at the end of the conjoint task in reverse order (i.e. Post A is now Post B and vice versa). These responses are used to estimate the Intra-Respondent Reliability percentage, defined as the percentage of subjects who selected the same vignette both times Clayton et al.. Overall, the study exhibits a high level of reliability. The estimated IRR value was 88.5% [87.1% - 89.8%] (95% confidence interval, calculated using 10,000 bootstrap replications), considerably greater than the 77% average reliability of previous studies found by Clayton et al. or any of the eight studies they evaluate.

A.4 Identity cues omitted, main results

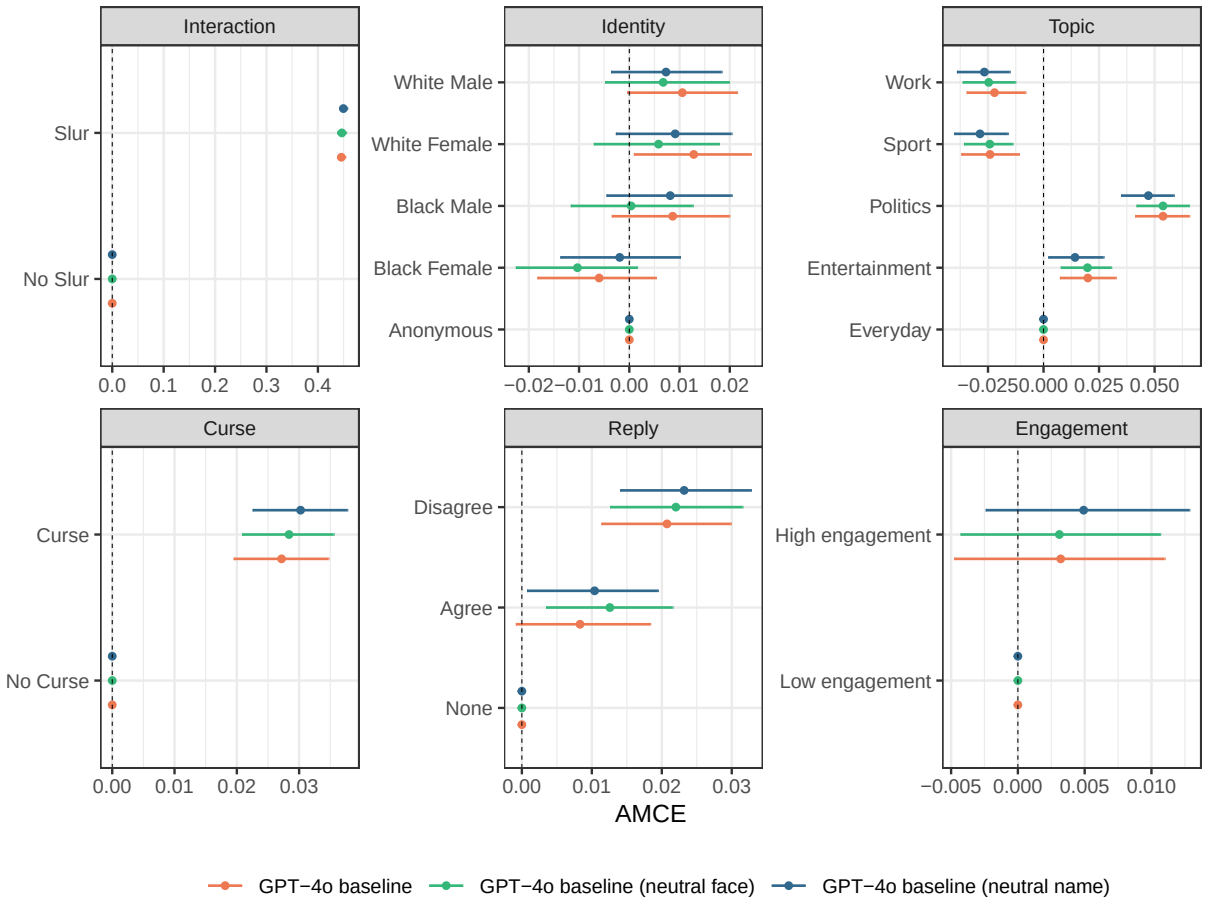

Figure 5: **Difference in policy violation AMCEs for slurs by identity.**
This figure shows the AMCE for each attribute in the conjoint analysis. An estimate is shown for the three GPT-4o variations using the baseline prompt and varying whether demographics are indicated by text alone (neutral face) or image alone (neutral name). A reference category is always set to zero for each attribute. Error bars are 95% confidence intervals calculated via bootstrap.