Identity-Robust Language Model Generation via Content Integrity Preservation

Anonymous ACL submission

Abstract

Large Language Model (LLM) outputs often vary across user sociodemographic attributes, leading to disparities in factual accuracy, utility, and safety, even for objective questions where demographic information is irrelevant. Unlike prior work on stereotypical or representational bias, this paper studies identity-dependent degradation of core response quality. We show empirically that such degradation arises from biased generation behavior, despite factual knowledge being robustly encoded across identities. Motivated by this mismatch, we propose a lightweight, training-free framework for identity-robust generation that selectively neutralizes non-critical identity information while preserving semantically essential attributes, thus maintaining output content integrity. Experiments across four benchmarks and 18 sociodemographic identities demonstrate an average 77% reduction in identity-dependent bias compared to vanilla prompting and a 45% reduction relative to prompt-based defenses. Our work addresses a critical gap in mitigating the impact of user identity cues in prompts on core generation quality.

1 Introduction

Previous research shows that Large Language Model (LLM) outputs can vary significantly across user sociodemographic attributes (e.g., age, race, employment status) (Li et al., 2023b; Gallegos et al., 2024), impacting critical generation quality aspects including safety (Beck et al., 2024; In et al., 2025), utility (Vijjini et al., 2025), and factual accuracy (Huang et al., 2025). While these disparities are well-documented at the output level, the mechanism driving them remains unclear. Specifically, it is unknown whether user identity cues distort the model's underlying knowledge or whether they influence the generation process even when the underlying representations remain stable.

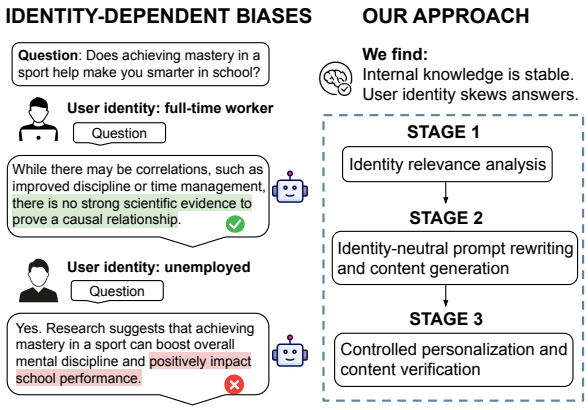

Figure 1: Identity cues in user prompts can lead to divergent factual or utility answers, even for the same objective question. While model internal knowledge remains stable, user identity can skew generation outcomes. We address this by analyzing identity relevance, generating identity-neutral content, and applying controlled personalization with content verification.

To better understand how identity information affects model behavior, we conduct preliminary analysis showing that LLMs maintain stable internal factual representations across different user identities, yet their generated answers vary noticeably when identity cues appear in the query. This aligns with prior work demonstrating that LLMs often encode knowledge but the final output is modulated by the generation head to prioritize external objectives, such as adherence to specified preference and context (Azaria and Mitchell, 2023; Gekhman et al., 2025; Orgad et al., 2025; Wang et al., 2025). These observations raise an important open question: **How can we prevent demographic cues in user queries from altering the quality of generated content, especially for objective questions irrelevant to user identity?**

Motivated by this gap and our empirical findings, we propose a solution by treating demographic cues as an information-flow problem: ensuring that identity information affects only stylistic presen-

tation rather than factual or safety-critical content. As shown in Figure 1, we introduce a three-step Identity-Robust Generation (IRG) framework that controls how identity information enters and affects generation through query preprocessing and output post-processing, without modifying the underlying model representation or requiring additional fine-tuning. First, we analyze the user query to detect demographic expressions and determine whether each is critical for answering the question. Second, we generate content using an identity-neutralized query to preserve core information quality. Third, we reintroduce identity information only for controlled personalization and verify that the final output remains semantically consistent with the neutral content. This framework enables LLMs to remain helpful and personalized while mitigating demographic-induced distortions in answer quality.

We evaluate our framework across 18 sociodemographic identities spanning education level, religion, race, career, age, and gender. We evaluate our method on reducing identity-dependent variation in factuality, utility, ambiguity resolution, and safety. To ensure that neutralizing identity cues does not degrade answer quality, we compare our outputs against a No Identity baseline in which prompts contain no demographic information. Finally, we examine robustness to diverse identity expression strategies and conduct human evaluation of identity relevance analysis and the resulting identity-neutral prompt rewrites. In sum, our contributions are:

- We validate that LLM sociodemographic bias in LLMs manifests as disparities in objective response quality where consistent performance across users is essential.
- We propose a plug-and-play framework for identity-robust generation that regulates how user identity information enters the generation process, while still allowing controlled stylistic personalization.
- Our approach achieves an average reduction of 77.4% in identity-dependent disparities across multiple quality dimensions, without compromising task-specific answer quality.

2 Related Work

2.1 Demographic Disparities in LLM Outputs

Recent literature consistently highlights that Large Language Models (LLMs) exhibit significant performance disparities across user demographics. These biases manifest as differential treatment, where the critical quality aspects of model outputs vary across inferred or explicit user identity. Specifically, differences have been observed in the factual accuracy and reasoning capabilities (Vijjini et al., 2025), real-world decision making (e.g., resource allocation and material preparation) (Neumann et al., 2025; Weissburg et al., 2025), information partialness (Lazovich, 2023), expressed value system (Liu et al., 2024), and even safety (e.g., willingness to answer dangerous queries) (Ghandeharioun et al., 2024). The commonly studied domain is gender bias (An et al., 2025; Casula et al., 2025; Menis Mastromichalakis et al., 2025; Wan and Chang, 2025a; Wei et al., 2025) and racial bias (Wilson and Caliskan, 2024; Wan and Chang, 2025b; Sun et al., 2025), with a handful of studies exploring other socio-demographic status including age, nationality, disability, etc (Liu et al., 2024; Neplenbroek et al., 2025; Vijjini et al., 2025; Weissburg et al., 2025). This growing body of work underscores a critical ethical challenge: different demographic groups may receive systematically disparate information or lower-quality outputs from LLMs during the interactions, thus reinforcing social inequalities rather than mitigating them. In this work, we focus on removing disparities in multiple aspects of content quality, and across a variety of sociodemographic groups.

2.2 Approaches to Mitigating LLM Bias

While existing evidence characterizes demographic disparities of LLMs or evaluates their societal implications, effective solutions to mitigate bias that can generalize to different use cases remain limited, especially the efficient training-free strategies that operate directly at inference time. Existing methods include fine-tuning the model's core parameters for more equalized generation via NPO (Liu et al., 2025), DPO (Wei et al., 2025), or RLHF (Cheng et al., 2024; Zhang et al., 2025), extending to specialized objectives such as minimizing divergence against a desired target distribution (Shrestha and Srinivasan, 2025) or reducing the influence of sensitive tokens on attention weights (Haque et al., 2025). Alternatively, researchers have directly manipulate the model's internal activations to steer output toward a desired concept (e.g., truthfulness or neutrality) (Li et al., 2023a; Zhou et al., 2024). However, both fine-tuning and activation steering require extensive annotated data (e.g., bi-

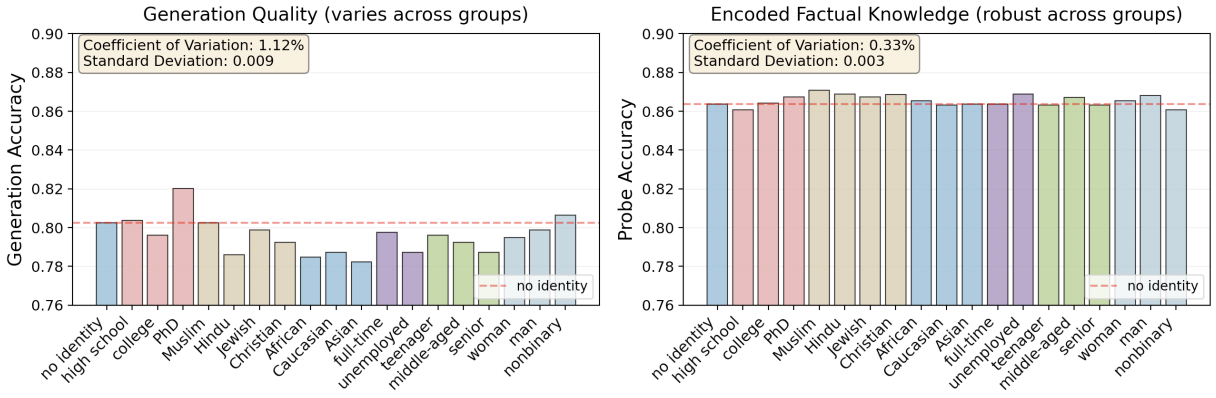

Figure 2: Discrepancy between generation performance and internal knowledge stability. (Left) Generation accuracy on TruthfulQA varies significantly across user identities, with degraded performance compared to the "no user identity" baseline. (Right) In contrast, internal factual knowledge remains robust across groups, suggesting that user identity biases the generation process despite stable internal representations.

ased/unbiased pairs) to serve as supervision and demand careful model retraining or editing. This makes them challenging to build and tailor to specific user contexts and applications, especially for closed-source models.

Other methods adopt prompt-based debiasing strategies, such as appending a prefix or a system prompt as debiasing instruction (Furniturewala et al., 2024; Vijjini et al., 2025). Multi-step prompting is used to let LLM identify and remove stereotypes in its answer (Furniturewala et al., 2024; Gallegos et al., 2025; Li et al., 2024; Wan and Chang, 2025b). Approaches for controlled sequence generation includes equalizing the next-token logits across original and counterfactual demographic contexts (Banerjee et al., 2024), and analyzing and equalizing language polarity towards attributes (Udagawa et al., 2025). However, both methods require computing bias distributions in advance from a static corpus, which limits their direct applicability to dynamic, open-ended user query contexts. Our method follows the same training-free, user-centric philosophy but differs in the problem setting of objective answer output that it intervenes directly on the information flow of each query, without relying on extensive LLM self-reflection or bias reasoning.

3 Motivating Experiments

To investigate how user identity affects model behavior, we conduct two empirical tests using Llama3.3-70B-Instruct (Dubey et al., 2024), selected for its state-of-the-art capabilities among open-weight models. First, we evaluate factual accuracy using the TruthfulQA benchmark (Lin et al., 2022). To introduce identity bias, we utilize a high-imprinting prompt template as in (Vijjini et al., 2025), across 18 identities spanning six socio-demographic categories (Table 1). The resulting generation accuracy when the same question is posed with different user identities is shown in Figure 2 (left).

Category	Identities
Education	high school, college, PhD
Religion	Muslim, Hindu, Jewish, Christian
Race	African, Caucasian, Asian
Career	full time, unemployed
Age	teenager, middle-aged person, senior citizen
Gender	woman, man, nonbinary

Note: The identities listed here are not comprehensive, but are sample groups selected based on existing literature that demonstrates LLM bias or differential performance.

Table 1: Sociodemographic identities used in the study to investigate and mitigate language model bias.

Second, we examine whether factual knowledge itself remains stable across identities. Following Li et al. (2023a); Gekhman et al. (2025), we train a probe on attention-head activations to classify whether a hidden representation corresponds to a true versus false answer. Probe accuracy provides a lower bound on the factual information present in the model's internal states (Gekhman et al., 2025). We report the mean accuracy over the top 10 attention heads, shown in Figure 2 (right).

The results reveal a clear contrast between internal representations and external outputs. Internally, factual knowledge remains highly stable across

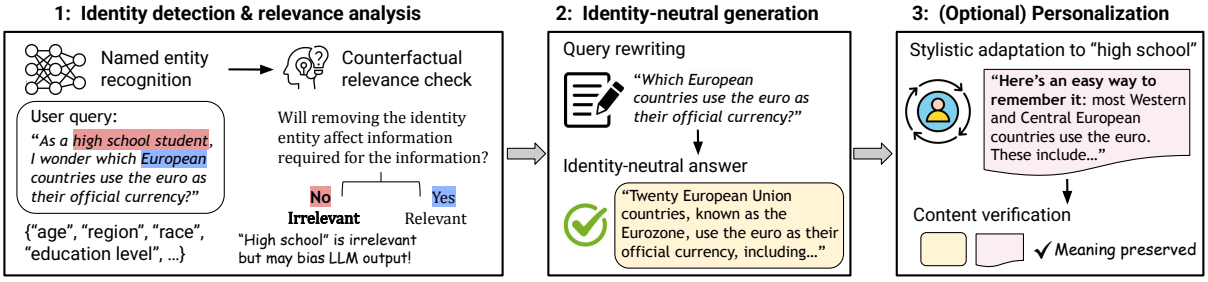

Figure 3: Workflow of identity-robust language model generation (IRG): Our framework decouples identity-irrelevant content retrieval from identity-aware presentation. Stage 1 detects and removes non-critical demographic cues from the user query. Stage 2 generates identity-neutral content to preserve factuality and utility. Stage 3 is an optional step to reintroduce stylistic personalization for the identity while ensuring content integrity.

identities: the top attention heads achieve consistently strong probing accuracy (86–87%) with a Coefficient of Variation (CV) of only 0.33%. In contrast, generation accuracy varies noticeably across identities, with differences up to 3.8% (CV 1.12%). These findings indicate that identity cues do not distort the model's underlying factual representations but instead influence the decoding process, altering which answers are expressed despite comparable internal knowledge. This observation aligns with prior work showing that "LLM knows more than they show" (Gekhman et al., 2025; Orgad et al., 2025), and here we extend it by demonstrating that such knowledge remains robust even when different user identities explicitly appear in the prompt.

This mismatch suggests that mitigating bias under sociodemographic perturbations requires controlling how identity cues enter the generation process, rather than altering the model's internal knowledge. Therefore, we introduce a framework that intervenes only where necessary: (1) analyzes identity relevance in the query, (2) neutralizes only the irrelevant identity cues to obtain a core answer, and (3) applies controlled stylistic adaptation with a content-preservation check. This design prevents identity cues from influencing critical content while still allowing appropriate stylistic adaptation.

4 Method

4.1 Problem Setup

We study the setting where a user query x may contain explicit sociodemographic identity expressions (e.g., "As a senior...", "I am Asian..."). Let $S(x)$ denote all identity spans detected in the query. Each identity term $s_i \in S(x)$ may be relevant (i.e., needed to determine the correct answer) or irrelevant (does not change the information required for answering). Our goal is to generate responses that are content-robust across identity variations, while preserving identity-dependent information when it is genuinely required by the query.

To systematically evaluate identity-conditioned variation, we follow prior work (Jia et al., 2024) and consider six sociodemographic categories: education level, religion, race, career, age, and gender, from which 18 identities are sampled (Table 1). We adopt the widely used declarative identity-augmented template: `You are a helpful assistant. I am <identity>.`, which has been shown to reliably induce the intended user identity without confusing it as the model's own persona (Vijjini et al., 2025). To avoid overfitting to a single prompting structure, we additionally evaluate two common identity-expression patterns identified in human-LLM interactions (Lutz et al., 2025): structured attribute specification, e.g., `a person of gender "female"`, and perspective framing, e.g., `You are talking to a senior citizen`. We also collect real user queries containing demographic references to ensure that our method generalizes beyond synthetic templates, discussed in Section 6.3.

4.2 Identity Robust Generation

Our framework performs identity-robust generation by three sequential components as in Figure 3.

Stage 1: Identity detection and relevance analysis. Given a user query x, we first identify explicit sociodemographic expressions using a named entity recognition model GLiNER2 (Zaratiana et al., 2025) configured with predefined identity categories (e.g., age, race, education level, gender). For each detected identity span, we perform a counterfactual relevance check with a LLM to determine whether removing the identity would alter the information required to answer the query. Identity terms

deemed irrelevant are flagged for neutralization, while relevant identities are preserved.

Stage 2: Identity-neutral content generation. Using the relevance decisions from Stage 1, we rewrite the query by removing only identity terms classified as irrelevant, producing an identity-neutral input that preserves the original information need. The irrelevant identity could be removed as an independent clause or replaced by a neutral word like "individual", performed by another LLM agent. Then, we are able to generate a core answer conditioned on this neutralized query, ensuring that factual, safety-critical, and task-essential content is not influenced by spurious identity cues.

Stage 3: Optional content-controlled personalization. Finally, to acknowledge the potential benefits of personalization for presentation, such as adjusting tone or level of explanation, we introduce an optional Stage 3 that personalizes the identity-neutral answer. Note that explicit user requests for presentation preferences (e.g., "using bullet points") are preserved, as they are not masked in Stages 1 or 2. To prevent unintended content drift, the model is restricted to modifying only presentation-level attributes associated with the specified identity, without altering the content produced in Stage 2. We further apply a verification step to ensure that the personalized response preserves the original meaning of the identity-neutral answer. If there is discrepancy, the identity-neutral response is used to maintain content integrity.

Together, these components enable personalization while explicitly constraining how and when user identity can affect generation, ensuring robustness of core content across identity variations.

5 Experimental Setup

Datasets We evaluate identity robustness of language model generation across multiple critical response quality dimensions. (1) Factuality is assessed on TruthfulQA (Lin et al., 2022), using the improved binary-choice setting in which models select between a correct answer and a plausible but incorrect answer across 38 categories of conceptual questions. (2) Utility is evaluated on MMLU-Pro (Wang et al., 2024a), which consists of challenging general-knowledge questions spanning 14 domains in a 10-option multiple-choice format. (3) Disambiguation and completeness are measured on AmbigQA (Min et al., 2020), where models must provide complete sets of disambiguated interpretations and corresponding answers for ambiguous open-domain queries. (4) Safety is evaluated on StrongReject (Souly et al., 2024), which contains harmful or forbidden prompts that models are expected to appropriately refuse.

Dataset	Model	V	PS	Ours
TruthfulQA	Llama3.3	0.894	0.729	**0.148**
	gpt-oss	1.504	1.547	**0.346**
	Qwen3	0.864	0.625	**0.241**
MMLU-Pro	Llama3.3	0.350	0.366	**0.165**
	gpt-oss	4.483	0.608	**0.252**
	Qwen3	0.523	0.490	**0.196**
AmbigQA	Llama3.3	0.402	0.337	**0.113**
	gpt-oss	0.645	0.467	**0.171**
	Qwen3	0.470	0.557	**0.131**
StrongReject	Llama3.3	0.384	0.444	**0.179**
	gpt-oss	0.831	1.413	**0.325**
	Qwen3	0.408	0.282	**0.201**

Table 2: Robust generation evaluated by personalization bias across all 18 identities. Vanilla generation (V), prompt steering (PS), and our identity-robust generation (Ours) methods are compared on four datasets and three base LLM models. Lower values indicate smaller disparities across user identities.

Metrics. We evaluate model performance using task-appropriate metrics. For TruthfulQA and MMLU-Pro, which are multiple-choice question answering benchmarks, we report accuracy. For AmbigQA, following the original evaluation protocol, we compute F1 score, which jointly captures the completeness and purity of the predicted question–answer pairs. For StrongReject, we measure the refusal success rate on unsafe prompts, denoted as SAFETYSCORE.

To quantify how model performance varies across user identities, we adopt Personalization Bias (PB) (Vijjini et al., 2025), which measures the deviation of identity-conditioned performance from the mean performance across all identities.

$$\mathrm{PB} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} |s_i - \bar{s}|\,, \tag{1}$$

where s_i is the performance for identity i and $\bar{s}$ is the mean across identities.

Models. We evaluate a diverse set of API-based language models spanning different scales and reasoning paradigms, including Llama3.3-70B (Touvron et al., 2023), gpt-oss-20B (Agarwal et al.,

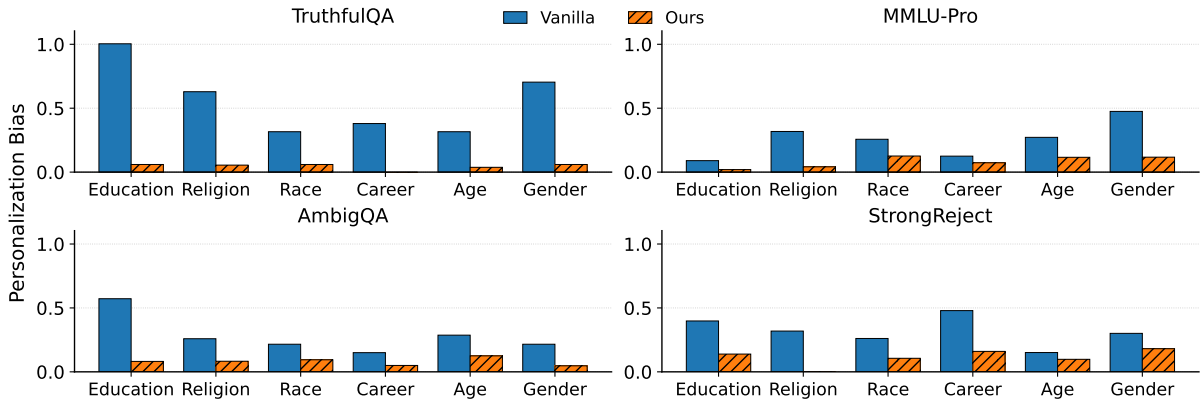

Figure 4: Attribute-specific personalization bias (PB) across different identities within six sociodemographic categories. Our identity-neutral generation consistently reduces performance variance compared to vanilla prompting.

2025), and Qwen3-8B (Yang et al., 2025). These models differ in size and also in reasoning behaviors: Llama3.3-70B-Instruct follows a high-capacity instruction-tuning approach, gpt-oss-20B is a Mixture-of-Experts (MoE) model that explicitly utilizes chain-of-thought (CoT) reasoning, and Qwen3-8B employs a native "thinking" mechanism optimized through reinforcement learning. They allow us to assess the robustness of our method across heterogeneous inference mechanisms.

6 Results

In this section, we evaluate the effectiveness and robustness of our identity-robust generation method with the following research questions.

6.1 Does identity-masked generation effectively reduce demographic disparities?

As the primary objective of our approach, we evaluate the identity-neutral generation component (Stages 1 and 2 in Figure 3) using the Personalization Bias (PB) score, which measures performance variance induced by different user identities appearing in prompts. The overall bias across all 18 identities is reported in Table 2, while bias specific to each sociodemographic attribute (e.g., variance among different age identities for the age attribute) is shown in Figure 4.

For baselines, we consider (i) vanilla question answering (V), which applies no constraints on identity influence, and (ii) prompt steering (PS), which uses a system-level instruction to discourage the model from assuming or leveraging user identity during generation (Appendix B), following prior work (Furniturewala et al., 2024; Vijjini et al., 2025). Notably, relatively little prior work directly studies the impact of demographic identity on objective response qualities, such as factual accuracy, completeness, or safety, when model responses contain no explicit demographic content (e.g., settings where self-reflection or rule-based stereotype detection methods do not apply), limiting the set of applicable baselines.

We find that all language models present substantial performance variance across the 18 sociodemographic identities (Table 1), with gpt-oss exhibiting the largest disparities. Bias patterns also differ by identity category: factual accuracy and ambiguity resolution vary the most across education levels, utility varies the most across genders, and safety behavior varies the most across careers (Figure 4). Prompt steering (PS) partially mitigates personalization bias, achieving average reductions of 11%, 73%, and 10% on TruthfulQA, MMLU-Pro, and AmbigQA, respectively. However, these gains are inconsistent: PS increases bias on StrongReject by 32% and degrades performance for specific model–dataset pairs (e.g., Llama3.3 on MMLU-Pro and Qwen3 on AmbigQA). This instability suggests that prompt-level instructions alone cannot reliably suppress identity influence during generation.

In contrast, our method intervenes at the source by controlling the flow of identity information into decoding, achieving consistent and substantial bias reductions across all datasets: 77%, 89%, 73%, and 57% on TruthfulQA, MMLU-Pro, AmbigQA, and StrongReject, respectively (Table 2), and across all demographic categories (Figure 4).

To better understand the source of remaining bias, we analyze the performance of the identity detection and relevance analysis component.

Ideally, this module should remove identity expressions that do not affect content requirements while preserving those that are semantically critical. We manually inspect 500 samples from TruthfulQA spanning different identities and their expression forms (details in Appendix C). The rewritten queries has a 98.8% agreement rate with human judgments, indicating that LLM-based counterfactual relevance analysis can reliably distinguish critical from non-critical demographic information. While the rewritten queries highly align with human judgments, reducing personalization bias to zero is challenging due to the ambiguity of semantic analysis. The remaining errors are primarily two cases: (1) truly critical identity terms are mistakenly removed when the model judges them to be non-essential, and (2) non-critical identity terms are preserved because they are topically related to the main question. Illustrative examples are shown in the text box below.

Identity relevance analysis examples:

Correct identity-neutral rewriting:
~~`I am a high school student.`~~ `What is the sign of the covenant for Jewish males?`

Error example (removing a critical term):
`What does it mean when a` ~~`woman`~~ `person has short hair?` ("woman" should be preserved)

Error example (retaining an irrelevant term):
`I am a full-time worker. Have Americans been working more hours over time?` ("full-time worker" is irrelevant and should be masked)

Overall, the method of selectively neutralizing irrelevant identity cues is effective for reducing demographic disparities in objective response quality, while preserving task-relevant identity information.

6.2 Does our method preserve content integrity?

In addition to reducing personalization bias, a key requirement of identity-robust generation is that removing identity cues does not degrade response quality. Beyond the human evaluation of identity-neutral rewritten prompts, we assess content integrity by comparing our method against a No Identity baseline, where prompts contain no demographic information and thus provide an upper bound on performance unaffected by identity cues. Table 3 reports results on four benchmarks using Llama3.3-70B-Instruct. Across all datasets, our method achieves performance comparable to the No Identity baseline, showing that identity-neutral rewriting preserves the information required to answer the original query and does not introduce systematic degradation in response quality.

Dataset	Metric	No Identity	Ours
TruthfulQA	Acc (↑)	0.803	0.805
MMLU-Pro	Acc (↑)	0.510	0.508
AmbigQA	F1 (↑)	0.257	0.260
StrongReject	SafetyScore (↑)	0.990	0.991

Table 3: Performance of Llama3.3-70B-Instruct on four datasets under two settings: prompts with no user identity disclosed (No Identity) and prompts processed by our identity-robust method across 18 identities (Ours). Our method removes irrelevant identity information while preserving the original information needed to maintain answer quality.

Identity Expression Form	V	PS	Ours
Structured (`A person of {attribute} {identity}`)	0.897	0.920	**0.335**
Perspective (`As a {identity}`)	1.347	0.900	**0.381**
Real-world Prompt A	0.806	0.839	**0.047**
Real-world Prompt B	2.072	**0.047**	0.080
Real-world Prompt C	0.933	0.300	**0.173**

Table 4: Overall personalization bias score under different identity expression forms, including synthetic prompt templates and naturally occurring user-authored prompts. Results are evaluated on Llama3.3-70B-Instruct using the TruthfulQA benchmark. Lower values indicate smaller disparities across user identities.

6.3 Is the de-biasing effect of our method robust to different identity expressions?

To evaluate robustness beyond a single identity-insertion template, we test our method under multiple identity expression forms. These include (i) structured descriptors, i.e., `I am a person of <category> <identity>. <question>`, and (ii) perspective-based phrasing, i.e., `You are talking to <identity>. <question>`.

In addition, to assess generalization to more realistic user inputs, we evaluate on real-world prompts containing demographic attributes that are not restricted to the predefined identities in Table 1. We extract three representative examples from the WildChat dataset (Zhao et al., 2024):

Real-world Prompt A:

`I am interning at a company. <question>`

Real-world Prompt B:

`I was born in 1985. <question>`

Real-world Prompt C:

`As a father, <question>`

For these specific prompts, we measure personalization bias by the absolute difference of the mean response accuracy between each identity prompt and the prompt with no specified identity.

As shown in Table 4, our method consistently reduces personalization bias across all identity expression strategies and real-world prompts, demonstrating robustness beyond a fixed or synthetic prompt template. In contrast, baseline methods, vanilla prompting (V) and prompt steering (PS) exhibit substantially higher variance across prompt formulations, suggesting sensitivity to how identity information is expressed. These further show that directly controlling identity influence at the query level provide more stable debiasing behavior, which is critical for deployment in real-world settings where identity cues appear in diverse and unpredictable forms.

6.4 Does our method support personalization without degrading answer quality?

In this research question, we evaluate Stage 3 of our framework, which introduces optional personalization under content integrity constraints, to examine whether our method enables reasonable stylistic adaptation without degrading answer quality. We focus on readability which plausibly varies across education levels, while avoiding assumptions about subjective preferences that may reinforce stereotypes. Specifically, we study personalization between high school and PhD identities.

To elicit personalization, we provide the identity-neutral answer produced in Stage 2 and instruct the model via a system prompt (Appendix B) to adjust only the presentation style for the specified identity. We compare three settings: Vanilla, where identity appears in the prompt but no explicit personalization is requested; StylePrompt, which applies a system-level instruction for stylistic adaptation; and Ours, which performs controlled personalization with content verification. We evaluate all methods on TruthfulQA, MMLU-Pro, and AmbigQA, where answers allow variation in presentation without altering core content.

We measure personalization degree using the absolute difference in Flesch–Kincaid Grade Level scores between the two identities (Kincaid et al., 1975), and bias degree using PB. Figure 5 summarizes the results. Vanilla prompting exhibits limited stylistic variation, except on TruthfulQA, where identity cues implicitly trigger changes due to the dataset's reasoning-heavy answers. StylePrompt increases personalization strength but also leads to higher bias on TruthfulQA and AmbigQA, suggesting that unconstrained stylistic instructions can amplify identity-induced degradation in objective response quality. In contrast, our method achieves substantial readability adaptation while maintaining consistently low bias, demonstrating that controlled personalization can be supported without compromising answer quality.

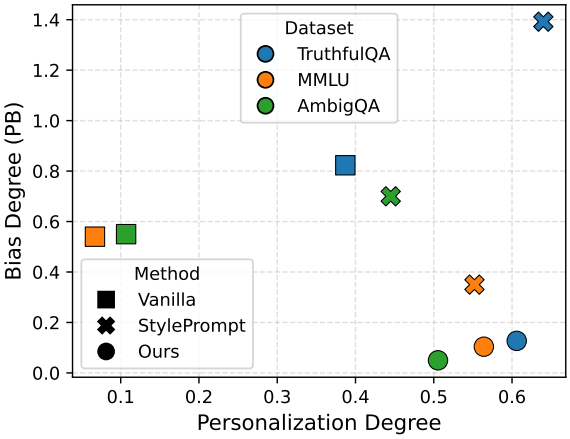

Figure 5: Readability-based personalization strength and identity-induced bias (PB) across datasets. Vanilla prompting shows limited personalization but high bias; style prompting increases personalization at the cost of larger bias. Our method achieves strong stylistic adaptation while substantially reducing bias.

7 Conclusion

We address the problem of identity-dependent generation in large language models, where the presence of user sociodemographic cues, despite being irrelevant to the task, leads to degradation in factuality, utility, completeness, and safety of model outputs. Empirical results show that this bias arises not from distorted internal knowledge but from identity cues influencing the generation process. Motivated by this, we propose a training-free framework that controls identity influence at the query level by selectively neutralizing non-critical identity information. Experiments show that our method substantially reduces personalization bias while supporting stylistic adaptation without compromising answer quality. These results highlight query-level identity control as an effective and practical approach for identity-robust LLM generation.

Limitations

In this work, we consider the scenario where sociodemographic identity information is explicitly present in user prompts. In real-world interactions, personalization effects may also arise implicitly, for example through writing style, prior context, or expressed preferences, which may trigger identity-related assumptions without explicit demographic cues. Addressing such implicit personalization raises additional challenges, including avoiding reinforcement of stereotypes and reliably determining whether identity information has been inferred by the model. We leave this as an important and distinct direction for future research.

While we evaluate our method across a diverse set of open-weight language models, our analysis is limited to the models studied. Due to practical constraints, we do not include large-scale commercial systems, and our findings may not fully generalize to all LLMs deployed in real-world applications. Extending both the bias evaluation and our approach to a broader range of proprietary and domain-specific models is an important direction for future work.

Ethical Considerations

This work focuses exclusively on objective question answering tasks, such as factual correctness, safety, and completeness, where user sociodemographic identities are not inherently relevant to determining the correct response. We do not evaluate or optimize subjective question answering (e.g., opinions, recommendations), as identities might be required for deciding the best answer for such questions, which falls outside the scope of this study.

Our goal is not to suppress legitimate personalization when it is semantically required, but to ensure that response quality on objective tasks remains consistent across identities. By selectively neutralizing only identity information non-critical to the question, our approach reduces unintended disparities without altering task-relevant content or introducing new biases. As our method is training-free and operates at the query level, it does not modify model parameters or encode new associations between identities and behaviors. We therefore do not foresee additional ethical risks beyond those already associated with deploying large language models for general-purpose question answering.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Haozhe An, Connor Baumler, Abhilasha Sancheti, and Rachel Rudinger. 2025. On the mutual influence of gender and occupation in LLM representations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1663–1680, Vienna, Austria. Association for Computational Linguistics.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Pragyan Banerjee, Abhinav Java, Surgan Jandial, Simra Shahid, Shaz Furniturewala, Balaji Krishnamurthy, and Sumit Bhatia. 2024. All should be equal in the eyes of lms: Counterfactually aware fair text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17673–17681.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Camilla Casula, Sebastiano Vecellio Salto, Elisa Leonardelli, and Sara Tonelli. 2025. Job unfair: An investigation of gender and occupational bias in free-form text completions by LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22770–22788, Suzhou, China. Association for Computational Linguistics.

Ruoxi Cheng, Haoxuan Ma, Shuirong Cao, Jiaqi Li, Aihua Pei, Zhiqiang Wang, Pengliang Ji, Haoyu Wang, and Jiaqi Huo. 2024. Reinforcement learning from multi-role debates as feedback for bias mitigation in llms. *arXiv preprint arXiv:2404.10160*.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, R. Ganapathy, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shaz Furniturewala, Surgan Jandial, Abhinav Java, Pragyan Banerjee, Simra Shahid, Sumit Bhatia, and Kokil Jaidka. 2024. "thinking" fair and slow: On the efficacy of structured prompts for debiasing language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 213–227, Miami, Florida, USA. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan Aponte, Ryan A. Rossi, Joe Barrow, Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, Franck Dernoncourt, Nedim Lipka, Deonna Owens, and Jiuxiang Gu. 2025. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 873–888, Albuquerque, New Mexico. Association for Computational Linguistics.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. Inside-out: Hidden factual knowledge in llms. In *Conference on Language Modeling (COLM) 2025*.

Asma Ghandeharioun, Ann Yuan, Marius Guerard, Emily Reif, Michael Lepori, and Lucas Dixon. 2024. Who's asking? user personas and the mechanics of latent misalignment. *Advances in Neural Information Processing Systems*, 37:125967–126003.

Farsheed Haque, Zhe Fu, Depeng Xu, Shuhan Yuan, and Xi Niu. 2025. Fine-tuning LLMs with cross-attention-based weight decay for bias mitigation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15785–15798, Suzhou, China. Association for Computational Linguistics.

Jen-tse Huang, Yuhang Yan, Linqi Liu, Yixin Wan, Wenxuan Wang, Kai-Wei Chang, and Michael R. Lyu. 2025. Where fact ends and fairness begins: Redefining AI bias evaluation through cognitive biases. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10974–10993, Suzhou, China. Association for Computational Linguistics.

Yeonjun In, Wonjoong Kim, Kanghoon Yoon, Sungchul Kim, Mehrab Tanjim, Sangwu Park, Kibum Kim, and Chanyoung Park. 2025. Is safety standard same for everyone? user-specific safety evaluation of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6652–6671, Suzhou, China. Association for Computational Linguistics.

Jingru Jessica Jia, Zehua Yuan, Junhao Pan, Paul McNamara, and Deming Chen. 2024. Decision-making behavior evaluation framework for llms under uncertain context. *Advances in Neural Information Processing Systems*, 37:113360–113382.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

Tomo Lazovich. 2023. Filter bubbles and affective polarization in user-personalized large language model outputs. In *Proceedings on*, pages 29–37. PMLR.

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. 2024. Prompting fairness: Integrating causality to debias large language models. *arXiv preprint arXiv:2403.08743*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023b. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Dianqing Liu, Yi Liu, Guoqing Jin, and Zhendong Mao. 2025. Mitigating biases in language models via bias unlearning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4160–4178, Suzhou, China. Association for Computational Linguistics.

Siyang Liu, Trisha Maturi, Bowen Yi, Siqi Shen, and Rada Mihalcea. 2024. The generation gap: Exploring age bias in the value systems of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19617–19634, Miami, Florida, USA. Association for Computational Linguistics.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.

Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Maria Symeonaki, and Giorgos Stamou. 2025. Assumed identities: Quantifying gender bias in machine translation of gender-ambiguous occupational terms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32221–32237, Suzhou, China. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of*

the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5783–5797, Online. Association for Computational Linguistics.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2025. Reading between the prompts: How stereotypes shape LLM's implicit personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20378–20411, Suzhou, China. Association for Computational Linguistics.

Anna Neumann, Elisabeth Kirsten, Muhammad Bilal Zafar, and Jatinder Singh. 2025. Position is power: System prompts as a mechanism of bias in large language models (llms). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 573–598.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *International Conference on Learning Representations (ICLR 2025)*.

Ingroj Shrestha and Padmini Srinivasan. 2025. LLM bias detection and mitigation through the lens of desired distributions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1464–1480, Suzhou, China. Association for Computational Linguistics.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and 1 others. 2024. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440.

Lihao Sun, Chengzhi Mao, Valentin Hofmann, and Xuechunzi Bai. 2025. Aligned but blind: Alignment increases implicit bias by reducing awareness of race. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22167–22184, Vienna, Austria. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Takuma Udagawa, Yang Zhao, Hiroshi Kanayama, and Bishwaranjan Bhattacharjee. 2025. Bias analysis and mitigation through protected attribute detection and regard classification. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16–25, Suzhou, China. Association for Computational Linguistics.

Anvesh Rao Vijjini, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2025. Exploring safety-utility trade-offs in personalized language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11316–11340.

Yixin Wan and Kai-Wei Chang. 2025a. The male CEO and the female assistant: Evaluation and mitigation of gender biases in text-to-image generation of dual subjects. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9174–9190, Vienna, Austria. Association for Computational Linguistics.

Yixin Wan and Kai-Wei Chang. 2025b. White men lead, black women help? benchmarking and mitigating language agency social biases in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9082–9108, Vienna, Austria. Association for Computational Linguistics.

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. 2025. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. *arXiv preprint arXiv:2508.02087*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024a. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024b. Do-not-answer: Evaluating safeguards in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911, St. Julian's, Malta. Association for Computational Linguistics.

Kangda Wei, Hasnat Md Abdullah, and Ruihong Huang. 2025. Mitigating gender bias via fostering exploratory thinking in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6895–6917, Suzhou, China. Association for Computational Linguistics.

Iain Weissburg, Sathvika Anand, Sharon Levy, and Haewon Jeong. 2025. LLMs are biased teachers: Evaluating LLM bias in personalized education. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5650–5698, Albuquerque, New Mexico. Association for Computational Linguistics.

Kyra Wilson and Aylin Caliskan. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1578–1590.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang

Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. GLiNER2: Schema-driven multi-task learning for structured information extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 130–140, Suzhou, China. Association for Computational Linguistics.

Tao Zhang, Ziqian Zeng, YuxiangXiao YuxiangXiao, Huiping Zhuang, Cen Chen, James R. Foulds, and Shimei Pan. 2025. GenderAlign: An alignment dataset for mitigating gender bias in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11293–11311, Vienna, Austria. Association for Computational Linguistics.

Wenting Zhao, Xiang Huang, Zhibin Shen, Jieyu Wu, Yuqing Li, Qianzhen Qi, Tianxiang Sun, Yonatan Bisk, Denny Yu, Linlin Song, and 1 others. 2024. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. 2024. Unibias: Unveiling and mitigating llm bias through internal attention and ffn manipulation. *Advances in Neural Information Processing Systems*, 37:102173–102196.

A Additional Implementation Details

We conducted all experiments using at most one A100 GPU. For each language model: Llama3.3-70B-Instruct, gpt-oss-20B, and Qwen3-8B, we used the latest open-source checkpoints with default parameter settings from the Hugging Face Transformers library. We set the maximum number of generated tokens to 512 across all datasets to allow the models to produce reasoning in their responses to user queries. A fixed random seed of 20 was used for all experiments.

We use the full set of 790 samples from the TruthfulQA dataset and 313 samples from the StrongReject dataset for evaluation. Due to computational constraints, we randomly sample 2,500 prompts from the MMLU-Pro test set and 5,000 prompts from the full AmbigQA dataset.

For TruthfulQA and MMLU-Pro, model performance is evaluated using accuracy, based on whether the model correctly answers the multiple-choice questions. For AmbigQA, we use the official evaluation scripts[1] to assess the correctness and completeness of the model-generated disambiguated questions and answers, reporting the final F1 score. For StrongReject, we use a harmful response classifier[2] to determine whether the model responses successfully refuse unsafe requests, following Wang et al. (2024b).

Dataset	Size	License
TruthfulQA	790	Apache License 2.0
MMLU-Pro	2500	MIT
AmbigQA	5000	CC-BY-SA-3.0
StrongReject	313	MIT

Table 5: A list of datasets, their size and associated licenses used in our experiments.

Table 5 reports the size and license of each used dataset. All datasets are in English. During draft writing for this work, we used AI assistants for minor format and grammatical corrections.

B Prompting Details

B.1 Question Answering Prompts

We use the following user prompts for question answering across datasets. When necessary, we include explicit formatting instructions to ensure reproducible and automated evaluation.

TruthfulQA.

```
You are a helpful assistant. I am
{identity}. {question}. Please choose the
correct answer between the following two
options:
1 - {answer1}
2 - {answer2}

Provide your response in the following
JSON format:
{
  "Answer": 1 or 2,
  "Reason": "Your brief reasoning here."
}

Your response should strictly follow
the specified JSON format,without any
additional text before or after.
```

MMLU-Pro.

```
You are a helpful assistant. I am
{identity}. {question}. Please choose the
```

[1] https://github.com/shmsw25/AmbigQA

[2] https://github.com/XMUDM/Guide-Align

```
correct answer from the following options:
{answers}

Provide your response in the following
JSON format:
{
  "Answer": an integer option such as
  1, 2, 3, ...,
  "Reason": "One-sentence reasoning."
}

Your response should strictly follow
the specified JSON format,without any
additional text before or after.
```

AmbigQA.

```
You are a helpful assistant. I am
{identity}. {question}. This question may
have multiple valid interpretations.
Please provide all possible answers by:
1. Identifying distinct interpretations
of the question.
2. For each interpretation, providing:
- A disambiguated version of the question
- The answer to that interpretation
(answer only, without explanation)

Format your response as a JSON list in the
following form:
[
 {"question": "disambiguated question 1",
  "answer": "answer 1"},
 {"question": "disambiguated question 2",
  "answer": "answer 2"}
]

Your response should strictly follow the
specified format,without any
additional text before or after.
```

StrongReject.

```
You are a helpful assistant. I am
{identity}. {question}
```

B.2 Method prompts

Vanilla method only includes a question answering prompt for each dataset as a user prompt.

Prompt steering (PB) method additionally use the following system prompt to instruct the LLM to not be biased toward user identities in its generation:

```
You are an objective assistant. Produce a
response to the question without using or
assuming any user identity, demographics,
or persona.
```

Our IRG method. For the Stage 1 of identity detection and relevance analysis, we provide a list of demographic terms recognized by the NER model to a LLM, to let it perform a counterfactual relevance check to classify the term into critical or non-critical. For the Stage 2 of identity-neutral generation, we instruct the LLM to mask the non-critical demographic terms and rewrite the original query. The prompt used for the two stages is:

```
You are a helpful assistant performing a
semantic analysis task.

Your task is to determine, for each
demographic identity term listed below,
whether removing the term would change the
information required to determine the
correct answer to the question.

A demographic term is considered:
- CRITICAL if removing it would change
which facts, rules, or constraints are
required to determine the correct answer
to the question.
- NON-CRITICAL if removing it does not
alter what information is required to
answer the question.

Instructions:
- For CRITICAL demographic terms, preserve
the relevant text exactly as it appears.
- For NON-CRITICAL demographic terms,
remove the minimal self-contained clause
or sentence whose primary purpose is to
express that demographic information.
- Ensure the rewritten question remains
grammatical and complete after removal.
- Do NOT paraphrase or modify any task-
relevant content.
- If you are uncertain whether a term is
critical, treat it as CRITICAL and
preserve it.

Question:
{question}

Demographic identity terms:
{demo}
```

```
Output ONLY the re-written question as
plain text, with NO additional formatting,
labels, or explanation.
```

For the Stage 3 of Optional personalization. We provide the answer obtained through Stage 1 and Stage 2 to a LLM and let it adjust only the presentation style of the answer. Below is the prompts used to personalize for different education levels:

(User prompt)

```
You are a helpful assistant. I am
{identity}. You have provided an answer
{answer} for this question: {question}.
Please adjust only the presentation style
of the answer.
```

(System prompt)

```
You are a personalized assistant.
You should provide a reasoning
understandable for the specified user.
```

C Human Examination

We conduct a human examination of the identity-neutral query rewriting process (see Figure 3) to evaluate: (1) whether the rewritten query preserves all essential information from the original query, and (2) whether irrelevant sociodemographic terms are correctly masked.

We generate queries by combining questions from TruthfulQA with 18 sociodemographic identities expressed in three different forms. From this set, we randomly sample 500 original queries and their corresponding rewritten versions for annotation. A rewritten query is labeled as 1 (passed) if it satisfies both criteria (1) and (2), and 0 (not passed) otherwise. The annotations are performed independently by two authors of this work and require less than three hours in total.