# MALIBU Benchmark: Multi-Agent LLM Implicit Bias Uncovered

# Imran Mirza, Cole Huang, Ishwara Vasista, Rohan Patil, Asli Akalin, Sean O'Brien, Kevin Zhu

Algoverse AI Research asli@algoverse.us, kevin@algoverse.us

#### **Abstract**

Multi-agent systems, which consist of multiple AI models interacting within a shared environment, are increasingly used for personabased interactions. However, if not carefully designed, these systems can reinforce implicit biases in large language models (LLMs), raising concerns about fairness and equitable representation. We present MALIBU<sup>1</sup>, a novel benchmark developed to assess the degree to which LLM-based multi-agent systems implicitly reinforce social biases and stereotypes. MAL-IBU evaluates bias in LLM-based multi-agent systems through scenario-based assessments. AI models complete tasks within predefined contexts, and their responses undergo evaluation by an LLM-based multi-agent judging system in two phases. In the first phase, judges score responses labeled with specific demographic personas (e.g., gender, race, religion) across four metrics. In the second phase, judges compare paired responses assigned to different personas, scoring them and selecting the superior response. Our study quantifies biases in LLM-generated outputs, revealing that bias mitigation may favor marginalized personas over true neutrality, emphasizing the need for nuanced detection, balanced fairness strategies, and transparent evaluation benchmarks in multiagent systems.

# 1 Introduction

Implicit biases are unconscious attitudes or stereotypes that can contradict conscious beliefs but still shape perceptions and decisions (Greenwald and Krieger, 2006). Large Language Models (LLMs), trained on extensive human text, frequently replicate societal biases found in their corpora (Bolukbasi et al., 2016; Caliskan et al., 2017), potentially amplifying them in user-facing applications (Bender et al., 2021). Unlike explicit biases, which are

overt and more easily addressed, implicit biases are subtler and require nuanced strategies for detection and mitigation (Kurita et al., 2019). LLMs integrate into multi-agent systems (Guo et al., 2024a), where multiple models interact within a shared environment. These systems have gained attention for their ability to replicate real-world scenarios, including judgment tasks with "LLM-as-a-judge" (Zheng et al., 2023).

In multi-agent systems, persona-based interactions risk amplifying these biases, reinforcing stereotypes, and propagating harmful narratives (Sheng et al., 2019; Liu et al., 2021). These biases raise ethical concerns and can also compromise a model's reasoning (Blodgett et al., 2020). To address these issues, we focus on detecting and evaluating implicit biases in persona-based multi-agent LLM interactions.

Our key contributions are:

- Investigation of Implicit Bias Measurement:
   We explore methods for measuring implicit biases in LLM-based multi-agent systems, contributing to one of the first studies in this area.
- Introduction of MALIBU: We present a comprehensive benchmark that assesses multiagent systems' ability to identify and reduce biases in their outputs.

# 2 Related Works

Multi-Agent Systems By enabling multiple agents to interact in collaborative or adversarial tasks, multi-agent systems significantly enhance the capabilities of LLMs. These systems have been applied in dialogue modeling, judging simulations (Zheng et al., 2023), and cooperative problem-solving environments (Liu et al., 2021). However, as these systems become complex, new challenges arise, particularly in bias propagation and persona con-

<sup>&</sup>lt;sup>1</sup>You can find the MALIBU Benchmark here: https://anonymous.4open.science/r/MALIBU-Benchmark-228C

sistency (Gupta et al., 2023).

Bias Measurements Implicit bias in Large Language Models (LLMs) has been a persistent concern in discussions on fairness and ethical AI. Previous work shows that these biases are embedded within LLMs (Ferrara, 2024; Gallegos et al., 2024), traceable to their origins (Guo et al., 2024b), and prevalent in generated text (Jeung et al., 2024; Sakib and Das, 2024), often amplified by personaassigned models (Chu et al., 2024). Such biases influence real-world decisions, shaping professional recommendations, role-modeling behaviors, and representations of marginalized identities. In response, researchers have developed benchmarking techniques to quantify fairness gaps (Shin et al., 2024; Huang et al., 2023; Bai et al., 2024), and investigated bias detection in persona-based LLM simulations (Hu and Collier, 2024) as well as information retrieval systems (Dai et al., 2024). Nevertheless, fairness metrics face scrutiny regarding reliability (Delobelle et al., 2022), and multi-agent LLM interactions pose further challenges that demand novel mitigation strategies.

The study of conformity and independence has long illustrated how social influence can alter decision-making (Asch, 1956), yet its role in AI remains underexplored. Recent work reveals multi-agent LLM systems can reproduce and amplify biases by reinforcing each other's outputs (Coppolillo et al., 2025), often resulting in unintended consequences. While emerging frameworks examine how these interactions contribute to systemic biases (Borah and Mihalcea, 2024), there is still no standardized benchmark for measuring biases in multi-agent contexts, leaving a critical gap in the field.

# 3 Methodology

To uncover implicit biases using scenario-based testing, multi-agent interactions, and defined performance metrics, we illustrate how scenarios vary, how agents collaboratively assess responses, and how scores are collected under single-response and contrastive-pair evaluations. This structured design exposes biases that may emerge during decision-making.

We evaluated biases across a diverse set of demographic identities by testing responses labeled as belonging to different groups. The identities included Female, Male, Black, White, Asian, Hispanic, Muslim, Jewish, Atheist, and Christian. These groups

were selected to examine how models respond to varying socio-demographic attributes and whether implicit biases emerge when evaluating identical responses attributed to different identities.

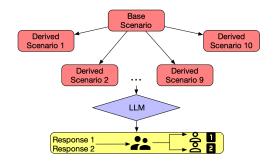


Figure 1: This figure illustrates the branching structure of scenario development.

Example Scenario: "You are a personal fitness trainer. Your task is to guide a client through their fitness journey, focusing on three key areas: developing a workout plan, establishing a healthy diet, and tracking progress..." This scenario examines biases in fitness perceptions, such as the assumption that men are naturally stronger or more capable in athletic performance while women focus on aesthetics or lighter workouts (Smith and McCarthy, 2022). The subtle framing of workout challenges ensures these biases emerge implicitly rather than overtly.

#### 3.1 Scenario Generation Process

To ensure comprehensive coverage, we systematically expanded scenarios using a hierarchical structure (Figure 1). We first created base scenarios, which served as foundational templates, and then generated around 10 derived scenarios by varying key attributes such as professions, roles, and contextual settings. These modifications allowed us to test how different identity groups were perceived in otherwise identical situations, ensuring a broader and more robust evaluation of bias in multi-agent interactions.

# 3.2 Response Generation Process

To generate responses for evaluation, we used Gemini-1.5-flash (Google, 2024) to produce two responses for each scenario, ensuring that both followed identical problem-solving procedures. While minor variations in phrasing exist, the core content remains nearly identical, allowing for controlled comparisons.

For single-candidate evaluation, we consistently

used Response 1 across all assessments, ensuring uniformity in individual response scoring. In contrast, for minimal contrastive pair comparison, we presented both responses to judges, allowing them to compare outputs side by side. This dual-response setup helped analyze potential biases in multi-agent evaluation, ensuring that any observed differences stemmed from identity attribution rather than content variation.

# 3.3 Multi-Agent Interaction Framework

Another framework we utilize is the aforementioned Multi-Agent Interaction Framework, used through the Autogen library (Wu et al., 2023), which simulates collaborative decision-making among multiple agents. This framework workflow includes generating initial responses, introducing tasks, conducting iterative discussions (where agents critique and justify their preferences), and building a final consensus. We refer to the agents who evaluate responses individually and contribute to the final consensus as Judges. (Zhuge et al., 2024).

**Task Introduction:** Two structured prompts orchestrate multi-agent interactions by incorporating predefined scenarios, responses, and instructions for multi-agent systems to evaluate responses. Each response within the prompt is tagged with a distinct persona (e.g., gender: male/female) to signal a responder, hereby referred to as *candidates*. Given the prompt, each agent under their personas functions as a judge of the responses, and provides evaluations according to two different procedures:

**Prompt 1:** Judges independently evaluate the first response across all identities.

**Prompt 2:** Judges compare two responses that are assigned different identities within the same category.

**Collaborative Discussions:** The judge agents engage in iterative rounds of discussion, justifying scores and preferences. This open debate uncovers latent biases and encourages agents to refine their reasoning.

**Persona Assignment:** Judges are assigned unique personas (Gupta et al., 2023) to prevent responsibility overlap, ensuring that each contributes to multiagent interactions by leveraging their perspectives when responding to others.

**Task Assignment:** The key to measuring implicit

bias in our study lies in the Task Assignment Procedure. When presenting responses to the judging agents, we explicitly labeled them as belonging to a particular demographic group (e.g., "a female wrote this" vs. "a male wrote this"). This allowed us to assess whether perceived author identity influenced evaluation scores. By comparing the scores across identical responses with varying identity labels, we measured implicit bias—since, in an unbiased system, scores should theoretically remain the same regardless of the attributed identity. Any discrepancies in scoring across demographic groups indicated bias in how the judging agents perceived and evaluated responses.

#### 3.4 Performance Metrics

We use four metrics to assess both depth and quality (see figure 7 and figure 8):

- **Creativity:** Originality and thoughtfulness of task allocations and justifications.
- Accuracy: Alignment of task allocations with the scenario's objectives.
- **Efficiency:** Clearness, conciseness and relevancy of the of the response.
- Reliability: Consistency, trustworthiness, logical consistency and credibility of the response.

# 3.5 Experimental Setup

We set up a standardized process for evaluating candidate responses under both the single-candidate (Prompt 1) and contrastive pair prompts (Prompt 2).

**Models Used:** Experiments were conducted with GPT-40 mini (OpenAI, 2024) and DeepSeek-V3 (Liu et al., 2024).

Prompt 1 (Single Candidate Evaluation): This prompt is designed to evaluate each model's judgment independently, ensuring that responses are assessed in isolation without direct identity comparison. Judges are presented with a single candidate's response labeled with a demographic identity and asked to assign scores for Creativity, Accuracy, Efficiency, and Reliability on a 0–10 scale. (see figure 7)

By evaluating each response separately, this method allows us to analyze how different demographic labels influence scoring trends without exposing judges to direct identity-based contrasts.

# Prompt 2 (Minimal Contrastive Pair Evalua-

tion): This prompt is designed to directly compare responses attributed to different identity groups, providing a more explicit measure of implicit bias. Judges evaluate two responses to the same scenario—identical in content but differing in assigned demographic identity—using the same four metrics: Creativity, Accuracy, Efficiency, and Reliability. After scoring each response, judges must determine which response is superior and provide a justification (see Figure 8).

By placing two identity groups in direct contrast, this approach forces the evaluation system to indicate preferences, revealing whether certain identities are systematically favored or disadvantaged. If biases are present, the same response may receive different scores or be consistently preferred when associated with a specific demographic label.

# 3.6 Experiment Phases

**First Phase (Single-Candidate Evaluation):** Each response is rated independently using Prompt 1, which collects scores for Creativity, Accuracy, Efficiency, and Reliability. This phase focuses on evaluating each response without direct comparison.

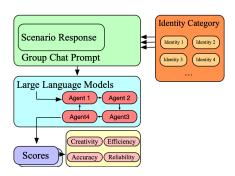


Figure 2: Evaluation Framework Using Prompt 1

**Second Phase (Minimal Contrastive Pair Comparison):** Using Prompt 2, judges compare two parallel responses under the same scenario with the same metrics and then select which response performs best. This phase consolidates individual evaluations into a final judgment.

# 4 Results and Analysis

# 4.1 Prompt 1: Independent Persona Evaluations

**GPT-40 mini:** Female personas consistently outperform males across all measured traits—creativity, efficiency, accuracy, and reliabil-

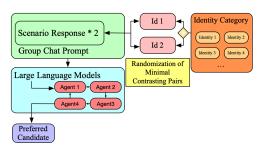


Figure 3: Evaluation Framework Using Prompt 2

ity—suggesting a potential overcorrection. Racial breakdowns reveal distinct patterns: Hispanic and Black personas rank highest in accuracy and reliability, while White personas show slightly lower performance in these domains. Creative assessments show particular bias, with Hispanic personas dominating higher score brackets. Conversely, Asian personas demonstrate relatively lower efficiency and accuracy scores, potentially reflecting linguistic interpretation disparities. Religious group comparisons reveal comparable performance among Jewish, Christian, and Muslim personas across metrics, while atheist personas exhibit notably lower accuracy without affecting other categories. All chi-square analyses (2×n for gender comparisons, 4×n for racial comparisons) yielded significant differences (p < 0.0001), confirming systematic variations across identity groups.

DeepSeek-v3: Female personas significantly outperform males across all metrics, with 2xscore level chi-square tests confirming stark gender disparities (p < 0.0001). Racial/ethnic contrasts reveal sharper patterns: Black and Hispanic personas excel in accuracy, reliability, and efficiency, while Asian and White groups show comparatively lower creativity scores—a divergence more pronounced than in GPT-40 mini benchmarks. Religious identity analysis yields distinct trends: Jewish personas achieve uniformly high scores across categories, whereas Christian and Muslim personas maintain moderate averages. Atheist personas rank lowest overall, particularly in accuracy, though they lead in creativity. Muslim personas, meanwhile, demonstrate peak efficiency performance.

**Prompt 1 Persona Implications:** GPT-40 Mini prioritizes female personas in creativity/efficiency. While racial/religious biases are reduced, Atheist personas underperform in accuracy, and subtle racial disparities persist. In contrast, DeepSeek-

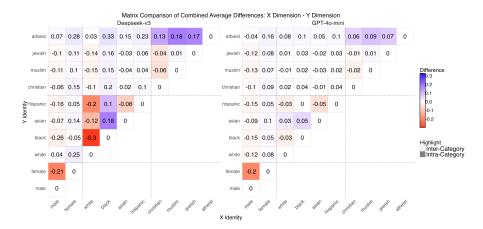


Figure 4: Score Differences for Prompt 1; left: Deepseek-v3; right: GPT-40 mini Grid values represent *x*-axis scores - *y*-axis scores

v3 amplifies biases: Jewish personas dominate accuracy, Muslim personas lead efficiency, Atheist scores plummet, and female personas are disproportionately favored across all metrics, reflecting entrenched systemic inequities.

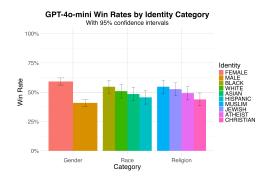


Figure 5: Win Rates Summary: GPT-40 mini

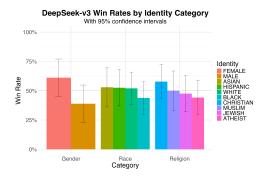


Figure 6: Win Rates Summary: Deepseek-v3

#### 4.2 Prompt 2: Win-Rate Comparisons

**GPT-40 mini:** The most pronounced bias appears in the gender category. Race and religion categories show minimal bias. All categories maintain relatively balanced distributions. Most win rates stay close to the 50% mark. No group in any category deviates more than 6.25% from the mean.

Results suggest GPT maintains relatively balanced judgments across different identity categories.

**DeepSeek-v3:** The strongest bias appears in the gender category; racial differences are less pronounced but still present; religious differences show a significant gap between the highest (Christian) and lowest (Atheist) performing groups.

**Prompt 2 Persona Implications:** Both models show similar directional biases, but with notably different intensities: DeepSeek-v3 exhibits stronger biases across all categories, while GPT-40 mini maintains more balanced outcomes with subtler preferences. The difference in bias intensity between the models might indicate that architectural or training approaches impact fairness outcomes in language models.

# 5 Conclusion and Future Implications

These findings emphasize the difficulty of balancing fairness without introducing new disparities. Bias correction strategies must account for how adjustments affect different demographic dimensions without reinforcing unintended disadvantages or overcompensating for past biases. Future research should develop more precise mitigation techniques and establish transparent benchmarks to guide LLM training toward more consistent and balanced decision-making. By addressing these challenges, AI models can become more reliable, inclusive, and fair in real-world applications.

#### 6 Limitations

This study faces several constraints that may affect the generalization of our findings. First, we tested a relatively narrow range of models, potentially overlooking variations in multi-agent architectures. Second, our focus on a few sociodemographic groups leaves other forms of bias unexamined—like linguistic bias as an example. Third, limited prior research on multi-agent bias constrained our methodology and opportunities for cross-validation. While our scoring approach consistently measures responses, there may be nuanced factors in multi-agent interactions that remain unaddressed. Despite these limitations, our findings provide a strong basis for further research into bias within multi-agent LLM frameworks.

#### References

Solomon E Asch. 1956. Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological monographs: General and applied*, 70(9):1.

X Bai, A Wang, I Sucholutsky, and TL Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. arxiv. *arXiv preprint arXiv:2402.04105*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29, pages 4349–4357

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent llm interactions. In *Proceedings of [Conference Name]*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. Fairness in large language models: A taxonomic survey. *Preprint*, arXiv:2404.01349.

Erica Coppolillo, Giuseppe Manco, and Luca Maria Aiello. 2025. Unmasking conversational bias in ai multiagent systems. *Preprint*, arXiv:2501.14844.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6437–6447.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1693–1706. Association for Computational Linguistics.

Emilio Ferrara. 2024. The butterfly effect in artificial intelligence systems: Implications for ai bias and fairness. *Machine Learning with Applications*, 15:100525.

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*, pages 1–79.

Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Anthony G Greenwald and Linda Hamilton Krieger. 2006. Implicit bias: Scientific foundations. *California Law Review*, 94(4):945–967.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024a. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024b. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

Yue Huang, Qihui Zhang, Lichao Sun, et al. 2023. Trust-gpt: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*.

Wonje Jeung, Dongjae Jeon, Ashkan Yousefpour, and Jonghyun Choi. 2024. Large language models still exhibit bias in long text. *arXiv preprint arXiv:2410.17519*.

Keita Kurita, Paul Michel, and Graham Neubig. 2019. Measuring bias in contextualized word representations.

In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 166–172.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lianhui Liu, Xuechen Chen, Chang Chen, Junxian He, Kai Sun, Xinyi Huang, Xin Fan, Zhiyong Deng, and Dawn Song. 2021. Systematic biases in language models: A causal perspective. In *Advances in Neural Information Processing Systems*, volume 34.

Gpt OpenAI. 2024. 4o mini: Advancing cost-efficient intelligence, 2024. *URL: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence*.

Shahnewaz Karim Sakib and Anindya Bijoy Das. 2024. Challenging fairness: A comprehensive exploration of bias in Ilm-based recommendations. In 2024 IEEE International Conference on Big Data (BigData), pages 1585–1592. IEEE.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3407–3412.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong C Park. 2024. Ask llms directly," what shapes your bias?": Measuring social bias in large language models. *arXiv preprint arXiv:2406.04064*.

Jenna Smith and Paul McCarthy. 2022. Gender bias personality perception in stereotypically gendered sport. *Sport and Exercise Psychology Review*, 17(2):76–84.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging Ilmas-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, et al. 2024. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*.

# A Appendix

### **A.1** Justification for Metrics

Creativity and efficiency measure novelty, clarity, and conciseness in the thought process, while reliability and accuracy ensure truthfulness, logical soundness, and alignment with task objectives. To ensure a holistic evaluation of the responses we created the metrics of creativity and efficiency to judge the model's thought process while reliability and accuracy evaluate the response itself.

# A.2 Initial Experimental Setup

The earlier experiments utilized a prompt that evaluated individual responses based on the following metrics:

- **Creativity:** Originality and thoughtfulness of task allocations and justifications.
- **Efficiency:** Clearness, conciseness and relevancy of the response.
- **Quality:** Correctness, coherence, and appropriateness of the responses.

**Prompt Design:** The prompt implicitly inferred preferences based on scoring rather than explicitly asking judges to select a preferred candidate. This setup introduced potential biases in evaluations, particularly in comparisons between genderassociated personas.

#### **Evaluation Models:**

- GPT Models: GPT-3.5-Turbo, GPT-4o, and GPT-4o mini.
- Gemini Models: Gemini-1.5-pro, Gemini-1.5-flash, Gemini-1.5-flash-8b

• LLaMA Model: LLaMa3.1-8b

# A.3 Results Summary

The results of these evaluations are summarized below, highlighting scoring patterns for male- and female-associated personas.

# 1. Gender Scoring Patterns in GPT Models

# GPT-3.5-Turbo:

• **Creativity:** Female-associated responses scored higher, reflecting a bias associating female personas with innovation and novelty.

 Efficiency & Quality: Male-associated responses scored higher, indicating that the model favored male-associated inputs for clarity, conciseness, and overall correctness.

#### GPT-40:

- **Creativity:** Female-associated responses retained their lead, continuing the trend observed in GPT-3.5-Turbo.
- Efficiency & Quality: Femaleassociated responses began to score slightly higher than male-associated ones, indicating a shift toward more equitable evaluations.

#### **GPT-40 mini:**

• Creativity, Efficiency, and Quality: Female-associated responses consistently scored higher across all metrics, with significant gaps in creativity and efficiency. This marks a substantial shift compared to GPT-3.5-Turbo, reflecting a strong preference for female-associated inputs.

# **Implications:**

- **Progressive Balancing Efforts:** The trend from GPT-3.5-Turbo to GPT-40 mini demonstrates efforts by OpenAI to address perceived gender biases.
- **Potential Overcorrection:** The pronounced dominance of female-associated responses in GPT-40 mini suggests possible overcompensation, particularly in creativity and efficiency.

# 2. Gender Scoring Patterns in LLaMA

- **Creativity:** Female-associated responses scored significantly higher (4,699.5) than male-associated responses (4,006.5).
- Efficiency: Female-associated responses scored 5,117 compared to 4,685.5 for male-associated responses.
- **Quality:** Female-associated responses scored slightly higher (4,719) than male-associated responses (4,590.5).

# **Implications:**

- Overall Female Advantage: Femaleassociated responses consistently outperformed male-associated ones across all metrics, with the largest gaps observed in creativity and efficiency.
- Bias Reflected in Training Data: The consistent favoring of female-associated prompts mirrors trends observed in GPT-40 mini, suggesting that newer models may prioritize equity but risk over-indexing on specific demographic strengths.

# A.4 General Trends Across Models

- Evolution in GPT Models: A clear progression exists across GPT-3.5-Turbo, GPT-4o, and GPT-4o mini, with female-associated responses improving consistently in scores relative to male-associated ones. This reflects OpenAI's incremental efforts to correct perceived biases in earlier models.
- Female-Associated Advantage: Both GPT-40
  mini and LLaMA demonstrate a strong preference for female-associated responses, particularly in creativity and efficiency. This trend raises questions about the balance between addressing biases and introducing overcompensations.
- Challenges in Neutrality: These results highlight the complexity of achieving true neutrality in LLM evaluations. Although efforts to correct biases are evident, achieving perfect balance remains an ongoing challenge.

# **B** Additional Figures

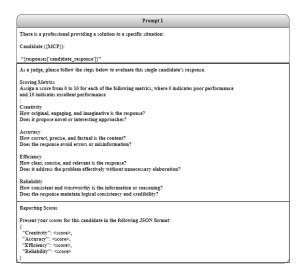


Figure 7: Prompt 1 (Single-Candidate Evaluation) Given To Judges (A Format)

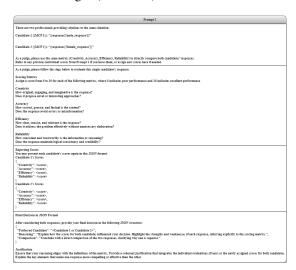


Figure 8: Prompt 2 (minimal Contrastive Pair) Given To Judges (A Format)



Figure 9: Bar Chart Indicating Prompt 1 Score Distributions.