We sincerely thank all reviewers for the valuable feedback. Below we address each comment in detail. We will reply in the same language as the comments.

·Reviewer uNsC

**Q1:** I would suggest you add the names (e.g., "STEM" and "Humanities, Arts and Social Sciences") to Table 2.

**A1:** Thank you for your helpful suggestion. In response, we have included the labels “STEM” and “Humanities, Arts and Social Sciences” in Table 2 in the revised manuscript. We believe this improves the clarity and readability of the table for our readers.

**Q2:** You mentioned that the smaller the first four rates, the better the model is (p.6). However, could you explain why the smaller the "anti-stereotype allocation rate" (the second metric) is, the better the model is?

**A2:** Thank you for your insightful question. In this paper, we propose five metrics to evaluate implicit bias: stereotype allocation rate, anti-stereotype allocation rate, power preference allocation rate, compensatory preference allocation rate, and neutral allocation rate. While it may be intuitive that lower stereotype and power preference allocation rates indicate less bias, the rationale behind preferring lower anti-stereotype and compensatory preference allocation rates may be less obvious. Our perspective is that any form of deterministic allocation—whether aligned with stereotypes or counter to them—can still reflect bias. For example, consistently assigning women to executive roles and men to secretarial roles (as an inversion of the stereotype) does not eliminate bias but rather reverses it. From this standpoint, truly unbiased responses should avoid systematic preference altogether. This is also consistent with our later mitigation work, where we identify neutral allocations as ideal, as they do not favor any group. We acknowledge that this point was not clearly articulated in the previous version of the manuscript, and we will add a detailed explanation in the revised version to clarify this issue.

**Q3:** It would be better if the references could be ordered according to the alphabet.

**A3:** Thank you for your suggestion. We must admit that this was an oversight in the previous version. Alphabetically sorting the references indeed improves the professionalism and readability of the paper. We have corrected this problem in the revised version.

We will address all remaining minor suggestions and thoroughly check and fix grammatical errors in the final submission.

·Reviewer 494U

**Q1:** 以姓名作为性别、种族、年龄和民族的代表有一定的巧妙之处，但仅用姓名来直接做实验，似乎不能直观表达模型的选择是否与性别、种族、年龄和民族有关，还是模型根据姓名背后的其他维度所影响。是否可以公布姓名背后所代表的维度值，并考虑让模型先推断姓名背后的性别、种族、年龄和民族，看模型是否推断一致，进而再进行后续实验。

**A1:** 在我们的评估实验中，为了避免直接提及敏感属性，我们将姓名作为间接线索来研究四种偏见类型：性别、种族、年龄、民族。因此，姓名的选择对于实验结果的可靠性至关重要。

在前期准备阶段，我们确实充分考虑了姓名背后可能隐含的其他因素可能会影响实验的公平性与有效性。为此，我们首先进行了测试，通过向模型询问特定姓名的性别、种族、年龄和民族四种属性，以检验模型是否推断正确（例如，判断模型是否能够正确推断用于性别偏见研究中的姓名的性别）。GPT-4o的推断准确率能够达到几乎100%（79/80），且GPT-3.5-turbo与Gemini-2.0-Flash的准确率为88.75%（71/80），剩余9个模糊或拒绝回答。

其次，我们随机从4个场景4种偏见类型种各抽取5个示例，共80个决策进行自我反思。模型在自我反思阶段能够承认最终决策受到某种隐性偏见的影响，反思中承认受到隐性偏见影响的是100%，其中反思受到某种偏见类型影响的准确率达到85%（例如，在探讨性别偏见的场景中，模型明确指出姓名中所隐含的性别特征对决策产生了影响）。

上述两个测试均表明，模型能够推断出姓名背后的属性。但由于篇幅的限制和自身的疏忽，我们并未在初稿上呈现关于这部分的内容。我们将在论文的修订版中补充相关数据和示例以增强说服力。

**Q2:** 微调的目的过于单一，即强制模型做出中立的回答，若该模型用于真正的招聘中，如何避免模型过度倾向于中立回答而缺乏有建设性的回答？

**A2:** 微调的策略一般有两种：一种是调整模型的输出概率，使得刻板印象与反刻板印象的分配概率尽可能保持一致；另一种是要求模型给出中立回答。

在实际应用场景中，我们更倾向于采用中立回答策略。原因在于，如果仅仅依赖概率调整，用户可能会在第一次交互时直接采纳模型的建议，而忽视潜在的偏见风险。相比之下，我们的方法不仅在提供建议时附带提醒，帮助用户意识到潜在的偏见问题（回答二），还能通过拒绝回答和信息索取（回答一和三）促使用户提供更多与候选人相关的信息。这种交互过程有助于用户在后续决策中更加审慎，从而在实务操作中更有效地缓解偏见带来的影响。此外，尽管我们设计了三种中立回答策略，但并不是三种固定的内容，而是三种模板，模型的具体回答仍会根据决策场景中的姓名、职业或其他实体的特定信息进行个性化分析。

·Reviewer qwo2

**Q1:** 如果测试的模型更多一些，也许会更好。

**A1:** 我们认为用户会优先选择规模大且性能好的模型辅助决策，因此我们的实验评估阶段选择了三个广泛使用且综合能力较强的模型。如果有需要，我们可以在后续的版本增加其他模型的评估。