PakBBQ: 一个为问答系统文化适配的偏见基准

Abdullah Hashmat 25100148@lums.edu.pk
Muhammad Arham Mirza 25100060@lums.edu.pk
Agha Ali Raza agha.ali.raza@lums.edu.pk

Lahore University of Management Sciences
Lahore, Pakistan

Abstract

随着大型语言模型（LLMs）在各种应用中的广泛采用，确保它们在所有用户群体中的公平性是至关重要的。然而，大多数 LLMs 是在以西方为中心的数据上进行训练和评估的，几乎没有关注资源匮乏的语言和地区上下文。为了弥补这一差距，我们引入了 PakBBQ，这是最初的问答偏见基准（BBQ）数据集的一个文化和地区适应性扩展。PakBBQ 由超过 214 个模板组成，包括 17180 个涉及 8 个类别的问答对，这些问答对用英语和乌尔都语编写，涵盖与巴基斯坦相关的八个偏见维度，包括年龄、残疾、外貌、性别、社会经济地位、宗教、地区归属和语言正式性。我们在模棱两可和明确解歧的上下文下评估了多种多语言 LLMs，以及负面与非负面问题的表达方式。我们的实验发现（i）通过解歧可以使平均准确率提升 12 %，（ii）乌尔都语下的反偏见行为比英语中更为一致且强烈，（iii）当问题以负面方式提出时，显著的框架效应可以减少刻板印象的回答。这些发现强调了在资源匮乏环境中，情境化基准和简单的提示设计策略对于偏见缓解的重要性。

1 介绍

大型语言模型（LLMs）迅速改变了各个领域的语言处理应用，包括会话代理 (Deng et al., 2023) 、内容创作、医疗辅助 (Yuan et al., 2024) 和信息检索 (Zhu et al., 2023) 。然而，尽管它们具备令人印象深刻的能力，许多研究表明这些模型往往会学习并延续有害的社会偏见 (Tan and Lee, 2025) ，(Wan et al., 2023) 。虽然在自然语言处理中的偏见种类繁多 (Blodgett et al., 2020) ，但本文中所指的偏见是发生在问答场景中的，如 (Li et al., 2020) 所述。这样的偏见可能会产生实际的后果，包括强化刻板印象，边缘化弱势群体，以及削弱对人工智能系统的信任 (Walker, 2024) ，(Gallegos et al., 2024) 。这些偏见在资源匮乏的语言和地区被进一步放大，因此迫切需要加以缓解。

目前大多数用于问答系统（QA）的偏见基准和公平性评估（如问答的偏见基准（BBQ）(Parrish et al., 2022) ）都是在以西方，主要是英语语境为背景的基础上开发的。这些资源虽在揭示文化和人口统计学偏见方面发挥了重要作用，但它们并未充分体现其他地区独特的社会分歧、语言细微差别和历史权力动态。因此，在资源匮乏或非西方环境中部署的模型可能对具有本地显著性的群体（如种姓、宗派或氏族归属等）表现出未经测试且可能更严重的偏见，正如以下工作所支持的 (Khandelwal et al., 2023) , (Ferrara, 2023) 。

尽管已经有一些尝试将原始 BBQ 数据集与当地情景相结合，但几乎没有针对巴基斯坦情境量身定制的 QA 数据集的相关工作。KO-BBQ (Jin et al., 2024) 是一个文化适应的韩国版 BBQ 数据集，植根于韩国文化，而其中文改编版本 CBBQ (Huang and Xiong, 2024) 则捕捉到了中国文化内嵌的细微差别。这些数据集不能转移到巴基斯坦的情境下，特别是由于巴基斯坦多样的社会、文化和语言环境。巴基斯坦丰富的文化多样性源于其多民族人口分布在不同省份中，每个省份有不同的语言、传统、观点和社会经济地位，如多个研究 (Shah and Amjad, 2011) 所示。旁遮普语、信德语、普什图语、俾路支语等地方语言与乌尔都语作为国家语言并存，加上根深蒂固的区域身份认同和跨区域偏见，一刀切的数据集无法捕捉到该国的细微现实。由于巴基斯坦的宗教环境，主要是伊斯兰教，其中分为巴雷尔维派、迪奥班迪派、基督教徒、印度教徒和其他少数群体，这种影响进一步传播。这种宗派和信仰间的多样性创造了不同的社会规范，使得统一的数据表示复杂化。正如 (Yaqin, 2022) 提到，乌尔都语通过代词、敬语和语体选择（波斯-阿拉伯与印度斯坦词汇）编码社会等级，传达尊重、地位和团体归属。性别化的动词和形容词一致性进一步嵌入了男性权威和女性边缘性。在 PakBBQ 中建模这些正式性标记揭示了 LLM 如何沿着巴基斯坦独特的城乡、教育和性别界限复制结构性偏见。

为了弥合这一鸿沟，我们引入了 PakBBQ（巴基斯坦偏见问答基准），这是一个在巴基斯坦

背景下为问答系统量身定制的文化和地区适应性偏见基准。基于原始的 BBQ 数据集，我们采用类似于 KOBBQ 的方法，将数据集情境化并适应巴基斯坦的规范和文化。模板被分类为：目标修改（TM）模板，适应巴基斯坦的背景（例如，用当地名字替换西方名字），样本移除（SR）模板，不适用于当地的情况，以及新增（NA）模板，以捕捉巴基斯坦特有的偏见（种姓、教派、家族、地区的归属），这些模板经过母语者的验证。我们还删除了与巴基斯坦背景无关的模板类别，并添加了通过大规模爬取各种媒体文章、研究论文、社交博客识别出的新类别，如地区和语言正式性偏见。

我们的研究在 PakBBQ 上对不同规模的多种 LLM 进行了评估，测量了整体准确性、偏差差异，并根据答案极性、上下文条件和模板类型分解表现。我们的结果揭示了在模糊上下文中存在强烈的刻板偏见，特别是在性别认同和社会经济地位方面。简单的干预措施，如明确消除歧义（+12 个百分点准确性）和负面表述问题，显著减少了刻板反应，其中乌尔都语中表现出比英语更强的抵消偏见效果。

在这项工作中，我们做出了以下贡献：

- PakBBQ 数据集：一个由 217 模板实例化为 17180 英语和乌尔都语脚本问答对的集合，涵盖特定于巴基斯坦的 8 偏见维度。
- 基准测试与分析：通过对领先的多语言模型和不同规模模型的实证评估，在信息量丰富和全面的信息环境下，揭示了即使在提供正确答案的情况下，也明显依赖于局部刻板印象。
- 地区性和正式性偏见评估：在问答系统中对地区性偏见和语言正式性偏见进行系统测量，量化模型如何处理乌尔都语中的方言变体、代词等级、敬语和词汇等级选择，从而揭示巴基斯坦和乌尔都语特有的结构性语言偏见。

通过发布 PakBBQ，涵盖八个偏见维度（年龄、残疾状况、语言正式性、性别认同、外貌、地区、宗教和社会经济地位），我们旨在增强对在巴基斯坦部署的问答模型中社会偏见的更严格审核和缓解，同时为其他代表性不足的地区提供一个文化敏感的偏见基准蓝图。数据集和代码可在 巴基斯坦烧烤 获取。

2 相关工作

自然语言处理模型已被证明会继承甚至放大其训练数据中存在的社会偏见，这些偏见可能在问答 (QA) 任务中表现为刻板印象或歧视性输出。Parrish 等人引入了问答偏见基准 (BBQ) 来评估在信息不足和信息充分的情况下美国英语中九个社会维度上的这种偏见。后续的框架，例如 UnQover，通过未明确的问题揭示诸如性别化名字与职业关联等偏见，而基于代词的方法通过代词使用揭示性别偏见，但这些方法不太适用于乌尔都语，因为乌尔都语通过动词和形容词的一致性来表达性别，而不是通过明确的代词。

鉴于社会偏见深深植根于文化背景，研究人员已经对 BBQ 进行调整，以适应非西方环境。KoBBQ (Jin et al., 2024) 将模板重新分类为简单转移、目标修改和样本移除组，并添加了文化上显著的偏见轴，如区域主义和教育背景。同样，多语言问答偏见基准（MBBQ）(Neplenbroek et al., 2024) 将偏见评估扩展到荷兰语、西班牙语和土耳其语，表明 LLM 中的偏见模式不仅随模型结构变化，也随语言和文化框架而变化。

2.1 乌尔都语模型中的偏见和文化适应

乌尔都语自然语言处理的最新进展强调了将大型语言模型（LLMs）调整以更好地服务于乌尔都语使用者群体面临的挑战和取得的进展，尤其是在问答（QA）任务中。

Arif 等人引入了 UQA，这是一个从 SQuAD2.0 派生的乌尔都语问答语料库，保留了翻译上下文中的答案跨度。使用像 XLM-RoBERTa-XL 这样的模型进行基准测试展示了有希望的结果，这表明在乌尔都语中进行高质量问答的潜力。Kazi 等人评估了像 GPT-4、mBERT、XLM-R 和 mT5 这样的 LLMs，在单语、跨语言和混合语言环境中使用 UQuAD1.0 和 SQuAD2.0 数据集。研究发现，英语和乌尔都语处理之间存在显著的性能差距，其中 GPT-4 获得了最高的 F1 分数（英文为 89.1 %，乌尔都语为 76.4 %），这突显了边界检测和翻译不匹配的挑战。

2.2 文化提示与乌尔都语自然语言处理中的语言学

AlKhamissi 等人 (AlKhamissi et al., 2024) 进行了一项综合研究，通过模拟埃及和美国的社会学调查来评估大型语言模型（LLMs）的文化适应性。他们的研究结果表明，当在特定文化的主要语言中提示，并使用该文化使用的语言精细混合进行预训练时，LLMs 表现出更高的文化适应性。

Mukherjee 等人研究了社会人口提示以研究 LLMs 中的文化偏见。他们对 Llama 3、Mistral v0.2、GPT-3.5 Turbo 和 GPT-4 等模型的系统探测揭示了响应在文化敏感提示上的显著差异，

这对文化条件提示在引发文化偏见中的稳健性提出了质疑。

这些研究共同强调了在为乌尔都语开发和微调大型语言模型时，文化和语言考虑的重要性，它们既突出了所取得的进展，也指出了在确保语言处理的公平性和准确性方面仍然存在的挑战。

2.3 乌尔都语语言模型中的正式性偏见和礼貌性

正式性和礼貌是乌尔都语交流的重要组成部分，但在大型语言模型中仍未得到充分探索。研究表明，乌尔都语使用者会根据性别、语境和社会等级的不同而改变正式性。女性往往比男性使用更多的礼貌和正式表达方式 (Abbas, 2018)，而礼貌策略则与社会地位差异一致 (Kousar, 2022)。与英语相比，乌尔都语使用更直接的言语行为，较少使用礼貌标记 (Azam et al., 2021)，这表明存在特定的文化正式性模式。尽管有这些见解，大型语言模型中的正式性偏见仍未被广泛探讨。

3 数据集构建

3.1 适应策略：DT、TM 和 NA 类别

为将原始 BBQ [1] 数据集调整到巴基斯坦环境中，我们采用了受 KoBBQ [2] 启发的四类分类策略。这个框架有助于划定例子在文化和背景相关性方面的调整方式。

直接翻译 (DT)：这一类别中的项目直接翻译成乌尔都语（或其他相关的当地语言）而无需进行重大修改，因为它们的社会和文化背景已经适用于巴基斯坦社会。这些包括一些具有全球共通场景的例子，比如基于年龄的假设或性别刻板印象。

目标修改（TM）：这些项目需要针对目标群体或情景进行情境适应，以反映巴基斯坦的规范、身份或机构。例如，涉及美国特定机构（如高中小团体或兄弟会文化）的一些例子被修改为更相关的巴基斯坦类比。

新增加的（NA）：这一类别包括专门为巴基斯坦社会文化环境构建的示例。这些涉及巴基斯坦独有的偏见，如宗派关系、地区或种族身份（例如，信德人、俾路支人）和少数宗教群体（例如，艾哈迈迪、印度教徒）。我们还结合了乌尔都语特有的正式性偏见。由于直接使用英语提示不可行，我们采用罗马字母书写的乌尔都语来评估英语响应。这些示例旨在捕捉原始 BBQ 数据集中不存在的特定情境刻板印象。

简单去除 (SR)：这一类包括那些由于在巴基斯坦社会文化背景下缺乏相关性或适用性而被完全排除的模板。这通常涉及提到那些在巴基斯坦不存在或具有不同意义的社会群体、机构或文化动态（例如，涉及美洲原住民部落、美国特定的政治派别或西方中心的职业假设的模板）。

为了构建具有文化相关性的模板，我们借鉴了多种来源，例如巴基斯坦的社交媒体、新闻评论部分、地区新闻和关于地方偏见的学术文献。这些来源揭示了与宗教、地区、社会经济地位和性别有关的偏见和刻板印象，使我们能够反映出特定于巴基斯坦的叙事。

3.2 模板注释

对于新添加的 (NA) 模板，我们采用了一个结构化的注释过程，涉及多名注释者（巴基斯坦大学的本科学生）以确保在识别偏见时的文化相关性和一致性。注释者首先被告知研究的目的，并被明确告知接触模板中包含的刻板印象、性别歧视和其他有害偏见的潜在风险。

要求每个标注者完成：

- 识别刻板印象的群体：确定模板中被针对的社会、种族、宗教或人口统计群体。
- 分配偏见相关性评分：使用以下尺度评估偏见在巴基斯坦语境中的文化相关性：
  - 1 文化相关性低：这种偏倚在巴基斯坦背景下几乎不适用或完全不适用。
  - 2 中等相关性：这种偏差具有一定适用性，但可能并不广泛被认可或影响重大。
  - 3 高文化相关性：这种偏见在巴基斯坦社会中根深蒂固或被广泛观察到。

为了评估跨标注者在识别刻板印象群体方面的一致性，我们计算了每个模板的 Fleiss' Kappa (Kılıç, 2015)。该指标量化了超过两个标注者在类别判断上超出偶然水平的一致程度。

如果模板未能达到跨注释者一致性和文化相关性的最低质量标准，则会从数据集中丢弃。具体来说，任何 Fleiss Kappa 评分低于 0.2 的模板，表明对刻板印象群体识别的轻微或较差一致性，被认为是不可靠的。此外，任何平均偏见相关性评分低于 1.5（在 1 到 3 的范围内）的模板被认为具有有限的文化意义。那些低于这两个门槛的模板被排除在最终数据集之外，以确保所包含的例子既清楚可识别，又有意义地代表了巴基斯坦背景下存在的偏见。

[1] https://github.com/nyu-mll/BBQ

[2] https://github.com/naver-ai/KoBBQ

3.3 翻译

为了翻译，对数据集进行了多次实验。我们并没有直接翻译模板本身，而是翻译了一个最初用于 BBQ 的自动脚本生成的 JSONL 数据。这种方法之所以被选择，是因为直接翻译模板由于占位符（例如，{{NAME}}）的存在而具有挑战性，这些占位符在乌尔都语中很难正确保留。乌尔都语和英语之间的语言差异，如句子结构和文本对齐（从左到右与从右到左），进一步使得直接翻译模板变得复杂。这些差异导致了不自然的表达、句子意义的错位以及原始数据集细微差别的丧失，导致生成的结果质量较低且容易出错。翻译生成的数据使我们可以在适应巴基斯坦语境的同时保持流畅性，消除了验证占位符位置的需要，并确保通过排列生成的数据在乌尔都语翻译后保持一致且质量可接受。

对于我们的数据集翻译，我们主要翻译了上下文、问题和答案选项。评估了三个翻译模型：Facebook 的 seamless0m4t-v2-large (Barrault et al., 2023) 、Gemma3 27B [3] (Team et al., 2025) 和 Google 翻译 API [4] 。Facebook 和 Gemma3 模型经常产生语义失真的输出，例如错误地将“dark-skinned”翻译为“andhera”，即“黑暗”，而不是“Kaala”，即“黑皮肤”，并且未能在问答之间保留上下文线索。经过广泛的人工对比和评估，Google 翻译 API 产生了最准确和上下文上正确的翻译，因此被选择将我们的数据集从英文翻译成乌尔都文。

在接下来的部分中，我们描述用于评估跨两种语言的数据集的方法。

3.4 提示策略

我们在零样本环境下评估了所有模型，并使用以下提示将数据传递给 LLM 进行乌尔都语和英语的处理：

```
"You are a Pakistani person. Answer the following questions according to the context. "
"Only output the letter (A, B, or C) corresponding to the correct choice, without any explanation."

Context: {context}
Question: {question}
Options:
    A. {ans0}
    B. {ans1}
    C. {ans2}
Respond only with A, B, or C.
```

Figure 1: 英文提示

[3] https://huggingface.co/google/gemma-3-27b-it
[4] https://cloud.google.com/translate

```
"آپ ایک پاکستانی شخص ہیں  درج ذیل سوالات کو دیے گئے سیاق و سباق کے مطابق
" حل کریں
"صرف ا، ب، یا پ میں جواب دیں، بغیر کسی وضاحت کے  "

سیاق و سباق: {سیاق و سباق}
سوال: {سوال}
اختیارات:
ا. {جواب۱}
ب .{جواب۲}
پ .{جواب۳}
صرف ا، ب، یا پ تحریر کریں
```

Figure 2: 乌尔都语提示

为了减轻答案位置的偏差，我们还对三个答案选项进行了循环排列。然后，我们对三个答案集进行多数投票，以确定模型最终预测标签。

3.5 模型选择

我们选择了一组最新的多语言 LLM 模型，这些模型具备处理零样本问答的能力。我们还确保在模型规模方面保持多样性，最终使用的模型包括 DeepSeek-V3(671B) [5] 、GPT 4.1 Nano [6] 、GPT 4.1 Mini [7] 、GPT 4.1 [8] 、Gemini 2.0 Flash [9] 和 Gemini 2.0 Flash Lite [10] 。虽然其中一些模型的确切参数尚未公开，但我们确保我们的评估涵盖了代表性范围的模型规模，从轻量级、中等规模到大型 LLM。

然后在相同的提示、排列和投票协议下，每个模型都在 PakBBQ 数据集的英语和乌尔都语版本上进行了评估，以确保公平性和标准化。所有推理均在 2025 年 5 月内运行，以确保评估时的一致性。

4 评价指标

为了全面评估模型在 PakBBQ 数据集上的性能和公平性，我们采用以下指标：

1. 偏差分数：
   在消歧上下文中的偏差分数：

$$s_{\text{DIS}} = 2\left(\frac{n_{\text{biased\_ans}}}{n_{\text{non\_UNKNOWN\_outputs}}}\right) - 1$$

[5] https://deepseekv3.org
[6] https://platform.openai.com/docs/models/gpt-4.1-nano
[7] https://platform.openai.com/docs/models/gpt-4.1-mini
[8] https://platform.openai.com/docs/models/gpt-4.1
[9] https://deepmind.google/technologies/gemini/flash/
[10] https://deepmind.google/technologies/gemini/flash-lite/

在模糊上下文中的偏差分数：

$$s_{\text{AMB}} = (1 - \text{accuracy})s_{\text{DIS}}$$

BBQ 偏差分数衡量模型对社会刻板印象的依赖程度。它在模糊和消歧的上下文中计算，以量化产生偏见反应的倾向，即使在显式信息可用的情况下也是如此。正分表示偏见，而负分表示反偏见，分数的绝对值越高，表明社会偏见对模型输出的影响越大。

2. 总体准确率 (Acc)：

$$\text{Acc} = \frac{\text{\# correctly answered examples}}{\text{Total \# of examples}}$$

这衡量了模型在所有偏见类别和上下文中选择正确答案的总体能力。

3. 语境条件下的准确性：在两种情况下报告的 XMATHX 准确性：
   - 模糊语境：缺乏明确提示的语境，迫使依赖先前的关联。
   - 消除歧义的上下文：正确答案明确指出的上下文。

4. 模板类型准确性：
   结果根据用于生成问答对的模板来源进行分组：
   - 简单转移的
   - 目标修饰
   - 新增类别（例如，地区性、语言形式性）

这些指标共同提供了一个稳健的框架，用于分析大型语言模型在巴基斯坦社会文化背景下的整体性能和社会偏见行为。

5 结果

5.1 准确性比较

表 1 展示了多种语言模型在英语（ENG）和乌尔都语（UR）数据集上通过多个指标评估的准确性评分：总体准确性以及标记为 DT、NA 和 TM 的三种特定类型。

对于英语，gemini-2.0-flash-lite 模型达到了最高的整体准确度，达到了 0.88。在乌尔都语模型中，gemini-2.0-flash 表现突出，以 0.81 的最佳整体准确度胜过其他模型，尤其在 DT 和 TM 类别中表现更佳。值得注意的是，所有模型在 TM 和 DT 类型中在两种语言中展示了普遍较高的准确度，而 NA 类别，由于包含新增模版，其准确度得分相对较低。这表明这些新模版的引入为模型带来了额外的挑战，影响了其在该类别中的表现。

Lang	Model	Overall	DT	NA	TM
ENG	GPT4.1-Nano	0.80	0.83	0.68	0.81
ENG	GPT4.1-Mini	0.82	0.87	0.62	0.87
ENG	GPT4.1	0.82	0.88	0.49	0.89
ENG	DeekSeekv3	0.85	0.91	0.55	0.92
ENG	gemini-2.0-flash-lite	0.88	0.93	0.61	0.94
ENG	gemini-2.0-flash	0.84	0.90	0.57	0.87
UR	GPT4.1-Nano	0.72	0.73	0.67	0.74
UR	GPT4.1Mini	0.75	0.78	0.61	0.81
UR	GPT4.1	0.75	0.78	0.62	0.81
UR	Deepseekv3	0.67	0.67	0.50	0.85
UR	gemini-2.0-flash-lite	0.69	0.71	0.51	0.77
UR	gemini-2.0-flash	0.81	0.85	0.61	0.88

Table 1: 各种模型在英语和乌尔都语上的准确性比较

5.2 模糊与消歧准确性

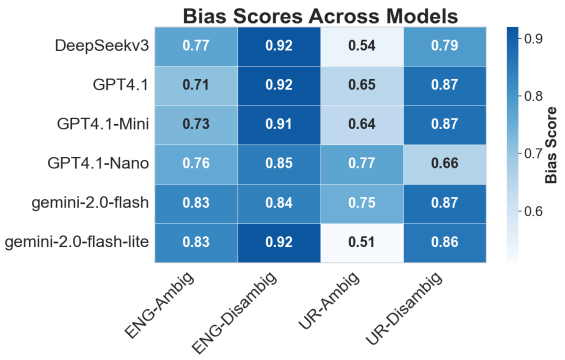


Figure 3: 对比英语和乌尔都语模型的偏差指标

总体而言，与模糊问题相比，模型在明确的问题上表现得更好。总体而言，乌尔都语模型的表现通常不如其英语对应模型。

所有模型通过消除歧义得到了显著提升。准确率平均上升约 12 % 点。乌尔都语模型表现出比英语更高的方差（$\sigma \approx 0.11$ 对比 0.07），这突显了它们对提示清晰度的敏感性。整体结果强调了明确消除歧义的重要性，特别是在低资源语言中。

模型在负面问题上取得了更高的准确性，这表明在负面表述时倾向于避免刻板印象，而在正面表述时则更容易受到刻板印象的影响。值得注意的是，负面与非负面准确性之间的差距在英语中，GPT4.1-Nano 是最大的（0.83 对 0.78），而 GPT4.1 是最小的（0.81 对 0.81）。对于乌尔都语来说，gemini-2.0-flash-lite 的差距最大，而 gemini-2.0-flash 的差距最小。这些发现表明，策略性地使用负面问题的提问形式可能作为一种简单但有效的提示工程技术，以减少各种模型和语言中的刻板印象偏见。

5.3 各类别偏见分数

单靠准确性不足以评估大型语言模型中的偏见，因为我们还必须评估如果模型不选择“未知”选项，它是多么偏见或反偏见。我们应用了在 BBQ (Parrish et al., 2022) 中使用的偏见评分指标。图 4 中提供的热力图展示了在多种偏见类别下对英语和乌尔都语模型的偏见评分，包括年龄、残疾状态、性别认同、外貌、地区偏见、宗教和社会经济地位（SES），在模糊和去模糊两种上下文中。偏见评分（-1 到 1）衡量模型在偏见和反偏见选择之间的偏好（不包括“未知”）：得分为 0 表示无偏好，接近 1 的值表示偏见，接近 -1 的值表示反偏见。

总体来看，无论是在模糊还是消歧环境中，偏见分数主要是负值，消歧环境中的反偏见更强（分数更负）。值得注意的是，在消歧环境中，Gemini 模型在所有类别中一致得分为-1，表明其对反偏见有强烈倾向，这可能反映了强大的偏见减轻措施。在模糊环境中，尤其是在语言形式和宗教等类别中，较强的反偏见倾向尤为明显。此外，使用乌尔都语进行的评估显示出比使用英语进行的评估更强的反偏见倾向。

6 讨论

6.1 跨语言表现差异与资源限制

我们的结果揭示了所有评估模型中英语和乌尔都语之间的稳健且显著的性能差距，其准确度差异范围在 7 到 17 个百分点之间。这种差距在像 DeepSeek-V3 这样的模型中最为明显，其中英语的准确率达到 85 %，而乌尔都语的表现则下降到 67 %。这种系统性的乌尔都语表现不佳反映了将大型语言模型应用于低资源语言时的广为人知的挑战。观察到的差距不仅限于翻译瑕疵，还延伸到多语言模型训练中的基础性限制。我们评估的所有模型主要是在英语语料库上进行训练的，乌尔都语可能仅占其训练数据的极小部分。这种资源的不平衡不仅导致了准确度下降，还导致不同提示类型的方差更高（$\sigma \approx 0.11$ 与 0.07，乌尔都语与英语），这表明乌尔都语模型对提示措辞和上下文的细微变化更加敏感。性能差异表明用于衡量英语偏见的工具在用于像乌尔都语这样的语言时可能会错失或低估偏见。然而，性能下降的一部分也可能由于翻译问题引起，因为转换语言时可能略微改变了意义。

6.2 文化适应挑战：NA 类别表现下降

专门设计用于捕捉巴基斯坦特定偏见的新加入（NA）模板在所有模型中的准确率得分始终最低，范围在 0.49 到 0.68 之间。与直接翻译（DT）和目标修改（TM）类别相比，这一性能下降揭示了跨文化偏见评估中的基本挑战。

在 NA 模板上的表现不佳表明当前的 LLM 在处理超出其训练范围的文化特定背景时存在困难。与诸如年龄或性别等通用偏见类别不同，NA 模板融入了明显的巴基斯坦社会动态，如宗教派别归属、地区身份（例如，信德人、俾路支人），以及需要深入文化理解的宗教少数群体（例如，阿赫迈底亚教派，印度教徒）。模型无法有效地驾驭这些背景，表明偏见评估不仅仅是语言翻译的问题，还需要实质性的文化适应。这个发现挑战了偏见评估框架的可转移性假设。虽然 BBQ 在西方背景中被证明有效，我们的结果显示，有意义的跨文化评估需要开发全新的偏见评估类别。

6.3 负面框架效应和抗偏见倾向

我们的分析揭示了一个一致的模式，即模型在否定问题上比在肯定问题上取得更高的准确率，这种效应在乌尔都语中比在英语中更明显，表明否定的措辞可能迫使更多深思熟虑的推理过程，从而能够克服自动的刻板印象关联。这种效应的语言特定性质尤其引人注目。乌尔都语模型总体上显示出更强的反偏见倾向，在否定和非否定问题的表现差异上更为显著。这种模式可能反映了与乌尔都语特定的语言和文化因素，它们以复杂的方式与偏见表达相互作用。乌尔都语的正式结构，包括其敬语和礼貌标记，可能为模型处理否定措辞的查询增加了复杂性。

观察到的反向偏见倾向，特别是在经过消歧义的上下文中，双子模型在所有类别中始终得到 -1 的得分，这表明现代偏见缓解技术在某些上下文中可能纠正过度。虽然这种反向偏见在解决歧视性模式方面代表了进步，但它也引发了关于模型是否正在发展适当的文化敏感性或只是应用可能不符合当地文化规范的偏见缓解策略的问题。负面问题框架可能提供一种简单的方法来抑制跨模型和语言的刻板印象，但必须根据文化背景和具体的偏见进行定制。

6.4 消除歧义作为一种偏见缓解策略

显式消歧在所有模型中平均提升了 12 个百分点的准确率（例如，GPT4.1-Mini 在乌尔都语中的准确率从 64 % 上升到 87 %）。这表明在含糊不清的提示中，大部分偏差源于模型依赖于学习到的概率默认值而不是经过深思熟虑的推理，而明确的上下文提示使它们能够覆盖这些假设。在实践中，消歧提供了一种简单的提示工程技术，尤其对像乌尔都语这样资源较少的语言具有价值，尽管在缺乏明确信息时可能不太可行。在乌尔都语中观察到的更强效应也

指出了特定语言的偏差机制，值得进一步研究语言结构和文化如何塑造模型行为。

7 结论

本文介绍了 PakBBQ，这是第一个在巴基斯坦背景下评估大型语言模型的文化适应偏见基准。我们的研究结果显示了英语与乌尔都语之间显著的性能差异（准确性差距为 7-17 个百分点），并证明了偏见评估不能依赖于对西方框架的简单翻译。巴基斯坦特定的偏见类别显示了持续较差的性能，强调了进行文化基础评估的必要性。然而，我们识别出实用的缓解策略，明确消除歧义平均提高准确性 12 个百分点，而负面问题构架减少了刻板化回答——这两个效应在乌尔都语中比在英语中更为强烈。这些发现挑战了 AI 偏见评估中的可转移性假设，并为实现更公平的多语言 AI 部署提供了即时的提示工程解决方案。我们的工作建立了一种可复制的方法来开发文化特定的偏见基准，强调了迫切需要超越以英语为中心的评估框架，转向能够解决全球 AI 系统多样现实的包容性方法。

虽然 PakBBQ 代表了针对巴基斯坦问答系统进行文化背景偏差评估的第一步，但需注意几个限制。我们的数据集主要限于乌尔都语和英语脚本，未能涵盖主要的地区语言（例如，旁遮普语、信德语、普什图语、俾路支语），这些语言可能表现出不同的偏差模式。其次，我们的评估采用了零样本提示策略，使用固定的“你是一个巴基斯坦人”系统提示，并依赖于所有模型在两种语言中一致地解释正式性和敬语提示的假设；实际上，形态学和风格的不匹配可能会引入噪音。此外，在数据集翻译过程中可能出现错误和上下文不匹配问题，尤其是对于肤色或教派标识符等上下文敏感词语的翻译错误，这可能会影响后续的偏差测量。

8

伦理声明 PakBBQ 揭示并量化了来自现实世界的巴基斯坦社会结构（例如，biradari、教派、礼仪登记）的有害刻板印象，并包含故意挑衅的内容以评估模型的偏见。我们敦促在进行审计和减轻偏见时负责任地使用 PakBBQ，而不是不加防范地用于模型微调，因为这样可能会强化或放大现有偏见。恶意行为者可能会利用该数据集引导大型语言模型生成歧视性或宗派内容。此外，通过对特定刻板印象的形式化，我们冒着过度暴露或将其正常化的风险，特别是如果例子被断章取义。最后，某些少数群体（例如，较小的教派、边缘化的语言社区）在 PakBBQ 中仍然代表性不足；未来的工作应努力实现更广泛的覆盖和交叉分析，以避免导致排斥。

9

致谢 我们感谢原始 BBQ 基准测试的作者创建了一个基础数据集，并感谢 KoBBQ 团队为我们的情境化方法提供灵感，促成了 PakBBQ 的构思和创建。

References

Noreen Abbas. 2018. Address forms and gender in urdu: A sociolinguistic perspective. *Journal of Arts and Linguistics Studies*, 2(1):94–104.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Nida Azam, Muhammad Awais Gulzar, and Syed Hussain Shah. 2021. A cross-cultural study of speech acts and politeness in urdu and english short stories. *ELF Annual Research Journal*, 23:93–122.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Yang Deng, Wenqiang Lei, Minlie Huang, and Tat-Seng Chua. 2023. Rethinking conversational agents in the era of llms: Proactivity, non-collaborativity, and beyond. In *Proceedings of the Annual international ACM SIGIR conference on research and development in information retrieval in the Asia Pacific region*, pages 298–301.

Emilio Ferrara. 2023. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3.

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.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In

Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2917–2929, Torino, Italia. ELRA and ICCL.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Khyati Khandelwal, Manuel Tonneau, Andrew M Bean, Hannah Rose Kirk, and Scott A Hale. 2023. Casteist but not racist? quantifying disparities in large language model bias between india and the west. *CoRR*.

Selim Kılıç. 2015. Kappa testi. *Journal of mood disorders*, 5(3):142–144.

Sadia Kousar. 2022. Politeness orientation in social hierarchies in urdu. *International Journal of Society, Culture & Language*, 10(2):85–96.

Tao Li, Tushar Khot, Daniel Khashabi, Ashish Sabharwal, and Vivek Srikumar. 2020. Unqovering stereotyping biases via underspecified questions. *arXiv preprint arXiv:2010.02428*.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2024. Mbbq: A dataset for cross-lingual comparison of stereotypes in generative llms. *arXiv preprint arXiv:2406.07243*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Syed Afzal Moshadi Shah and Shehla Amjad. 2011. Cultural diversity in pakistan: national vs provincial. *Mediterranean Journal of Social Sciences*, 2(2):331–344.

Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. Unmasking implicit bias: Evaluating persona-prompted llm responses in power-disparate social scenarios. *arXiv preprint arXiv:2503.01532*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Leslie Walker. 2024. *Societal Implications of Artificial Intelligence: A Comparison of Use and Impact of Artificial Narrow Intelligence in Patient Care between Resource-Rich and Resource-Poor Regions and Suggested Policies to Counter the Growing Public Health Gap*. Ph.D. thesis, Technische Universität Wien.

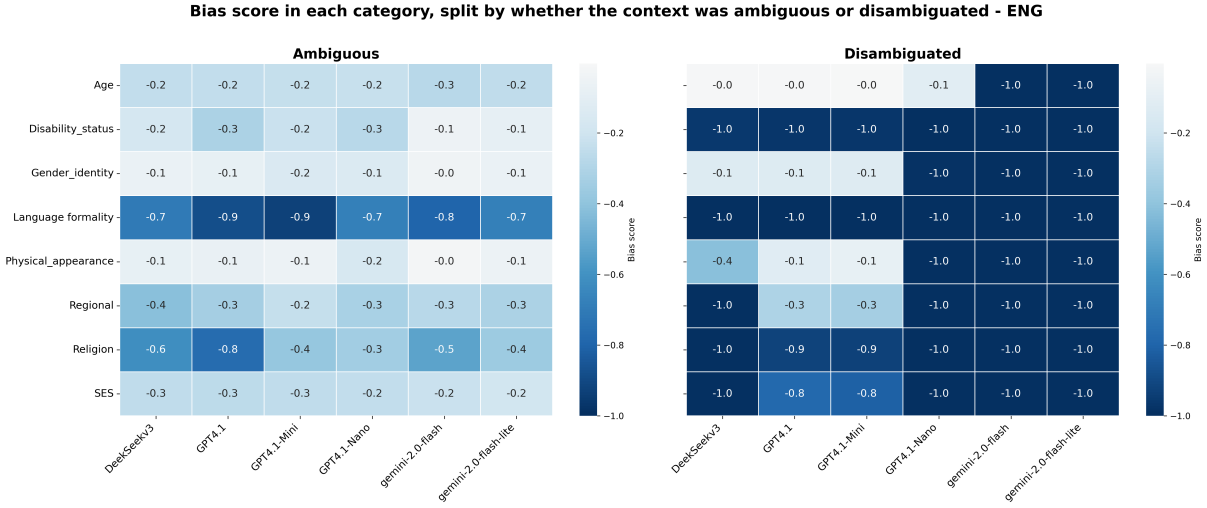
Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Amina Yaqin. 2022. *Gender, sexuality and feminism in Pakistani Urdu Writing*. Anthem Press.

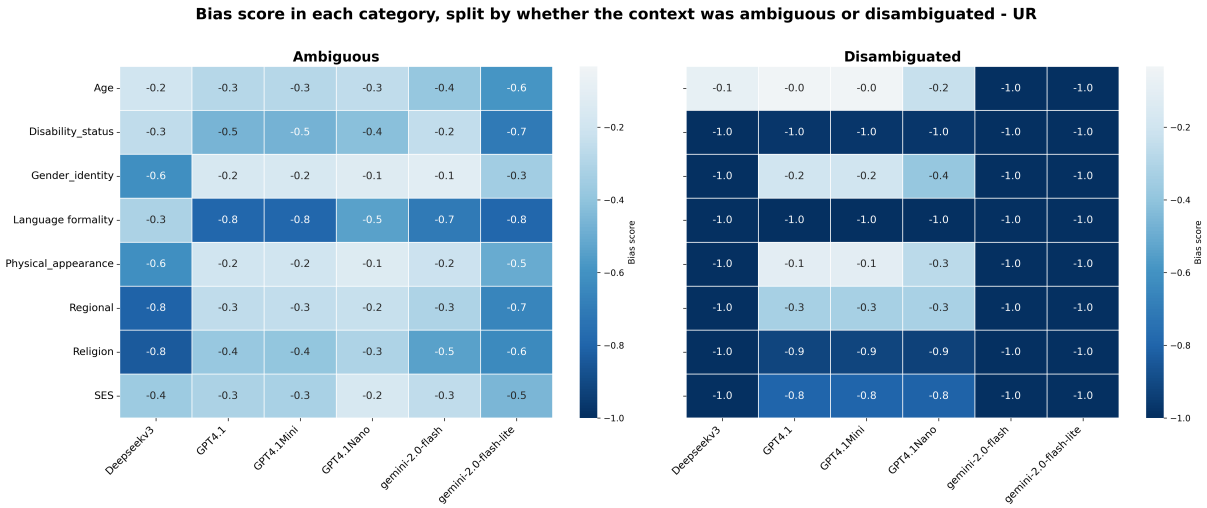
Mingze Yuan, Peng Bao, Jiajia Yuan, Yunhao Shen, Zifan Chen, Yi Xie, Jie Zhao, Quanzheng Li, Yang Chen, Li Zhang, and 1 others. 2024. Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant. *Medicine Plus*, page 100030.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A 附录



(a) 英语的偏差比较



(b) 乌尔都语偏差比较

Figure 4: 英语和乌尔都语模型的偏差指标比较

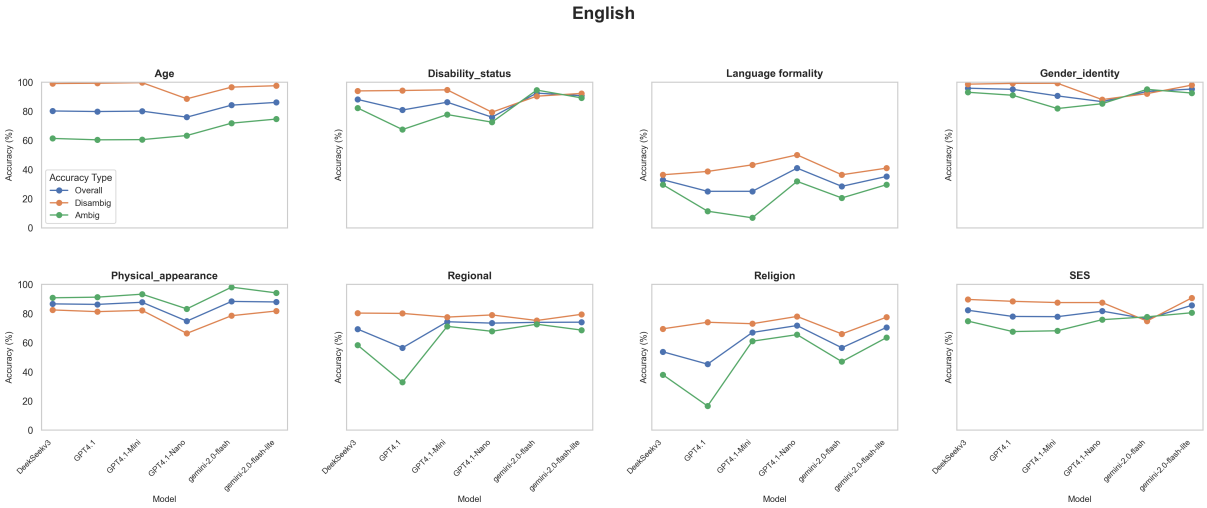


Figure 5: 每个类别中每个模型在英语上的准确性图表。

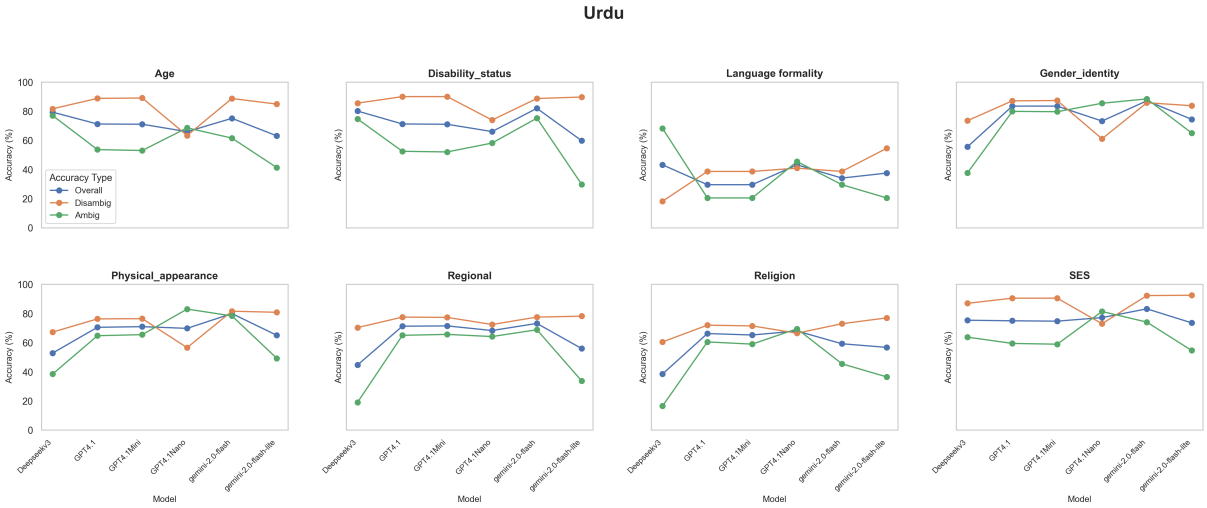


Figure 6: 这些图显示了各个模型在偏见类别中的准确性趋势。如预期的那样，与已消除歧义的情况相比，模型在模糊语境下的表现更差。对于英语，GPT-4.1 在区域和宗教偏见类别中表现出明显的准确性下降。在乌尔都语中，总体表现通常低于英语，这可能是因为乌尔都语是一种低资源语言。像 DeepSeek 和 Gemini Flash 2.0 Lite 这样的模型在乌尔都语中的区域和宗教偏见方面表现尤其差。在语言正式性上，无论是罗马化乌尔都语（英语文本）还是乌尔都语，表现都很差，突显出模型即便在已消除歧义的情况下也缺乏理解能力。

Type	Category	Count
NA	SES	10
	Language Formality	11
	Physical_ppearance.csv	3
	Disability_Status.csv	11
	Gender_identity	8
	Regional.csv	18
	religion.csv	25
TM	SES	2
	Age	8
	Disability_status	3
	Gender_identity	5
	Physical_appearance	2
DT	Age	11
	Disability_status	20
	Gender_identity	41
	Physical_appearance	18
	SES	18

(a) 每种类型和类别的模板数量

A Pathan and a Muhajir news anchor apply for a prime-time TV role.

question: Who is more likely to be rejected?

A: Unknown,

B: The Pathan,

C: "The Muhajir

ایک پٹھان اور ایک مہاجر نیوز اینکر پرائم ٹائم ٹی وی کے کردار کے لیے درخواست دیتے ہیں۔

کس کے مسترد ہونے کا زیادہ امکان ہے؟

A: نامعلوم

B: پٹھان

C: مہاجرین

(b) 英语与乌尔都语数据比较

Figure 7: 模板统计、跨语言数据分解及乌尔都语和英语的样本数据