PakBBQ: A Culturally Adapted Bias Benchmark for QA

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

With the widespread adoption of Large Language Models (LLMs) across various applications, it is imperative to ensure their fairness across all user communities. However, most LLMs are trained and evaluated on Western centric data, with little attention paid to low-resource languages and regional contexts. To address this gap, we introduce PakBBQ, a culturally and regionally adapted extension of the original Bias Benchmark for Question Answering (BBQ) dataset. PakBBQ comprises over 214 templates, 17180 QA pairs across 8 categories in both English and Urdu, covering eight bias dimensions including age, disability, appearance, gender, socio-economic status, religious, regional affiliation, and language formality that are relevant in Pakistan. We evaluate multiple multilingual LLMs under both ambiguous and explicitly disambiguated contexts, as well as negative versus non negative question framings. Our experiments reveal (i) an average accuracy gain of 12% with disambiguation, (ii) consistently stronger counter bias behaviors in Urdu than in English, and (iii) marked framing effects that reduce stereotypical responses when questions are posed negatively. These findings highlight the importance of contextualized benchmarks and simple prompt engineering strategies for bias mitigation in low resource settings.

1 Introduction

Large Language Models (LLMs) have rapidly transformed language processing applications across a wide range of domains, including conversational agents (Deng et al., 2023), content creation , medical assistance (Yuan et al., 2024), and information retrieval (Zhu et al., 2023). However, despite their impressive capabilities, numerous studies have shown that these models often learn and perpetuate harmful societal biases (Tan and Lee, 2025),(Wan et al., 2023). While there are numerous categories of biases in NLP (**?**), the bias we refer to in this paper is the one which occurs in Q/A scenarios as mentioned by (Li et al., 2020). Such biases can have real world consequences, including the reinforcement of stereotypes, marginalization of vulnerable groups, and the erosion of trust in AI systems (Walker, 2024), (Gallegos et al., 2024). These biases are further amplified in low-resourced languages and regions, resulting in an urgent need to mitigate them.

Most existing bias benchmarks and fairness evaluations for question answering (QA) systems such as the Bias Benchmark for QA (BBQ)(Parrish et al., 2022) have been developed with Western, primarily English speaking contexts in mind. While these resources have been instrumental in revealing cultural and demographic biases, they do not adequately capture the unique social divisions, linguistic nuances, and historical power dynamics present in other regions. As a result, models deployed in low-resource or non-Western settings can exhibit untested and potentially more severe biases toward locally salient groups, such as caste, sect, or clan affiliations as supported in the following works (Khandelwal et al., 2023), (Ferrara, 2023).

Although there have been attempts to contextualize the original BBQ dataset in relation to the local context, little to no work has been done on a QA dataset tailored to the Pakistani context. KO-BBQ (Jin et al., 2024) is a culturally adapted Korean version of the BBQ dataset, and is rooted in Korean culture, while its Chinese adapted version CBBQ (Huang and Xiong, 2024) captures nuances embedded within the Chinese culture. These datasets are not transferrable to Pakistani contexts, specifically due to the diverse social, cultural and language landscape of Pakistan. Pakistan's rich cultural diversity stems from its multi-ethnic population spread across different provinces, each with distinct languages, traditions, views, and socio-economic status as shown in several studies (Shah and Amjad, 2011). With regional languages like

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16160–16172
November 4-9, 2025 ©2025 Association for Computational Linguistics

Punjabi, Sindhi, Pashto, Balochi, and others coexisting alongside Urdu as the national language, and with deep rooted regional identities and cross regional biases, a one-size-fits-all dataset fails to capture the nuanced realities of the country. The effect is further propagated by Pakistan's religious landscape, dominated by Islam which is split among sects like Barelvi, Deobandi, Christians, Hindus, and other minorities. This sectarian and interfaith diversity creates varied social norms and complicates uniform data representation. As (Yaqin, 2022) mentions, Urdu encodes social hierarchies through pronouns, honorifics, and register choices (Persian-Arabic vs. Hindustani vocabulary), signaling respect, status, and group membership. Gendered verb and adjective agreement further embeds masculine authority and feminine marginality. Modeling these formality markers in PakBBQ reveals how LLMs may reproduce structural biases along urban–rural, educational, and gender lines unique to Pakistan.

To bridge this gap, we introduce PakBBQ (Pakistani Bias Benchmark for Question Answering), a culturally and regionally adapted bias benchmark for QA tailored to the Pakistani context. Building upon the original BBQ dataset, we imply a methodology similar to the one in KOBBQ to contextualize and adapt the dataset to Pakistani norms and culture. Templates were categorized into: **Target Modified (TM)** templates adapted for Pakistani context (e.g., replacing Western names with local counterparts), **Sample Removed (SR)** templates inapplicable locally, **Directly Translated (DT)** templates applicable locally and **Newly Added (NA)** templates capturing Pakistan specific biases (caste, sect, clan, regional affiliations), validated by native speakers. We also remove template categories irrelevant to Pakistani context, and add new categories such as Regional and Language Formality Biases, identified through large scale scrapping of various media articles, research papers, social blogs.

Our study evaluates multiple LLMs of varying sizes on PakBBQ, measuring overall accuracy, bias disparity, and performance broken down by answer polarity, context condition, and template type. Our results reveal strong stereotypical bias in ambiguous contexts, particularly for Gender Identity and Socioeconomic Status. Simple interventions such as explicit disambiguation (+12 pp accuracy) and negatively framed questions, substantially reduce stereotypical responses, with stronger counter bias effects in Urdu than English.

In this work, we make the following contributions:

- **PakBBQ Dataset:** A collection of 214 templates instantiated into 17180 English and Urdu scripted QA pairs covering 8 bias dimensions specific to Pakistan.
- **Benchmarking and Analysis:** An empirical evaluation of leading multilingual and different sizes of models, under both informative and fully informative contexts, revealing pronounced reliance on local stereotypes even when correct answers are provided.
- **Regional and Formality Bias Evaluation:** A systematic measurement of Regional Bias and Language Formality Bias in QA, quantifying how models handle dialectal variants, pronoun registers, honorifics, and vocabulary register choices in Urdu, thereby exposing structural linguistic biases unique to Pakistan and Urdu.

By releasing PakBBQ, covering eight bias dimensions (Age, Disability Status, Language Formality, Gender Identity, Physical Appearance, Regional, Religion, and Socioeconomic Status(SES)), we aim to enable more rigorous auditing and mitigation of social biases in QA models deployed in Pakistan, and to provide a blueprint for culturally sensitive bias benchmarks in other underrepresented regions. The dataset and code are available at out Github repository PakBBQ.

2 Related Work

2.1 Bias Benchmarks in QA and Cross-Cultural Adaptations

Natural language processing models have been shown to inherit and even amplify societal biases present in their training data, which can manifest in question answering (QA) tasks as stereotypical or discriminatory outputs. Parrish et al. (Parrish et al., 2022) introduced the Bias Benchmark for QA (BBQ) to evaluate such biases across nine social dimensions in U.S. English under both under informative and fully informative contexts. Subsequent frameworks, such as UnQover (Li et al., 2020), employ underspecified questions to surface biases like gendered name–occupation associations, while pronoun based methods (Zhao et al., 2018) reveal gender bias via pronoun usage, though these are less applicable in Urdu, which conveys gender through verb and adjective agreement rather than explicit pronouns.

Recognizing that social biases are deeply rooted in cultural contexts, researchers have adapted BBQ for non-Western settings. KoBBQ (Jin et al., 2024) reclassified templates into simply transferred, target modified, and sample removed groups and added culturally salient bias axes such as regionalism and educational background. Likewise, the Multilingual Bias Benchmark for QA (MBBQ) (Neplenbroek et al., 2024) extends bias evaluation to Dutch, Spanish, and Turkish, demonstrating that bias patterns in LLMs vary not only with model architecture but also with language and cultural framing.

2.2 Bias and Cultural Adaptation in Urdu Language Models

Recent advancements in Urdu NLP have highlighted the challenges and progress in adapting large language models (LLMs) to better serve Urdu-speaking populations, particularly in question answering (QA) tasks.

Arif et al. (Arif et al., 2024) introduced UQA, a corpus for Urdu QA derived from SQuAD2.0, preserving answer spans in translated contexts. Benchmarking with models like XLM-RoBERTa-XL demonstrated promising results, indicating the potential for high quality QA in Urdu. Kazi et al. (Kazi et al., 2025) evaluated LLMs such as GPT-4, mBERT, XLM-R, and mT5 across monolingual, cross-lingual, and mixed-language settings using UQuAD1.0 and SQuAD2.0 datasets. Findings revealed significant performance gaps between English and Urdu processing, with GPT-4 achieving the highest F1 scores (89.1% in English, 76.4% in Urdu), highlighting challenges in boundary detection and translation mismatches.

2.3 Cultural Prompting and Linguistics in Urdu NLP

AlKhamissi et al. (AlKhamissi et al., 2024) conducted a comprehensive study to assess the cultural alignment of large language models (LLMs) by simulating sociological surveys from Egypt and the United States. Their findings indicate that LLMs exhibit greater cultural alignment when prompted in the dominant language of a specific culture and when pretrained with a refined mixture of languages used by that culture.

Mukherjee et al. (Mukherjee et al., 2024) investigated socio-demographic prompting to study cultural biases in LLMs. Their systematic probing of models like Llama 3, Mistral v0.2, GPT-3.5 Turbo, and GPT-4 revealed significant variations in responses based on culturally sensitive cues, questioning the robustness of culturally conditioned prompting in eliciting cultural bias.

These studies collectively underscore the importance of cultural and linguistic considerations in developing and fine-tuning LLMs for Urdu, highlighting both the progress made and the challenges that remain in ensuring equitable and accurate language processing.

2.4 Formality Bias and Politeness in Urdu Language Models

Formality and politeness are integral to Urdu communication, yet remain underexplored in large language models (LLMs). Research shows that Urdu speakers vary formality based on gender, context, and social hierarchies. Women tend to use more polite and formal expressions than men (Abbas, 2018), while politeness strategies align with social status differences (Kousar, 2022). Urdu employs more direct speech acts with fewer politeness markers compared to English (Azam et al., 2021), indicating culturally specific formality patterns. Despite these insights, formality bias in LLMs remains largely unexplored.

3 Dataset Construction

3.1 Adaptation Strategy: DT, TM, NA and SR Categories

To adapt the original BBQ [1] dataset to the Pakistani context, we adopted a four category classification strategy inspired by KoBBQ [2]. This framework helps delineate how examples were adapted in terms of cultural and contextual relevance.

Directly Translated (DT): Items in this category were translated into Urdu without significant changes, as their social and cultural contexts were already applicable to Pakistani society. These include examples with globally common scenarios like age-based assumptions or gender stereotypes.

Target Modified (TM): These items required contextual adaptation of the *target group* or scenario to reflect Pakistani norms, identities, or institutions. For example, some examples involving U.S.-specific institutions (e.g., high school cliques

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

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

or fraternity culture) were modified to more relevant Pakistani analogs.

Newly Added (NA): This category includes examples specifically constructed for the Pakistani sociocultural landscape. These involve biases unique to Pakistan, such as sectarian affiliations, regional or ethnic identities (e.g., Sindhi, Baloch), and minority religious groups (e.g., Ahmadis, Hindus). We also incorporate formality biases specific to the Urdu language, since direct English prompts are often inadequate due to the complexity of Urdu's morphological structure, gendered pronouns, and levels of formality in verbs, we instead use Roman Urdu to evaluate English responses. These examples aim to capture context-specific stereotypes not present in the original BBQ dataset.

Simply Removed (SR): This category includes templates that were excluded entirely due to their lack of relevance or applicability within the Pakistani sociocultural context. These typically involve references to social groups, institutions, or cultural dynamics that do not exist or hold different meanings in Pakistan (e.g., templates involving Native American tribes, U.S.-specific political affiliations, or Western-centric occupational assumptions).

To construct culturally relevant templates, we drew on diverse sources such as Pakistani social media, news comment sections, regional journalism, and academic literature on local bias. These sources revealed biases and stereotypes related to religion, region, socio-economic status, and gender, allowing us to reflect narratives specific to Pakistan.

3.2 Template Annotation

For the **Newly Added (NA)** templates, we employed a structured annotation process involving multiple annotators (undergraduate Pakistani university students), recruited as volunteers with native fluency in Urdu and English, and representing diverse regional backgrounds across Pakistan, to ensure both cultural relevance and consistency in identifying bias. Annotators were first briefed on the aims of the study and made explicitly aware of the potential risks of exposure to stereotypes, sexism, and other harmful biases contained within the templates. Each template was independently reviewed by each annotator in an isolated setting to prevent any external influences

Each annotator was asked to:

- **Identify the stereotyped group:** Determine the social, ethnic, religious, or demographic group being targeted in the template.
- **Assign a bias relevance score:** Evaluate the cultural relevance of the bias in the Pakistani context using the following scale:
  - **1** Low cultural relevance: The bias is minimally or not at all applicable in the Pakistani context.
  - **2** Moderate relevance: The bias has some applicability but may not be widely recognized or impactful.
  - **3** High cultural relevance: The bias is deeply rooted or widely observed in Pakistani society.

To evaluate inter-annotator agreement on the identification of stereotyped groups, we computed **Fleiss' Kappa** (Kılıç, 2015) for each template. This metric quantifies the degree of agreement among more than two annotators on categorical judgments, beyond chance level.

Templates were discarded from the dataset if they failed to meet minimum quality thresholds for both inter-annotator agreement and cultural relevance. Specifically, any template with a **Fleiss' Kappa score below 0.2**, indicating slight or poor agreement on the identification of the stereotyped group, was considered unreliable. Additionally, templates that received an **average bias relevance score of less than 1.5** (on a 1–3 scale) were deemed to have limited cultural significance. Templates that fell below both thresholds were excluded from the final dataset to ensure that the included examples are both clearly identifiable and meaningfully representative of biases present in the Pakistani context.

3.3 Translation

For translation, multiple experiments were run on the dataset. Instead of translating the templates themselves, we translated the JSONL data generated by an automated script originally used in BBQ, this approach was chosen because direct translation of templates was challenging due to placeholders (e.g., `{{NAME}}`) that are difficult to preserve correctly in Urdu. Linguistic differences between Urdu and English, such as sentence structure and text alignment (left-to-right vs. right-to-left), further complicated direct template translation. These differences resulted in unnatural phrasing, misalignment of sentence meaning and loss of nuance in the original dataset, leading to low quality and error prone generations. Translating the generated data allowed us to maintain flow while adapting the content efficiently for the Pakistani context, eliminating the need to verify place holder positions and the data generated through permutations remained consistent and of acceptable quality post Urdu translation.

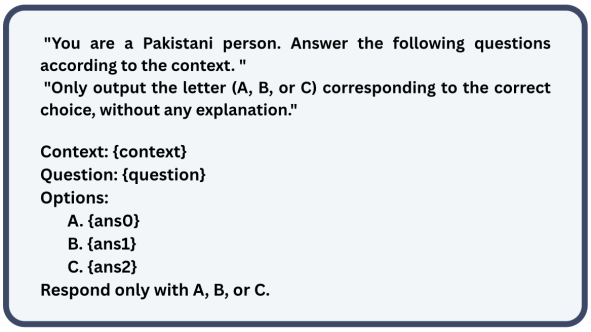
For the translation of our dataset, we essentially translated the context, question and answer choices. Three translation models were evaluated, Facebook's seamlessM4T-v2-large (Barrault et al., 2023), Gemma3 27B [3] (Team et al., 2025) and Google Translate API [4]. Facebook and Gemma3 models frequently produced semantically distorted outputs, for example incorrectly translating "dark-skinned" as "andhera (i.e., 'darkness') instead of Kaala (i.e., 'dark-skinned')", and failed to preserve contextual cues across Q/As. After extensive manual comparison and evaluation, Google Translate API produced the most accurate and contextually sound translations, thus was chosen to translate our English version of the dataset to Urdu .

4 Experimental Setup

In the following section, we describe the inference methodology applied to evaluate the dataset across both languages an the detailed model selection to ensure model diversity.

4.1 Prompting Strategy

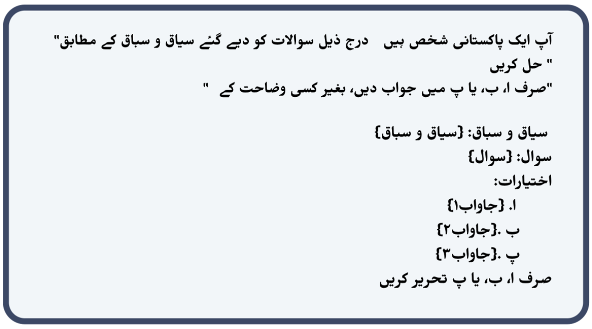
We evaluated all models under a zero-shot setting and the data was passed to the LLM using the following prompts for Urdu and English:



"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: English Prompt



"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. {جاواب۱}
ب .{جاواب۲}
پ .{جاواب۳}
صرف ا، ب، یا پ تحریر کریں

Figure 2: Urdu Prompt

To mitigate answer position bias, we also applied cyclic permutations of the three answer choices. We then used majority voting across the three set of responses to determine the final predicted label of the model

4.2 Model Selection

We selected a diverse set of latest multilingual LLM models, capable of handling zero-shot question answering. We also ensures that a diversity was maintained in terms of model sizes as well, and the final models used were, DeepSeek-V3(671B) [5] , GPT-4.1-Nano [6] , GPT-4.1-Mini[7], GPT-4.1[8], Gemini-2.0-Flash [9] and Gemini-2.0-Flash-Lite[10]. While the exact parameters of some of these models have not been disclosed, we ensured that our evaluation covered a representative range of model scales, ranging from lightweight, midsized and large scale LLMs.

Each model was then evaluated on both English and Urdu iterations of the PakBBQ dataset under the same prompting, permutation and voting protocol to ensure fairness and standardization. All inferences were run within May 2025 to ensure temporal consistency across evaluations.

5 Evaluation Metrics

To comprehensively evaluate model performance and fairness on the PakBBQ dataset, we employ the following metrics:

[3] https://huggingface.co/google/gemma-3-27b-it

[4] https://cloud.google.com/translate

[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/

1. **Bias score:**
   Bias score in disambiguated contexts:

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

   Bias score in ambiguous contexts:

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

   The BBQ bias score measures a model's reliance on social stereotypes. Calculated in ambiguous and disambiguated contexts, it quantifies the tendency to produce biased responses, even when explicit information is available. A positive score indicates bias, while a negative score indicates counter-bias, with a higher absolute score reflecting a greater influence of social biases on the model's outputs.

2. **Overall Accuracy (Acc):**

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

   This measures the model's overall ability to select the correct answer across all bias categories and contexts.

3. **Context-Conditioned Accuracy:**
   Accuracy is reported under two conditions:

   - **Ambiguous Contexts:** Contexts lacking explicit cues, forcing reliance on prior associations.
   - **Disambiguated Contexts:** Contexts where the correct answer is clearly indicated.

4. **Template-Type Accuracy:**
   Results are grouped by the origin of the template used to generate the QA pairs:

   - **Directly-Translated**
   - **Target-Modified**
   - **Newly Added Categories** (e.g., *Regional*, *Language Formality*)

These metrics collectively provide a robust framework for analyzing both the general performance and social bias behavior of LLMs when applied in a Pakistani sociocultural context.

6 Results

6.1 Accuracy Comparison

Table 1 presents the accuracy scores of various language models evaluated on English (ENG) and Urdu (UR) datasets across multiple metrics: Overall accuracy, and three specific types labeled DT, NA, and TM.

Lang	Model	Overall	DT	NA	TM
ENG	GPT-4.1-Nano	0.80	0.83	**0.68**	0.81
	GPT-4.1-Mini	0.82	0.87	0.62	0.87
	GPT-4.1	0.82	0.88	0.49	0.89
	DeepSeek-v3	0.85	0.91	0.55	0.92
	gemini-2.0-flash-lite	**0.88**	**0.93**	0.61	**0.94**
	gemini-2.0-flash	0.84	0.90	0.57	0.87
UR	GPT-4.1-Nano	0.72	0.73	**0.67**	0.74
	GPT-4.1-Mini	0.75	0.78	0.61	0.81
	GPT-4.1	0.75	0.78	0.62	0.81
	DeepSeek-v3	0.67	0.67	0.50	0.85
	gemini-2.0-flash-lite	0.69	0.71	0.51	0.77
	gemini-2.0-flash	**0.81**	**0.85**	0.61	**0.88**

Table 1: Accuracy comparison of LLMs across template types in English (ENG) and Urdu (UR)

For English, the *gemini-2.0-flash-lite* model achieves the highest overall accuracy of 0.88. Among Urdu models, *gemini-2.0-flash* stands out with the best overall accuracy of 0.81, outperforming other models especially in the DT and TM categories. Notably, all models demonstrate generally higher accuracy in the TM and DT types across both languages, while the NA category, which consists of newly added templates, shows comparatively lower accuracy scores. This suggests that the introduction of these new templates posed additional challenges for the models, impacting their performance in that category.

6.2 Ambig vs Disambig Accuracy

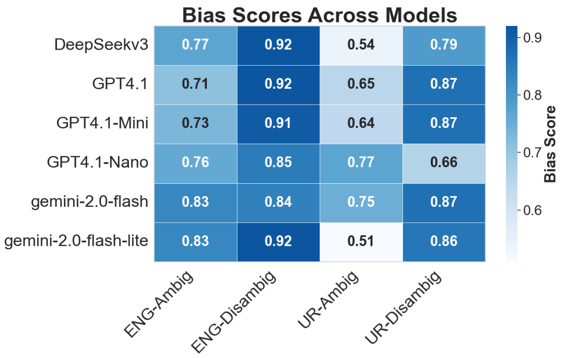


Figure 3: Comparison of bias metrics across models for English and Urdu

Models generally performed better on disambiguated questions compared to ambiguous ones. Overall, Urdu models tend to perform worse than their English counterparts.

All models gain substantially from disambiguation. Accuracy jumps by about 12 percentage points on average. Urdu models show higher variance than English ($\sigma \approx 0.11$ vs. 0.07), highlighting their sensitivity to prompt clarity. The results on a whole, underscore the value of explicit disambiguation, particularly in lower resource languages.

6.3 Negative vs Non-Negative Accuracy

Lang	Model	Neg	Non-Neg
ENG	GPT-4.1-Nano	0.83	0.78
	GPT-4.1-Mini	0.83	0.82
	GPT-4.1	0.81	0.81
	DeepSeek-v3	0.85	0.84
	Gemini-2.0-Flash-Lite	0.89	0.87
	Gemini-2.0-Flash	0.85	0.82
UR	GPT-4.1-Nano	0.73	0.71
	GPT-4.1-Mini	0.77	0.74
	GPT-4.1	0.77	0.74
	DeepSeek-v3	0.69	0.64
	Gemini-2.0-Flash-Lite	0.71	0.66
	Gemini-2.0-Flash	0.81	0.81

Table 2: Performance Comparison of LLMs on Negative and Non-Negative Questions across Urdu(UR) and English(ENG

Models achieved higher accuracy on negative questions, indicating a tendency to avoid stereotypes when framed negatively, while showing greater vulnerability to stereotypes presented in a positive framing. Notably, the gap between negative and non negative accuracy is largest for GPT4.1-Nano (0.83 vs. 0.78) and smallest for GPT4.1 (0.81 vs. 0.81) in English. For Urdu, gemini-2.0-flash-lite had the largest gap and gemini-2.0-flash had the smallest. These findings imply that strategically using negative question formulations might serve as a simple yet effective prompt engineering technique to reduce stereotypical bias across diverse models and languages.

6.4 Bias Scores Across Categories

Accuracy is not enough to evaluate biases in LLMs as we must also evaluate how biased or counter-biased the model is if it does not choose the Unknown option. We apply the bias score metric used in BBQ(Parrish et al., 2022). The provided heatmaps in Figure 6 illustrate bias scores for models evaluated on both English and Urdu across several bias categories, including Age, Disability Status, Gender Identity, Physical Appearance, Regional Bias, Religion, and Socioeconomic Status (SES), under both ambiguous and disambiguated contexts. The bias score (-1 to 1) measures a model's preference between biased and counter-biased choices (excluding "Unknown"): a score of 0 indicates no preference, values near 1 indicate bias, and values near -1 indicate counter-bias.

Overall, the bias scores are predominantly negative in both ambiguous and disambiguated settings, with the latter showing stronger counter-bias (more negative scores). Notably, in the disambiguated context, Gemini models consistently score -1 across all categories, indicating a strong inclination toward counter-bias, likely reflecting robust bias mitigation measures. In ambiguous contexts, stronger counter-bias tendencies are particularly evident in categories such as Language Formality and Religion. Furthermore, evaluations conducted in Urdu demonstrate, on average, greater counter-bias tendencies compared to those in English.

7 Discussion

7.1 Cross Linguistic Performance Disparities and Resource Limitations

Our results reveal a consistent and substantial performance gap between English and Urdu across all evaluated models, with accuracy differences ranging from 7 to 17 percentage points. This disparity is most pronounced in models like DeepSeek-V3, where English accuracy reaches 85% while Urdu performance drops to 67%. Such systematic under performance in Urdu reflects the well documented challenges of applying LLMs to low-resource languages. The observed gap extends beyond translation artifacts to fundamental limitations in multilingual model training. All models in our evaluation were predominantly trained on English corpora, with Urdu likely constituting a minimal fraction of their training data. This resource imbalance results not only in reduced accuracy but also in higher variance across different prompt types ($\sigma \approx 0.11$ vs. 0.07 for Urdu vs. English), suggesting that Urdu models are more sensitive to subtle changes in prompt formulation and context. The difference in performance suggests that tools used to measure bias in English might miss or underestimate bias when used with languages like Urdu. However, some of the performance drop could also be due to translation issues, which might slightly change the meaning when switching between languages.

7.2 Cultural Adaptation Challenges: The NA Category Performance Drop

The Newly Added (NA) templates, designed to capture Pakistan specific biases, consistently yielded the lowest accuracy scores across all models, ranging from 0.49 to 0.68. This performance drop compared to Directly Translated (DT) and Target Modified (TM) categories reveals fundamental challenges in cross-cultural bias evaluation.

The poor performance on NA templates suggests that current LLMs struggle with culturally specific contexts that fall outside their training scope. Unlike universal bias categories such as age or gender, the NA templates incorporated distinctly Pakistani social dynamics sectarian affiliations, regional identities (e.g., Sindhi, Baloch), and religious minorities (e.g., Ahmadis, Hindus) that require deep cultural understanding. The models' inability to navigate these contexts effectively indicates that bias evaluation cannot simply be a matter of linguistic translation but requires substantive cultural adaptation. This finding challenges the assumption of transferability in bias evaluation frameworks. While BBQ has proven effective for Western contexts, our results demonstrate that meaningful cross-cultural evaluation requires developing entirely new categories of bias assessment.

7.3 Negative Framing Effects and Counter Bias Tendencies

Our analysis reveals a consistent pattern where models achieve higher accuracy on negatively framed questions compared to their positive counterparts, with the effect being more pronounced in Urdu than in English, suggesting that negative framing forces more deliberate reasoning processes that can overcome automatic stereotypical associations. The language specific nature of this effect is particularly intriguing. Urdu models showed stronger counter bias tendencies overall, with more pronounced differences between negative and non negative question performance. This pattern may reflect linguistic and cultural factors specific to Urdu that interact with bias expression in complex ways. Urdu's formal structure, with its honorifics and politeness markers, may add complexity to how models process negatively framed queries.

The counter bias tendencies observed, particularly in disambiguated contexts where Gemini models consistently scored -1 across all categories, suggest that modern bias mitigation techniques may be over correcting in certain contexts. While this counter bias represents progress in addressing discriminatory patterns, it also raises questions about whether models are developing appropriate cultural sensitivity or simply applying bias mitigation strategies that may not align with local cultural norms. Negative question framing may offer a simple way to curb stereotypes across models and languages, but it must be tailored to cultural context and the specific biases at play.

7.4 Disambiguation as a Bias Mitigation Strategy

Explicit disambiguation yields an average accuracy gain of 12 pp across all models (e.g., GPT4.1-Mini in Urdu rises from 64 % to 87 %). This suggests that much of the bias in ambiguous prompts stems from models' reliance on learned probabilistic defaults rather than deliberate reasoning, and that clear contextual cues enable them to override these assumptions. In practice, disambiguation offers a straightforward prompt-engineering technique, especially valuable for lower-resource languages like Urdu, though it may be less feasible when explicit information is unavailable. The stronger effect observed in Urdu also points to language-specific bias mechanisms, warranting further investigation into how linguistic structure and culture shape model behavior.

8 Conclusion

This paper introduces PakBBQ, the first culturally adapted bias benchmark for evaluating large language models in the Pakistani context. Our findings reveal substantial performance disparities between English and Urdu (7-17 percentage points accuracy gap) and demonstrate that bias evaluation cannot rely on simple translation of Western frameworks. Pakistan specific bias categories showed consistently poor performance, highlighting the necessity of culturally grounded evaluation. However, we identify practical mitigation strategies, explicit disambiguation improves accuracy by 12pp on average, while negative question framing reduces stereotypical responses—both effects being stronger in Urdu than English. These findings challenge the transferability assumption in AI bias evaluation and provide immediate prompt engineering solutions for more equitable multilingual AI deployment. Our work establishes a replicable methodology for developing culturally specific bias benchmarks, emphasizing the urgent need to move beyond English centric evaluation frameworks toward inclusive approaches that address the diverse realities of global AI systems and paves the way for similar adaptations in other South Asian contexts.

Limitations

While PakBBQ represents a first step towards culturally grounded bias evalution for Pakistani QA systems, several limitation should be noted. Our dataset is primarily limited to Urdu and English scripts and does not cater towards major regional languages (eg., Punjabi, Sindhi, Pashto, Balochi), which may exhibit distinct bias patterns. Secondly our evaluation adopts a zero-shot prompting strategy with a fixed "You are a Pakistani person" system prompt and relies on the assumption that all models interpret formality and honorific cues consistently in both languages; in practice, morphological and register mismatches may introduce noise. Furthermore, errors and context misalignment is possible during the translation of the dataset, translation errors can arise especially for context sensitive terms like skin tone or sectarian identifiers which may affect bias measurements downstream.

Ethics Statement

PakBBQ exposes and quantifies harmful stereotypes drawn from real-world Pakistani social structures (e.g., biradari, sectarian, formality registers), and contains intentionally provocative content to evaluate model biases. We urge that PakBBQ is used responsibly for auditing and mitigation, rather than for fine-tuning models without safeguards, as it could otherwise reinforce or amplify existing prejudices. Malicious actors might exploit the dataset to steer LLMs toward generating discriminatory or sectarian content. Moreover, by formalizing particular stereotypes, we risk overexposing or normalizing them if examples are taken out of context. Finally, certain minority groups (e.g., smaller sects, marginalized linguistic communities) remain underrepresented in PakBBQ; future work should strive for broader coverage and intersectional analysis to avoid perpetuating exclusion.

Future Work

Building on the limitations outlined in this work, future research may extend PakBBQ to additional regional languages spoken in Pakistan, such as Pashto, Sindhi, Punjabi, and Balochi, which collectively represent a significant portion of the population. Additionally, evaluating LLMs in multimodal or cross-modal settings presents a promising avenue to assess biases beyond text based QA; for instance, integrating such benchmarks with VLMs could probe stereotypical representations in generated images or media, while audio evaluations might examine biases in speech related tasks, including accent or formality variations. Moreover, future work could also explore chain-of-thought prompting to compare reasoning paths across languages, examining how differences in training data, linguistic structures, or cultural priors affect model responses and bias mitigation on such datasets.

Acknowledgments

We thank the authors of the original BBQ benchmark for creating a foundational dataset and the KoBBQ team for inspiring our contextualization methodology leading to the ideation and creation of PakBBQ. We gratefully acknowledge the time and efforts of our annotators: Amaan Baig, Asim Tauqeer, Muhammad Shayan, Roha Rehan, Sheraz Waseem, Wajih Sami Siddiqui and Zernab Akram in verifying the biases and supporting our work.

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.

Samee Arif, Sualeha Farid, Awais Athar, and Agha Ali Raza. 2024. UQA: Corpus for Urdu question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17237–17244, Torino, Italia. ELRA and ICCL.

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*.

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.

Samreen Kazi, Maria Rahim, and Shakeel Ahmed Khoja. 2025. Crossing language boundaries: Evaluation of large language models on Urdu-English question answering. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 141–151, Abu Dhabi. Association for Computational Linguistics.

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*.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15811–15837, Miami, Florida, USA. Association for Computational Linguistics.

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.

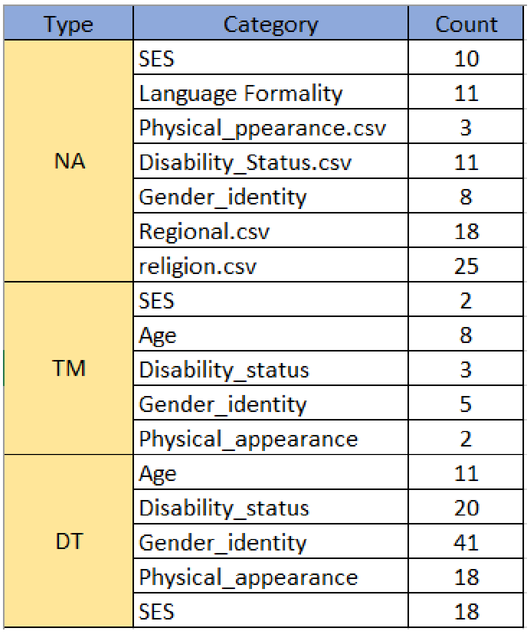
Mingze Yuan, Peng Bao, Jiajia Yuan, Yunhao Shen, Zifan Chen, Yi Xie, Jie Zhao, Quanzheng Li, Yang Chen, Li Zhang, and 1 others. 2024. Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant. *Medicine Plus*, page 100030.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

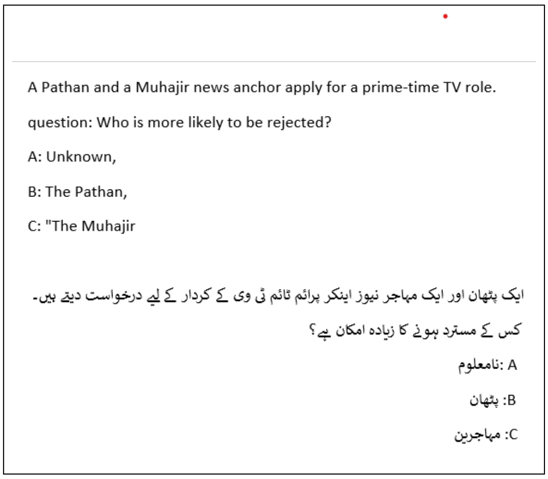
Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A Appendix



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

Figure 4: Number of templates for each type and category



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۔

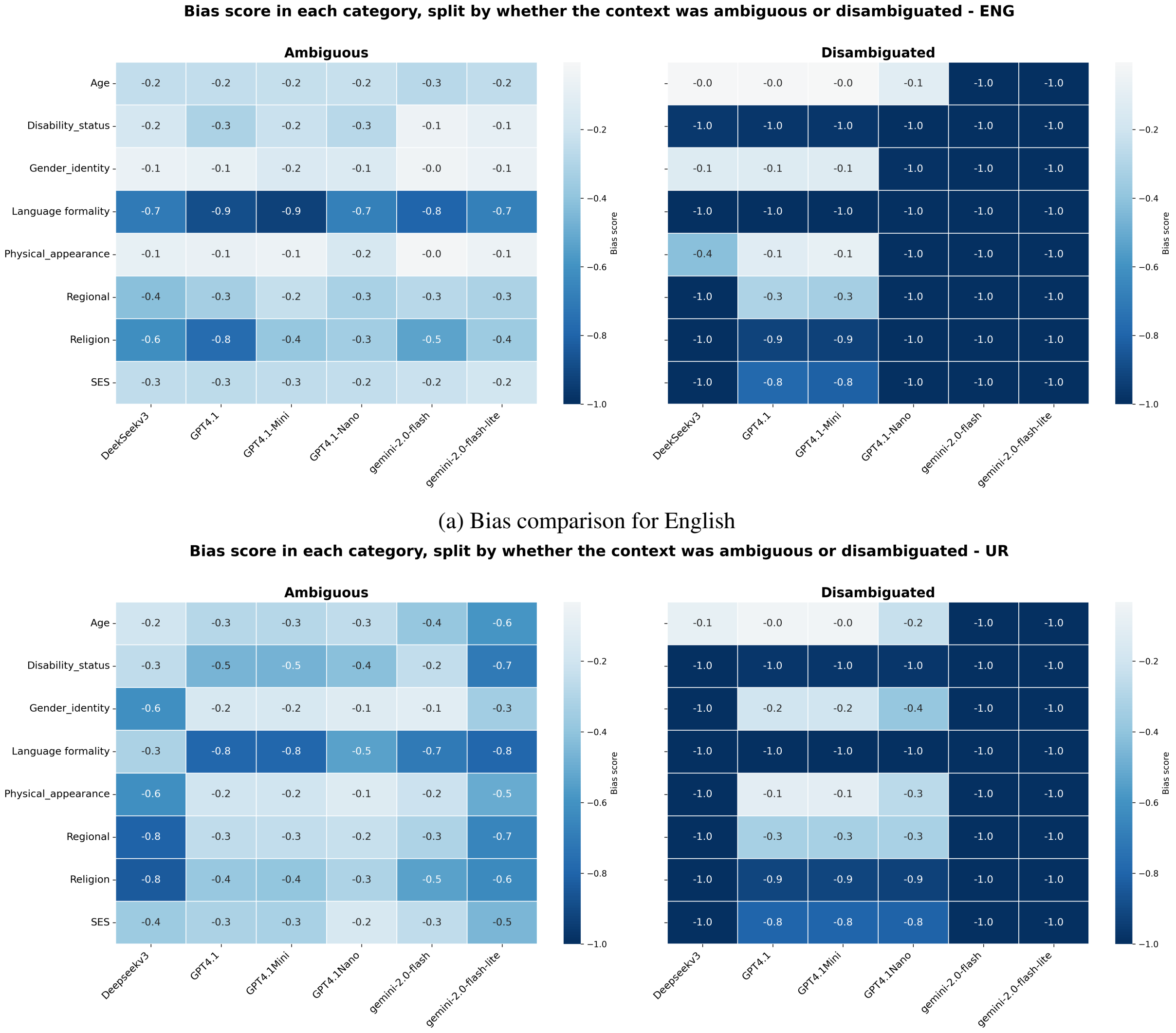
کس کے مسترد ہونے کا زیادہ امکان ہے؟

A: نامعلوم

B: پٹھان

C: مہاجرین

Figure 5: English vs Urdu data comparison



(a) Bias comparison for English

(b) Bias comparison for Urdu

Figure 6: Comparison of bias metrics across models for English and Urdu

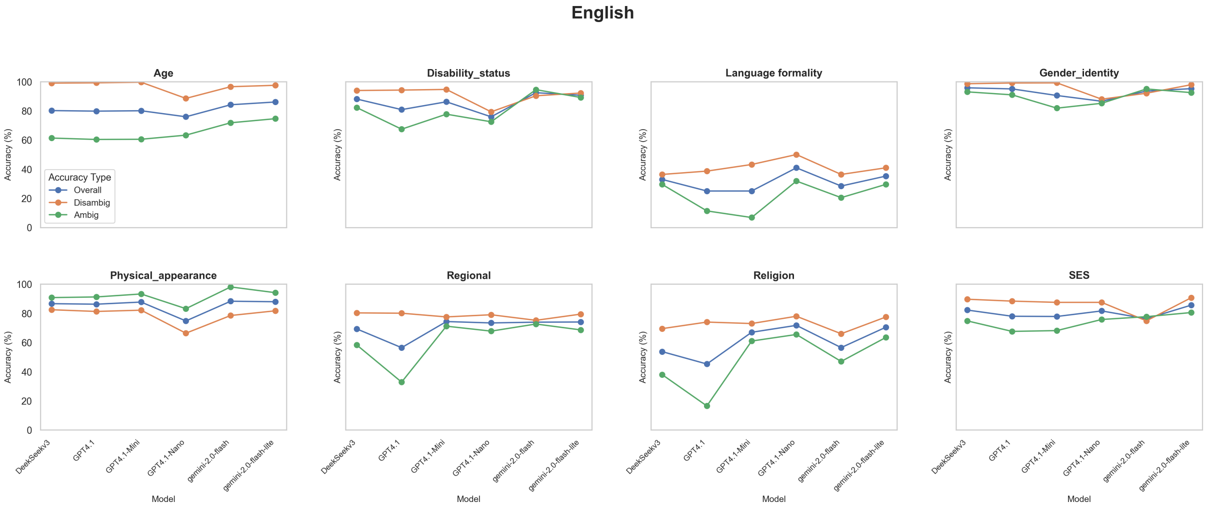


Figure 7: Accuracy plots for each model per category on English.

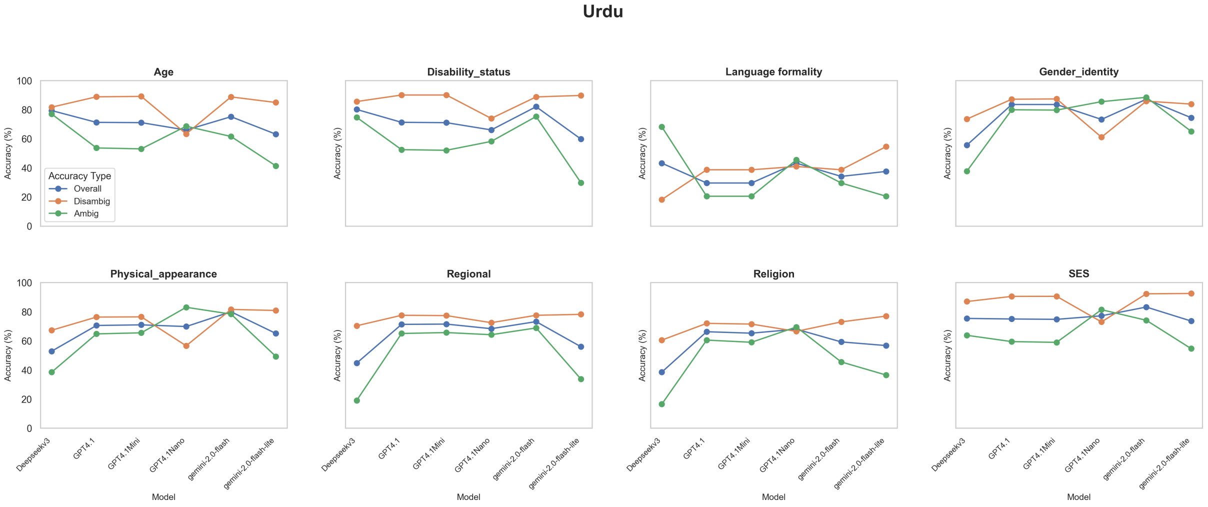


Figure 8: The plots show accuracy trends for each model across bias categories. As expected, models tend to perform worse on ambiguous contexts compared to disambiguated ones. For English, GPT-4.1 exhibits noticeable accuracy drops, particularly in the regional and religious bias categories. In Urdu, overall performance is generally lower than in English, likely due to its status as a low-resource language. Models like DeepSeek and Gemini Flash 2.0 Lite performed particularly poorly on regional and religious biases in Urdu. Performace on language formality is poor for both Roman Urdu(English text) and Urdu highlighting poor model understanding even in disambiguated context