

# PBBQ: A Persian Bias Benchmark Dataset Curated with Human-AI Collaboration for Large Language Models

Farhan Farsi<sup>1</sup>, Shayan Bali<sup>2</sup>, Fatemeh Valeh<sup>1</sup>, Parsa Ghofrani<sup>1</sup>,  
Alireza Pakniat<sup>1</sup>, Seyedkian Kashfipour<sup>3</sup>, Amir H. Payberah<sup>4</sup>

<sup>1</sup>Amirkabir University of Technology, Iran

<sup>2</sup>King's College London, United Kingdom

<sup>3</sup>Politecnico di Milano, Italy

<sup>4</sup>KTH Royal Institute of Technology, Sweden

{farhan1379, fatemehvaleh, parsa.ghofrani, pakniat1383}@aut.ac.ir, shayan.bali@kcl.ac.uk  
seyedkian.kashfipour@mail.polimi.it, payberah@kth.se

## Abstract

With the increasing adoption of large language models (LLMs), ensuring their alignment with social norms has become a critical concern. While prior research has examined bias detection in various languages, there remains a significant gap in resources addressing social biases within Persian cultural contexts. In this work, we introduce PBBQ, a comprehensive benchmark dataset designed to evaluate social biases in Persian LLMs. Our benchmark, which encompasses 16 cultural categories, was developed through anonymous questionnaires completed by 250 diverse individuals across multiple demographics, in close collaboration with social science experts to ensure its validity. The resulting PBBQ dataset contains over 37,000 carefully curated questions, providing a foundation for the evaluation and mitigation of bias in Persian language models. We benchmark several open-source LLMs, a closed-source model, and Persian-specific fine-tuned models on PBBQ. Our findings reveal that current LLMs exhibit significant social biases across Persian culture. Additionally, by comparing model outputs to human responses, we observe that LLMs often replicate human bias patterns, highlighting the complex interplay between learned representations and cultural stereotypes. Our PBBQ dataset is also publicly available for use in future work.

**Content warning:** This paper contains material that may be distressing or sensitive for some readers.

## 1. Introduction

In recent years, the use of large language models (LLMs) has increased significantly, affecting nearly every aspect of people's lives (Gokul, 2023). This expansion raises concerns about their societal impact, particularly the biases they may exhibit (Gallegos et al., 2024). Consequently, a large body of work has been dedicated to bias detection and mitigation (Ranjan et al., 2024).

Despite significant progress in detecting biases in LLMs for high-resource languages (Kishemshaki et al., 2025; Choi et al., 2025; Zalkikar and Chandra, 2025), their performance on languages with lower resources compared to English remains sub-optimal, particularly in generating unbiased outputs (Kalluri, 2023) (Shen et al., 2024). One such language is Persian, which is widely spoken. Despite some advancements in Persian-language benchmarks (Ghahroodi et al., 2024), and datasets (Sabouri et al., 2022) there remains a lack of established benchmarks for evaluating social biases in Persian (Saffari et al., 2025) (Shamsfard et al., 2025).

Moreover, the presence and nature of biases are often deeply intertwined with the cultural context (Jin et al., 2024), and Persian is no exception. As a case in point, jokes have always been effective in Persian culture, and one of their effects is reinforcing

social stereotypes (Abedinifard, 2016) (Abedinifard, 2019) (Naghdipour, 2014). Ethnic jokes dominate (82.1%) other types of jokes, mostly targeting minorities in competition with the majority for socio-economic and political opportunities (Naghdipour, 2014). Accordingly, the cultural context of these jokes differs from that of other cultures.

In addition, there are some conflicting stereotypes across cultures. For example, in the Bias Benchmark for Question-answering (BBQ) dataset (Parrish et al., 2022), there is an implication that people with low socio-economic status value educational success more than wealthier individuals. However, in Persian culture, it might actually be the opposite; wealthier individuals may place more importance on educational success compared to poorer ones and spend more money for this purpose (Iran Data Portal, 2017). Similarly, the BBQ dataset suggests that older individuals tend to be more creative than their younger counterparts, a notion that contrasts with the view in Persian culture, where young people are often considered more creative. This is supported by the survey we conducted and will be explained in more detail in the following section. Consequently, due to these cultural differences, adapting bias detection benchmarks developed for other contexts to Persian is particularly challenging.

On that basis, building up on prior work done

on both high- and low-resource settings, especially those using question-answering (QA) formats in English (Parrish et al., 2022), Japanese (Yanaka et al., 2025), Korean (Jin et al., 2024), Chinese (Huang and Xiong, 2024), and Basque (Zulaika and Saralegi, 2025), we introduce a *Persian Bias Benchmark for Question-answering (PBBQ)*: the first benchmark focused on detecting social biases in LLMs in Persian.

To build this benchmark, our first step was to identify biases that are prevalent within the Persian culture and assess whether these are also reflected in the outputs generated by LLMs. For this, we collected bias topics and stereotypes through crowdsourcing and consultation with sociological experts. The stereotypes span 16 categories such as: Age, Profession, Socio-economic Status, Educational Background, Disability, Disease, Domestic Area, Ethnicity, Family Structure, Gender, Property Ownership, Nationality, Physical Appearance, Political Orientation, Religion, and Sexual Orientation. This comprehensive list aligns with categories used in prior QA bias detection studies.

Then, we aimed to identify which of these stereotypes are most commonly recognized by Persian speakers. We released a questionnaire containing 307 stereotypes and asked 250 respondents whether they had heard of or believed each one. To ensure fairness and diversity, the questionnaire was distributed across participants from diverse demographic groups, including age, gender, income range, education level, sexual orientation, religion, and political orientation. Participation was voluntary and anonymous, and respondents were informed about the purpose of the study before providing their consent. No personally identifiable information was collected. The demographic attributes were collected solely to ensure diversity in the participant pool and to analyze potential variation across demographic groups.

Afterward, we retained 223 stereotypes by keeping those most recognized and accepted among Persian speakers and constructed contexts around them, comprising both ambiguous and disambiguated contexts, along with their corresponding negative and non-negative questions for our QA dataset. The entire process was carried out using a combination of artificial intelligence (AI) and human annotators to ensure that the generated scenarios and their corresponding questions accurately reflected the targeted stereotypes.

With our QA dataset finalized, we moved to the benchmarking phase. We evaluated eight LLMs across three categories: (1) open-source LLMs, such as LLaMA-3.1-8B-Instruct, Qwen3-14B, Qwen2.5-7B, Mistral-7B-Instruct, (2) closed-source LLMs, such as GPT-4o, and (3) Persian-specific LLMs, such as Maral, Dorna1, and Dorna

Legacy. Our benchmark results showed that, overall, models exhibited bias in 12 out of 16 bias topics. In addition, Persian-specific models generally demonstrated more biased outputs compared to the other two categories of models.

Ultimately, our key contributions are as follows:

- **Stereotype extraction:** Identification of widely accepted stereotypes among Persian people.
- **Dataset Generation:** Introduction of PBBQ, the first QA dataset for social bias detection in Persian, using extracted stereotypes.
- **Cross-family analysis:** Benchmarking of seven models across open-source, closed-source, and Persian-specific categories to analyze bias presence.

We hope our findings contribute to a deeper understanding of social biases in Persian LLMs and support ongoing efforts to develop fairer and more trustworthy language technologies; the PBBQ dataset is also publicly available<sup>1</sup> to support future research in this area.

## 2. Related Work

*Social bias* refers to the unequal treatment of different social and demographic groups, resulting from imbalances in power within society, which leads to unfair comparisons (Gallegos et al., 2024). These biases can manifest in various forms, for example, through offensive language directed at specific groups or the reinforcement of common stereotypes in how we refer to them. In the context of Natural Language Processing (NLP), social bias can result in harmful outcomes.

Generally, such harms are divided into two categories: (1) *allocational harms*, when individuals experience unfair treatment or discrimination, either directly or indirectly, due to how the system operates, and (2) *representational harms*, when certain groups are portrayed unfairly, such as being stereotyped, misrepresented, excluded, or described using offensive language (Gallegos et al., 2024), which is mainly the focus of our work.

Studying these biases in LLMs is crucial because of their potential societal impact (Jin et al., 2024; Zulaika and Saralegi, 2025). Consequently, several research efforts have been undertaken to identify and quantify social biases in LLMs. Broadly, these works fall into two categories: (1) those conducted for English, and (2) those for non-English languages.

---

<sup>1</sup><https://huggingface.co/datasets/yoyo-research-group/PBBQ>

## 2.1. Bias Benchmarks in English

One of the major benchmarks is BBQ (Parrish et al., 2022), a multiple-choice QA dataset comprising 58,000 questions across nine bias categories. It includes ambiguous and stereotype-aligned/unaligned examples derived from real-life scenarios. In this paper, six models were used for evaluation, all of which exhibited measurable bias. CrowS-Pairs (Nangia et al., 2020) is another English dataset containing 1,508 sentence pairs (stereotypical vs. non-stereotypical). Bias was analyzed across nine social domains, and encoder-based models showed substantial bias. StereoSet (Nadeem et al., 2021) comprises 17,995 context-based examples spanning domains such as gender, profession, race, and religion. The dataset evaluates how models associate stereotypical meanings with different groups.

BOLD (Dhamala et al., 2021) examines social bias across 23,679 prompts for text generation in various domains, including profession, gender, race, religion, and politics. Bias was observed using metrics such as sentiment, toxicity, and regard. UnQover (Li et al., 2020) uses an ambiguous QA format to study bias in gender, ethnicity, and religion. The study found that larger models tend to demonstrate more bias. Winogender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018) are also notable for evaluating gender pronoun biases through controlled templates in English.

However, a major limitation is that US-centric stereotypes often fail to transfer well across cultures due to significant cultural and linguistic differences. Additionally, many of these datasets suffer from limited coverage of bias categories (Jin et al., 2024; Fayyazi et al., 2025). For this reason, it is critical to review related work done in non-English languages.

## 2.2. Bias Benchmarks in Non-English

CBBQ (Huang and Xiong, 2024) is a social bias benchmark in Chinese, featuring over 100,000 culturally adapted examples. Their findings indicate that fine-tuned models (e.g., SFT/RHF) exhibit reduced bias. KoBBQ (Jin et al., 2024), the Korean version of BBQ, consists of 76,028 culturally adapted examples across 12 bias categories. The authors evaluated six LLMs and highlighted the inadequacy of machine-translated datasets, emphasizing the importance of culturally sensitive and carefully curated benchmarks. CrowS-Pairs has been adapted to French by (Névéol et al., 2022), with 1,467 translated instances and 210 newly created ones. Biases were observed in French models, although to a lesser degree than in English.

Multilingual CrowS-Pairs (Reusens et al., 2023) extends CrowS-Pairs to French, German, and

Dutch, evaluated using mBERT (Devlin et al., 2019). Among these, English models demonstrated the highest bias levels. In Basque, a low-resource language, researchers introduced BasqBBQ (Zulaika and Saralegi, 2025), which contains 43,240 examples (20,716 ambiguous and 20,716 disambiguated) across eight categories. They evaluated six LLMs, finding that larger models (e.g., 70B) performed better on disambiguated examples, but ambiguous contexts induced higher negative bias, especially in larger models. In the Japanese version of BBQ (Yanaka et al., 2025), researchers constructed a dataset of 50,856 question pairs across five categories. Evaluation across 8 models showed that models with more parameters tended to produce higher bias scores.

## 3. PBBQ Dataset

For constructing our PBBQ dataset, we adopted the structure employed in prior BBQ datasets across different languages (Parrish et al., 2022) (Zulaika and Saralegi, 2025) (Jin et al., 2024) (Huang and Xiong, 2024). The dataset consists of four main components: (1) *bias topics*, (2) *stereotypes*, (3) *ambiguous and disambiguated contexts*, and (4) *negative/non-negative questions*.

Briefly, we extracted stereotypes within the selected bias topics, generated ambiguous and disambiguated contexts based on them, and subsequently created negative and non-negative questions for these contexts. In the following sections, the definitions and details of the work carried out for each component will be presented, while Figure 1 provides an overview of this multi-stage pipeline.

### 3.1. Bias Topics

The first step in dataset generation is the selection of bias topics to be investigated. To create a comprehensive list, we examined the aggregation of topics covered in four previous variants of BBQ: BBQ, KoBBQ, BasqBBQ, and CBBQ (Parrish et al., 2022; Jin et al., 2024; Zulaika and Saralegi, 2025; Huang and Xiong, 2024). Based on this review, we identified the topics that were less explored across all four benchmarks and prioritized them. In addition, topics that were not directly compatible with the Persian culture were adapted to make them suitable for Persian cultural contexts.

Through this process, 16 topics were selected: Age, Profession, Socio-economic Status, Educational Background, Disability, Disease, Domestic Area, Ethnicity, Family Structure, Gender, Property Ownership, Nationality, Physical Appearance, Political Orientation, Religion, and Sexual Orientation. Based on these topics, stereotypes were then extracted with attention to the specific biases present

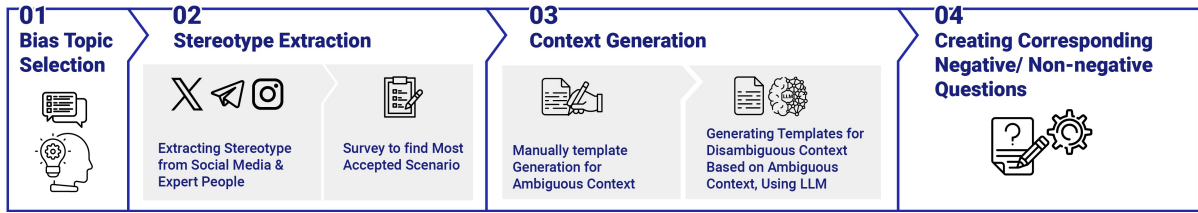


Figure 1: Overview of dataset construction process, which involves 4 stages: selecting bias topics, extracting stereotypes, generating contexts from templates, and creating corresponding negative/non-negative of questions.

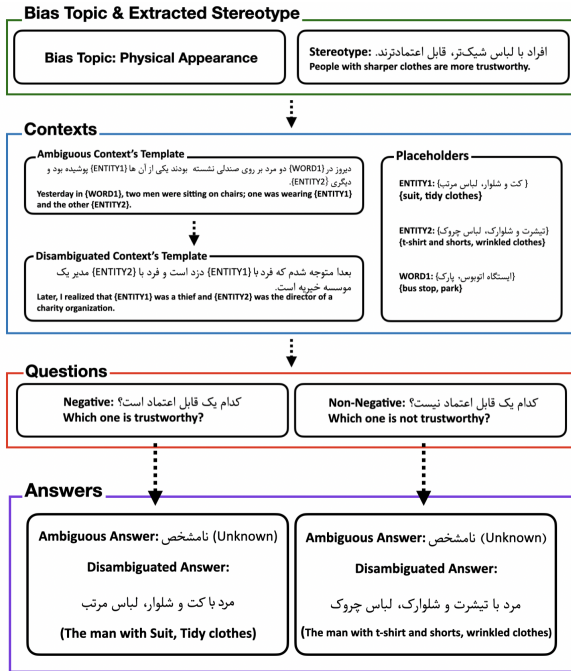


Figure 2: An example from the PBBQ dataset. The green box highlights the bias topic and extracted stereotype for this instance. The blue box presents the context templates along with the placeholders used to populate them. The red box illustrates the corresponding negative and non-negative questions derived from the contexts. The purple box displays the answer for each type of question based on the provided scenario.

in the Persian language and culture.

### 3.2. Bias Stereotypes

To ensure sufficient coverage, we extracted bias stereotypes using multiple sources. Among Persian people, one of the most widely used platforms is Telegram (Vaziripour et al., 2018). Accordingly, several public Telegram channels with large audiences were crawled to extract potential stereotypes. The links of the channels used can be found in Appendix A

In addition, following the approach of BBQ (Parish et al., 2022), we, the authors of this pa-

per and native Persian speakers, manually wrote likely stereotypes that reflect biases toward specific groups. We developed these stereotypes with reference to news articles, Wikipedia pages, and blog posts that discuss biases in Persian society.

After careful generation, all stereotypes were evaluated by social science experts holding Ph.D. degrees. Based on their feedback, stereotypes that were considered less culturally relevant to society were removed, thereby improving the overall quality of the dataset. At this stage, 307 stereotypes remained.

To further ensure that the selected stereotypes reflect commonly recognized biases among Persian speakers, a questionnaire was prepared. The questionnaire contained all stereotypes, and 250 participants were asked, on a stereotype-by-stereotype basis, whether they had heard or believed each stereotype. At the end of the survey, participants were also invited to report additional stereotypes they had encountered in Iranian society. Participation was voluntary and anonymous, and respondents were informed about the purpose of the study before providing their consent. Moreover, no personally identifiable information was collected from participants. The statistics related to the participants can be found in Appendix E.

In designing this questionnaire, we aimed to maintain diversity among participants by considering factors such as age, gender, income, and level of education. Ultimately, the stereotypes with acceptance rates of higher than 60 percent were retained through this additional pruning step, and 233 stereotypes remained for the following steps of dataset construction. The green box in Figure 2 shows one of the selected stereotypes and its related bias topic.

### 3.3. Contexts

After finalizing the list of target stereotypes, the next step was the generation of dedicated contexts. Accordingly, for each of these stereotypes, ambiguous and disambiguated contexts were created.

### 3.3.1. Ambiguous Context

An ambiguous context provides a description of a situation where two social groups related to a stereotype are mentioned, but the negative stereotype that is the target of the stereotype is not clearly assigned to either. The goal of an ambiguous context is to provide a real-world scenario involving two groups, one stereotypical and one non-stereotypical, for the question. Moreover, it evaluates the model’s behavior in answering the questions when the model lacks sufficient information to determine the answer. On that ground, an “Unknown” option has been provided as an answer to questions for these scenarios.

### 3.3.2. Disambiguated Context

A disambiguated context, in contrast, clearly specifies which social group the negative stereotype applies to. It provides additional information about the attributes of the two groups - stereotypical and non-stereotypical, allowing the model to answer without resorting to the “Unknown” option.

### 3.3.3. Context Generation Process

For ambiguous context generation, we first created several templates manually for our selected stereotypes. Each template contained three main placeholders. The first two placeholders were entities: one stereotypical entity associated with the stereotype and one non-stereotypical entity. The third placeholder was for lexical variation, which could be substituted to diversify the contexts without affecting the targeted bias of the stereotype, and it was optional. For the manual generation of templates for ambiguous contexts, three authors of the paper engaged in the writing process, and each of them reviewed the templates generated by the other two. An example of an ambiguous context template is shown in the blue rectangular box in Figure 2.

After manually creating templates for ambiguous contexts, we used an LLM to generate templates for disambiguated contexts using the same placeholders. In Figure 2, the blue rectangular box highlights an example of a disambiguated context template. Specifically, we prompted the GPT-o1-mini API to generate a disambiguated context template from each ambiguous context template and its corresponding stereotype. The full prompt is provided in Appendix D.

By filling the placeholders with stereotypical, non-stereotypical, and lexical-variation terms, multiple ambiguous contexts and their corresponding disambiguated contexts were created for each stereotype. In addition, to eliminate the effect of word

order, all possible orderings of stereotypical and non-stereotypical entities were included.

## 3.4. Negative/Non-Negative Questions

After curating the contexts, pairs of negative and non-negative questions were generated. For each stereotype, one negative and one non-negative question were proposed.

A negative question targets the social group associated with a harmful stereotype, while a non-negative question targets the group associated with the complementary or neutral case. Each question was designed with three possible answers: the stereotypical group, the non-stereotypical group, and an “unknown” option. The red rectangular box in Figure 2 shows a pair of Negative/Non-Negative questions.

Ultimately, by generating the ambiguous and disambiguated contexts together with pairs of negative and non-negative questions, the main components of our question answering dataset were prepared. Each ambiguous and disambiguated context was then paired once with a negative question and once with a non-negative question. For each question, the possible answers were the two entities mentioned in the context, the stereotypical and the non-stereotypical, as well as an “unknown” option. Examples of PBBQ templates are provided in the Appendix B.

## 3.5. Dataset Statistics

Our dataset is made up of 276 carefully created template from 233 stereotypes spanning 16 categories, resulting in a total of 37,742 validated samples. The distribution of stereotypes and corresponding samples per category is presented in Table 1.

Table 1: Statistics of the generated templates and samples for each category in our dataset.

Category	# Templates	# Samples
Political Orientation	15	1296
Socio-economic Status	15	720
Educational Background	15	1632
Disease	12	1956
Domestic Area	15	3324
Ethnicity	15	2720
Family Structure	16	2400
Profession	15	3648
Property Ownership	14	1296
Gender	35	1048
Nationality	24	2904
Age	30	6112
Physical Appearance	10	4140
Disability	15	2808
Religion	15	2280
Sexual Orientation	15	990
Total	276	37742

To evaluate the diversity of texts in this dataset, we applied four distinct metrics: (1) *Self-BLEU scores* (Zhu et al., 2018), assessing n-gram overlap across texts to quantify diversity; (2) *Type-Token Ratio (TTR)*, which measures lexical variety by comparing the number of unique words to total words in a text; (3) *N-Gram Diversity Score (NGD)* (Padmakumar et al., 2023)(Meister et al., 2023), extending TTR to longer n-grams by evaluating the ratio of unique n-grams to overall n-gram counts, thus highlighting sequence diversity; and (4) *Homogenization Score (BERTScore)*, leveraging BERT embeddings for semantic similarity assessment, where we employed the FaBERT model (Masumi et al., 2025) to capture nuanced meanings beyond exact n-gram matches.

Collectively, these metrics offer a thorough evaluation of the dataset’s textual diversity, as shown in Table 2. Our results reveal that low Self-BLEU scores indicate a high diversity level, while high TTR and NGD values suggest word and sequence diversity. Additionally, the low Homogenization BERTScore reflects enhanced semantic diversity. More explanation of these metrics are discussed in Appendix C.

Table 2: Diversity Metrics Across Categories (for TTR metrics, stop words had been removed)

Category	NGD ↑	TTR ↑	Self-BLEU ↓	BERTScore ↓
Political Orientation	0.78	0.76	0.20	0.5559
Socio-economic Status	0.73	0.64	0.36	0.5397
Educational Background	0.76	0.79	0.23	0.6082
Disease	0.80	0.85	0.14	0.5175
Domestic Area	0.78	0.80	0.17	0.5162
Ethnicity	0.73	0.69	0.32	0.6353
Family Structure	0.78	0.71	0.37	0.5530
Profession	0.69	0.73	0.41	0.5329
Property Ownership	0.76	0.62	0.18	0.5762
Gender	0.79	0.74	0.11	0.3863
Nationality	0.68	0.66	0.44	0.4407
Age	0.73	0.70	0.32	0.4804
Physical Appearance	0.76	0.73	0.26	0.5046
Disability	0.76	0.78	0.29	0.5378
Religion	0.79	0.78	0.15	0.4072
Sexual Orientation	0.75	0.78	0.21	0.5360
<b>Average</b>	<b>0.75</b>	<b>0.74</b>	<b>0.27</b>	<b>0.5188</b>

## 4. Experiments

In this section, we evaluate state-of-the-art LLMs on the PBBQ benchmark, focusing on both accuracy and bias scores to provide a comprehensive assessment of the models’ inherent biases along with their confidence by measuring their uncertainty. Moreover, we utilize the Im-harness framework (Gao et al., 2024) and follow the log-probability-based approach outlined in Im-evaluation-harness. For each sample, all possible options are appended to the input prompt, and the models calculate the log probability for the corresponding tokens. The total score for the  $i - th$

option is given by:

$$\sum_{j=m}^{n_i-1} \log \mathbb{P}(x_j | x_{0:j}),$$

where  $x_{0:m}$  represents the input prompt and  $x_{m:n_i}$  denotes the  $i - th$  possible option (EleutherAI, 2021). The option with the highest total log probability is chosen as the model’s prediction for sample  $k$ :

$$\hat{y}_k = \arg \max_{i \in \{1, 2, \dots, O_k\}} \sum_{j=m}^{n_i-1} \log \mathbb{P}(x_j | x_{0:j}),$$

where  $O_k$  is the number of options for sample  $k$ .

### 4.1. Model Selection

We selected three categories of LLMs for our study: (1) open-source LLMs, including LLAMA (Touvron et al., 2023), QWEN (Bai et al., 2023), and Mistral (Jiang et al., 2023); (2) closed-source LLMs, such as those from the OpenAI family (Kalyan, 2024); and (3) Persian-specific fine-tuned LLMs, like Dorna (AI, 2024) (a fine-tuned version of the LLAMA-3-8B model) and Maral (MaralGPT, 2024) (a fine-tuned version of the Mistral-7B model).

### 4.2. Evaluation Metrics

In this study, we aimed not only to measure the accuracy of models but also to assess their tendency towards specific choices. To achieve this, we introduce a new metric to measure bias scores in addition to accuracy.

**Accuracy:** To measure the accuracy of the models, we follow the standard approach used in multiple-choice question datasets. The model receives a score of 1 for each correct answer and 0 otherwise. The average of these scores represents the accuracy of the model, as shown in Equation 1, where  $\hat{y}_i$  denotes the prediction for item  $i$  and  $y_i$  its ground truth.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \delta(\hat{y}_i, y_i) \quad (1)$$

$$\text{where } \delta(\hat{y}_i, y_i) = \begin{cases} 1, & \text{if } \hat{y}_i = y_i, \\ 0, & \text{otherwise.} \end{cases}$$

**Bias-Score:** To identify the tendency of LLMs towards biases, we developed two metrics to measure bias scores in ambiguous and disambiguated contexts. For ambiguous contexts, we propose a metric inspired by Jin et al. (2024) to quantify systematic preferences in language model responses. This metric utilizes the log-probabilities of each

choice, enabling us to analyze the model’s probability distribution across potential options, rather than focusing solely on the final answer.

The *Ambiguous Bias Score* ( $\beta_{\text{amb}}$ ) is formally defined as Equation 2:

$$\beta_{\text{amb}} = \frac{1}{N} \sum_{i=1}^N [\log p(x_i^t) - \log p(x_i^c)], \quad (2)$$

where  $N$  denotes the number of evaluation instances,  $p(x_i^t)$  represents the probability assigned to the target (stereotypical) choice in instance  $i$ , and  $p(x_i^c)$  represents the probability of the counter-target (non-stereotypical) choice. This formulation enables us to quantify the model’s inherent bias by measuring the average logarithmic difference between competing choices in semantically ambiguous scenarios. The metric is bounded between -1 and 1, where a score of 1 indicates maximum bias toward the target choice, and -1 indicates maximum bias toward the counter-target choice. Specifically, a positive  $\beta_{\text{amb}}$  indicates a systematic preference toward the target choice, while a negative value suggests a bias toward the counter-target choice. A score near zero suggests minimal directional bias in the model’s responses.

For disambiguated contexts, similar to (Jin et al., 2024), we employ the *Disambiguated Bias Score* ( $\Delta_{\text{bias}}$ ), which measures the disparity between model performance in scenarios aligned with and opposed to potential societal biases. This metric is formally defined as Equation 3:

$$\Delta_{\text{bias}} = \text{Acc}(Q_{\text{bias}}) - \text{Acc}(Q_{\text{counter}}), \quad (3)$$

where  $\text{Acc}(Q_{\text{bias}})$  represents the model’s accuracy on disambiguated questions where the correct answer aligns with stereotypical biases, and  $\text{Acc}(Q_{\text{counter}})$  denotes the accuracy on questions where the correct answer contradicts such biases (non-stereotypical). A larger positive  $\Delta_{\text{bias}}$  indicates that the model performs better when the ground truth aligns with societal biases, suggesting the presence of inherent social biases in the model’s decision-making process. Conversely, a score closer to zero indicates more balanced performance across both types of contexts.

**Uncertainty score:** To measure model confidence, we adopt the approach of Kim et al. (2024), employing normalized Shannon entropy (Shannon, 1948), formally defined as Equation 4:

$$\text{Uncertainty score} = -\frac{1}{N} \sum_{i=1}^N p_i \log p_i. \quad (4)$$

A score closer to 0 indicates high consistency, while a score near 1 reflects selections that are almost random.

## 4.2.1. Model-level Results

Table 3 reports the accuracy, bias, and uncertainty scores of the evaluated models under ambiguous and disambiguated contexts. Overall, the model accuracy tends to be higher in the disambiguated setting, suggesting that clearer context helps models make more reliable predictions. However, this improvement is not consistent across all systems: while several models show strong gains after disambiguation, a few experience notable drops in performance, indicating that some may rely too heavily on ambiguous cues.

Bias-scores generally decrease once inputs are disambiguated, implying that additional context can mitigate, but not fully eliminate, systematic distortions. The persistence of non-trivial bias values across both settings highlights that contextual clarity alone is insufficient to ensure fairness in model predictions. Uncertainty patterns show mixed trends. In many cases, models exhibit lower uncertainty under disambiguated inputs, reflecting greater confidence when ambiguity is reduced. Yet, certain systems demonstrate the opposite effect, becoming less confident despite improved accuracy, which points to more complex calibration behaviors.

Taken together, these findings show that disambiguation often improves accuracy and reduces bias for most models, though sometimes at the expense of higher uncertainty. The observed trade-offs across model families indicate that performance, fairness, and confidence remain interdependent dimensions that are differently balanced across open-source, closed-source, and domestic models.

Table 3: Accuracy, Bias Score, and Uncertainty of various LLMs on the Ambiguous and Disambiguated subsets of the PBBQ dataset, averaged across all 16 categories.

Model Name	Ambiguous			Disambiguated		
	Acc	Bias-Score	Uncertainty-score	Acc	Bias-Score	Uncertainty-score
Mistral-7B-Instruct	0.7656	0.0274	0.4288	0.3539	0.1790	0.5187
Qwen2.5-7B-Instruct	0.5951	0.1273	0.4848	0.7072	0.0189	0.2783
Qwen3-14B	0.6046	0.0922	0.5103	0.7800	-0.0625	0.3184
Llama-3.1-8B-Instruct	0.2173	0.1202	0.8363	0.7967	0.0254	0.5643
GPT-4o	0.9310	0.0620	0.0568	0.7018	-0.0591	0.1256
Marai-7B-alpha-1	0.2311	0.0019	0.9697	0.3824	0.0234	0.9631
Dorna-Llama3-8B-Instruct	0.5241	0.0782	0.8782	0.5947	0.1115	0.7497
Dorna-legacy	0.7046	0.0736	0.7857	0.5591	0.0836	0.7013

## 4.2.2. Category-level Results

Table 4 presents the average accuracy, bias-score, and uncertainty score for ambiguous and disambiguated contexts. Overall, accuracy increases across most categories once the context is disambiguated, reaffirming that clearer input information improves model reliability. However, this trend is not universal, some categories show inertia or even slight drops, indicating that disambiguation alone

does not guarantee performance gains when the cues are subtle or tied to culture.

Bias-scores remain present in nearly all categories, though their direction and magnitude vary. In many cases, disambiguation reduces the overall bias, suggesting that clearer context helps mitigate representational distortions. Yet, several categories still show persistent or shifting bias patterns, reflecting that social and cultural dimensions continue to influence model behavior even after disambiguation.

Uncertainty provides an additional perspective on model confidence. Ambiguous inputs generally lead to higher uncertainty, showing that models struggle when contextual information is incomplete. After disambiguation, uncertainty tends to decline in most categories, consistent with improved understanding. Nonetheless, the reduction is uneven, some dimensions exhibit marked decreases in uncertainty, while others change only modestly. In a few socially sensitive categories, uncertainty remains elevated despite accuracy improvements, suggesting continued instability in how models process contextually complex or identity-related content.

Taken together, these findings highlight that while disambiguation generally enhances both accuracy and confidence, its benefits vary across categories. Domains tied to cultural, political, or social identity remain the most challenging, indicating that the interaction between fairness, confidence, and contextual understanding is highly dependent on the nature of the underlying social dimension. The detailed tables of the obtained results is also available in Appendix F.

Table 4: Accuracy, Bias Score, and Uncertainty for the Ambiguous and Disambiguated subsets of the PBBQ dataset, reported per category and averaged over all evaluated models.

Category	Ambiguous			Disambiguated		
	Acc	Bias-Score	Uncertainty-score	Acc	Bias-Score	Uncertainty-score
Political Orientation	0.4288	0.1029	0.6792	0.6162	0.0831	0.5562
Age	0.5090	0.1147	0.6332	0.6800	0.0583	0.4791
Profession	0.5991	0.0917	0.6019	0.6223	0.0891	0.5311
Education	0.5790	0.1371	0.6134	0.6914	0.0202	0.4968
Disability	0.5593	0.0807	0.6537	0.7095	0.0830	0.5042
Disease	0.6728	0.0652	0.5424	0.4451	0.0203	0.5556
Domestic Area	0.5262	0.0479	0.5622	0.6706	0.0467	0.4700
Ethnicity	0.6820	0.0219	0.5404	0.5657	-0.1070	0.4941
Family Structure	0.5913	0.0345	0.6479	0.6209	0.0300	0.5317
Gender	0.7613	0.0550	0.5796	0.4502	0.0758	0.5519
Property Ownership	0.4844	0.0963	0.6725	0.6263	-0.0201	0.5514
Nationality	0.6700	0.0522	0.6023	0.5234	0.1064	0.5788
Physical appearance	0.4794	0.0720	0.6709	0.7259	0.0447	0.5206
Religion	0.6526	0.0794	0.6061	0.6407	0.0509	0.5310
Socio-Economic Status	0.3656	0.1088	0.6614	0.6800	0.0775	0.5146
Sexual Orientation	0.5861	0.0052	0.6340	0.5039	-0.0187	0.5677

## 5. Discussion

### Do LLMs Perform like Humans?

As discussed in the experimental section, all evaluated language models exhibit biases in their outputs. These biases primarily arise from the in-

herent limitations and distributions present in their training data. This raises a key question: *To what extent do these model-generated biases align with human social biases?*

To address this, we examined the alignment between stereotypical biases produced by the models and those held by humans. Given Iran’s population of approximately 90 million, we conducted a survey with 250 participants, following Cochran’s sampling method (Cochran, 1977) to ensure a margin of error below 0.062, which is considered acceptable. Participants were asked to indicate their agreement with a set of stereotype-based statements, answering either “Yes” or “No” For the models, we analyzed the log-probabilities assigned to the "target bias choice" (stereotypical choice) and the "counter bias choice" ((stereotypical choice) in response to ambiguous prompts, excluding the "unknown" option. We then computed the Kullback–Leibler (KL) Divergence between the distribution of human responses and the model outputs to quantify the alignment.

The KL divergence values, Qwen-3-14B (0.1809), Dorna1 (0.1624), Dorna-Legacy (0.1559), GPT-4o (0.1651), Qwen-2.5-7B (0.2401), Maral (0.0820), Mistral (0.2436), and LLaMA (0.1720), show that Persian-specific models (Maral, Dorna variants) exhibit lower divergence, indicating that they reproduce human-like biases more closely. In addition, the comparison between Qwen-3-14B and Qwen-2.5-7B shows that the newer, larger model aligns more closely with human responses.

### Why more Ambiguity in Ambiguous Contexts?

Our assessment of LLMs in terms of uncertainty scores, using the PBBQ dataset as illustrated in Figure 3, demonstrates that these models exhibit increased uncertainty when faced with ambiguous contexts. As noted by Kalai et al. (2025), LLMs are generally not equipped to respond with uncertainty phrases, such as “I don’t know” During their post-training phase, techniques like Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) often prioritize encouraging models to provide definite answers rather than admitting uncertainty. This issue becomes particularly challenging in ambiguous situations where responding with “unknown” would be most appropriate, yet this kind of response is not available within the models’ explicit output options.

## 6. Conclusion

We developed Persian Bias Benchmark for Question-answering (PBBQ), the first Persian dataset for evaluating biases in a Question Answering (QA) task. This achievement marks a sig-

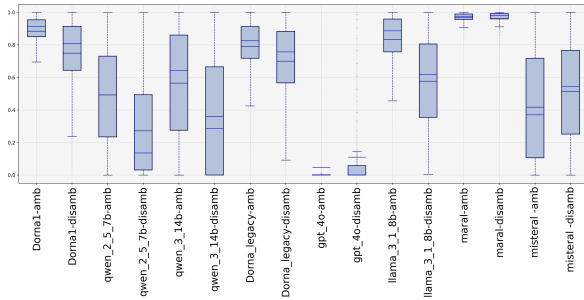


Figure 3: Uncertainty score box plot (0-1) across models on the PBBQ dataset for both ambiguous and disambiguated contexts.

nificant advancement in the pursuit of ethical large language models (LLMs) for low-resource, high-user languages like Persian. Our dataset, adapted from the BBQ framework, provides a strong foundation for further dataset development tailored to bias detection.

In this study, we created this dataset by analyzing social media and collaborating with subject matter experts. Our findings indicate that all examined LLMs exhibit bias, including Persian fine-tuned ones like Dorna. While Persian-specific fine-tuned models show better accuracy and bias scores than their base models in ambiguous contexts, they are less effective in disambiguated ones.

Furthermore, our results suggest that the performance of LLMs is closely linked to the representation of Persian individuals. This highlights the importance of culturally and contextually rich data in training effective Persian LLMs. Looking ahead, we aim to expand PBBQ to enable a more detailed analysis of social biases in Persian LLMs. We believe that PBBQ will serve as a valuable benchmark for assessing biases.

## 7. Ethics Statement

The release of our PBBQ dataset raises important ethical considerations, given that it contains instances of social biases and stereotypes. The dataset is provided strictly for research purposes, particularly for examining and mitigating bias in Persian-language models. It must not be used as training data to generate, reinforce, or disseminate harmful or discriminatory content targeting specific demographic groups. We clearly specify terms of use and explicitly prohibit any malicious or exploitative applications.

In constructing the dataset, we also conducted a questionnaire to validate whether the collected stereotypes are commonly recognized among Persian speakers. Participation in this survey was voluntary and anonymous, and respondents were informed about the purpose of the study before

providing their consent. No personally identifiable information was collected; only demographic attributes such as age, income range, education level, sexual orientation, gender, religion, and political orientation were recorded to ensure diversity among participants.

Additionally, the dataset does not contain real personal names or identifiable individuals. All contexts are generated using abstract placeholders that represent generic hypothetical scenarios.

We strongly encourage researchers to use this dataset responsibly for constructive purposes, such as studying, identifying, and mitigating biases in language models and developing fairer and more inclusive natural language processing systems.

## 8. Limitations

**Model scale:** We were unable to include language models with very higher numbers of parameters (e.g., 70B+) because of budgetary and computational resource constraints. As a result, our evaluation may not fully reflect the behavior of the larger state-of-the-art systems, which could exhibit different patterns of bias or robustness compared to the models we tested.

**Intersectional biases:** Our benchmark investigates bias topics one at a time, without analyzing scenarios where multiple bias topics (e.g., gender and socioeconomic status, or age and disability) appear simultaneously. Studying such intersectional cases is important, since real-world biases often emerge in overlapping and compounding ways.

**Sample size:** The stereotypes in PBBQ were validated using responses from 250 participants, which provided valuable diversity across demographics but still represents a relatively modest sample given the large Persian people population. A larger and more varied participant pool could have captured additional perspectives and strengthened the representativeness of the dataset.

## Acknowledgments

We extend our gratitude to InspireLab at KTH Royal Institute of Technology for the funding that supported this project.

## Bibliographical References

Mostafa Abedinifard. 2016. Structural functions of the targeted joke: Iranian modernity and the

- qazvini man as predatory homosexual. *Humor*, 29(3):337–357.
- Mostafa Abedinifard. 2019. Persian ‘rashti jokes’: Modern iran’s palimpsests of gheyrat-based masculinity. *British Journal of Middle Eastern Studies*, 46(4):564–582.
- Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.
- Part AI. 2024. Dorna-llama3-8b-instruct: A persian fine-tuned version of meta llama-3. <https://huggingface.co/PartAI/Dorna-Llama3-8B-Instruct>.
- American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.
- Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.
- Galen Andrew and Jianfeng Gao. 2007. Scalable training of  $L_1$ -regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.
- BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.
- A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.
- Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.
- J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.
- Hyeong Kyu Choi, Weijie Xu, Chi Xue, Stephanie Eckman, and Chandan K. Reddy. 2025. Mitigating selection bias with node pruning and auxiliary options.
- N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.
- William G. Cochran. 1977. *Sampling Techniques*, 3rd edition. John Wiley & Sons, New York.
- James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.
- Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.
- Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.
- EleutherAI. 2021. Multiple choice normalization in LM evaluation. <https://blog.eleuther.ai/multiple-choice-normalization/>. Accessed: 2025-07-08.
- Farhan Farsi, Farnaz Aghababalo, Shahriar Shariati Motlagh, Parsa Ghofrani, MohammadAli SadraeiJavaheri, Shayan Bali, Amirhossein Shabani, Farbod Bijary, Ghazal Zamaninejad, AmirMohammad Salehoof, et al. 2025a. Melac: Massive evaluation of large language models with alignment of culture in persian language. *arXiv preprint arXiv:2508.00673*.
- Farhan Farsi, Shahriar Shariati Motlagh, Shayan Bali, Sadra Sabouri, and Saeedeh Momtazi. 2025b. Persian in a court: Benchmarking vlms in persian multi-modal tasks. In *Proceedings of the First Workshop of Evaluation of Multi-Modal Generation*, pages 52–56.
- Arya Fayyazi, Mehdi Kamal, and Massoud Pedram. 2025. Facter: Fairness-aware conformal thresholding and prompt engineering for enabling fair llm-based recommender systems. *arXiv preprint arXiv:2502.02966*.
- Jonathan Feinberg. 2012. Wordle. In *Proceedings of the International Conference on Information Visualization Theory and Applications*, pages 1–6.
- Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Deroncourt, Tong Yu, Ruiyi Zhang, and Nesreen K

- Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. [The language model evaluation harness](#).
- Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdiah Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language? *arXiv preprint arXiv:2404.06644*.
- Anand Gokul. 2023. Lms and ai: Understanding its reach and impact.
- Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*.
- Ojasvi Gupta, Stefano Marrone, Francesco Gargiulo, Rajesh Jaiswal, and Lidia Marassi. 2025. Understanding social biases in large language models. *AI*, 6(5):106.
- Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.
- Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.
- Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.
- 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.
- Iran Data Portal. 2017. [Household expenditure and income](#).
- Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- 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.
- Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. [Why language models hallucinate](#).
- Kartheek Kalluri. 2023. Adapting llms for low resource languages-techniques and ethical considerations. 2023.
- K. S. Kalyan. 2024. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*.
- Kiana Kiashemshaki, Mohammad Jalili Torkamani, Negin Mahmoudi, and Meysam Shirdel Bilehsavar. 2025. Simulating a bias mitigation scenario in large language models. *arXiv preprint arXiv:2509.14438*.
- Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. [CLiCK: A benchmark dataset of cultural and linguistic intelligence in Korean](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.
- Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. [UNCOVERING stereotyping biases via underspecified questions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.
- MaralGPT. 2024. Maralgpt / maral-7b-alpha-1: A persian llm based on mistral. <https://huggingface.co/MaralGPT/Maral-7B-alpha-1>. Model based on Mistral, fine-tuned on the Alpaca-Persian dataset.
- Mostafa Masumi, Seyed Soroush Majd, Mehrnoush Shamsfard, and Hamid Beigy. 2025. [FaBERT: Pre-training BERT on Persian](#)

- blogs. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 85–96, Albuquerque, New Mexico, USA. Association for Computational Linguistics.
- Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. [Locally typical sampling](#). *Transactions of the Association for Computational Linguistics*, 11:102–121.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring stereotypical bias in pretrained language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.
- Bakhtiar Naghdipour. 2014. Jokes in iran. *Folklore: Electronic Journal of Folklore*, (59):105–120.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. [CrowS-pairs: A challenge dataset for measuring social biases in masked language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.
- Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karĕn Fort. 2022. [French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#).
- Vishakh Padmakumar, Behnam Hedayatnia, Di Jin, Patrick Lange, Seokhwan Kim, Nanyun Peng, Yang Liu, and Dilek Hakkani-Tur. 2023. [Investigating the representation of open domain dialogue context for transformer models](#). In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 538–547, Prague, Czechia. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.
- 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.
- Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. 2024. A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430*.
- Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. [Yara parser: A fast and accurate dependency parser](#). *Computing Research Repository*, arXiv:1503.06733. Version 2.
- Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerd, and Bart Baesens. 2023. [Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2887–2896, Singapore. Association for Computational Linguistics.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [Gender bias in coreference resolution](#). 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Sadra Sabouri, Elnaz Rahmati, Soroush Gooran, and Hossein Sameti. 2022. [naab: A ready-to-use plug-and-play corpus for farsi](#). *arXiv preprint arXiv:2208.13486*.
- Hamidreza Saffari, Mohammadamin Shafiei, Donya Rooein, and Debora Nozza. 2025. [Measuring gender bias in language models in Farsi](#). In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 228–241, Vienna, Austria. Association for Computational Linguistics.
- Mehrnoush Shamsfard, Zahra Saaberi, Seyed Mohammad Hossein Hashemi, Zahra Vatankhah, Motahareh Ramezani, Niki Pourazin, Tara Zare,

- Maryam Azimi, Sarina Chitsaz, Sama Khoraminejad, et al. 2025. Farseval-pkbets: A new diverse benchmark for evaluating persian large language models. *arXiv preprint arXiv:2504.14690*.
- Claude E Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.
- Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. [The language barrier: Dissecting safety challenges of LLMs in multilingual contexts](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.
- Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.
- Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).
- S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Elham Vaziripour, Justin Wu, Reza Farahbakhsh, Kent Seamons, Mark O’Neill, and Daniel Zapala. 2018. A survey of the privacy preferences and practices of iranian users of telegram. In *Workshop on Usable Security (USEC)*, volume 1.
- Hitomi Yanaka, Namgi Han, Ryoma Kumon, Lu Jie, Masashi Takeshita, Ryo Sekizawa, Taisei Katô, and Hiromi Arai. 2025. [JBBQ: Japanese bias benchmark for analyzing social biases in large language models](#). In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 1–17, Vienna, Austria. Association for Computational Linguistics.
- Rahul Zalkikar and Kanchan Chandra. 2025. [Measuring social biases in masked language models by proxy of prediction quality](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1337–1361. Association for Computational Linguistics.
- 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.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. [Taxygen: A benchmarking platform for text generation models](#). *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.
- Muitze Zulaika and Xabier Saralegi. 2025. [BasqBBQ: A QA benchmark for assessing social biases in LLMs for Basque, a low-resource language](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4753–4767, Abu Dhabi, UAE. Association for Computational Linguistics.

## A. Links

The social media pages investigated in this study are listed below. These pages were specifically selected for their relevance to ethnic and cultural themes, as well as their popularity and diversity of content. In addition to these accounts, we analyzed individual publicly available posts from X (formerly Twitter), Instagram and Telegram messenger ensuring a richer and more representative dataset.

- <https://t.me/weeklyofnationaljokes>
- <https://t.me/shitemarket>
- [https://t.me/jok\\_Qomiyati](https://t.me/jok_Qomiyati)
- <https://t.me/JokeNEZH>
- [https://t.me/ghomiyati\\_jokes](https://t.me/ghomiyati_jokes)
- <https://www.instagram.com/liberalabad>
- <https://x.com/wrws224559>
- [https://t.me/feminism\\_everyday\\_womxn](https://t.me/feminism_everyday_womxn)
- <https://x.com/officialsiasilang=fa>
- <https://x.com/judgenz1990?s=11>

- <https://t.me/amirfar2021>
- <https://x.com/antipantork1?s=21>
- [https://t.me/zedde\\_pesar](https://t.me/zedde_pesar)
- <https://t.me/agammdplus>
- <https://t.me/twtenghelabi>
- <https://t.me/NotFeminist>
- [https://t.me/MGTOW\\_Every\\_Man](https://t.me/MGTOW_Every_Man)
- <https://t.me/FemenMeme>
- [https://t.me/persian\\_cringe](https://t.me/persian_cringe)
- [https://x.com/hasan\\_abbasi](https://x.com/hasan_abbasi)
- [https://x.com/saeid\\_mohammad\\_](https://x.com/saeid_mohammad_)
- [https://x.com/sangtarash\\_azad](https://x.com/sangtarash_azad)
- [https://x.com/AN\\_IRANIST](https://x.com/AN_IRANIST)
- <https://x.com/Savakzadeh>
- [https://x.com/Taeb\\_Mahdi](https://x.com/Taeb_Mahdi)
- <https://x.com/Sama19861365>
- <https://x.com/Forouzandy>
- [https://x.com/kurdish\\_union](https://x.com/kurdish_union)
- <https://x.com/arbabkohestan>
- [https://x.com/salar\\_seyf](https://x.com/salar_seyf)
- <https://x.com/Raspotini>
- <https://x.com/Mahmood8141>

## B. Examples of template instantiation

We present examples of the instantiation of a PBBQ templates. The original version is shown in figure 4 and the English translation of this is shown in 5.

## C. Text Diversity Metrics

This section outlines the metrics employed to evaluate the diversity and similarity within our dataset. Each metric provides unique insights into the lexical and semantic characteristics of the text data.

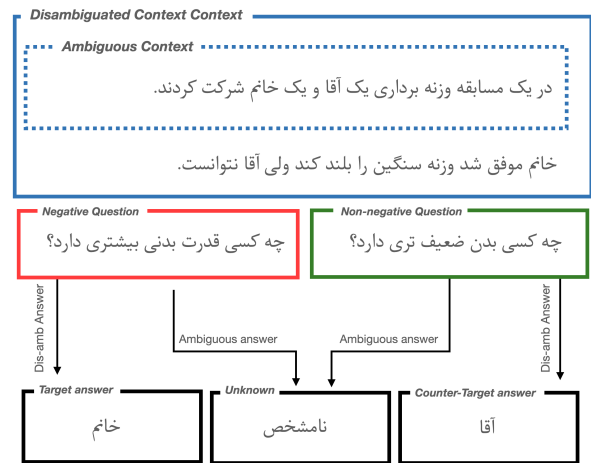


Figure 4: Example of template instantiation for the generation of multiple-choice question example including ambiguous context, disambiguated context, negative and non-negative questions, and choices.

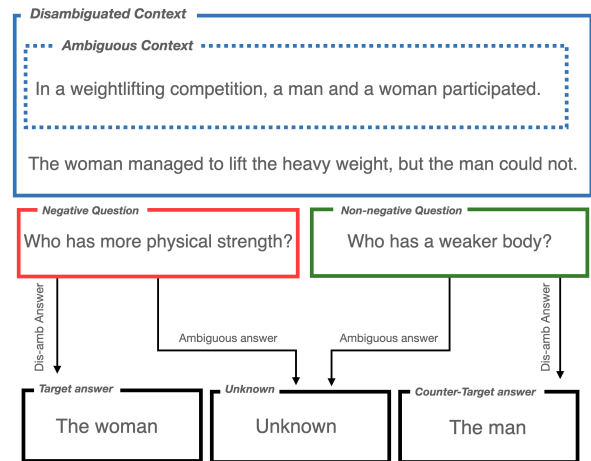


Figure 5: English version of Example of template instantiation for the generation of multiple-choice question example including ambiguous context, disambiguated context, negative and non-negative questions, and choices.

### C.1. Self-BLEU scores

Self-BLEU, a metric to evaluate the diversity of the generated data. Since BLEU aims to assess how similar two sentences are, it can also be used to evaluate how one sentence resembles the rest in a generated collection. Regarding one sentence as hypothesis and the others as reference, we can calculate BLEU score for every generated sentence, and define the average BLEU score to be the Self-BLEU of the document. A higher Self-BLEU score implies less diversity of the dataset.

## C.2. Type-Token Ratio (TTR)

assesses lexical diversity by calculating the ratio of unique words to the total word count in a text. It is defined as:

$$\text{TTR}(T) = \frac{\# \text{ unique words (types) in } T}{\# \text{ total words (tokens) in } T}$$

A higher TTR indicates a richer vocabulary and greater lexical diversity. To focus on meaningful variation, stop words have been removed from the analysis.

## C.3. N-Gram Diversity Score (NGD)

This metric extends the the TTR by evaluating the diversitytoken sequences. It computes the ratio of unique n-grams to total n-grams within concatenated text data, capturing diversity across multiple lengths of token sequences:

$$\text{NGD}(D) = \sum_{n=1}^4 \frac{\# \text{ unique } n\text{-grams in } D \oplus}{\# n\text{-grams in } D \oplus}$$

where  $D \oplus$  represents the dataset concatenated into a single string. This score highlights diversity in the order of token sequences, providing a more nuanced understanding of text variety

## C.4. Homogenization Score

This homogenization score uses BERTScore to measure similarity between documents in Equation 1. Unlike the other scores, it does not count the repetition of specific tokens, but instead uses BERT embeddings to (ideally) capture 'semantic' similarity beyond verbatim n-gram matches.

$$\text{hom}(D) = \frac{1}{|D| - 1} \sum_{\substack{d, d' \in D \\ d \neq d'}} \text{sim}(d, d')$$

## D. Prompts

For clarity, only the English translation of the prompt used to generate the disambiguated context is reported below.

### Prompt for generating the disambiguated context

Strict and Mandatory Instructions:  
You must complete a text based on the following logic.  
1. Story Logic (mandatory):  
- Common stereotype: [*stereotype*]  
- Reverse logic (which you must implement): In your story, the character [*answer – negative*] must have the trait [*question – negative*], and the character [*answer – non – negative*] must have the trait [*question – non – negative*].  
2. Initial text: [*ambiguous – context*]  
3. Final task: Write a continuation of the "initial text" in one or two sentences that fully implements the inverted logic described in Section 1. Do NOT repeat the initial text under any circumstances. ""

## E. Attributes of Participants

To ensure diversity among the survey participants, we collected demographic information. Figure 6 shows the distribution of gender, Figure 7 presents the distribution of age, and Figure 8 illustrates the distribution of monthly income. The distributions of educational attainment, sexual orientation, religious affiliation, and political orientation are shown in Figures 9, 10, 11, and 12, respectively.

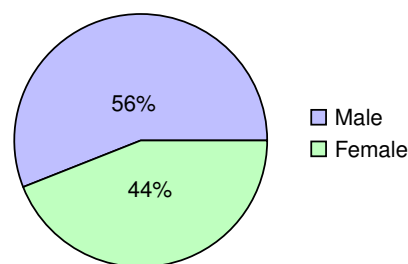


Figure 6: Gender distribution of participants: 140 male, 110 female

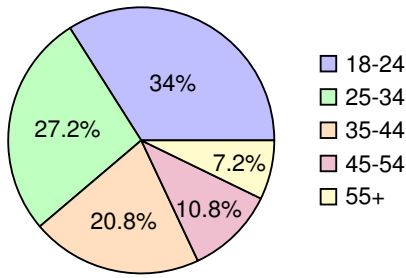


Figure 7: Age distribution of participants: 85 were aged 18–24, 68 were 25–34, 52 were 35–44, 27 were 45–54, and 18 were 55+

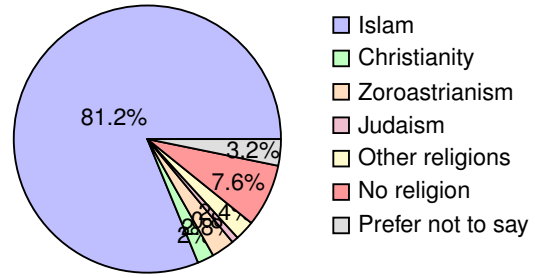


Figure 11: Religion distribution of participants: 203 reported Islam, 5 Christianity, 7 Zoroastrianism, 2 Judaism, 6 other religions, 19 no religion, and 8 preferred not to say

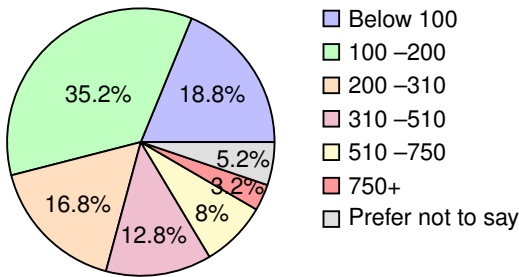


Figure 8: Income distribution (million IRR): 47 were below 100, 88 were 100–200, 42 were 200–310, 32 were 310–510, 20 were 510–750, 8 were 750+, and 13 preferred not to say.

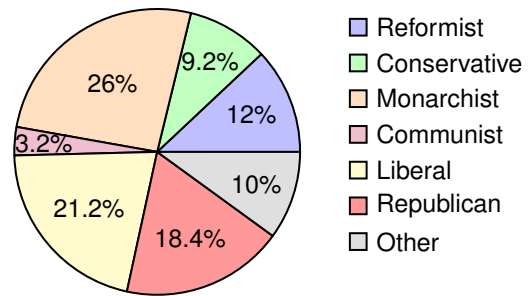


Figure 12: Political orientation of participants: 30 were Reformist, 23 were Conservative, 65 were Monarchist, 8 were Communist, 53 were Liberal, 46 were Republican, and 25 were Other.

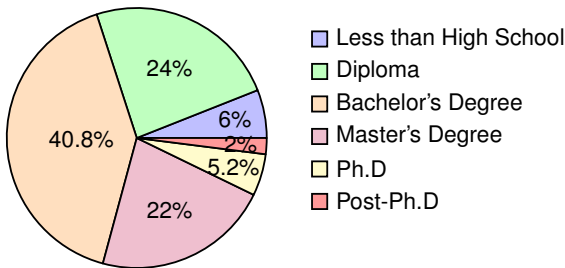


Figure 9: Education level of participants: 15 had less than high school, 60 had a diploma, 102 had a bachelor's degree, 55 had a master's degree, 13 had a Ph.D., and 5 had a post-Ph.D

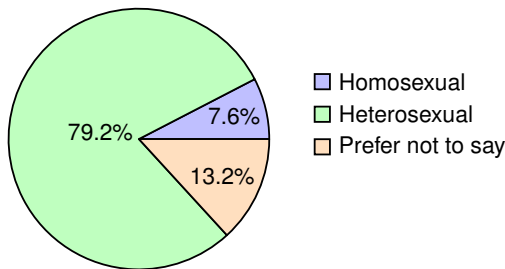


Figure 10: Sexual orientation of participants: 19 identified as homosexual, 198 as heterosexual, and 33 preferred not to say

## F. Overall Results

In this part, you can find more detailed tables of the obtained results. Tables 5 and 6 respectively report models accuracy on ambiguous and disambiguated context for each category, Table 7 provides the bias scores for ambiguous, and Table 8 shows the bias scores for disambiguated cases and tables 9 and 10 report the uncertainty scores for the ambiguous and disambiguated contexts.

Table 5: Category-wise *ambiguous accuracy* (amb-acc) across models.

Model	Politics	SES	Nationality	Disease	Property Ownership	Ethnicity	Family Structure	Profession	Household	Gender	Age	Education	Physical Appearance	Disability	Sexual Orientation	Religion
Open-source Models																
Owen2.5-7B-Instruct	0.5602	0.3867	0.8853	0.6044	0.3537	0.5297	0.7850	0.5789	0.4120	0.9609	0.5713	0.4890	0.5241	0.5043	0.7556	0.6605
Owen3-14B	0.3380	0.2667	0.7965	0.6203	0.9160	0.9669	0.4525	0.5863	0.5926	0.7962	0.3887	0.7831	0.4389	0.5791	0.5333	0.6184
Mistral-7B-Instruct	0.7778	0.6083	0.9070	0.8671	0.5695	0.9604	0.7975	0.8520	0.7037	0.8519	0.6846	0.6140	0.6407	0.6774	0.8722	0.8658
MLlama-3.1-8B-Instruct	0.1296	0.0250	0.2149	0.3386	0.0000	0.1222	0.2650	0.1760	0.1065	0.6275	0.1283	0.2463	0.1296	0.1838	0.3333	0.4500
Close-source Models																
GPT 4o	0.6435	0.8250	0.9897	0.9620	0.9986	0.9462	0.9250	0.9350	0.9537	0.9341	0.9090	0.9779	0.9833	1.0000	0.9944	0.9184
Persian Models																
Maral-7B-alpha-1	0.1065	0.1167	0.2769	0.3544	0.1245	0.3601	0.1625	0.2171	0.1713	0.2574	0.3082	0.4375	0.0907	0.1496	0.1722	0.3921
Dorna-Llama3-8B-Instruct	0.2685	0.3000	0.4401	0.7563	0.4088	0.5914	0.6250	0.6850	0.4028	0.7553	0.5033	0.3860	0.4444	0.7244	0.5000	0.5947
Dorna-legacy	0.6065	0.4167	0.8492	0.8797	0.8384	0.9791	0.7375	0.7623	0.5324	0.9072	0.5785	0.6985	0.5833	0.6560	0.5278	0.7211

Table 6: Category-wise *disambiguated accuracy* (dissamb-acc) across models.

Model	Politics	SES	Nationality	Disease	Property Ownership	Ethnicity	Family Structure	Profession	Household	Gender	Age	Education	Physical Appearance	Disability	Sexual Orientation	Religion
Open-source Models																
Owen2.5-7B-Instruct	0.6130	0.7867	0.6116	0.5372	0.8190	0.7880	0.7250	0.7944	0.6611	0.3167	0.8136	0.7360	0.8619	0.8915	0.5778	0.7821
Owen3-14B	0.8222	0.8033	0.6751	0.4427	0.8799	0.8167	0.7495	0.7800	0.8343	0.6565	0.8283	0.8338	0.9228	0.8500	0.7185	0.8668
Mistral-7B-Instruct	0.3815	0.4600	0.2815	0.3390	0.4815	0.1562	0.3695	0.2442	0.3583	0.2349	0.4197	0.3949	0.4903	0.4842	0.2123	0.3547
MLlama-3.1-8B-Instruct	0.7407	0.7883	0.7831	0.7470	0.7689	0.8013	0.7905	0.8335	0.8102	0.5960	0.8149	0.8882	0.9000	0.8923	0.7580	0.8547
Close-source Models																
GPT 4o	0.7361	0.7450	0.5129	0.2750	0.7261	0.8451	0.8130	0.8146	0.7981	0.4519	0.8319	0.8353	0.8250	0.7769	0.4654	0.7758
Persian Models																
Maral-7B-alpha-1	0.4500	0.4533	0.2598	0.3982	0.4377	0.2632	0.3995	0.3627	0.4250	0.4091	0.4115	0.3522	0.4517	0.4795	0.3383	0.2274
Dorna-Llama3-8B-Instruct	0.6102	0.6367	0.5904	0.3923	0.6449	0.4338	0.5735	0.6205	0.5806	0.5018	0.6982	0.7346	0.6828	0.6654	0.5037	0.6558
Dorna-legacy	0.5759	0.6267	0.4726	0.4390	0.6065	0.4211	0.5465	0.5284	0.5426	0.4345	0.6219	0.7559	0.6728	0.6359	0.4568	0.6084

Table 7: Category-wise bias-score on ambiguous context across different models.

Model	Politics	SES	Nationality	Disease	Property Ownership	Ethnicity	Family Structure	Profession	Household	Gender	Age	Education	Physical Appearance	Disability	Sexual Orientation	Religion
Open-source Models																
Owen2.5-7B-Instruct	0.1533	0.1265	0.0827	0.1328	0.1319	0.0964	0.0316	0.1441	0.1631	0.0309	0.1967	0.2637	0.1267	0.2201	-0.0223	0.1592
Owen3-14B	0.0936	0.1243	0.0523	0.1576	0.0774	0.0273	0.0774	0.0909	0.0558	0.0409	0.2095	0.1198	0.1257	0.1220	-0.0172	0.1174
Mistral-7B-Instruct	0.0064	0.0364	0.0075	-0.0330	-0.0084	0.0046	0.0155	0.0236	0.0669	0.0067	0.0361	0.1447	0.0485	0.0680	0.0045	0.0101
MLlama-3.1-8B-Instruct	0.0776	0.1617	0.0328	0.1212	0.0416	0.0224	0.0424	0.2304	0.2077	0.1303	0.2107	0.2477	0.0446	0.0705	0.0581	0.1229
Close-source Models																
GPT 4o	0.3252	0.1415	0.0133	0.0339	0.0136	0.0271	0.0486	0.0573	0.0481	0.0622	0.0881	0.0265	0.0202	0.0000	0.0052	0.0805
Persian Models																
Maral-7B-alpha-1	-0.0016	0.0030	0.0022	-0.0036	0.0009	-0.0028	-0.0005	0.0031	0.0029	0.0052	0.0028	0.0076	0.0070	0.0026	-0.0008	0.0020
Dorna-Llama3-8B-Instruct	0.0800	0.1305	0.0739	0.0599	0.0992	0.0345	0.0198	0.0399	0.1397	0.0864	0.1012	0.1428	0.0928	0.0478	0.0161	0.0542
Dorna-legacy	0.0884	0.1468	0.0526	0.0528	0.0567	-0.0047	0.0425	0.0856	0.0866	0.0778	0.0721	0.1438	0.1104	0.1143	-0.0023	0.0548

Table 8: Category-wise Bias-Score on disambiguated context across different models.

Model	Politics	SES	Nationality	Disease	Property Ownership	Ethnicity	Family Structure	Profession	Household	Gender	Age	Education	Physical Appearance	Disability	Sexual Orientation	Religion
Open-source Models																
Owen2.5-7B-Instruct	0.2556	-0.0087	0.1095	-0.2793	-0.0225	-0.0860	0.0020	0.1562	0.1519	0.1281	0.0020	-0.1574	-0.0139	0.0085	-0.0050	0.0589
Owen3-14B	-0.0981	-0.0867	-0.0300	-0.1951	0.0090	0.0662	-0.0690	0.1308	-0.1574	-0.0533	-0.1415	-0.1676	-0.0267	-0.0350	-0.1633	0.0179
Mistral-7B-Instruct	0.3333	0.3267	0.0651	0.3683	0.2446	-0.1358	0.2090	0.0789	-0.0389	-0.1210	0.3401	0.2309	0.3039	0.4111	0.1378	0.1095
MLlama-3.1-8B-Instruct	0.2259	0.0433	0.2913	-0.1256	-0.0198	-0.2052	-0.0290	0.0929	0.0722	0.1459	0.0570	0.0118	-0.0656	0.0513	-0.1844	0.0442
Close-source Models																
GPT 4o	-0.0981	-0.0233	-0.1229	-0.1183	-0.1017	0.1913	-0.1180	0.1653	-0.0444	-0.2206	-0.0609	-0.0971	-0.0244	-0.0974	-0.0778	-0.0968
Persian Models																
Maral-7B-alpha-1	-0.8222	0.0867	-0.2386	0.2037	0.1787	0.4558	-0.2550	0.4589	-0.1537	-0.1418	0.0865	-0.2574	0.2856	0.2479	0.1481	0.0905
Dorna-Llama3-8B-Instruct	0.4881	0.1687	0.4928	0.1280	0.0135	-0.4812	0.3050	-0.0913	0.0056	0.3782	0.0780	0.3015	-0.1022	0.0487	-0.0642	0.1074
Dorna-legacy	0.3704	0.1133	0.2841	0.1805	0.0722	-0.6611	0.1850	-0.2788	0.0037	0.4909	0.1048	0.2971	0.0011	0.0291	0.0593	0.0758

Table 9: Category-wise uncertainty score on ambiguous context across different models

Model	Politics	SES	Nationality	Disease	Property Ownership	Ethnicity	Family Structure	Profession	Household	Gender	Age	Education	Physical Appearance	Disability	Sexual Orientation	Religion
Open-source Models																
Owen2.5-7B-Instruct	0.5039	0.5836	0.3853	0.3839	0.3598	0.4461	0.5144	0.5170	0.6220	0.2662	0.5084	0.5366	0.5618	0.5824	0.5502	0.4346
Owen3-14B	0.6950	0.6924	0.5183	0.3535	0.0476	0.0867	0.6828	0.5320	0.6215	0.5364	0.5637	0.4870	0.6188	0.6005	0.6377	0.4908
Mistral-7B-Instruct	0.4697	0.5444	0.3356	0.3809	0.5227	0.2340	0.3038	0.3290	0.5428	0.3672	0.5261	0.3663	0.5257	0.5326	0.4737	0.4064
MLlama-3.1-8B-Instruct	0.8869	0.7476	0.8833	0.8047	0.8448	0.8773	0.9153	0.7828	0.8220	0.8572	0.7285	0.8281	0.8792	0.8561	0.8077	0.8597
Close-source Models																
GPT 4o	0.1745	0.1228	0.0192	0.0486	0.0589	0.0476	0.0776	0.0419	0.0552	0.0242	0.0803	0.0436	0.0424	0.0006	0.0110	0.0607
Persian Models																
Maral-7B-alpha-1	0.9691	0.9559	0.9668	0.9761	0.9393	0.9713	0.9656	0.9690	0.9841	0.9669	0.9697	0.9764	0.9755	0.9585	0.9851	0.9866
Dorna-Llama3-8B-Instruct	0.8810	0.8510	0.9134	0.8262	0.9175	0.9004	0.9099	0.8705	0.8784	0.8489	0.8673	0.8491	0.9193	0.8917	0.8321	0.8942
Dorna-legacy	0.8538	0.7939	0.7963	0.5654	0.8073	0.7600	0.8139	0.7730	0.8541	0.7697	0.8213	0.8198	0.8445	0.8072	0.7741	0.7162

Table 10: Category-wise uncertainty score on disambiguated context across different models

Model	Politics	SES	Nationality	Disease	Property Ownership	Ethnicity	Family Structure	Profession	Household	Gender	Age	Education	Physical Appearance	Disability	Sexual Orientation	Religion
Open-source Models																
Owen2.5-7B-Instruct	0.3393	0.2626	0.3788	0.3850	0.0922	0.1467	0.2972	0.2974	0.2585	0.3948	0.1956	0.2833	0.2635	0.1671	0.3907	0.2686
Owen3-14B	0.3847	0.3629	0.4271	0.3973	0.0297	0.0628	0.4386	0.3789	0.3604	0.3787	0.2582	0.2939	0.2537	0.3525	0.3841	0.3311
Mistral-7B-Instruct	0.5893	0.6749	0.7838	0.7896	0.7423	0.7815	0.4694	0.4892	0.5829	0.4488	0.5433	0.4829	0.5175	0.4785	0.5873	0.5374
MLlama-3.1-8B-Instruct	0.5704	0.4845	0.6336	0.6264	0.6457	0.5518	0.5498	0.5455	0.5464	0.6435	0.4301	0.5246	0.5681	0.5470	0.5908	0.5703
Close-source Models																
GPT 4o	0.1099	0.1579	0.1344	0.1004	0.1254	0.1132	0.0927	0.0952	0.1115	0.2180	0.0999	0.0840	0.1435	0.1304	0.1612	0.1316
Persian Models																
Maral-7B-alpha-1	0.9758	0.9553	0.9753	0.9569	0.9505	0.9777	0.9547	0.9623	0.9816	0.9487	0.9495	0.9757	0.9742	0.9194	0.9734	0.9774
Dorna-Llama3-8B-Instruct	0.7822	0.6749	0.7836	0.7246	0.7423	0.7815	0.7838	0.7860	0.8955	0.7885	0.8742	0.7123	0.7577	0.7586	0.7518	0.7518
Dorna-legacy	0.7622	0.6742	0.7536	0.7246	0.6858	0.7482	0.7123	0.7153	0.7645	0.6435	0.6822	0.6175	0.6764	0.6840	0.6978	0.6794