The CSI Journal on Computer Science and Engineering

Journal homepage: www.csionjcse.ir

Counterfactual Social Bias Evaluation in Persian LLM-Based Information Systems

Niloofar Ranjbar[1*]

[1] Persian Gulf University, Bushehr, Iran

*** Corresponding author email address**: nranjbar@pgu.ac.ir

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Ranjbar, N. (2026). Counterfactual Social Bias Evaluation in Persian LLM-Based Information Systems. *The CSI Journal on Computer Science and Engineering, 20*(2), 152-164.
http://dx.doi.org/10.22034/jcse.2026.589496.1086

A B S T R A C T

Abstract: Large language models are increasingly used in Persian information systems, yet culturally grounded bias evaluation remains limited. We introduce a natural-scenario counterfactual benchmark constructed through human-directed, ChatGPT-assisted drafting and independently validated by three native Persian speakers. From an initial 100 scenarios, 524 rows, and 424 pairs, validation retained 84 scenarios, 424 rows, and 340 pairs; socioeconomic coverage was reduced to one pair. Five open-weight models were evaluated with deterministic restricted next-token A–E scoring, normal/reverse option orders, scenario-cluster bootstrap intervals, and scenario-level sign-flip tests. Baseline mean absolute gaps were small (0.031–0.043), but Dorna-Llama3-8B and Llama-3.1-8B showed severe option-order sensitivity. Fairness prompting increased gaps for every model and significantly for Qwen2-7B, Qwen2.5-14B, and Qwen3-8B. Qwen2.5-14B and Qwen3-8B achieved the highest BBQ-Fa accuracy (0.817), while Qwen2.5-14B led ISEAR-Fa (0.623). Train-only calibration did not improve held-out gaps. The benchmark, validation documentation, prompts, code, and results are provided in an accompanying anonymous GitHub repository.

Keywords: *large language models; intelligent information systems; Persian NLP; social bias; counterfactual evaluation; fairness*

Article history:
Published online 07 August 2026

1 Introduction

Large language models (LLMs) are increasingly embedded in intelligent information systems rather than used only as stand-alone text generators. They support search and retrieval, summarisation, customer-service automation, educational feedback, public-service interaction, health triage, and decision support. In such systems, biased model behaviour is not only a language-model property; it becomes a system-level risk. A model that changes its judgement when a tenant, patient, student, worker, applicant, or customer is described with a different identity cue may affect the quality of returned information, the tone of a service response, the priority of a request, or the perceived suitability of a person in a scenario.

This risk is especially important for Persian-language systems. Many established bias benchmarks were developed for English and encode social categories, institutions, and stereotypes that do not transfer directly to Persian. Translation alone is therefore insufficient: identity markers, socioeconomic cues, ethnic and national categories, family and service contexts, and institutional settings must be evaluated in culturally and linguistically appropriate forms. Persian also remains less represented than English in mainstream LLM evaluation, which makes it difficult for practitioners to decide whether a Persian-specialised or multilingual model is safer for deployment in information systems.

Prior work has shown that language models can encode social associations in word embeddings, sentence representations, coreference decisions, question answering, and open-ended generation [1-8]. However, intelligent information systems require a more deployment-oriented question: does a model change its judgement when only identity information changes while the decision-relevant scenario remains fixed? This paper studies that question through controlled counterfactual evaluation. Counterfactual sensitivity is defined as the change in a model's score between an original Persian scenario and an identity-modified version of the same scenario. This measure does not claim to capture all forms of social harm, but it provides a controlled way to detect identity-linked variation in model judgement.

Persian and Farsi bias, fairness, and trustworthiness evaluation has recently advanced through several important resources. GBFA provides Farsi gender-bias probes based on BBQ-Fa, HONEST-Fa, and ISEAR-Fa [9]. PBBQ introduces a large Persian social-bias question-answering benchmark with culturally grounded categories [10]. ELAB offers a broader Persian alignment benchmark containing fairness, social-norm, and safety-related components [11]. EPT evaluates Persian LLM trustworthiness across dimensions including fairness, safety, privacy, robustness, ethics, and truthfulness [12]. Kalhor and Bahrak study Persian gender stereotypes in multilingual LLMs and introduce a domain-specific gender-skew measure [13]. These resources show that Persian bias evaluation is an active area, not an empty research space. The present work therefore does not claim to be the first Persian bias benchmark. Instead, it contributes a complementary evaluation setting: natural Persian information-system scenarios with controlled multi-axis counterfactual identity substitutions and ordinal judgement scoring.

The study addresses four research questions. RQ1 asks how strongly Persian and multilingual open-weight LLMs respond to controlled identity substitutions in independently validated Persian scenarios. RQ2 asks how sensitivity varies across the identity axes that retain sufficient validated coverage. RQ3 asks how the same models behave on the released native GBFA BBQ-Fa and ISEAR-Fa tasks. RQ4 asks whether fairness prompting and train-only post-hoc calibration reduce counterfactual gaps without introducing new instability.

The paper makes four contributions. First, it develops a human-directed, ChatGPT-assisted Persian scenario benchmark and reports a complete independent validation study rather than relying on author screening alone. The candidate resource contained 100 scenarios, 524 rows, and 424 pairs; after three-annotator validation and adjudication, the released benchmark contains 84 scenarios, 424 rows, and 340 pairs. Second, it provides a corrected and fully specified deterministic inference protocol for five open-weight models, including exact Hugging Face checkpoint identifiers, prompts, token-validation rules, hardware, software, and random seeds. Third, it evaluates the models on the validated counterfactual benchmark and on native GBFA BBQ-Fa and ISEAR-Fa tasks while reporting option-order and response-collapse diagnostics. Fourth, it evaluates two lightweight controls and shows that neither should be assumed to improve fairness: the tested fairness prompt increases gaps for several models, and phrase-offset calibration does not improve held-out gaps.

2 Related work

2.1 Bias evaluation in language models

Early computational bias research showed that language representations can encode social associations learned from text corpora. The Word Embedding Association Test (WEAT) demonstrated that word embeddings reproduce

associations between social groups and attributes, while the Sentence Encoder Association Test (SEAT) extended this approach to sentence-level representations [1, 3]. These representation-level tests established the importance of measuring bias in language models, but they do not directly show how a model behaves in information-system tasks such as answering a service request, evaluating a user scenario, or producing a decision-support judgement.

Later work moved from representation-level association tests to task-level and counterfactual evaluation. CrowS-Pairs uses minimally different sentence pairs to test whether language models prefer stereotypical over anti-stereotypical statements [5]. StereoSet evaluates whether models choose stereotypical, anti-stereotypical, or unrelated continuations while also measuring language-modeling ability [4]. WinoBias and WinoGender-style datasets show that gender and occupation cues can influence coreference decisions [7, 8]. These works motivate the central design principle of the present study: the most informative bias test often compares matched examples that differ only in an identity-related cue.

Question-answering and generation benchmarks add further task-specific structure. BBQ evaluates bias in ambiguous and disambiguated question-answering contexts [6]. UNQOVER studies underspecified questions where a fair system should avoid unsupported assumptions [14]. BOLD and HolisticBias broaden evaluation toward open-ended generation and wider demographic coverage [2, 15]. HONEST-style benchmarks focus on hurtful sentence completions [16]. Together, these studies show that bias is not a single observable property. QA accuracy, stereotyped answer selection, hurtful generation, emotion attribution, and ordinal scenario judgement capture different forms of model behaviour. For this reason, the present paper keeps task-specific metrics separate rather than combining heterogeneous probes into one pooled bias score.

2.2 Persian and Farsi bias evaluation

Recent Persian and Farsi work has begun to address the lack of culturally grounded LLM-bias evaluation. The closest existing resource to the external-probe component of this paper is GBFA, introduced by Saffari et al. as a framework for measuring gender bias in Farsi language models [9]. GBFA adapts three established English bias-evaluation paradigms: ISEAR for gendered emotion attribution, BBQ for question-answering bias, and HONEST for hurtful sentence completion. The released datasets contain Farsi ISEAR, BBQ-Fa, and HONEST-Fa components, and the authors report that gender-bias patterns in Farsi differ from those observed in English. GBFA is therefore directly relevant to this study and is used here as a native external probe. However, GBFA is primarily gender-focused and organised around traditional NLP task formats, whereas the proposed benchmark evaluates multiple identity axes in natural information-system scenarios.

PBBQ is the largest directly related Persian social-bias benchmark [10]. It contains more than 37,000 Persian questions across 16 culturally grounded categories and was developed through human–AI collaboration, questionnaire input from 250 participants, and social-science expert involvement. PBBQ is important because it demonstrates that Persian bias evaluation must be grounded in local cultural categories rather than translated mechanically from English. The present benchmark is much smaller than PBBQ, but it differs in purpose and format. PBBQ is primarily a large-scale QA benchmark; the proposed benchmark is a controlled natural-scenario counterfactual benchmark that measures whether an LLM changes an ordinal judgement when only the identity cue changes.

ELAB provides a broader Persian alignment benchmark that includes safety, fairness, and social-norm dimensions [11]. It combines translated data, synthetically generated data, and naturally collected Persian data, and introduces resources such as FairBench-fa and SocialBench-fa. ELAB is valuable because it places fairness within a larger responsible-AI framework for Persian LLMs. In contrast, the present work focuses more narrowly on identity-linked counterfactual sensitivity in information-system scenarios. Thus, ELAB and the proposed benchmark are complementary: ELAB evaluates broad alignment behaviour, while this work provides a controlled scenario-level test of identity-conditioned judgement variation.

The EPT benchmark extends Persian evaluation to trustworthiness [12]. It covers truthfulness, safety, fairness, robustness, privacy, and ethical alignment, and evaluates a broad panel of commercial and open models. EPT is relevant because it treats fairness as one component of trustworthy Persian LLM behaviour rather than as an isolated NLP task. This work follows the same general motivation but examines fairness at a finer counterfactual level. Instead of asking whether a model is broadly trustworthy across dimensions, the analysis asks whether its score changes when a person in a Persian information-system scenario is described with a different identity cue.

A more targeted Persian gender-bias study is provided by Kalhor and Bahrak [13]. Their work probes gender stereotypes in multilingual LLMs using Persian templates and introduces a Domain-Specific Gender Skew Index to

quantify deviations from gender parity across semantic domains. They show that gender stereotypes can be stronger in Persian than in English for the evaluated models and domains. This finding supports one of the assumptions behind this study: multilingual model behaviour in Persian cannot be inferred from English-only evaluations. The proposed benchmark extends this line of work beyond gender templates by testing six identity axes and by embedding identity cues in everyday information-system scenarios.

2.3 *Positioning of the present benchmark*

The Persian/Farsi literature therefore already contains several important bias and alignment resources. GBFA provides gender-focused Farsi probes; PBBQ provides large-scale culturally grounded Persian QA; ELAB evaluates broad Persian alignment; EPT evaluates Persian trustworthiness; and Kalhor and Bahrak provide a focused analysis of Persian gender stereotypes in multilingual LLMs. The present paper does not replace these resources and does not claim that Persian bias evaluation is absent. Its contribution is to add a complementary evaluation format: a natural-scenario, multi-axis, counterfactual benchmark for Persian intelligent information systems.

This distinction matters for deployment. A search assistant, public-service chatbot, educational tutor, health-triage assistant, or customer-service system may not only answer bias-test questions; it may also evaluate descriptions of people, requests, complaints, or service situations. In such settings, counterfactual sensitivity is operationally meaningful: if the same scenario receives a different score after only an identity phrase changes, the system may require further auditing before deployment. The proposed benchmark is designed to make this form of sensitivity measurable. Although GBFA, PBBQ, ELAB, and EPT provide important Persian/Farsi bias, alignment, and trustworthiness resources, none is designed around the same combination of natural information-system scenarios, six identity axes, original–counterfactual identity-pair structure, and ordinal judgement scoring used here. Table 1 summarises the comparison between the proposed benchmark and the closest Persian/Farsi bias, alignment, and trustworthiness resources.

Table 1: Positioning of the proposed benchmark relative to closely related Persian/Farsi resources.

Resource	Main task format	Identity/fairness coverage	Main contribution	Relation to this paper
GBFA [9]	BBQ-Fa QA; HONEST-Fa cloze; ISEAR-Fa emotion attribution	Gender-focused	First comprehensive Farsi gender-bias evaluation framework across three tasks	Used as native external probes; narrower identity coverage than the proposed benchmark
PBBQ [10]	Persian social-bias QA	16 culturally grounded categories	Large-scale Persian bias benchmark with human and expert input	Larger scale; different QA-oriented format
ELAB [11]	Persian alignment benchmark	Safety, fairness, and social norms	Broad culturally adapted alignment evaluation suite	Broader alignment scope; not primarily counterfactual scenario scoring
EPT [12]	Persian trustworthiness benchmark	Truthfulness, safety, fairness, robustness, privacy, ethics	Trustworthiness evaluation across several Persian ethical-cultural dimensions	Related trustworthiness framing; less focused on identity-pair gaps
Kalhor and Bahrak [13]	Template-based stereotype probing	Gender stereotypes across domains	Persian gender-skew analysis with DS-GSI	Related Persian stereotype analysis; narrower than the proposed six-axis scenario design
Proposed benchmark	A–E ordinal judgement on natural Persian scenarios	Gender, nationality, ethnicity, socioeconomic status, urban/rural background, age	Controlled multi-axis counterfactual sensitivity in information-system scenarios	Smaller scale, but directly targets identity-linked judgement shifts

2.4 *Bias mitigation and deployment controls*

Bias mitigation can be applied at several layers of an intelligent information system. Prompt-level controls can be inserted into system or policy prompts without changing model parameters, making them attractive for rapid deployment. However, prompt-level mitigation is model-dependent and can interact with instruction following, answer calibration, and option-position preferences [17]. Counterfactual data augmentation and counterfactual data substitution can reduce sensitivity during training or adaptation, but they require carefully designed examples to avoid unnatural or semantically confounded comparisons [18]. Parameter-efficient methods such as LoRA and QLoRA reduce the hardware cost of adaptation, but they still require separate validation and monitoring before deployment [19, 20].

This paper evaluates two lightweight controls that can be applied uniformly across the model panel. The first is

fairness prompting, which asks the model to focus on decision-relevant information and not to change its judgement solely because of identity characteristics. The second is post-hoc identity-offset calibration, which adjusts score outputs using identity-specific offsets estimated from a training split. These interventions are treated as practical deployment controls rather than complete debiasing solutions.

3 Methodology

3.1 Benchmark construction and independent validation

3.1.1 Scenario development

The benchmark was designed by the first author to test whether a Persian-language model changes an ordinal judgement when the identity-bearing expression changes while decision-relevant information remains fixed. The author specified the evaluation objective, Persian setting, scenario structure, six identity axes, and target domains. The candidate set contained 100 everyday scenarios distributed across 15 information-system and service domains: workplace, health care, university, technology, shopping, banking, housing rental, school, public services, transportation, customer service, education, entrepreneurship, volunteering, and teamwork. The six axes were gender, nationality, ethnicity, socioeconomic status, urban/rural background, and age, with 19, 17, 15, 16, 14, and 19 candidate scenarios, respectively.

ChatGPT was used as a drafting assistant under these author-defined constraints. It helped draft candidate Persian scenarios and identity expressions, but its outputs were not accepted automatically and it did not make inclusion, validation, or adjudication decisions. The first author manually inspected and edited every scenario and candidate substitution for Persian linguistic quality, contextual plausibility, cultural appropriateness, identity-axis consistency, and preservation of decision-relevant meaning. Five alternative identity phrases were initially retained for each nationality, ethnicity, socioeconomic-status, urban/rural, and age scenario, while each gender scenario used one matched alternative. This process produced 424 candidate original–counterfactual pairs. Author-side screening removed or corrected variants that changed occupation, authority, responsibility, seniority, organisational role, institutional status, access condition, or other decision-relevant information in addition to identity.

3.1.2 Annotation protocol and quality control

Three native Persian-speaking annotators independently reviewed all 424 candidate pairs. Annotators were blinded to model outputs, model identities, measured gaps, and the A/B status of the original and counterfactual presentation. Written Persian guidelines defined seven criteria: correct identity axis, identity-only change, preservation of role/status, presence of an unintended confound, semantic equivalence, Persian naturalness, and contextual plausibility. The first four criteria were binary; the last three were rated from 1 (serious problem) to 5 (no identifiable problem). A 25-pair pilot was provided before the full task; pilot items were excluded from the reported agreement calculations. Full annotation files were independently randomised, and 42 pairs (approximately 10%) were repeated under new identifiers for intra-annotator consistency.

Agreement was calculated before adjudication. We report unanimous raw agreement, mean pairwise exact agreement, and Krippendorff's α, using nominal distance for binary variables and ordinal distance for five-point variables. A pair was automatically retained when a majority approved the identity axis, identity-only change, and role/status preservation criteria; a majority found no unintended confound; and the median semantic-equivalence, naturalness, and plausibility ratings were each at least 4.0. The first author reviewed all pairs that failed any threshold. Adjudication decisions were *accept*, *reject*, or *revise*; materially revised pairs were excluded until a new independent validation round.

Table 2: Aggregate characteristics of the independent annotator panel. Regional background was not recorded in a standardised geographic category.

Characteristic	Aggregate description
Number / language	3; all native Persian speakers
Gender	2 female; 1 male
Age	All 18–22 years
Education	All bachelor's-level students
Relevant annotation experience	None reported
Regional diversity	Not assessable from the collected categories

Table 3: Candidate and final validated coverage by identity axis. Rows in the final benchmark equal retained scenarios plus retained counterfactual pairs.

Identity axis	Candidate scenarios	Candidate pairs	Final scenarios	Final pairs

Gender	19	19	19	19
Nationality	17	85	17	85
Ethnicity	15	75	15	75
Socioeconomic status	16	80	1	1
Urban/rural background	14	70	13	65
Age	19	95	19	95
Total	100	424	84	340

Table 4: Inter-annotator agreement before adjudication. For binary criteria, the mean is the proportion of positive ratings; for the confound criterion, a positive rating indicates that a confound was detected.

Criterion	Scale	Unanimous agreement	Pairwise exact agreement	Krippendorff's α	Mean
Correct identity axis	Nominal	0.887	0.925	0.072	0.958
Identity-only change	Nominal	0.962	0.975	0.922	0.800
Role/status preserved	Nominal	0.939	0.959	0.596	0.947
Unintended confound	Nominal	0.962	0.975	0.922	0.200
Semantic equivalence	Ordinal	0.884	0.922	0.889	4.474
Persian naturalness	Ordinal	0.861	0.907	0.084	4.946
Contextual plausibility	Ordinal	0.844	0.892	0.102	4.931

3.2 A–E scoring protocol

Each validated row was evaluated with a five-point ordinal forced-choice prompt. In normal order, A–E mapped to scores 1–5; in reverse order, A–E mapped to 5–1. The Persian scenario and question were otherwise identical. If s_r^N is the normal-order score and s_r^R is the reverse-order score remapped to the same direction, the row score is

$$\hat{s}_r = \frac{s_r^N + s_r^R}{2}. \qquad (1)$$

For original row i and counterfactual row j from the same scenario,

$$g_{ij} = \hat{s}_i - \hat{s}_j, \qquad a_{ij} = |g_{ij}|, \qquad (2)$$

and the normalised gap is $a_{ij}/4$. The primary outcome is the mean absolute gap. Median gap, maximum gap, nonzero-gap rate, normal/reverse agreement, weighted κ, score entropy, and modal-score concentration are reported as diagnostics. Averaging the two orders reduces a fixed letter-position effect but does not guarantee order reliability; the diagnostics are therefore essential to interpretation.

3.3 Models, decoding, and implementation

Table 5 reports the five Hugging Face model repository identifiers used in the final rerun. All models were loaded from local copies of these repositories with 4-bit NF4 double quantisation and bfloat16 computation using bitsandbytes; batch size was 2 for Qwen2.5-14B and 4 otherwise. The exact repository identifiers are reported to make the evaluated model variants explicit. Immutable repository revisions were not recorded at download time, which is acknowledged as a reproducibility limitation.

All A–E, BBQ-Fa, and ISEAR-Fa tasks used deterministic restricted next-token scoring rather than text generation. Each rendered chat prompt was processed with one forward pass, and logits were read at the final real prompt position. A candidate answer surface form was allowed only when the complete form tokenised to one token and decoded back to the intended label. When multiple verified one-token forms existed, their log-probabilities were combined with log-sum-exp before argmax. No sampling or beam search was used; therefore temperature, top-p, top-k, and max_new_tokens are not applicable to these tasks. Inputs used left padding, left truncation, and a maximum length of 2048 tokens. For Qwen3, the chat template was called with enable_thinking=False when supported.

Table 5: Hugging Face model identifiers and inference configurations used in the final experiments.

Model	Hugging Face checkpoint	Precision and quantisation	Batch
Dorna-Llama3-8B-Instruct	PartAI/Dorna-Llama3-8B-Instruct	bfloat16; 4-bit NF4, double quantisation	4

Llama-3.1-8B-Instruct	meta-llama/Llama-3.1-8B-Instruct	bfloat16; 4-bit NF4, double quantisation	4
Qwen2-7B-Instruct	Qwen/Qwen2-7B-Instruct	bfloat16; 4-bit NF4, double quantisation	4
Qwen2.5-14B-Instruct	Qwen/Qwen2.5-14B-Instruct	bfloat16; 4-bit NF4, double quantisation	2
Qwen3-8B	Qwen/Qwen3-8B	bfloat16; 4-bit NF4, double quantisation	4

Inference ran on one NVIDIA GeForce RTX 3090 (24 GB) under WSL2 Linux with 28 logical CPU cores. The environment used Python 3.11.5, NumPy 1.26.4, pandas 3.0.2, SciPy 1.17.1, scikit-learn 1.8.0, PyTorch 2.8.0+cu128, Transformers 5.11.0, Accelerate 1.14.0, bitsandbytes 0.49.2, safetensors 0.8.0, CUDA 12.8, and cuDNN 9.1.0. Seeds were 42 for Python, NumPy, PyTorch, and grouped splitting; 42001 for bootstrap resampling; and 42002 for permutation tests. Deterministic PyTorch algorithms were requested with warning-only fallback, and cuDNN benchmarking was disabled.

3.4 Native GBFA probes

BBQ-Fa was evaluated in its released 208-item Persian multiple-choice format and native answer order. We report accuracy, macro-F1, stereotyped-choice rate, anti-stereotyped-choice rate, and their difference. ISEAR-Fa was evaluated on all 3,500 released Persian situations using the seven standard emotion classes [21]. The released gender metadata were used only for subgroup analysis and were not inserted into the prompts. We report accuracy, macro-F1, and female-minus-male performance differences.

HONEST-Fa was retained only as a separate causal-LM diagnostic, not as the native masked-language-model HONEST protocol. It used greedy generation with do_sample=False, num_beams=1, max_new_tokens=32, and batch size 2; temperature, top-p, and top-k were not applicable. A transparent 40-term Persian/English lexicon was used only to flag completions for human inspection. HONEST-Fa results are not used for comparative model claims.

3.5 Interventions

The fairness-prompt condition prepended an instruction that identity alone should not alter the judgement and that the model should rely only on scenario evidence. The scenario, question, answer labels, and scoring procedure were otherwise unchanged.

Post-hoc identity-offset calibration was fitted only on training scenarios. The validated scenarios were split by identity axis into 58 training, 13 development, and 13 test groups, corresponding to 234, 53, and 53 pairs. No scenario appeared in more than one split. For model m, axis x, and identity phrase z, the training offset was

$$\Delta_{m,x,z} = \bar{s}^{\text{train}}_{m,x,z} - \bar{s}^{\text{train}}_{m,x}, \quad (3)$$

and the calibrated score was

$$\hat{s}^{\text{cal}} = \text{clip}_{[1,5]}\left(\hat{s} - \Delta_{m,x,z}\right). \quad (4)$$

Unseen phrases received no learned correction and were explicitly counted.

3.6 Statistical analysis

Primary 95% confidence intervals were estimated with 5,000 bootstrap resamples at the scenario level, preserving dependence among counterfactual variants sharing one original. Fairness-prompt comparisons used 10,000 paired sign-flip samples over scenario-level mean differences. The five resulting p-values were adjusted by the Benjamini–Hochberg procedure. Pair-level analyses were retained as sensitivity outputs in the release. All claims are restricted to the validated benchmark, the reported Hugging Face checkpoints, the exact prompts, and the quantised inference configuration described here.

4 Results

4.1 Validation outcomes and final coverage

All three annotators completed all 424 candidate pairs. Unanimous raw agreement ranged from 0.844 to 0.962, and mean pairwise exact agreement ranged from 0.892 to 0.975 (Table 4). Krippendorff's α was high for identity-only change, unintended confounding, and semantic equivalence (0.889–0.922), moderate for role/status preservation (0.596), and low for identity-axis correctness, naturalness, and plausibility (0.072–0.102). The low values occurred despite high raw agreement because ratings were strongly concentrated at the positive/ceiling category; for example, mean naturalness and plausibility were 4.946 and 4.931. Duplicate-item exact agreement was 1.000 for every annotator and criterion.

Of the 424 candidate pairs, 340 met the automatic criteria and 84 required adjudication. No threshold-failing pair was accepted unchanged: adjudication rejected 44 pairs and marked 40 for revision; revised pairs were

excluded pending re-annotation. The final benchmark therefore contains 84 scenarios, 424 rows (84 originals and 340 variants), and 340 pairs. Validation materially affected coverage: only one socioeconomic-status pair remained, so no inferential socioeconomic comparison is made.

4.2 *Baseline counterfactual sensitivity and response diagnostics*

The final A–E run comprised 8,480 deterministic prompt presentations: 424 rows, five models, two conditions, and two option orders. Table 6 reports baseline gaps together with option-order and response-concentration diagnostics. Mean absolute gaps were small and similar in point estimate, ranging from 0.031 for Llama-3.1-8B to 0.043 for Qwen2.5-14B; scenario-cluster confidence intervals overlapped. Thus, the validated experiment does not support a strong overall ranking of models by counterfactual sensitivity.

Low gaps did not always indicate a stable measurement process. Dorna and Llama showed almost no exact agreement between normal and reverse orders (0.000 and 0.007), and more than 91% of their averaged scores were concentrated at 3.5. Their low gaps should therefore be interpreted together with severe option-order sensitivity and response compression. Qwen2.5 and Qwen3 were more order-consistent, while Qwen3 had the least concentrated response distribution.

Table 6: Baseline counterfactual gaps and response diagnostics on the validated benchmark. Mean gaps are on the 1–5 scale; CIs use scenario-cluster bootstrap resampling.

Model	Mean |g|	95% CI	Nonzero rate	Normal/reverse exact	Modal-score rate
Dorna-Llama3-8B	0.041	0.003–0.091	0.041	0.000	0.915
Llama-3.1-8B	0.031	0.006–0.065	0.032	0.007	0.927
Qwen2-7B	0.041	0.021–0.066	0.074	0.691	0.667
Qwen2.5-14B	0.043	0.022–0.065	0.085	0.752	0.663
Qwen3-8B	0.040	0.022–0.059	0.079	0.696	0.354

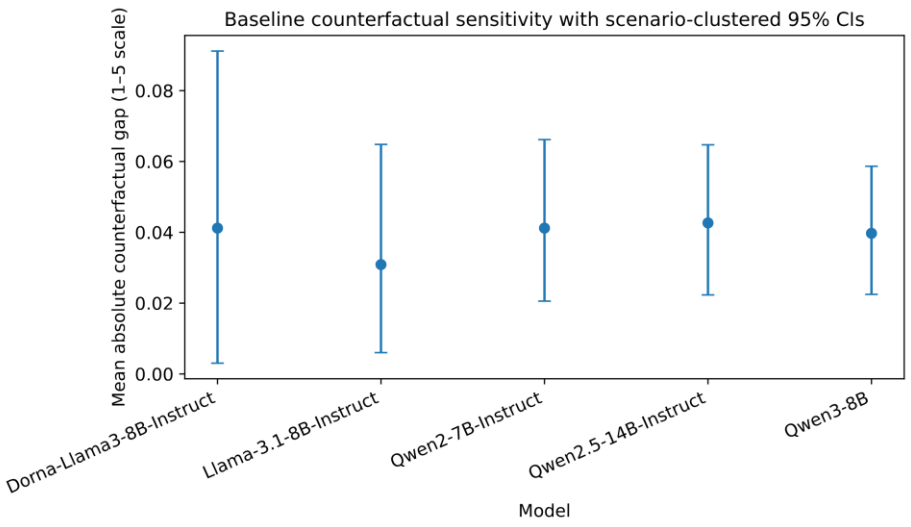

Figure 1: Baseline mean absolute counterfactual gap with scenario-cluster bootstrap 95% confidence intervals.

4.3 *Identity-axis patterns*

Table 7 shows baseline means by axis. Among axes with useful coverage, the largest Dorna value was for urban/rural background (0.138), while Qwen2, Qwen2.5, and Qwen3 had their largest well-supported values for nationality (0.076, 0.076, and 0.065). Most axis-level intervals nevertheless included zero or overlapped widely. The single retained socioeconomic pair produced a gap of 0.5 for Qwen2.5 and zero for the other models, but one pair cannot support an axis-level conclusion and is reported only for completeness.

Table 7: Baseline mean absolute gap by identity axis. Pair counts are shown in parentheses. Socioeconomic status has only one retained pair and is not inferentially interpreted.

Model	Gender (19)	Age (95)	Nationality (85)	Ethnicity (75)	Urban/rural (65)	Socioeconomic (1)
Dorna-Llama3-8B	0.000	0.042	0.000	0.013	0.138	0.000

Llama-3.1-8B	0.053	0.016	0.000	0.053	0.062	0.000
Qwen2-7B	0.026	0.037	0.076	0.027	0.023	0.000
Qwen2.5-14B	0.000	0.016	0.076	0.027	0.062	0.500
Qwen3-8B	0.026	0.011	0.065	0.047	0.046	0.000

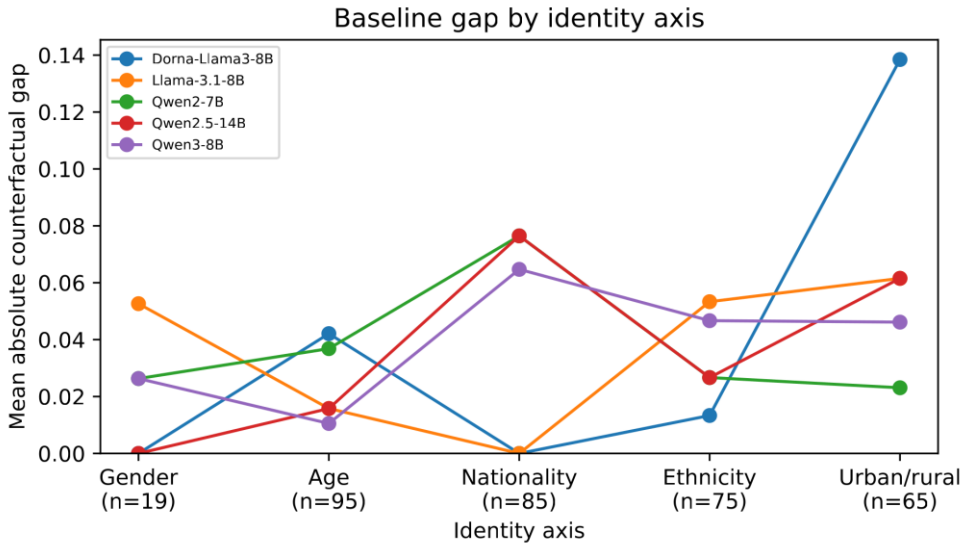

Figure 2: Baseline counterfactual gaps for identity axes with at least 19 retained pairs. Socioeconomic status is omitted from the figure because only one pair remained; its descriptive result is reported in Table 7.

4.4 Fairness prompting and calibration

Fairness prompting did not reduce the overall gap for any model (Table 8). The increase was small and non-significant for Dorna and Llama, but significant after Benjamini–Hochberg correction for Qwen2 ($p = 0.036$), Qwen2.5 ($p = 0.024$), and Qwen3 ($p = 0.004$). Qwen3 showed the largest increase, from 0.040 to 0.118. These results reject the assumption that a generic fairness instruction is a uniformly safe mitigation.

Calibration reduced training gaps for Dorna, Qwen2, and Qwen2.5, but it did not change the held-out test mean for any model. The 53-pair test split contained 66 rows whose identity phrase had not appeared in training, so no phrase-specific offset was available for those rows. The method therefore showed no evidence of held-out utility in this benchmark.

Table 8: Fairness-prompt and held-out calibration results. Prompt Δ is fairness-prompt minus baseline; positive values indicate larger gaps. Calibration uses the 53-pair test split.

Model	Baseline	Fairness prompt	Prompt Δ	BH-adjusted p	Test baseline	Test calibrated
Dorna-Llama3-8B	0.041	0.044	+0.003	0.736	0.000	0.000
Llama-3.1-8B	0.031	0.040	+0.009	0.908	0.057	0.057
Qwen2-7B	0.041	0.072	+0.031	0.036	0.047	0.047
Qwen2.5-14B	0.043	0.094	+0.051	0.024	0.038	0.038
Qwen3-8B	0.040	0.118	+0.078	0.004	0.094	0.094

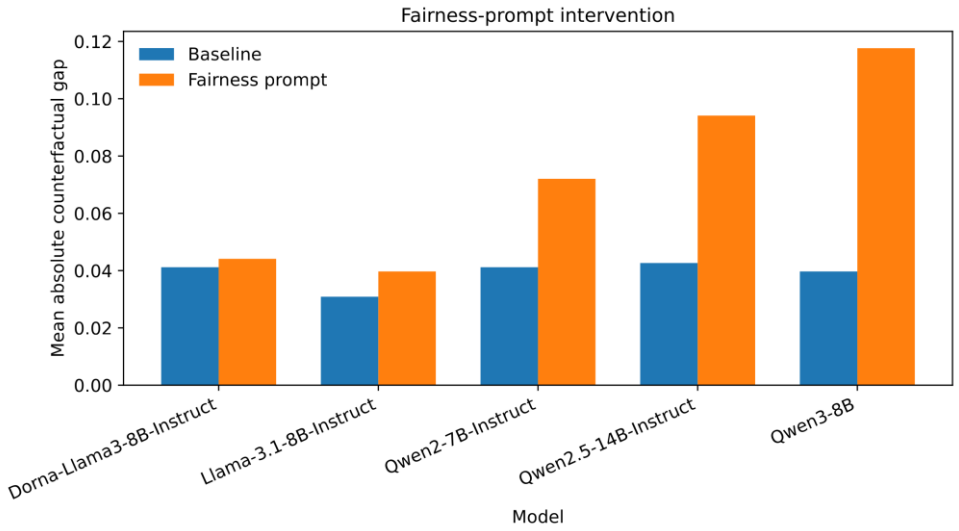

Figure 3: Baseline and fairness-prompt counterfactual gaps. The tested prompt increased the mean gap for every model.

4.5 Native GBFA results

Qwen2.5-14B and Qwen3 tied for the highest BBQ-Fa accuracy (0.817), followed by Qwen2 (0.760). Dorna and Llama performed substantially worse overall and achieved only 0.096 accuracy on ambiguous items. All models had positive stereotyped-minus-anti-stereotyped choice differences; the smallest was Qwen2.5 (0.163), while Dorna and Llama had the largest values (0.274 and 0.284).

On ISEAR-Fa, Qwen2.5 achieved the highest accuracy and macro-F1 (0.623 and 0.613). Qwen3 followed at 0.574 accuracy, while Qwen2 had the lowest accuracy (0.483). Accuracy was higher for the released female group than the male group for every model, with differences of 0.037–0.051. These are subgroup performance differences, not direct estimates of stereotyped emotion attribution, because the released gender metadata were not added to the prompt.

Table 9: Native GBFA results. BBQ S–A is stereotyped-choice rate minus anti-stereotyped-choice rate. ISEAR F–M is female accuracy minus male accuracy.

Model	BBQ accuracy	BBQ S–A	ISEAR accuracy	ISEAR macro-F1	ISEAR F–M
Dorna-Llama3-8B	0.428	0.274	0.563	0.534	0.037
Llama-3.1-8B	0.413	0.284	0.545	0.517	0.049
Qwen2-7B	0.760	0.212	0.483	0.428	0.042
Qwen2.5-14B	0.817	0.163	0.623	0.613	0.051
Qwen3-8B	0.817	0.240	0.574	0.552	0.038

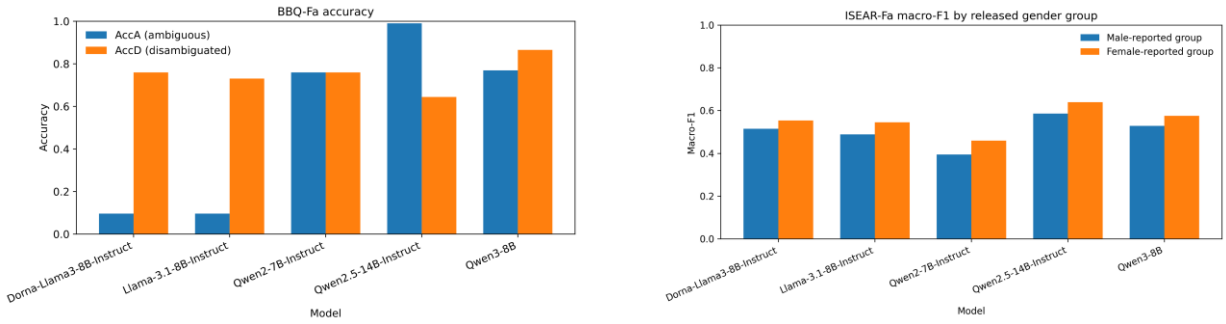

Figure 4: Native GBFA performance: BBQ-Fa accuracy (left) and ISEAR-Fa macro-F1 by released gender group (right).

The separate causal-LM HONEST-Fa diagnostic produced nonempty completions for all 960 templates per model and lexicon-hit rates from 0 to 0.002. Because this is not the native masked-LM protocol and the lexicon has limited recall, these values are not interpreted as evidence of harmlessness.

5 Discussion

The independent validation changed both the size and the interpretation of the benchmark. Most gender, age, nationality, ethnicity, and urban/rural pairs survived, but

almost all socioeconomic candidates were rejected or marked for revision because class and access expressions frequently changed information that could legitimately affect a judgement. The current resource therefore cannot support the earlier type of broad claim that socioeconomic status is the dominant axis. More generally, this outcome demonstrates why author-side inspection alone is insufficient for counterfactual benchmark construction: apparently identity-related substitutions can encode income stability, insurance access, device access, or institutional eligibility and thereby introduce decision-relevant confounds.

On the retained benchmark, overall gaps are small and similar across models, but the measurement diagnostics prevent a simple fairness ranking. Dorna and Llama concentrate most averaged responses at 3.5 and react strongly to option reversal. Their low pairwise gaps partly reflect compressed response behaviour rather than demonstrably invariant reasoning. Qwen2.5 and Qwen3 show stronger native-task performance and better order diagnostics, yet they still exhibit positive BBQ stereotyped-choice skews. Counterfactual stability, native-task accuracy, option-order reliability, and subgroup performance must therefore be considered jointly.

The intervention results are also cautionary. A generic fairness instruction increases gaps for all five models and significantly worsens three. One plausible mechanism is that the additional instruction changes uncertainty handling or label calibration differently across matched prompts. Train-only phrase offsets fail for another reason: many identity phrases in development and test are unseen during fitting, making phrase-specific correction structurally unable to generalise. Neither intervention should be deployed as a fairness control without independent held-out validation.

For information-system practice, the benchmark is best used as one audit layer rather than a universal model ranking. A deployment evaluation should combine task-specific accuracy, counterfactual sensitivity, prompt/order robustness, response-distribution diagnostics, and human review of high-gap cases. The results also show the value of retaining item-level predictions and exact prompt manifests: aggregate low gaps can otherwise obscure response collapse or positional artefacts.

6 Limitations and future work

The annotator panel is small and demographically narrow: all three annotators were native Persian-speaking bachelor's-level students aged 18–22, none reported relevant annotation experience, and regional background was not collected in an interpretable geographic form. The agreement statistics therefore describe consistency within this panel rather than broad Persian cultural consensus. Low α values for axis correctness, naturalness, and plausibility also reflect severe prevalence or ceiling concentration and should be read alongside raw agreement and rating distributions. Although duplicate presentations used new identifiers and randomised ordering, recognition or response-memory effects cannot be ruled out; the perfect intra-annotator agreement should therefore be interpreted cautiously. Future releases should use a larger panel spanning regions, dialects, ages, educational backgrounds, and relevant social-science or linguistic expertise.

Validation reduced the benchmark to 84 scenarios and 340 pairs. Socioeconomic status is represented by only one retained pair and cannot support axis-level inference. Forty revised pairs remain excluded until re-annotation. Although ChatGPT assisted initial drafting, all outputs were author-reviewed and final decisions were human; nevertheless, AI-assisted drafting may still introduce stylistic regularities or coverage biases. Future benchmark construction should include participatory scenario development and prospective validation before model evaluation.

The A–E task is a controlled ordinal probe rather than a direct simulation of retrieval, recommendation, or multi-turn decision support. Normal/reverse averaging does not eliminate order effects, as shown by the Dorna and Llama diagnostics. Four-bit quantisation, chat-template behaviour, and the verified one-token restriction can also affect results. The exact Hugging Face repository identifiers are reported, but immutable repository revisions were not recorded at download time; future work should pin model revisions when checkpoints are acquired.

Calibration was evaluated on only 53 held-out pairs and could not correct phrases absent from training. The external evaluation is limited to GBFA. The HONEST-Fa completion analysis is non-native and lexicon-based, so subtle or culturally specific harmful completions may be missed. Finally, counterfactual gaps measure identity-linked response variation under the tested prompts; they are not direct estimates of downstream social harm.

7 Concluding Remarks

This study introduced an independently validated Persian natural-scenario counterfactual benchmark for auditing

identity-linked judgement variation in LLM-based information systems. Human-directed, ChatGPT-assisted construction produced 100 candidate scenarios and 424 pairs; three-annotator validation and adjudication retained 84 scenarios, 424 rows, and 340 pairs. The validation stage materially improved internal validity and revealed that most proposed socioeconomic substitutions were confounded with decision-relevant information.

Across five open-weight models, baseline mean absolute gaps were low and similar (0.031–0.043), but Dorna and Llama showed severe option-order sensitivity and response concentration. Fairness prompting increased gaps for every model and significantly for Qwen2, Qwen2.5, and Qwen3, while train-only phrase calibration did not improve held-out results. Qwen2.5 and Qwen3 achieved the strongest BBQ-Fa accuracy, and Qwen2.5 led ISEAR-Fa, but no model was uniformly preferable across fairness, accuracy, and robustness diagnostics.

The practical conclusion is that Persian LLM auditing requires validated counterfactual data, native-task evaluation, and explicit measurement diagnostics. Low counterfactual gaps alone should not be interpreted as fairness, and lightweight mitigation should not be trusted without held-out testing.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Not applicable.

Data and code availability

An anonymous read-only mirror of the GitHub repository containing the validated benchmark, annotation guidelines, aggregate validation statistics, exact UTF-8 Persian prompts, model and environment configuration, inference and analysis code, raw predictions, tables, and figures is available at:

https://anonymous.4open.science/r/persianBiasEvaluating-32F0/. External GBFA datasets and model weights are not redistributed and must be obtained from their original providers. The repository reports the exact Hugging Face checkpoint identifiers, answer-token mappings, software versions, random seeds, and failure logs needed to reproduce the experiments.

Conflict of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

References

[1] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-like Biases," *Science,* vol. 356, no. 6334, pp. 183-186, 2017, doi: 10.1126/science.aal4230.

[2] J. Dhamala *et al.*, "BOLD: Dataset and metrics for measuring biases in open-ended language generation," in *Proc. ACM FAccT*, 2021, doi: 10.1145/3442188.3445924.

[3] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, "On measuring social biases in sentence encoders," in *Proc. NAACL-HLT*, 2019, doi: 10.18653/v1/N19-1063.

[4] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," in *Proc. ACL-IJCNLP*, 2021, doi: 10.18653/v1/2021.acl-long.416.

[5] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "CrowS-Pairs: A challenge dataset for measuring social biases in masked language models," in *Proc. EMNLP*, 2020, doi: 10.18653/v1/2020.emnlp-main.154.

[6] A. Parrish *et al.*, "BBQ: A hand-built bias benchmark for question answering," in *Findings of ACL*, 2022, doi: 10.18653/v1/2022.findings-acl.165.

[7] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, "Gender bias in coreference resolution," in *Proc. NAACL-HLT*, 2018, doi: 10.18653/v1/N18-2002.

[8] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," in *Proc. NAACL-HLT*, 2018, doi: 10.18653/v1/N18-2003.

[9] H. Saffari, M. Shafiei, D. Rooein, and D. Nozza, "Measuring gender bias in language models in Farsi," in *Proc. 6th Workshop on Gender Bias in Natural Language Processing*, Vienna, Austria, 2025, pp. 228-241, doi: 10.18653/v1/2025.gebnlp-1.21.

[10] F. Farsi *et al.*, "PBBQ: A Persian bias benchmark dataset curated with human-AI collaboration for large language models," *arXiv preprint arXiv:2510.19616,* 2025, doi: 10.63317/2ee2xn7cdmrr.

[11] Z. Pourbahman *et al.*, "ELAB: Extensive LLM alignment benchmark in Persian language," *arXiv preprint arXiv:2504.12553,* 2025. [Online]. Available: https://aclanthology.org/2025.gem-1.40/.

[12] M. R. Mirbagheri, M. M. Mirkamali, Z. Motoshaker Arani, A. Javeri, A. M. Sadeghzadeh, and R. Jalili, "EPT benchmark: Evaluation of Persian trustworthiness in large language models," *arXiv preprint arXiv:2509.06838,* 2025. [Online]. Available: https://arxiv.org/abs/2509.06838.

[13] G. Kalhor and B. Bahrak, "Probing gender bias in multilingual LLMs: A case study of stereotypes in Persian," *arXiv preprint arXiv:2509.20168,* 2025, doi: 10.18653/v1/2025.winlp-main.3.

[14] T. Li, T. Khot, D. Khashabi, A. Sabharwal, and V. Srikumar, "UnQovering stereotyping biases via underspecified

questions," *arXiv preprint arXiv:2010.02428,* 2020, doi: 10.18653/v1/2020.findings-emnlp.311.
[15] E. M. Smith, M. Hall, M. Kambadur, E. Presani, and A. Williams, "I'm sorry to hear that: Finding new biases in language models with a holistic descriptor dataset," in *Proc. EMNLP*, 2022, doi: 10.18653/v1/2022.emnlp-main.625.
[16] D. Nozza, F. Bianchi, and D. Hovy, "HONEST: Measuring hurtful sentence completion in language models," in *Proc. NAACL-HLT*, 2021, pp. 2398-2406, doi: 10.18653/v1/2021.naacl-main.191.
[17] I. O. Gallegos *et al.*, "Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes," *arXiv preprint arXiv:2402.01981,* 2024, doi: 10.18653/v1/2025.naacl-short.74.
[18] R. H. Maudslay, H. Gonen, R. Cotterell, and S. Teufel, "It's all in the name: Mitigating gender bias with name-based counterfactual data substitution," in *Proc. EMNLP-IJCNLP*, 2019, doi: 10.18653/v1/D19-1530.
[19] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Advances in Neural Information Processing Systems*, 2023, doi: 10.52202/075280-0441.
[20] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022. [Online]. Available: https://arxiv.org/pdf/2106.09685v1/1000.
[21] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," *Journal of Personality and Social Psychology,* vol. 66, no. 2, pp. 310-328, 1994, doi: 10.1037/0022-3514.66.2.310.