Mind the Gap: Multilingual Divide in LLM Bias Detection and Reasoning

Medha Hira[*1] **Prachi Goyal**[*1] **Raj Maheshwari**[*1] **Arnav Goel**[1]
[1]Carnegie Mellon University
{mhira, prachigo, rajmahes, arnavgoe}@cs.cmu.edu

Abstract

Large Language Models (LLMs) are increasingly deployed in multilingual settings, yet most bias evaluation remains English-centric and overlooks how bias manifests within reasoning. We present a systematic study of social bias in both predictions and chain-of-thought reasoning across English, Dutch, Spanish, and Turkish using the MBBQ benchmark. We evaluate instruction-tuned, CoT-prompted, and reasoning-native models under supervised fine-tuning and preference optimization, using accuracy, F1, bias metrics, and a novel reasoning-level language drift measure. We find that (1) bias varies substantially across languages, with consistent degradation in non-English settings, (2) reasoning traces often introduce additional stereotype-driven signals beyond final outputs, and (3) English-trained debiasing methods fail to generalize reliably, with preference optimization introducing cross-lingual trade-offs. We further show that performance gains in multilingual settings are frequently driven by implicit reliance on English-centric reasoning, revealed through increased language drift. Together, our results demonstrate that multilingual fairness cannot be inferred from English performance and requires reasoning-aware, language-specific evaluation and alignment.

1 Introduction

Large Language Models (LLMs) increasingly mediate information access, decision support, and content generation across diverse linguistic and cultural contexts. As their global deployment expands, concerns about social bias and representational harm have intensified. Prior work has shown that LLMs encode stereotypes related to gender, age, disability, socioeconomic status, and nationality (Nadeem et al., 2021; Nangia et al., 2020; Parrish et al., 2021). However, the vast majority of this research focuses exclusively on English (Yong et al., 2025).

Emerging multilingual studies suggest that bias is not language-invariant: stereotypes may strengthen, weaken, or even reverse depending on linguistic structure, cultural context, and training data distribution (Raza et al., 2024; Levy et al., 2023). At the same time, modern reasoning-centric models expose explicit chain-of-thought (CoT) traces (Wei et al., 2023), creating new pathways for bias to manifest—not only in final predictions, but also within intermediate reasoning.

Despite this progress, three key gaps remain:

1. Existing work primarily evaluates *outputs*, leaving reasoning traces largely unexplored.
2. Debiasing and bias-detection methods are almost entirely English-specific.
3. Cross-lingual transfer of debiasing remains poorly understood.

To address these gaps, we conduct a systematic study of multilingual social bias in both predictions and reasoning traces. Using the MBBQ dataset (Parrish et al., 2021; Neplenbroek et al., 2024), we evaluate instruction-tuned, CoT-prompted, and reasoning-native models across English, Dutch, Spanish, and Turkish.

Research Questions

- **RQ0:** How and why does social bias differ across languages?
- **RQ1:** Does debiasing in one language generalize to others?
- **RQ2:** Can multilingual training improve cross-lingual robustness?
- **RQ3:** What metrics best capture reasoning-level bias in multilingual settings?

Contributions We make the following contributions:

- We provide the first systematic evaluation of **reasoning-level bias** across multiple languages using MBBQ.

[*] denotes equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1305–1316

- We extend the **BiasGuard framework** to a multilingual setting and analyze the behavior of SFT and preference optimization under cross-lingual transfer.
- We introduce a **reasoning-level language drift metric** that quantifies language switching in CoT traces, revealing a new failure mode in multilingual reasoning.

Our results show that bias varies significantly across languages, reasoning-native models can amplify stereotypes in intermediate steps, and English-trained debiasing methods fail to generalize reliably. These findings highlight the need for culturally adaptive and reasoning-aware approaches to fairness in multilingual LLMs.

2 Related Work

2.1 Multilingual Bias in LLMs

Recent work has shown that bias in LLMs is both language- and culture-dependent. Zhao et al. (2024) evaluate gender bias across multiple languages, finding that while some bias patterns persist, their strength and manifestation vary across linguistic contexts. Similarly, Mihaylov and Shtedritski (2024) and Mukherjee et al. (2023) demonstrate that biases reflect localized socio-cultural patterns rather than being purely Western-centric.

These studies highlight the importance of multilingual evaluation, but primarily focus on *final model outputs*, leaving the role of intermediate reasoning largely unexplored.

2.2 Bias in Chain-of-Thought Reasoning

A growing body of work shows that bias can propagate through intermediate reasoning steps. Shaikh et al. (2023) and Bajaj et al. (2024) demonstrate that Chain-of-Thought (CoT) prompting can amplify stereotypical or toxic reasoning, even when final outputs appear neutral. Wu et al. (2025) further show that reasoning traces can systematically encode bias on benchmarks such as BBQ.

Several approaches attempt to mitigate reasoning-level bias. Wu et al. (2024) model internal knowledge as a latent confounder and apply causal interventions, while Wan et al. (2025) show that CoT can exacerbate cognitive biases such as confirmation bias. However, these works are largely restricted to English, leaving open questions about how reasoning bias manifests across languages.

2.3 Reasoning-Native Models

Recent reasoning-native models, such as DeepSeek-R1 (DeepSeek-AI, 2025), Magistral (Mistral AI, 2025), and Llama 3 reasoning variants (Grattafiori et al., 2024), are explicitly trained to produce structured and interpretable reasoning traces. While these models improve step-wise reasoning quality, their behavior with respect to social bias and in multilingual settings remains underexplored.

2.4 Debiasing Methods for LLMs

A range of methods have been proposed to detect and mitigate bias in LLMs. Fan et al. (2025) introduce **BiasGuard**, which leverages CoT reasoning and reinforcement learning to detect bias in model outputs. Other approaches, such as prompt-based debiasing (Yang et al., 2025), aim to reduce biased outputs through instruction design, but often fail to address bias embedded in reasoning.

Importantly, most existing debiasing techniques are developed and evaluated in English-only settings, raising concerns about their effectiveness under cross-lingual transfer.

2.5 Bias Measurement and Evaluation

Bias in LLMs has been evaluated using a variety of metrics, including classification-based measures, bias scores (e.g., BiasAmb and BiasDis), and distributional divergence methods. Recent work also uses LLM-as-a-judge frameworks to assess bias in reasoning traces (Wu et al., 2025).

While these metrics capture different aspects of bias, they are typically designed for monolingual settings and do not explicitly account for multilingual reasoning behavior or language-dependent bias dynamics.

2.6 Positioning of This Work

Prior work has explored multilingual bias, reasoning-level bias, and debiasing methods largely in isolation. In contrast, our work jointly studies **multilingual bias, reasoning traces, and debiasing** within a unified framework. We extend BiasGuard to a multilingual setting and introduce a reasoning-level language drift metric to capture language switching in CoT traces, enabling a more comprehensive analysis of bias in multilingual LLMs.

3 Data

3.1 Dataset and Task Formulation

We use the Multilingual BBQ (MBBQ) dataset (Neplenbroek et al., 2024), a cross-lingual extension of BBQ (Parrish et al., 2021) for evaluating social bias in question-answering. MBBQ provides parallel instances across multiple languages, enabling controlled cross-lingual evaluation. We focus on four languages: English (EN), Dutch (NL), Spanish (ES), and Turkish (TR).

Each instance is represented as a tuple $(c, q, \mathcal{A}, y, m)$, where c is a context describing a social scenario, q is a question about an individual in that context, and $\mathcal{A} = \{a_i\}_{i=0}^{2}$ denotes three answer options. The label y identifies the correct answer, while metadata m specifies whether an option corresponds to an "unknown" response.

A key attribute of MBBQ is the `context_condition` variable, which partitions instances into *ambiguous* cases (correct answer is "unknown") and *disambiguated* cases (sufficient evidence is provided). This structure enables evaluation of stereotype-driven reasoning under underspecified versus evidence-supported conditions.

We consider four bias categories:

Category	Instances
Disability Status	1296
Physical Appearance	1176
Socioeconomic Status	3600
Sexual Orientation	152

3.2 BiasGuard Conversion

To enable reasoning-based bias detection, we convert MBBQ into a binary classification dataset following the BiasGuard framework.

Conversion Procedure. Given an instance $(c, q, \mathcal{A}, y, m)$ with $\mathcal{A} = \{a_i\}_{i=0}^{2}$, we derive three instances by pairing the context c with each answer option a_i. Each instance is assigned a binary label: the option corresponding to the ground-truth answer y is labeled *unbiased* (0), and the remaining options are labeled *biased* (1).

For each a_i, we construct a declarative statement by concatenating c with a templated assertion of the form `<answer choice> is the person <question>`, converting the task into statement-level classification.

Handling "Unknown" Responses. Answer options corresponding to "unknown" are excluded, as they cannot be unambiguously assigned to either class. Identification is performed using metadata (`answer_info`) with additional heuristic matching when required.

Filtering. Training is restricted to disambiguated instances, excluding ambiguous cases to maintain consistency with the binary labeling scheme. Ambiguous examples are retained for evaluation when computing bias-sensitive metrics (e.g., BiasAmb) using the original annotations.

We refer to the resulting dataset as `MBBQ-BiasGuard`.

3.3 Data Splits and Training Setup

We study cross-lingual transfer between English (EN) and Turkish (TR), and evaluate generalization to additional languages.

Training Regimes. We consider two settings: (i) **EN-only**, trained exclusively on English data, and (ii) **EN+TR**, trained on a balanced mixture of English and Turkish. In the multilingual setting, we subsample to ensure equal representation, mitigating bias toward the higher-resource language.

Validation Protocol. For each setting, we construct a validation set via a 10% stratified split of the training data. Validation sets are maintained separately for each regime.

Test Set Construction. We construct a multilingual test set spanning EN, TR, Dutch (NL), and Spanish (ES) by aligning semantically equivalent instances across languages. This minimizes distributional variation and enables controlled cross-lingual comparison. The resulting test set consists of parallel examples across all four languages and will be released publicly.

4 Methodology

We investigate how reasoning-based bias detection methods behave under cross-lingual transfer. To this end, we propose `MBBQ-BiasGuard`, a multilingual extension of the BiasGuard framework (Fan et al., 2025). Our approach separates reasoning generation from final classification, allowing us to analyze how bias emerges within intermediate reasoning traces.

The pipeline consists of two stages: (i) supervised fine-tuning (SFT) on chain-of-thought (CoT) reasoning traces, and (ii) preference-based optimization using reinforcement learning.

4.1 Task Formulation

Given a context–question pair from MBBQ, the model is trained to generate a structured reasoning trace followed by a binary classification (biased vs. unbiased). Unlike standard approaches that evaluate only final predictions, our formulation explicitly models intermediate reasoning, enabling analysis of bias at the reasoning level.

Extending this setup to multilingual settings introduces additional challenges, including differences in linguistic structure, uneven pretraining coverage, and instability of CoT generation across languages. Our framework is designed to control for these factors through consistent data construction and evaluation.

4.2 Supervised Fine-Tuning

We adopt a teacher–student distillation setup to obtain high-quality reasoning traces. A teacher model generates step-by-step reasoning and a final verdict for each input. Generated traces are filtered to ensure (i) consistency between reasoning and label, (ii) presence of a valid conclusion marker, and (iii) sufficient reasoning depth.

The filtered dataset is used to train a student model via next-token prediction over the full reasoning sequence. We consider two training regimes: English-only (SFT-EN) and bilingual (SFT-EN+TR), where the latter uses a balanced mixture of English and Turkish examples.

4.3 Preference Optimization

Following SFT, we refine the model using preference-based learning. For each input, multiple candidate reasoning traces are sampled and filtered for structural validity and correctness. A high-quality trace is selected as the preferred response, while alternative valid traces are treated as rejected responses.

We evaluate two optimization methods: Direct Preference Optimization (DPO) (Rafailov et al., 2024) and Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024). Both methods optimize relative likelihoods of preferred over rejected responses, but differ in their treatment of normalization and stability. We apply both methods under monolingual and multilingual training settings to assess their robustness to cross-lingual variation in preference signals.

4.4 Evaluation Metrics

We evaluate models along three dimensions. First, task performance is measured using accuracy and F1 score on disambiguated examples. Second, we use standard BBQ/MBBQ bias metrics: BiasAmb measures stereotype activation in ambiguous contexts, while BiasDis captures errors when disambiguating information is present.

Finally, we introduce a **Language Drift Rate** to quantify cross-lingual instability in reasoning:

$$\text{Drift} = 1 - \frac{N_{\text{inp}}}{N_{\text{total}}} \tag{1}$$

where N_{inp} is the number of tokens in the input language and N_{total} is the total number of tokens in the reasoning trace. This metric captures the extent to which reasoning traces divert to another language(s) (typically English). We treat this metric as a diagnostic signal: lower drift indicates better language alignment and is desirable in a fully multilingual system.

5 Experimental Setup

5.1 Model Selection

We use DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI, 2025) as both teacher and student model. This choice reflects a balance between reasoning capability and computational efficiency. A scaling comparison across model sizes is provided in Appendix A.1.

For baseline comparisons, we additionally evaluate instruction-tuned short-answer models (Llama-2-7b and Llama-3.1-8B) (Touvron et al., 2023; Grattafiori et al., 2024) to contextualize performance across architectures.

5.2 Training Configuration

We evaluate both monolingual (EN) and multilingual (EN+TR) training regimes. In the multilingual setting, training data is balanced across languages to avoid over-representation of English.

We train SFT, DPO, and ORPO variants under both settings, resulting in six models. Full hyperparameter details and implementation specifics are provided in Appendix A.2.

5.3 Preference Construction

Preference pairs are constructed by sampling multiple reasoning traces per input and filtering them based on correctness and structural validity. The highest-quality trace is selected as the preferred

response. Additional filtering criteria and sampling details are provided in Appendix A.2.

5.4 Evaluation Protocol

All models are evaluated on a matched multilingual test set spanning English, Spanish, Dutch, and Turkish. Metrics are computed per language.

We assess statistical significance of cross-lingual differences using Kruskal–Wallis tests; details are provided in Appendix A.2.

6 Results and Discussion

We present results across multiple model variants and training regimes, summarized in Table 1, Table 2, and Figure 1. These results are analyzed with respect to our research questions.

6.1 RQ0: Cross-Lingual Variation in Bias

We observe substantial and systematic variation in both accuracy and bias across languages. As shown in Table 1 (By Language), English consistently achieves the highest accuracy across all models, while Turkish remains the most challenging language. Dutch and Spanish exhibit intermediate performance.

Figure 1 further highlights that these trends are consistent across training paradigms (SFT, DPO, ORPO): languages that are difficult under one method remain difficult under others. This suggests that cross-lingual differences are not an artifact of training strategy but instead reflect deeper issues in representation and reasoning.

Bias metrics further reveal that stereotype activation is consistently highest in ambiguous settings (BiasAmb). As shown in Appendix A.9, Dutch and Spanish exhibit particularly high BiasAmb values despite relatively strong representation in pretraining corpora, indicating that bias is not purely driven by data availability but also by linguistic structure. In particular, languages with stronger lexical commitments (e.g., gender marking or explicit descriptors) tend to produce more stereotypical completions.

We further observe that the gap between BiasAmb and BiasDis is largest in Dutch and Spanish (Appendix A.9), suggesting that models rely more heavily on culturally patterned stereotypes when contextual information is underspecified. Statistical analysis confirms that these differences are significant ($p < 10^{-47}$; Appendix A.5).

Key Insight. Bias is strongly language-dependent and cannot be inferred from English performance alone.

6.2 RQ1: Cross-Lingual Generalization of Debiasing

We next evaluate whether debiasing learned in English transfers to other languages. Table 1 shows that English-only SFT achieves strong performance in English (0.713 accuracy) but significantly lower performance in Turkish (0.433), with similar gaps for Dutch and Spanish.

Preference optimization further complicates this behavior. While DPO and ORPO sometimes improve performance in non-English languages, they often degrade English performance (e.g., DPO: -0.066 F1 in Table 2). This suggests that preference-based learning does not reliably improve cross-lingual generalization and may introduce trade-offs across languages.

Importantly, Figure 1 shows that performance gaps persist across all training strategies, reinforcing that these effects are not method-specific.

Key Insight. Debiasing learned in English transfers only partially and inconsistently across languages.

6.3 RQ2: Effect of Multilingual Training

To assess whether multilingual supervision improves performance, we compare monolingual (EN) and bilingual (EN–TR) training. Table 2 shows that multilingual SFT yields consistent improvements across all languages, with particularly large gains for Turkish (+0.395 accuracy) and Dutch (+0.385 accuracy).

In contrast, DPO and ORPO exhibit mixed behavior. While both improve performance in lower-resource languages, they often introduce regressions in English. This trade-off is also visible in Table 1, where EN–TR models improve Turkish performance at the cost of slight degradation in English.

These results highlight a fundamental distinction between training paradigms. SFT benefits from exposure to diverse linguistic patterns, leading to improved shared representations. In contrast, preference optimization appears to overfit to language-specific reasoning traces, resulting in conflicting optimization signals across languages.

Further analysis of this behavior is provided in Appendix A.4.

	Base Acc	Base F1	Acc (EN)			Acc (EN+TR)			F1 (EN)			F1 (EN+TR)		
Group	**Deepseek**	**Deepseek**	**SFT**	**DPO**	**ORPO**	**SFT**	**DPO**	**ORPO**	**SFT**	**DPO**	**ORPO**	**SFT**	**DPO**	**ORPO**
By Language														
EN	0.600	0.540	0.713	0.666	0.641	0.776	0.643	0.632	0.702	0.688	0.661	0.783	0.622	0.626
TR	0.367	0.465	0.433	0.400	0.479	0.381	0.479	0.508	0.540	0.500	0.581	0.444	0.586	0.614
NL	0.457	0.303	0.464	0.437	0.433	0.391	0.440	0.425	0.442	0.320	0.357	0.303	0.270	0.365
ES	0.508	0.241	0.512	0.578	0.539	0.523	0.570	0.539	0.450	0.437	0.352	0.247	0.337	0.289
By Category														
Disability	0.414	0.316	0.504	0.428	0.437	0.468	0.445	0.419	0.485	0.407	0.4	0.438	0.396	0.368
Phys. Appearance	0.526	0.516	0.515	0.608	0.598	0.600	0.624	0.649	0.584	0.666	0.636	0.577	0.646	0.691
SES	0.496	0.382	0.538	0.525	0.536	0.507	0.531	0.524	0.521	0.447	0.498	0.427	0.428	0.456
Sexual Orientation	0.375	0.545	0.625	0.500	0.25	0.429	0.625	0.572	0.727	0.666	0.4	0.600	0.727	0.666
Overall	0.482	0.401	0.527	0.519	0.522	0.515	0.531	0.524	0.532	0.496	0.505	0.464	0.478	0.498

Table 1: Comparison of DeepSeek (baseline) with models trained using Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO), and Odds Ratio Preference Optimization (ORPO) on the MBBQ-BiasGuard dataset. DeepSeek metrics are reported in the first two columns.

	SFT		DPO		ORPO	
Language	ΔAcc	ΔF1	ΔAcc	ΔF1	ΔAcc	ΔF1
EN	+0.063	+0.081	−0.023	−0.066	−0.009	−0.035
TR	+0.395	+0.339	+0.164	+0.036	+0.124	+0.012
NL	+0.385	+0.480	+0.203	+0.352	+0.207	+0.261
ES	+0.253	+0.536	+0.073	+0.285	+0.093	+0.337

Table 2: Changes in accuracy and F1 when moving from EN-only training to bilingual EN–TR training. Positive values indicate performance gains from multilingual training.

Key Insight. Multilingual supervision improves cross-lingual performance, but RL-based optimization introduces instability and trade-offs.

6.4 Why Does Preference Optimization Fail Cross-Lingually?

While multilingual SFT yields consistent gains, preference-based methods (DPO, ORPO) (Rafailov et al., 2024; Hong et al., 2024) exhibit unstable behavior across languages. In particular, they improve performance in lower-resource languages while degrading English performance (Table 2), indicating poor cross-lingual transfer of preference signals.

We attribute this to weak and inconsistent supervision. In multilingual settings, "chosen" and "rejected" reasoning traces often differ only superficially (e.g., phrasing or structure), rather than in substantive reasoning quality, resulting in noisy preference signals. This limitation aligns with prior work showing that preference-based optimization is highly sensitive to the quality and separability of feedback signals (Ouyang et al., 2022). Cross-lingual variation further exacerbates this issue: semantically equivalent reasoning traces can differ substantially in surface form across languages, reducing the reliability of pairwise comparisons.

More fundamentally, preference optimization assumes a shared notion of "better" reasoning across training examples. However, prior work shows that both bias expression and reasoning behavior vary across linguistic and cultural contexts (Raza et al., 2024; Yong et al., 2025). As a result, preferences that are meaningful in one language may not transfer cleanly to others.

As discussed in Section 6.5, models frequently rely on English-centric reasoning templates even when generating non-English outputs. This creates conflicting supervision signals: preference optimization encourages alignment with language-specific reasoning patterns, while the model's internal representations remain anchored in English. These competing objectives lead to unstable updates and performance trade-offs across languages.

Consistent with this interpretation, DPO—being more sensitive to relative likelihood differences—often degrades high-resource performance, while ORPO yields more stable but still limited improvements. Overall, these results suggest that standard preference optimization is ill-suited for multilingual debiasing without stronger, language-aware supervision or more semantically grounded

preference signals.

6.5 RQ3: Reasoning Drift and Multilingual Bias

While output metrics capture final predictions, they do not reveal how models arrive at their decisions. To address this, we analyze language drift in reasoning using the proposed Language Drift Rate as a diagnostic signal.

Table 3 shows that multilingual training significantly increases drift for Turkish (+0.24) and Spanish (+0.14), while Dutch remains relatively stable. Interestingly, these changes align closely with performance gains observed in Table 2: languages with larger increases in drift also show larger improvements in accuracy and F1.

Step-wise analysis (Figure 2) reveals that code-switching begins early in the reasoning process and intensifies in intermediate steps. However, in multilingual models, drift decreases in later steps, suggesting that the model returns to the target language when producing the final answer.

This behavior suggests that models rely on English-centric reasoning templates to perform the task, even when generating outputs in other languages. The final output masks this effect by switching back to the input language, creating the illusion of fully multilingual reasoning.

Key Insight. Improved performance in multilingual settings is often achieved by reverting to English reasoning, revealing a hidden dependence on English-centric representations.

6.6 Category-Level Trends

Table 1 (By Category) shows that performance varies significantly across bias categories. Physical Appearance consistently yields higher accuracy and F1 scores, while Disability and SES remain more challenging.

These differences suggest that certain bias types are easier to detect due to clearer lexical cues, while others require more nuanced reasoning. Notably, multilingual training improves performance across all categories, but the relative difficulty ordering remains consistent.

Key Insight. Bias detection difficulty is category-dependent, and multilingual training improves performance without eliminating these differences.

Language	SFT-EN Drift (↓)	SFT-EN-TR Drift (↓)
TR	0.60	0.84
NL	0.77	0.79
ES	0.81	0.95

Table 3: Mean drift rates across languages

6.7 Model Class Differences

We observe systematic differences across model classes. Instruction-tuned short-answer models (e.g., Llama variants) exhibit more stable cross-lingual behavior but lower reasoning fidelity, while reasoning-native models (e.g., DeepSeek-R1) achieve stronger performance in English but degrade more sharply in lower-resource languages.

Notably, reasoning-native models tend to introduce additional stereotype-driven signals within intermediate reasoning steps, suggesting that explicit reasoning can amplify bias rather than mitigate it. Supporting quantitative comparisons are provided in Appendix A.9.

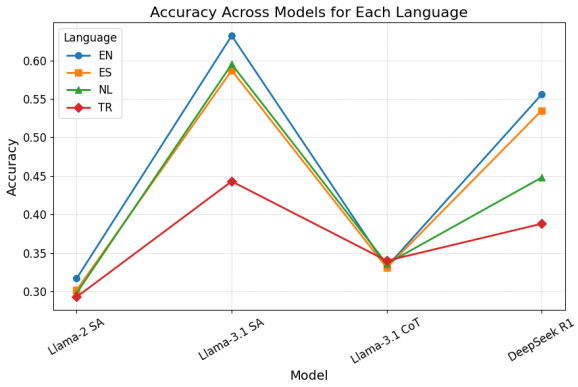

Figure 1: Trends in Accuracy Across Models for Each Language

6.8 Qualitative Analysis of Reasoning Failures

To complement quantitative results, we analyze model-generated reasoning traces and identify recurring failure modes that are not captured by aggregate metrics.

First, models frequently reinforce stereotypes through unsupported causal reasoning. In ambiguous contexts, reasoning traces often introduce implicit assumptions linking protected attributes (e.g., mental health status or socioeconomic background) to behavioral outcomes, even when such relationships are not entailed by the input.

Second, models exhibit limited sensitivity to implicit bias. In several cases, reasoning traces fail to recognize subtle or indirect forms of stereotyping,

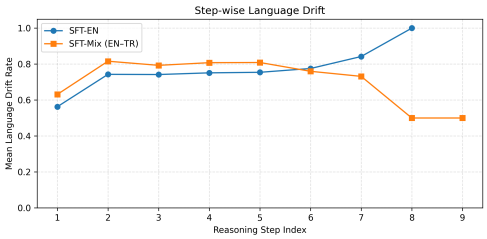

Figure 2: Mean language drift per reasoning step for SFT-EN vs. SFT-Mix (EN–TR). Lower values indicate better language alignment.

particularly when bias is expressed through contextual framing rather than explicit descriptors. This leads to false negatives despite seemingly coherent reasoning.

Third, we observe over-triggering behavior, where models classify inputs as biased based on surface-level cues such as emotionally charged language or informal phrasing, rather than substantive evidence. This suggests an over-reliance on lexical signals rather than deeper contextual understanding.

These failure modes are more pronounced in non-English languages, where reasoning traces tend to be more assertive and exhibit stronger lexical commitments. Overall, the qualitative analysis highlights that reasoning traces not only expose bias, but can also actively introduce and amplify it. We also observe that model-generated confidence scores are poorly calibrated and often uninformative, indicating that models lack reliable internal uncertainty estimates for bias-related reasoning. This further highlights the gap between fluent reasoning generation and actual epistemic awareness.

6.9 Summary

Across all experiments, we identify three consistent patterns. **First,** bias varies significantly across languages and is influenced by both linguistic structure and training data. **Second,** multilingual training improves performance but introduces trade-offs, particularly under preference optimization. **Third,** reasoning-level analysis reveals that models often rely on English-centric reasoning even in multilingual settings, highlighting a key limitation of current approaches.

7 Conclusion

Our findings have several implications for multilingual evaluation and alignment of large language models.

First, **English-centric evaluation is insufficient for assessing multilingual safety**. We observe substantial variation in both accuracy and bias across languages, indicating that English benchmarks systematically underestimate cross-lingual risks.

Second, our **results highlight limitations of current debiasing approaches**. While supervised fine-tuning benefits from multilingual exposure, preference-based optimization is brittle under cross-lingual transfer due to weak and inconsistent supervision signals. This suggests that alignment methods developed in monolingual settings may not generalize without explicit language-aware design.

Third, **reasoning-based analysis reveals a previously underexplored failure mode**: models often rely on English-centric reasoning templates even when operating in other languages. This hidden dependence is not visible in final outputs, but emerges through reasoning traces and is captured by our language drift metric. As a result, apparent multilingual performance may mask underlying reliance on a single dominant language.

Finally, our results suggest that reasoning introduces both opportunities and risks for bias evaluation. While chain-of-thought enables fine-grained analysis of model behavior, it also creates additional pathways for bias to emerge and propagate. This underscores the need for evaluation frameworks that jointly consider outputs and intermediate reasoning.

Overall, these findings point toward the importance of culturally and linguistically grounded approaches to evaluation and alignment, as well as the need for methods that explicitly model cross-lingual consistency in reasoning.

Acknowledgements

This work was completed as part of the Advanced Natural Language Processing course (11-711, Fall 2025) at Carnegie Mellon University. We thank Professor Sean Welleck for his guidance, feedback, and mentorship throughout the project.

Limitations

While our work provides a cross-lingual analysis of reasoning-level debiasing on MBBQ, several limitations remain. We evaluate four languages, which offers meaningful coverage but does not capture the full linguistic diversity of MBBQ. Our multilingual CoT traces come from a single teacher model

and may reflect its stylistic preferences, leaving room for future work to incorporate more diverse reasoning sources. Finally, because MBBQ is a multiple-choice benchmark and we do not include human evaluation or larger model scales, our findings should be interpreted as controlled empirical evidence rather than a complete characterization of real-world bias behavior.

Ethical Considerations

Our work raises ethical considerations regarding data quality, cultural representation, and deployment. MBBQ and teacher-generated multilingual CoT traces may contain historical or culturally specific biases, leading to incomplete representations of sensitive groups, especially in low-resource languages. Using multiple-choice questions to measure bias overlooks complex real-world harms like contextual stereotyping or microaggressions. While our method reduces measurable bias under controlled settings, it does not produce "bias-free" models. Future work should include human-centered evaluations, culturally informed annotation, and caution in high-stakes applications.

References

Divij Bajaj, Yuanyuan Lei, Jonathan Tong, and Ruihong Huang. 2024. Evaluating gender bias of LLMs in making morality judgements. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15804–15818, Miami, Florida, USA. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhiting Fan, Ruizhe Chen, and Zuozhu Liu. 2025. Biasguard: A reasoning-enhanced bias detection tool for large language models. *Preprint*, arXiv:2504.21299.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *Preprint*, arXiv:2403.07691.

Sharon Levy, Neha John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. 2023. Comparing biases and the impact of multilingual training across multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10260–10280, Singapore. Association for Computational Linguistics.

Viktor Mihaylov and Aleksandar Shtedritski. 2024. What an elegant bridge: Multilingual llms are biased similarly in different languages. *Preprint*, arXiv:2407.09704.

Mistral AI. 2025. Magistral. *arXiv preprint arXiv:2506.10910*.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global Voices, local biases: Socio-cultural prejudices across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15828–15845, Singapore. Association for Computational Linguistics.

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.

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.

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

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. *Preprint*, arXiv:2203.02155.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *Preprint*, arXiv:2110.08193.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Shaina Raza, Ananya Raval, and Veronica Chatrath. 2024. MBIAS: Mitigating bias in large language models while retaining context. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 97–111, Bangkok, Thailand. Association for Computational Linguistics.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yue Wan, Xiaowei Jia, and Xiang Lorraine Li. 2025. Unveiling confirmation bias in chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3788–3804, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. 2024. DeCoT: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14073–14087, Bangkok, Thailand. Association for Computational Linguistics.

Xuyang Wu, Jinming Nian, Ting-Ruen Wei, Zhiqiang Tao, Hsin-Tai Wu, and Yi Fang. 2025. Evaluating social biases in llm reasoning. *Preprint*, arXiv:2502.15361.

Xinyi Yang, Runzhe Zhan, Shu Yang, Junchao Wu, Lidia S. Chao, and Derek F. Wong. 2025. Rethinking prompt-based debiasing in large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26538–26553, Vienna, Austria. Association for Computational Linguistics.

Zheng-Xin Yong, Beyza Ermis, Marzieh Fadaee, Stephen H. Bach, and Julia Kreutzer. 2025. The state of multilingual llm safety research: From measuring the language gap to mitigating it. *Preprint*, arXiv:2505.24119.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages. *Preprint*, arXiv:2403.00277.

A Appendix

A.1 Model Selection

We use *deepseek-ai/DeepSeek-R1-Distill-Qwen-14B* as both teacher and student model. To justify this choice, we evaluate model scaling behavior on the GPQA benchmark.

As shown in Table 4, scaling from 7B to 14B yields a substantial improvement (+10%), while the gain from 14B to 32B is comparatively small (+3%).

A 32B-parameter model also exceeds the memory capacity of a single 40 GB GPU. At 16-bit precision, parameter storage alone requires:

$$32\text{B} \times 2 \text{ bytes} = 64\text{GB},$$

excluding optimizer states, gradients, and activations.

Therefore, the 14B model provides the best trade-off between performance and computational feasibility.

Model	GPQA Accuracy (%)
DeepSeek-7B	49.1
DeepSeek-14B	59.1
DeepSeek-32B	62.1

Table 4: Scaling performance of DeepSeek models on GPQA.

A.2 Implementation Details

All models are implemented using `Transformers`, `Datasets`, `PEFT`, and `vLLM`.

Supervised Fine-Tuning. We train using a learning rate of 2×10^{-4} for 3 epochs. Training uses a per-device batch size of 1 with gradient accumulation steps of 4 (effective batch size of 4). We apply 100 warmup steps and use the `paged_adamw_8bit` optimizer with FP16 precision.

LoRA is configured with rank 16, $\alpha = 32$, and dropout 0.05. The maximum sequence length is 2048 tokens, with a maximum prompt length of 1024.

Preference Optimization. DPO and ORPO models are trained for 2 epochs with a learning rate of 2×10^{-4}. For DPO, we use $\beta = 0.1$ with a sigmoid-based loss. All models use the same optimizer and precision settings as SFT.

Model	EN			ES			NL			TR		
	Acc	BiasAmb	BiasDis	Acc	BiasAmb	BiasDis	Acc	BiasAmb	BiasDis	Acc	BiasAmb	BiasDis
Llama-2 SA	0.317	0.847	0.361	0.302	0.867	0.368	0.298	0.876	0.379	0.293	0.979	0.048
Llama-3.1 SA	**0.632**	0.673	0.426	**0.587**	0.686	0.432	**0.595**	0.622	0.469	**0.443**	0.737	0.376
DeepSeek R1	0.556	0.710	0.387	0.535	0.622	0.434	0.448	0.660	0.443	0.388	0.704	0.378

Table 5: BiasAmb and BiasDis across models and languages.

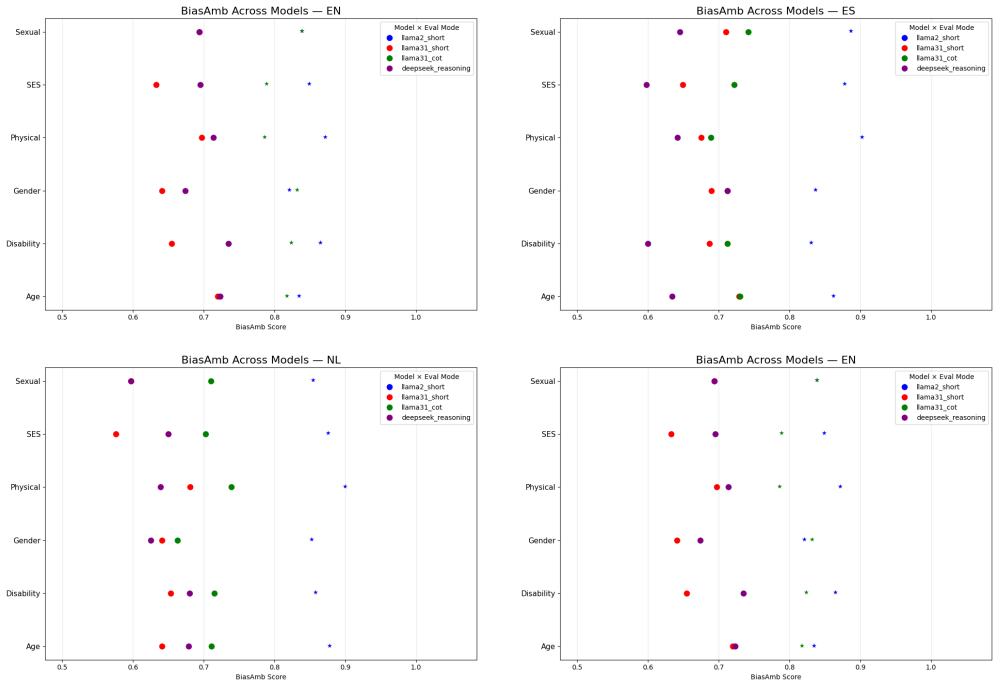

Figure 3: BiasAmb scores across models and languages. Higher values indicate stronger stereotype activation in ambiguous contexts.

Tokenizer. We use the tokenizer corresponding to *DeepSeek-R1-Distill-Qwen-14B*.

Figure 4: Training loss during preference optimization.

A.3 Preference Construction

For each input, we sample 20 candidate reasoning traces. Outputs are filtered based on:

1. correctness of the final label,
2. presence of a valid reasoning structure,
3. minimum reasoning length.

Among valid candidates, the highest-quality completion is selected as the preferred response, while remaining valid alternatives are treated as rejected responses.

A.4 DPO vs ORPO Behavior

DPO struggles in this setting due to weak preference signals. Chosen and rejected reasoning traces are often semantically similar, differing only in phrasing, which reduces effective learning signal. Additionally, rejected samples may contain incomplete reasoning, leading the model to penalize superficial artifacts rather than true errors. This results in unstable optimization and degraded performance.

ORPO mitigates this issue by integrating preference learning into the supervised objective. Its odds-ratio formulation applies softer penalties to dispreferred outputs, resulting in more stable training dynamics and smoother loss curves (Figure 4). This makes ORPO more robust in settings with

Section	Model	Evaluated	Accuracy	Correct	Skipped
Overall	SFT	485	0.515	250	8
	RL	484	0.531	257	9
By Language	EN (SFT)	116	0.776	90	1
	EN (RL)	112	0.643	72	5
	TR (SFT)	113	0.381	43	7
	TR (RL)	119	0.479	57	1
	NL (SFT)	128	0.391	50	0
	NL (RL)	125	0.440	55	3
	ES (SFT)	128	0.523	67	0
	ES (RL)	128	0.570	73	0

Table 6: Detailed breakdown of SFT vs RL performance across languages and categories.

overlapping reasoning traces.

A.5 Statistical Analysis

We evaluate cross-lingual differences using Kruskal–Wallis tests (Table 7). Results show that language identity has a highly significant effect on both metrics.

Table 7: Kruskal–Wallis test results for cross-lingual significance.

Metric	H-statistic	*p*-value
BiasAmb	224.01	2.72×10^{-48}
BiasDis	534.47	1.62×10^{-115}

A.6 Further Analysis (RQ2)

Across both DPO and ORPO, English performance tends to degrade slightly. Since English representations are already highly optimized, preference-based updates introduce relatively weak signals, increasing the risk of:

1. overfitting to preference data,
2. misalignment with the original task objective.

In contrast, lower-resource languages (Turkish, Dutch, Spanish) benefit more from preference-based updates due to stronger corrective signals.

A.7 Reasoning Drift Analysis

We analyze language drift in reasoning traces. Table 3 shows average drift rates.

Drift emerges early in reasoning and persists across intermediate steps, indicating reliance on English-centric reasoning templates.

A.8 Extended Error Analysis

We identify three dominant failure modes:

(1) Stereotype Completion in Ambiguous Contexts. In underspecified scenarios, models frequently infer demographic attributes or outcomes based on prior stereotypes, leading to systematic bias amplification. This behavior is strongly correlated with high BiasAmb scores.

(2) Cross-Lingual Reasoning Degradation. In non-English languages, models exhibit reduced logical coherence, including hallucinated attributes and inconsistent reasoning chains. This suggests weaker alignment between latent representations and surface forms.

(3) Surface-Level Over-Triggering. Models often classify inputs as biased based on lexical cues (e.g., emotional tone or informal phrasing) rather than contextual evidence, leading to false positives.

These failure modes are significantly more pronounced in Turkish and Dutch, aligning with lower accuracy and higher drift rates observed in the main results.

A.9 Bias Metrics Across Languages

We report BiasAmb and BiasDis scores across models and languages to complement accuracy-based evaluation.