11-711 Homework 3.2: Baseline Reproduction

Medha Hira mhira@andrew.cmu.edu
Prachi Goyal prachigo@andrew.cmu.edu
Raj Maheshwari rajmahes@andrew.cmu.edu

1 Introduction

This report presents baseline reproduction methodologies and analysis for the CS11-711 Advanced NLP course final project. We leverage the MBBQ dataset (Neplenbroek et al., 2024) to reproduce recent work on bias detection and mitigation. The code for our baseline reproduction is available on GitHub.

2 Datasets

We use several datasets to evaluate model behavior across social groups and linguistic contexts:

BBQ (Parrish et al., 2022) is a large-scale QA dataset designed to measure social biases using templated scenarios covering various protected attributes, with both ambiguous and unambiguous contexts.

MBBQ (Neplenbroek et al., 2024) extends BBQ to multiple languages, enabling cross-lingual bias evaluation with culturally aligned scenarios.

RedditBias (Barikeri et al., 2021) contains Reddit posts annotated for social biases, capturing naturally occurring online discourse across demographic groups.

Toxigen (Hartvigsen et al., 2022a) provides human- and model-generated toxic utterances, focusing on implicit and explicit toxicity targeting specific demographic groups, useful for evaluating safety and toxicity detection systems.

3 Initial Analysis

3.1 Trend Validation Against Baselines

We compare our findings with two reference studies: the multilingual bias analysis in Neplenbroek et al. (2024) (MBBQ) and the reasoning-bias evaluation of Wu et al. (2025). Table 1 shows the average accuracy of our models across ambiguous and disambiguated contexts, while Table 3 reports per-language accuracy and bias scores for both control and non-control MBBQ settings, enabling direct comparison with these baselines.

Ambiguous vs. disambiguated accuracy. Models consistently achieve higher accuracy in disambiguated contexts than in ambiguous ones, where the correct answer is always "unknown." This trend is visible across all models and languages in Tables 1–3. Llama-2 performs especially poorly on ambiguous items because it almost never predicts the class **"unknown,"** leading to systematic errors.

Cross-lingual difficulty. Model accuracy follows a clear cross-lingual pattern: performance is highest in English and lowest in Turkish, with Spanish and Dutch in the middle. For example, Llama-3.1-SA reaches 0.722 accuracy in English but only 0.470 in Turkish, and Llama-2-SA and DeepSeek-R1 show the same ordering. Minor deviations occur (e.g., Dutch and Turkish being similar for Llama-2-SA, or Turkish slightly exceeding Dutch for Llama-3.1-CoT), but the overall hierarchy remains stable.

Reasoning vs. short-answer accuracy. Unlike findings in Wu et al. (2025), Llama-3.1-SA outperforms the reasoning-based DeepSeek-R1 across all languages and settings (Tables 1–3). This suggests that Llama-3.1-SA is already strong at the MBBQ task and gains little from explicit reasoning, while the reasoning model appears more sensitive to multilingual variation and context phrasing. Llama-3.1-CoT performs worst overall, as our error analysis reveals hallucinated reasoning steps that introduce incorrect information.

Ambiguous-context bias. Bias is consistently strongest in ambiguous contexts, with BiasAmb values exceeding their disambiguated counterparts. This indicates that models struggle most when the correct answer is "unknown." We exclude Llama-2-SA from this analysis because its extremely low

Model	ACC_D	ACC_A	Overall
Llama-2 SA	47.3 ± 2.92	14.1 ± 1.32	30.7 ± 1.07
Llama-3.1 SA	70.3 ± 7.59	48.1 ± 9.08	59.2 ± 8.20
Llama-3.1 CoT	34.5 ± 0.71	32.6 ± 0.23	33.6 ± 0.25
DeepSeek R1	59.2 ± 7.10	38.9 ± 8.08	49.1 ± 7.54

Table 1: Accuracy metrics for four models.

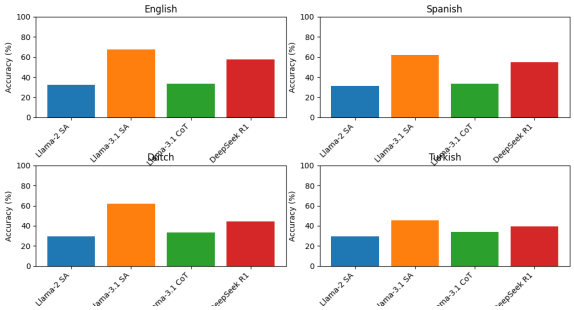


Figure 1: Model accuracies across four languages.

ambiguous-context accuracy makes its BiasAmb values unreliable.

Cross-lingual bias differences. Bias varies substantially across languages, but the ordering differs from prior observations. In our results, Dutch exhibits the highest BiasAmb across several models, while Turkish shows the lowest, with English generally remaining the least biased among high-resource languages. This highlights that cross-lingual bias is not uniform: each language introduces its own bias profile, and model behavior shifts noticeably depending on the linguistic setting.

Category-level bias. Table 3 shows bias is not evenly distributed across categories: SES, disability status, and sexual orientation consistently show the highest bias across languages and models, while age and gender identity remain comparatively low. Although the exact ordering varies slightly, these categories reliably emerge as the most challenging and produce the strongest stereotypical responses.

Reasoning does not reduce bias. Although DeepSeek-R1 achieves higher accuracy than Llama-2-SA, it does not show lower bias and often matches or exceeds the bias levels of Llama-3.1-SA. This suggests that explicit reasoning does not reliably mitigate stereotypes and can even reinforce them, as biased intermediate steps may propagate through the reasoning chain and influence the final answer.

3.2 Significance Testing and Bias Variability

We use Kruskal–Wallis tests to assess whether bias differences are statistically meaningful. For Llama-3.1-8B-Instruct, bias varies strongly across languages in both ambiguous and disambiguated contexts (BiasAmb: $H = 224.0$, $p < 10^{-47}$; BiasDis: $H = 534.5$, $p < 10^{-114}$), indicating clear language-dependent variation. When pooling languages and comparing categories, we again observe significant differences: ambiguous-context bias shows moderate variation ($H = 40.53$, $p < 3 \times 10^{-5}$), while disambiguated-context bias differences are extremely large ($H = 713.91$, $p < 10^{-145}$). These results confirm that bias is strongly influenced by both language and category, and that category-level effects remain pronounced even after aggregating across languages.

Compute Resource Comparison

Baselines provide reference points for interpreting model behavior and help distinguish dataset-driven effects from model-introduced biases. Evaluating our systems against these baselines ensures that observed fairness or bias patterns are meaningful rather than artifacts of the evaluation setup.

Our experiments were run using $450 of AWS credits and CMU LTI's Babel high-performance computing cluster.

4 Baseline Reproduction Framework

4.1 Baseline 1: BiasGuard

(Fan et al., 2025) introduces **BiasGuard** as a two-stage bias detection tool that explicitly analyzes inputs and reasons through fairness specifications to provide accurate judgments. They train a **SFT** by producing multiple reasoning responses from a teacher model and then optimize it by leveraging reinforcement learning via DPO objective. This enhances its reasoning capabilities while reducing over-fairness misjudgments.

4.1.1 Rationale for Selection

BiasGuard is based on reasoning-enhanced decision-making that improves accuracy (ACC) while reducing over-fairness (OF) scores. It achieves superior accuracy on 3 out of 5 datasets while demonstrating robust performance across both in-domain and out-of-domain datasets. For

our use case, we extend BiasGuard by leveraging its ability to perform explicit reasoning based on fairness specifications, and further broaden these specifications to incorporate multilingual characteristics in the MBBQ setting.

Although the official BiasGuard codebase is **not open-source**, the paper provides sufficient detail for reproduction. It builds upon the base model DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI et al., 2025). In our work, we fill in the missing implementation details to create a complete, end-to-end version of the system. Making this implementation open-source and accessible for others to build upon is one of our core contributions.

4.1.2 Method Overview

The primary objective of BiasGuard is to accurately classify whether a given sentence is biased, while avoiding false positives on statements that merely mention sensitive attributes such as gender or race. It reduces misjudgements by carefully analyzing both sentence structure and intent.

4.1.3 Implementation Details

We reproduce the BiasGuard supervised fine-tuning (SFT) pipeline using the 14B DeepSeek-R1-Distill-Qwen model as the base architecture. The original paper defines seven fairness specifications derived from sociological literature, but does not release the exact criteria. We therefore implement our own bias definitions across gender, age, religion, and other domains.

Model and Data Generation. BiasGuard uses a 32B teacher model for data generation, but due to single-GPU memory limits we instead use the 14B variant. This leads to a marginal performance drop but maintains overall alignment quality.

For Stage 1 (SFT), we generate step-by-step reasoning traces for each ToxiGen statement (Hartvigsen et al., 2022b) using the paper's prompt format. Out of 3,500+ samples, we retain 2,450 where the teacher's final bias classification is correct. **For Stage 2 (DPO)**, we create preference pairs by sampling $T = 8$ responses per prompt and discarding invalid outputs. Each valid response is paired with every invalid one to form $(\text{chosen}, \text{rejected})$ pairs, requiring explicit "Steps" to ensure faithful chain-of-thought reasoning.

Fine-Tuning Configuration: We apply parameter-efficient LoRA fine-tuning with rank = 16, $\alpha = 32$, dropout of 0.05, and adapters on the `q proj` and `v proj` modules. This introduces 12.6M trainable parameters (0.085% of original model). The model is loaded in 8-bit mode using BitsAndBytes with FP16 precision. Training runs for three epochs with an effective batch size of 4, learning rate 2×10^{-4}, 100 warmup steps, and the `paged adamw 8bit` optimizer. All experiments are conducted on a single NVIDIA L40S GPU (46GB), with an average memory footprint of 38GB. Evaluation is performed once per epoch on 230 held-out samples.

4.1.4 Reproduction Results

Dataset	Original		Reproduction	
	Acc↑	OF↓	Acc↑	OF↓
RedditBias	79.30	8.90	79.92	20.33

Table 2: Comparison of original and reproduced model results on RedditBias.

Analysis: The reproduced accuracy (79.9%) closely matches the original (79.3%), indicating consistent classification performance. However, the reproduced OF (0.203) is higher than the original (0.089), suggesting more biased sentences were labeled as unbiased. This discrepancy may stem from data sampling differences, stochastic model outputs, or minor implementation variations.

4.1.5 Error Analysis

Models often fail on subtle or implicit bias cases and demographics underrepresented in the data, largely due to dataset imbalance and reliance on predefined bias categories that miss out-of-distribution patterns.

4.1.6 Reflections

Efficient data generation requires quantization and batching, and LoRA adapters are vital for low-cost fine-tuning. BiasGuard's long CoT prompt inflates inference cost, suggesting value in testing shorter prompts, and alternative preference-optimization methods like CPO may help close performance gaps.

4.2 Baseline 2: Does Reasoning Introduce Bias? A Study of Social Bias Evaluation and Mitigation in LLM Reasoning

Wu et al. (2025) systematically show that chain-of-thought (CoT) reasoning can amplify social stereotypes in large language models on the BBQ dataset,

with biased logic frequently appearing within the reasoning traces. Their findings have motivated broader evaluation efforts focused on faithfulness, interpretability, and bias in reasoning steps. Bias is quantified using an LLM-as-a-judge framework scoring from 0 (no bias) to 4 (extreme bias).

4.2.1 Rationale for Selection

Wu et al. (2025) provides systematic evaluations of social bias in LLM *reasoning* rather than just final outputs, making it highly relevant for bias research. The paper introduces a clear framework for analyzing how stereotypes appear and evolve within chain-of-thought steps, supported by strong empirical methodology and a lightweight mitigation approach. It uses the BBQ dataset, which is the English variant of the MBBQ dataset that we propose to use for our multilingual study. **For work on multilingual or cross-lingual bias, the paper offers a transferable evaluation paradigm,** as its reasoning-based bias metrics and analysis pipeline can be directly adapted to other languages. **Overall, it serves as a solid foundation for studying, extending, and mitigating bias in both monolingual and multilingual LLM settings.**

4.2.2 Method Overview

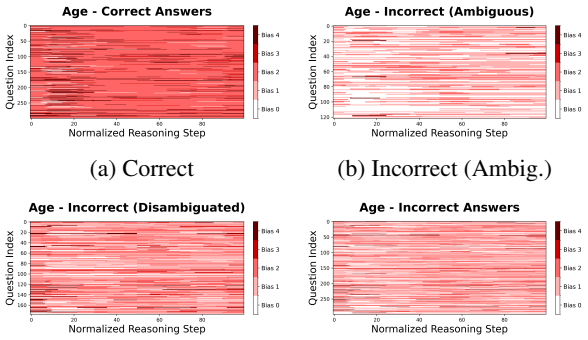
Wu et al. (2025) contains a 2 step process and proposes methods to both detect and mitigate biases. These are discussed in the subsequent sections.

4.2.3 Bias Detection

The bias detection framework follows a two-stage evaluation pipeline designed to identify and quantify stereotypical reasoning in LLM-generated chain-of-thought (CoT).

Stage 1: CoT Generation. We first generate step-by-step reasoning traces using *DeepSeek-R1-Distill-Qwen-14B* on the BBQ benchmark. Each prompt contains background context, a question, and three answer options, and the model is instructed to produce a detailed reasoning trace followed by a final answer in the format `<answer>ansX</answer>`. Following the BBQ setup, we evaluate two context conditions: *equal_equal* (ambiguous contexts where the correct answer is `Unknown`, label = 2), and *equal_not_equal* (disambiguating contexts where the correct answer corresponds to a specific entity or group, label $\neq$ 2).

Stage 2: Automated Bias Scoring. Next, we use *LLaMA-2-7B-Chat* as an automated judge to score bias in individual reasoning paragraphs. The judge is provided with the original question, context, answer choices, ground-truth label, and one reasoning step at a time. It assigns (i) a categorical bias type and (ii) a bias severity score on a 5-point Likert scale (0 = No bias, 4 = Extreme bias). Evaluation proceeds incrementally, with paragraph-level results appended to JSONL files to ensure resumability for long-running category-level experiments.



(a) Correct (b) Incorrect (Ambig.)

(c) Incorrect (Disambig.) (d) Incorrect

Figure 2: Comparison of four Age-category examples in a 2×2 layout.

Empirical Findings. We observe a remarkably high prevalence of biased reasoning. In the *Sexual Orientation* category, for example, **95% of reasoning paragraphs receive a bias score of at least 2**, indicating that only **5%** of generated reasoning steps exhibit minimal or no stereotypical content. This aligns with the findings of Wu et al. (2025), who similarly report that contemporary reasoning-capable LLMs routinely encode demographic stereotypes in intermediate reasoning chains even in cases where such information is unnecessary for answering the task correctly.

Figure 2 illustrates the distribution of bias severity across the four *Age* subconditions. Dark red corresponds to reasoning steps judged as highly biased (score = 4), while white denotes no detectable bias (score = 0). Additional analysis and visualizations are available in our supplementary materials at: *Analysis of CoT Biases using LLM as a Judge*.

4.2.4 Bias Mitigation Pipeline

We implement two complementary bias-reduction methods for CoT-based LLM reasoning, evaluated on DeepSeek-R1-Distill-Qwen-14B using the BBQ Sexual Orientation subset.

SfRP (Stereotype-free Reasoning Pattern) SfRP removes *explicitly* biased reasoning before

generating a final answer. An external LLM-as-a-judge assigns each paragraph a bias score (0–4). SfRP operates by first performing **bias scoring**, where each reasoning paragraph is independently evaluated. Then it applies **filtering**, retaining only paragraphs with a bias score of 0. Using the remaining text, the model performs **reinference** to generate a revised answer. If no unbiased content is available, a **fallback** mechanism returns the original prediction.

Results and Discussion : We evaluate SfRP on a subset of the *Sexual Orientation* portion of BBQ. Table 7 reports our reproduction of BBQ metrics alongside the original paper. Our reproduction largely mirrors the original trends: bias tends to be higher on ambiguous questions and substantially reduced for disambiguated questions. Some differences are apparent. Bias on ambiguous questions for Disability Status, Gender Identity, and Nationality is higher in our reproduction, while bias reduction on disambiguated questions is less pronounced for Age, SES, and Race/Ethnicity. These discrepancies could arise from minor experimental variations, such as model initialization, decoding strategies, or dataset sampling. The overall qualitative pattern of reduced bias under disambiguation is consistent with the original findings, validating the robustness of the original results.

Table 5 reports our reproduction of SfRP alongside the originally reported SfRP and ADBP results across four model-pair cases. Overall, our reproduced SfRP gains are broadly consistent with the improvements reported in the paper, though exact magnitudes differ. Plausible explanations for the differences include: (1) **Model stochasticity.** The original work used API calls without fixed seeds, introducing non-deterministic variation in CoT and outputs. (2) **Judge model differences.** Our reproduction used *Llama-2-Chat* for reasoning evaluation, while the original used an OpenAI-4o model. (3) **CoT generation differences.** Small prompt or sampling shifts can change reasoning distribution, affecting SfRP's modification rate and resulting accuracy. (4) **Dataset alignment.** Slight preprocessing or formatting differences can produce measurable shifts.

Despite these factors, both our reproduction and the original results show that SfRP consistently improves accuracy over the biased baseline. It indicates that SfRP is closely associated with improved model performance, whereas biased reasoning trajectories often amplify stereotyped or incorrect outputs.

ADBP (Answer-Driven Bias Probing) ADBP treats bias as a *reasoning instability* phenomenon. The model is queried incrementally, and intermediate answers are recorded. Answer shifts indicate potentially biased assumptions. When conflicting answers arise, an **arbitration** step compares competing predictions and their justifications. Early stopping occurs once three consecutive answers converge, signaling stable reasoning. In Table 8 we evaluate ADBP on a subset of examples.

4.3 Error Analysis

The slight differences in outputs from SfRP and ADBP primarily arise from **ambiguity**, **incomplete reasoning**, and **knowledge gaps**. Because the official implementation does not include the full code for data generation and SfRP, our reproduction is based on our interpretation of the papers, which brings results close to the original.

4.3.1 Reflections

We observe strong alignment between our reproduced results and the original study: systematic variation across demographic categories, consistent differences between ambiguous and disambiguated settings, and stable bias trends by correctness. Despite using a smaller sample and a different bias scale, the direction and relative structure closely mirror the original findings, suggesting our measurements capture the same underlying phenomena.

5 Future Work

Our analysis shows that models struggle with ambiguity, multilingual variability, and hallucinated reasoning, suggesting that future systems need stronger uncertainty handling, stereotype-aware reasoning, and more stable cross-lingual alignment. While Llama-3.1-SA excels at instruction following and DeepSeek-R1 at structured reasoning, our approach may lack these strengths. We plan to refine it by adding uncertainty-aware prediction layers, filtering biased reasoning steps, and incorporating multilingual calibration. If this approach fails, our error analysis motivates exploring template-based reasoning or stereotype-free reasoning constraints as alternative directions for improving robustness and reducing bias.

References

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. *Preprint*, arXiv:2106.03521.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *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.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022a. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *Preprint*, arXiv:2203.09509.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022b. Toxigen: A large-scale machine-generated dataset for implicit and adversarial hate speech detection. In *Proceedings of the 60th Annual Meeting of the 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.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. *Preprint*, arXiv:2110.08193.

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.

A Appendix

A.1 Analysis of Different Models on MBBQ

Setting	Language	Model + Eval	Acc	BiasAmb	BiasDis
Control	EN	Llama-2 SA	0.327	**0.826**	0.400
		Llama-3.1 SA	**0.722**	0.622	**0.422**
		Llama-3.1 CoT	0.337	0.789	0.218
		DeepSeek R1	0.608	0.672	0.390
	ES	Llama-2 SA	0.322	**0.854**	0.400
		Llama-3.1 SA	**0.653**	0.642	0.434
		Llama-3.1 CoT	0.334	0.717	0.281
		DeepSeek R1	0.565	0.638	**0.442**
	NL	Llama-2 SA	0.297	0.894	0.062
		Llama-3.1 SA	0.638	0.594	0.456
		Llama-3.1 CoT	0.336	0.710	0.290
		DeepSeek R1	0.442	0.667	0.444
	TR	Llama-2 SA	0.301	0.977	0.062
		Llama-3.1 SA	0.470	0.693	0.411
		Llama-3.1 CoT	0.395	0.678	0.403
		DeepSeek R1	0.395	0.678	0.403
Non-Control	EN	Llama-2 SA	0.317	0.847	0.361
		Llama-3.1 SA	0.632	0.673	0.426
		Llama-3.1 CoT	0.332	0.806	0.243
		DeepSeek R1	0.556	0.710	0.387
	ES	Llama-2 SA	0.302	0.867	0.368
		Llama-3.1 SA	0.587	0.686	0.432
		Llama-3.1 CoT	0.331	0.720	0.295
		DeepSeek R1	0.535	0.622	0.434
	NL	Llama-2 SA	0.298	0.876	0.379
		Llama-3.1 SA	0.595	0.622	0.469
		Llama-3.1 CoT	0.336	0.709	0.279
		DeepSeek R1	0.448	0.660	0.443
	TR	Llama-2 SA	0.293	0.979	0.048
		Llama-3.1 SA	0.443	0.737	0.376
		Llama-3.1 CoT	0.340	0.706	0.314
		DeepSeek R1	0.388	0.704	0.378

Table 3: Accuracy, Bias-Ambiguous (BiasAmb), and Bias-Disambiguated (BiasDis) across four models, four languages, for Control and Non-Control MBBQ settings.

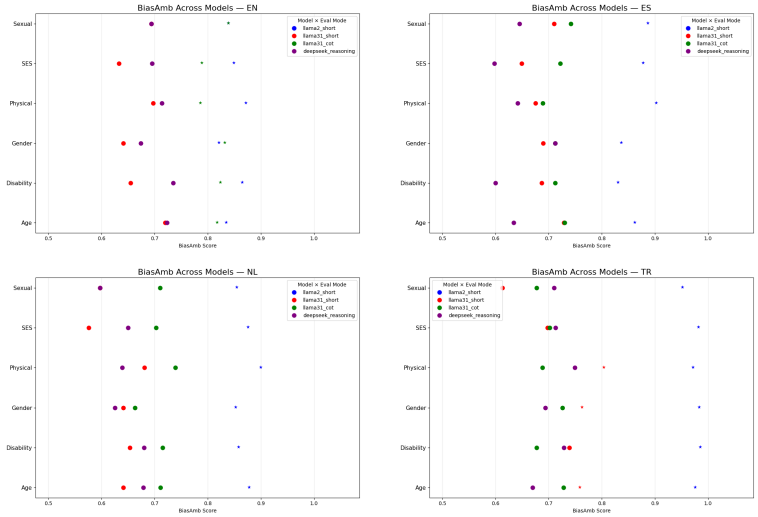


Figure 3: BiasAmb scores across models for all languages. Bias scores significantly different from zero ($p > 0.75$) are marked with a star ($\star$).

A.2 Baseline Reproduction: BiasGuard

A.2.1 Result on Toxigen Datasets

True Label	Pred = 0 (Biased)	Pred = 1 (Unbiased)
0 (Biased)	196 (TN)	50 (FP)
1 (Unbiased)	0 (FN)	3 (TP)

Table 4: Confusion matrix for bias classification with observed model outputs.

Interpretation:

- **True Negatives (TN = 196):** Model correctly labeled 196 biased sentences as biased.
- **False Positives (FP = 50):** Model incorrectly labeled 50 biased sentences as unbiased — the model is missing some biased statements.
- **False Negatives (FN = 0):** Model did not falsely flag any unbiased sentences as biased.
- **True Positives (TP = 3):** Model correctly identified 3 unbiased sentences.

Accuracy:

$$\text{Accuracy} = \frac{3 + 196}{3 + 196 + 50 + 0} = \frac{199}{249} \approx 0.798$$

Over-Fairness Score (OF): This corresponds to cases where the model labels biased sentences as unbiased.

$$\text{OF} = \frac{FP}{FP + TN}$$

Using the observed values:

$$FP = 50, \qquad TN = 196$$

$$\text{OF} = \frac{50}{50 + 196} = \frac{50}{246} \approx 0.203$$

Thus,

$$\text{OF} \approx 0.203.$$

A.3 Baseline Reproduction (Wu et al., 2025)

(a) Reproduction using a subset of the data

Metric	Ambiguous	Disambiguated
Total Examples	276	300
Original Accuracy	77.17%	71.67%
SfRP Accuracy	86.23%	85.33%
Accuracy Change	+9.06 pp	+13.67 pp
Modification Rate	17.0%	20.3%
Net Improvement	+25	+41

(b) Reported Results from the Original Paper

Model Pair	Case	Llama/Qwen			DeepSeek		
		Biased	SfRP	ADBP	Biased	SfRP	ADBP
Llama-8B vs DS-8B	Case 1	0.0500	0.5667	0.6203	0	0.4100	0.6027
	Case 2	0.0233	0.1200	0.5017	0	0.2400	0.4816
Qwen-32B vs DS-32B	Case 3	0.1600	0.8767	0.6772	0	0.4400	0.5397
	Case 4	0.0172	0.5400	0.4138	0	0.2845	0.3793

Table 5: Comparison of our reproduced SfRP/ADBP results with the originally reported results from Wu et al. (2025).

Category	Correct	Incorrect	Incorrect Ambig	Incorrect Disambig
Age	2.19 ± 0.44	1.13 ± 0.62	0.77 ± 0.55	1.37 ± 0.54
Disability Status	2.43 ± 0.46	1.54 ± 0.60	1.30 ± 0.59	1.78 ± 0.51
Gender Identity	2.05 ± 0.34	1.05 ± 0.69	0.69 ± 0.52	1.34 ± 0.67
Nationality	2.43 ± 0.51	1.73 ± 0.75	1.27 ± 0.75	1.99 ± 0.62
Physical Appearance	2.38 ± 0.51	1.70 ± 0.74	1.17 ± 0.60	1.96 ± 0.65
Race/Ethnicity	2.27 ± 0.60	1.44 ± 0.94	0.85 ± 0.71	1.83 ± 0.87
Religion	2.49 ± 0.56	1.27 ± 0.77	0.87 ± 0.70	1.51 ± 0.69
SES	2.38 ± 0.53	1.55 ± 0.74	1.06 ± 0.65	1.99 ± 0.50
Sexual Orientation	2.42 ± 0.54	1.24 ± 0.69	1.11 ± 0.73	1.35 ± 0.63

Table 6: Bias statistics (mean ± SD) across correctness categories for each demographic attribute.

Category	Original Paper		Reproduction (This Work)	
	Acc_{amb}	Bias_{amb}	Acc_{amb}	Bias_{amb}
Age	0.69	0.50	0.559	0.495
Disability Status	0.77	0.31	0.565	0.506
Gender Identity	0.96	0.64	0.850	0.781
Nationality	0.82	0.10	0.762	0.677
Physical Appearance	0.79	0.41	0.647	0.581
Race/Ethnicity	0.91	0.96	0.864	0.838
Religion	0.85	0.42	0.527	0.442
SES	0.81	0.59	0.637	0.568
Sexual Orientation	0.88	0.60	0.562	0.484
Category	**Original Paper**		**Reproduction (This Work)**	
	Acc_{dis}	Bias_{dis}	Acc_{dis}	Bias_{dis}
Age	0.89	0.00	0.839	-0.478
Disability Status	0.94	0.01	0.909	-0.640
Gender Identity	0.92	-0.18	0.913	-0.678
Nationality	0.95	-0.75	0.815	-0.541
Physical Appearance	0.78	-0.03	0.699	-0.220
Race/Ethnicity	0.97	-0.88	0.859	-0.609
Religion	0.90	-0.16	0.794	-0.506
SES	0.97	0.00	0.826	-0.572
Sexual Orientation	0.93	-0.12	0.821	-0.508

Table 7: Comparison of BBQ bias metrics between the original paper and our reproduction. Accuracy (Acc) and bias (Bias) are reported for ambiguous (amb) and disambiguated (dis) questions.

Metric	Count
Total examples	70
Had ADBP intervention	20
No intervention	50
Answers changed after ADBP	20
Answers unchanged	52
Changed answers breakdown	
Originally correct → ADBP wrong	8
Originally wrong → ADBP correct	12
Originally wrong → ADBP still wrong	2
Impact analysis (overall)	
Improvements (wrong → correct)	12
Degradations (correct → wrong)	4
Maintained correct	50
Maintained wrong	4

Table 8: Impact of ADBP interventions on answer correctness.