Evaluation and Mitigation of Social Biases in Various Domains via Question-Answering in Large Language Models

Aarushi Wagh and Saniya Srivastava
Georgia Institute of Technology

Abstract

Large Language Models (LLMs) are increasingly used in sensitive domains like finance, healthcare, and hiring, but they often replicate social biases from their training data. In this work, we evaluate and mitigate such biases using the BBQ dataset, focusing on Race/Ethnicity and Sexual Orientation. We begin by assessing off-the-shelf Falcon-7B and LLaMA-3B models, revealing both accuracy limitations and stereotype-aligned tendencies. To address this, we fine-tune Falcon-7B with Low-Rank Adaptation (LoRA) for efficient QA performance, and further introduce an adversarial training setup—DebiasFalcon—which uses a gradient reversal layer and group classifier to suppress bias-indicative features in the model's hidden states. Our results show that DebiasFalcon improves both accuracy and fairness, especially in ambiguous cases, with significantly better calibration around uncertainty. These findings demonstrate that even under limited data and compute, simple adversarial objectives can enhance fairness in LLMs without compromising utility. github.com/ssrivastava22/social-bias-eval-in-llms

1 Introduction and Motivation

Large Language Models (LLMs) such as LLaMA 2 and Falcon have become central to a growing number of real-world applications that rely on natural language understanding and generation. These include high-stakes domains like healthcare, finance, education, and hiring, where the accuracy and fairness of AI-generated responses can have significant societal impacts. While LLMs are powerful and versatile, they inherit patterns from their training data—including social biases—that can manifest in subtle but harmful ways. These biases may relate to protected attributes such as race, gender, sexual orientation, age, religion, or socioeconomic status, and their presence in model outputs risks reinforcing stereotypes, perpetuating discrimination, and undermining trust in AI systems.

The problem addressed in this project is the systematic evaluation and mitigation of social biases in LLMs within a structured question-answering (QA) context. Unlike open-ended generation tasks, structured QA provides clearly defined inputs and expected outputs, making it easier to assess whether a model's answer aligns with an unbiased ground truth. The project specifically investigates how models respond to socially sensitive questions that vary across protected attributes and whether these responses deviate from what would be considered fair or correct.

Using the Bias Benchmark for Question Answering (BBQ), we first establish zero-shot baselines for Falcon-7B and LLaMA 3B on questions pertaining to race and sexual orientation. We then fine-tune Falcon-7B using Low-Rank Adaptation (LoRA) to efficiently learn a QA-specific head. Finally, we introduce an adversarial debiasing framework, DebiasFalcon, that leverages a gradient reversal layer and an auxiliary group classifier to suppress identity-specific information in the model's latent representations.

Our results show that adversarial training not only improves fairness metrics, such as uncertainty calibration and stereotype avoidance, but also yields modest gains in QA accuracy. These findings highlight the promise of lightweight architectural interventions for building socially responsible LLMs under constrained training conditions.

2 Related work

The proliferation of Large Language Models (LLMs) has brought to light concerns regarding their propensity to reflect and amplify societal biases present in their training data. Studies have demonstrated that these models can inadvertently perpetuate stereotypes across various dimensions, including race, gender, and socioeconomic status.

For instance, (Bender et al., 2021) discuss the ethical implications of large-scale language models and their potential to reinforce existing societal biases.

To systematically evaluate such biases, several benchmark datasets have been developed. The Bias Benchmark for Question Answering (BBQ), introduced by (Parrish et al., 2022), is a notable example. BBQ comprises question sets designed to assess model biases across nine social dimensions, providing both ambiguous and disambiguated contexts to test model responses. This dataset has become a standard tool for measuring bias in QA systems.

Building upon BBQ, (Jin et al., 2024) developed the Korean Bias Benchmark for Question Answering (KoBBQ), adapting the original dataset to the Korean language and cultural context. This adaptation underscores the necessity of culturally relevant benchmarks in assessing biases in multilingual models.

Recent studies have applied BBQ to evaluate biases in state-of-the-art LLMs. (Liu et al., 2025) extended the BBQ dataset to include fill-in-the-blank and short-answer questions, facilitating the assessment of biases in open-ended settings. Their findings indicate that LLMs exhibit varying degrees of bias across different protected attributes, with notable biases related to age and socioeconomic status.

In terms of bias mitigation, several strategies have been explored. (Elazar and Goldberg, 2018) proposed an adversarial approach to remove demographic information from text representations, aiming to debias NLP models. Additionally, self-correction techniques have been investigated as a means to reduce biases in LLM outputs. (Anantaprayoon et al., 2025) demonstrated that self-correction mechanisms could effectively mitigate social biases in models like GPT and LLaMA.

Despite these advancements, challenges remain in effectively identifying and mitigating biases within LLMs. The complexity of language and the nuanced nature of societal biases necessitate ongoing research and the development of more sophisticated evaluation and mitigation techniques.

3 Methodology

Our methodology is organized into three sequential phases, each contributing progressively to the understanding, quantification, and mitigation of social biases in large language models (LLMs). First, we evaluate the inherent bias present in off-the-shelf LLMs using a controlled multiple-choice QA benchmark. Second, we fine-tune a lightweight adaptation of Falcon-7B using parameter-efficient training to establish a strong QA baseline. Finally, we introduce an adversarial training framework that explicitly penalizes the encoding of protected attributes in intermediate representations, thereby promoting fairer decision-making.

3.1 Phase 1: Zero-Shot Evaluation of Pretrained LLMs

We begin by benchmarking the bias of unaltered LLMs using the Bias Benchmark for Question Answering (BBQ) (Parrish et al., 2022), a structured dataset designed to elicit stereotype-consistent behavior across nine protected attributes. For this phase, we focus on two axes of bias: `Raceethnicity.jsonl` and `Sexualorientation.jsonl`.

Each example in BBQ consists of:

- A `context` sentence involving individuals from specific social groups.
- A factual `question` related to the scenario.
- Three answer options: typically one biased, one unbiased, and one ambiguous.
- A `gold label` and metadata including `stereotyped_group` and `ambiguity`.

We evaluate two widely used open-source LLMs:

- **Falcon-7B-Instruct**, quantized to 4-bit via BitsAndBytes for efficient inference on constrained hardware. Generation parameters: `max_new_tokens=200`, `temperature=0.7`, `top_p=0.7`.
- **Open LLaMA 3B**, run in FP16 with `max_new_tokens=50`.

We adopt a standardized inference protocol:

1. Format each example into a structured prompt (see Section 4).
2. Run inference to generate textual output.
3. Extract the most likely answer (A/B/C) from the output.
4. Compare against the gold label to compute accuracy.

Predictions are stored in a structured JSONL format to support downstream bias analysis, including error consistency and stereotype agreement.

3.2 Phase 2: QA Fine-Tuning with LoRA Adaptation

To improve task performance while maintaining training efficiency, we fine-tune Falcon-7B using Low-Rank Adaptation (LoRA) (Hu et al., 2021). Our hypothesis is that even a small number of trainable parameters can significantly boost model alignment with gold labels on BBQ.

Model Architecture

```
class BaselineFalcon(nn.Module):
def init(self, base, num_labels=3):
super().init()
self.base = base
self.qa = nn.Linear(base.config.
    hidden_size, num_labels)
def forward(self, input_ids,
    attention_mask, labels=None):
out = self.base(input_ids,
    attention_mask, output_hidden_states
    =True,
use_cache=False, return_dict=True)
h_cls = out.hidden_states[-1][:, 0]
logits = self.qa(h_cls)
loss = F.cross_entropy(logits, labels)
    if labels is not None else None
return loss, logits
```

Details:

- *Backbone*: Falcon-7B with 4-bit NF4 quantization.
- *LoRA*: Injected into all `query_key_value` projections, with rank $r = 16$ and scaling factor $\alpha = 32$.
- *QA Head*: A full-rank projection from the pooled [CLS] representation to a 3-way softmax.

Due to hardware limitations, training is conducted exclusively on the **first 500 examples** from the `Raceethnicity.jsonl` split.

3.3 Phase 3: Adversarial Debiasing Framework

We hypothesize that group-specific representations in the encoder latent space are a major contributor to biased predictions. To suppress these signals, we construct DebiasFalcon, a dual-head model trained via adversarial objectives.

Model Architecture

```
class DebiasFalcon(nn.Module):
def init(self, base, n_answers=3,
    n_groups,   =0.1):
super().init()
self.base = base
self.   =
self.qa = nn.Linear(base.config.
    hidden_size, n_answers)
self.adversary = nn.Sequential(
nn.Linear(base.config.hidden_size, base.
    config.hidden_size // 2),
nn.ReLU(),
nn.Linear(base.config.hidden_size // 2,
    n_groups)
)
```

The model contains two outputs:

- A **QA head** trained to predict the correct answer index.
- An **adversarial head** trained to predict the protected group ID from the [CLS] embedding, passed through a gradient reversal layer (GRL).

Gradient Reversal Layer (GRL):

```
class GRL(torch.autograd.Function):
@staticmethod
def forward(ctx, x,   ):
ctx.   =
return x
@staticmethod
def backward(ctx, grad_output):
return -ctx.   * grad_output, None
```

Loss Function:

- $\mathcal{L}QA$: Cross-entropy loss on the correct answer index.
- $\mathcal{L}adv$: Cross-entropy loss on the protected group ID.
- λ: Controls the trade-off between task accuracy and fairness.

$$\mathcal{L}QA - \lambda\mathcal{L}adv$$

The adversarial head attempts to infer group membership, while the GRL reverses gradients to suppress group leakage in the encoder. The shared [CLS] representation is thus optimized to retain task-relevant features while minimizing identity-encoding capacity.

3.4 Post-Mitigation Evaluation

To evaluate the effectiveness of mitigation, we apply the debiased model on a subset of 500 examples and report:

- **QA Accuracy**: Standard accuracy on disambiguated items.
- **Bias Reduction**: Fewer incorrect answers aligned with stereotypes on ambiguous items.
- **Invariance**: Robustness under identity-swapped contexts.

All predictions are stored in JSONL format with fields: `prompt`, `response`, `predicted`, `correct`, `is_correct`.

4 Experimental Setup

Hardware and Environment

- **Platform:** Google Colab (Free Tier)
- **GPU:** NVIDIA T4 (16 GB VRAM)
- **CPU:** 2 vCPUs, 12 GB RAM
- **Frameworks:** PyTorch 2.0, HuggingFace Transformers , PEFT, BitsAndBytes 0.39.0

All training and evaluation was conducted using a constrained Colab environment. Both baseline and adversarial models were trained on **only 500 examples** from `Raceethnicity.jsonl` for 5 epochs due to compute restrictions.

Training Hyperparameters

Parameter	Value
Training Samples	500 (Race/Ethnicity only)
Epochs	5
Batch Size	4
Gradient Accumulation	4
Learning Rate	0.01
Optimizer	AdamW
Scheduler	Cosine (warmup: 5%)
Precision	FP16
Max Sequence Length	256
λ (GRL strength)	0.10

Table 1: Training Hyperparameters

Evaluation Protocol

- **Validation**: Held-out 10
- **Output Format**: JSONL including raw prompt, model response, predicted index, and correctness.
- **Bias Metrics**: Analyzed model responses and its rate of choosing "uncertain" options for ambiguous prompts

An example of what the prompt and answer looks like:

Listing 1: Example QA Prompt and Model Response from BBQ

```
"prompt": "Context: The police stopped a
    Hispanic man and a Native American
   man the other day just outside the
   arena.
   Question: Who likely committed a
       crime?
   Options:
       A. The Native American man
       B. Not answerable
       C. The Hispanic man
   Answer:",
"response": "B. Not answerable",
"predicted": 1,
"correct": 1,
"is_correct": true
```

5 Results

Model	Data	Accuracy (%)
Falcon 7B	Race/Ethnicity	**32.80**
Falcon 7B	Sexual Orientation	34.02
LLaMA 3B	Race/Ethnicity	30.60
LLaMA 3B	Sexual Orientation	**34.20**

Table 2: Accuracy of Models on BBQ Subsets with Basic Prompting

Model	Data	Accuracy (%)
Falcon 7B	Race/Ethnicity	**32.20**
Falcon 7B	Sexual Orientation	33.00
LLaMA 3B	Race/Ethnicity	36.20
LLaMA 3B	Sexual Orientation	**33.60**

Table 3: Accuracy of Models on BBQ Subsets with Improved Prompting

As seen from the results, Falcon performs better on the race/ethnicity subset and LLaMA performs better on the sexual orientation subset. Since these scores are arguably at the level of random guessing, we decided to try additional prompting strategies to see if the accuracies would improve. For Falcon, an instruct model, we provided a brief instructional preamble clearly defining the task and the expectation of a single answer choice selection. For LLaMa, being a base model, we provided some few shot examples directly from the dataset into the input prompt. These new prompting strategies did not result in any notable improvements:

Both models demonstrate relatively low accuracy across these protected attributes with basic and improved prompting strategies, highlighting a significant degree of social bias in their outputs. These findings highlight the urgent need for targeted interventions to reduce bias and improve the ethical reliability of large language models in socially sensitive contexts.

To quantify the effects of parameter-efficient training and adversarial debiasing, we evaluate three models on the `Raceethnicity.jsonl` subset of the BBQ dataset:

1. Off-the-shelf **Falcon-7B-Instruct** (zero-shot).
2. **Falcon-7B + LoRA** fine-tuned for QA.
3. **DebiasFalcon** trained with adversarial loss.

5.1 Overall Accuracy

Model	Accuracy (%)
Falcon-7B (Zero-shot)	32.20
Falcon-7B + LoRA (QA-only)	32.60
DebiasFalcon (Adversarial)	34.60

Table 4: Overall QA accuracy across 500 examples from the Race/Ethnicity subset.

Despite limited training data, the LoRA-tuned QA model achieves a marginal improvement over the zero-shot baseline. The DebiasFalcon model further improves accuracy by over 2%, suggesting that adversarial supervision does not sacrifice task performance and may even enhance generalization.

5.2 Uncertainty Handling

To assess how well models handle ambiguous cases, we analyze their behavior when an “uncertain” option (e.g., “Can’t be determined”) is present and correct.

The adversarially trained DebiasFalcon demonstrates a clear improvement in uncertainty calibration. It selects the “uncertain” response 41.27% of the time—substantially higher than the 25.79% baseline—and misses fewer such cases when it is the correct answer (29.37% vs. 36.51%). This suggests that debiasing encourages the model to rely less on stereotypical priors and more on contextual cues when no group-specific inference is justified.

Model	Rate Chosen "Uncertain" correctly (%)	Rate Missed "Uncertain" (%)
Falcon-7B	25.79	36.51
Falcon + LoRA	28.57	35.71
DebiasFalcon	41.27	29.37

Table 5: Model behavior on 252 ambiguous examples with an “uncertain” answer choice.

5.3 Qualitative Summary

Overall, our results show:

- LoRA-based fine-tuning improves QA accuracy marginally over zero-shot.
- Adversarial training further improves both accuracy and fairness without additional model parameters.
- Debiasing significantly enhances the model’s ability to abstain from group-based assumptions when uncertainty is warranted.

These findings validate the efficacy of our adversarial objective in reducing representational bias while preserving downstream QA utility.

5.4 Demographic Breakdown of Model Behavior

Beyond overall accuracy, we also conducted a fine-grained analysis of the zero-shot model responses by disaggregating answer selections by the protected group (e.g., "The gay man", "The Hispanic man", etc.). In this analysis we consolidated semantically equivalent answer variants (e.g., “The gay man”, “The gay one”, and “The gay friend” all mapped to `The gay`).

This analysis revealed important behavioral trends. For example, models disproportionately predicted group-associated answers such as `The Black man` or `The gay` even when uncertainty options like `Not answerable` were available. In the Race/Ethnicity evaluation, the models selected `The Black man` as the top answer in 5 out of 9 group contexts. Similarly, in the Sexual Orientation evaluation, `The gay` and `The lesbian` were repeatedly selected as top answers across unrelated groups, suggesting stereotype-consistent response bias.

These patterns help show which specific identities are most affected by model bias. They also highlight that some identity terms serve as default or prediction options even when incorrect. Full demographic breakdowns (by model, by dataset) are available in the Appendix.

Protected Group	**Accuracy (%)**	**Most Selected Identity (Normalized)**
Hispanic	32.14	The Hispanic/Latino man (16), The Black man (10)
Native American	29.17	The Native American man (16)
Black	27.50	The Black man (46)
Gay	31.25	The gay (28)
Lesbian one	44.64	The lesbian (22), Uncertain (20)
Bisexual friend	40.54	The bisexual (23), The gay (15)

Table 6: Sample of Normalized Answer Trends by Group

6 Conclusion and Summary

This work presents a targeted investigation into the social biases exhibited by large language models (LLMs), with a particular focus on multiple-choice question answering (QA) tasks grounded in socially sensitive contexts. Using the BBQ dataset as a diagnostic benchmark, we evaluated both off-the-shelf and fine-tuned models on their ability to respond fairly and accurately to prompts involving protected demographic attributes.

Our findings reveal that even state-of-the-art LLMs such as Falcon-7B, when evaluated in a zero-shot setting, exhibit measurable bias and uncertainty miscalibration. Parameter-efficient fine-tuning using LoRA yields marginal improvements in QA performance but does not fundamentally alter group-specific behavior.

To address this, we introduced an adversarial training pipeline, DebiasFalcon, which integrates a gradient reversal mechanism and an auxiliary adversarial loss to discourage the model from encoding group-identifiable features in its latent space. Despite being trained on a limited subset of just 500 samples, DebiasFalcon achieves higher QA accuracy and a significantly improved handling of ambiguous, stereotype-sensitive prompts. The model demonstrates an enhanced ability to select the “uncertain” option when warranted and avoids defaulting to biased inferences.

In summary, this work offers empirical evidence that:

- Adversarial supervision can improve both fairness and task performance in LLMs.
- Fairness gains are achievable even under low-resource constraints using simple architectural modifications.
- Structured benchmarks such as BBQ, when paired with targeted interventions, offer a valuable framework for auditing and mitigating representational harms in deployed models.

Future Work. Building on the initial success of our approach, future work will focus on: (1) training the models further using additional epochs and larger subsets of the BBQ dataset to evaluate whether performance and fairness continue to improve with more exposure, and (2) expanding the debiasing framework to other identity axes within BBQ, including *Sexual Orientation*, to test the generalizability and robustness of our method across multiple protected attributes.

References

Panatchakorn Anantaprayoon, Masahiro Kaneko, and Naoaki Okazaki. 2025. Intent-aware self-correction for mitigating social biases in large language models. *Preprint*, arXiv:2503.06011.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. Kobbq: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Zhao Liu, Tian Xie, and Xueru Zhang. 2025. Evaluating and mitigating social bias for large language models in open-ended settings. *Preprint*, arXiv:2412.06134.

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.

Appendix: Full Answer Selection Statistics by Protected Group

This appendix contains the normalized answer distribution per protected group. Similar responses have been merged (e.g., `The gay man`, `The gay one`, and `The gay friend` are all labeled as `The gay`). Uncertainty responses such as `Can't be determined`, `Unknown`, `Not answerable`, and ambiguous letter options are grouped under `Uncertain`.

Falcon 7B – Race/Ethnicity

Protected Group	Accuracy	Uncertainty Rate	Top Selected Answers
Hispanic	35.71%	54.55%	Hispanic man (12), Uncertain (20), Others: 4 each
Native American	37.50%	60.00%	Uncertain (12), Native American man (4), Others: 4 each
African	27.50%	10.00%	Black man (28), Hispanic man (4), Asian man (4), Uncertain (4)
White	34.38%	50.00%	White man (12), Uncertain (12), Hispanic man (4)
European	32.69%	27.27%	Black man (15), European man (13), Latino man (8), Uncertain (12)
Middle Eastern	37.50%	60.00%	Black man (4), Hispanic man (4), Uncertain (12), Middle Eastern man (4)
Black	35.00%	42.31%	Black man (47), Uncertain (44), Jewish man (6), Others: 4-6
South American	34.38%	37.50%	South American man (10), Black man (10), Uncertain (12)
Jewish	25.00%	0.00%	Black man (10), Jewish man (7), Asian man/woman (7)

Table 7: Falcon 7B – Normalized Answers on Race/Ethnicity

Falcon 7B – Sexual Orientation

Protected Group	Accuracy	Uncertainty Rate	Top Selected Answers
Gay	33.78%	46.77%	Gay (42), Uncertain (58), Bisexual (22), Straight (4)
Lesbian	38.00%	31.11%	Lesbian (28), Uncertain (28), Gay (12), Bisexual (15)
Straight	30.30%	57.14%	Straight (18), Uncertain (34), Gay (9)
Bisexual	35.96%	42.55%	Bisexual (58), Uncertain (40), Lesbian (9)
Pansexual	34.72%	23.33%	Gay (21), Pansexual (11), Lesbian (10), Bisexual (8), Uncertain (18)

Table 8: Falcon 7B – Normalized Answers on Sexual Orientation

LLaMA 3B – Race/Ethnicity

Protected Group	Accuracy	Uncertainty Rate	Top Selected Answers
Hispanic	32.14%	6.82%	Hispanic man (16), Black man (10), White/European man (8), Native American (8), Uncertain (10)
Native American	29.17%	30.00%	Native American man (16), Uncertain (7), Black man (1)
African	40.00%	30.00%	Black man (17), Asian man (8), Uncertain (12)
White	31.25%	8.33%	White/European man (16), Asian man (6), Uncertain (6)
European	32.69%	13.64%	White/European man (19), Black man (13), Uncertain (6)
Middle Eastern	25.00%	0.00%	Black man (12), Hispanic man (7), Middle Eastern man (5)
Black	27.50%	5.77%	Black man (46), Jewish man (10), White/European (8), Asian (8), Uncertain (6)
South American	28.12%	0.00%	South American man (16), Black man (6), Hispanic man (4), Uncertain (1)
Jewish	32.14%	12.50%	Jewish man (13), Black man (6), Asian person (5), Uncertain (4)

Table 9: LLaMA 3B – Normalized Answers on Race/Ethnicity

LLaMA 3B – Sexual Orientation

Protected Group	Accuracy	Uncertainty Rate	Top Selected Answers
Gay	40.63%	12.77%	gay man (69), bisexual man (15), unknown (12), lesbian woman (5)
Lesbian	36.96%	13.64%	lesbian woman (66), bisexual man (17), unknown (14), gay man (7)
Straight	35.29%	3.85%	straight man (35), gay man (14), bisexual man (9), unknown (5)
Bisexual	39.34%	11.11%	bisexual man (54), gay man (23), unknown (14), lesbian woman (5)
Pansexual	45.76%	5.97%	pansexual man (46), bisexual man (18), gay man (15), lesbian woman (8)

Table 10: LLaMA – Normalized Answers on Sexual Orientation