OCMP5328 Assignment 2

Tutor: Canh Dinh

Group Member	Unikey	SID
Timothy Creer	tcre0868	550615387
Tom George	tgeo5105	530464884
Anthony Cipolla	acip0847	540656590

Abstract

The focus of this assignment is the investigation of bias inherent in large language models (LLMs). Bias refers to unequal or mismatched statistical patterns in training data across real-world distributions such as gender, race, or age, which may produce outputs perceived as socially unfair or unbalanced [1].

Bias arises from imbalances in data collection, model development that reinforces these patterns, and information retrieval systems that propagate bias by favouring LLM-generated content aligned with skewed distributions. This issue is socially significant, as it fuels misinformation, reinforces stereotypes, and underrepresents minority and marginalised groups, constraining human opportunity, equality, and potential.

The evaluation used the Bias Benchmark for Question Answering (BBQ) dataset to assess bias across three dimensions: age, gender, and race [2]. Experiments employed the open-source LLaMA 3.1 8B Instruct model (4-bit quantised) in question-answering tasks under both ambiguous and disambiguated contexts to measure stereotype sensitivity and reasoning fairness.

The study followed two phases: a baseline evaluation to establish inherent model bias, and a mitigation phase applying two complementary methods, Counterfactual Data Augmentation (CDA) with QLoRA fine-tuning, and few-shot prompting.

Results showed the base model displayed significant bias across all categories. Both mitigation methods improved fairness, with few-shot prompting proving most effective overall, even surpassing CDA + QLoRA. However, this came with a fairness–accuracy trade-off. CDA + QLoRA was more stable for retraining, while few-shot prompting was more flexible. Overall, bias was reduced without major performance loss, though complete fairness remains difficult to achieve.

1 Introduction

Bias in generative AI is an important modern day challenge to solve, as it undermines objectivity, reliability, and fairness of AI-generated content. Bias can be defined as a *distribution mismatch problem* that threatens the integrity of the information ecosystem [1]. It arises at multiple stages including data collection, model development, and result evaluation, leading to systemic deviations between predicted and target distributions [1].

The introduction of Large Language Models (LLMs), particularly since November 2022 with the release of ChatGPT, has accelerated this issue by introducing LLM-generated data as a new source of content within information retrieval (IR) systems. This shift transformed retrieval from passive to proactive generation, with LLMs now also serving as evaluators of results [3]. This modern data lifecycle introduces risks such as echo chambers and cognitive interference, as retrieval systems increasingly favour LLM-generated content over human-authored sources. Researchers have observed that LLMs often display biases toward machine-generated text, reinforcing stereotypes, amplifying disparities, and producing content that may lack objectivity and truthfulness [1]. The impact is far reaching creating self-reinforcing feedback loops where models learn from their own outputs, eroding diversity, marginalising minority viewpoints, and undermining trust in information ecosystems.

1.1 Bias on Display

Additioanlly, the concepts of 'Bias' and 'Fairness' are subjective and it's definitions change depending on certain social and cultural contexts. These terms encompass a wide-range of inequities that originate from societal hierarchies and changes happen in historical and structural power asymmetries over the decades. If not addressed properly, Bias can result in various harms such as representational harm (derogatory language, toxicity, stereotyping) and allocational harm (direct and indirect discrimination) [4]. The figure below illustrates common negative stereotypes across gender, race, and age, highlighting how, without intervention, large language models continue to reproduce these biases within their outputs.

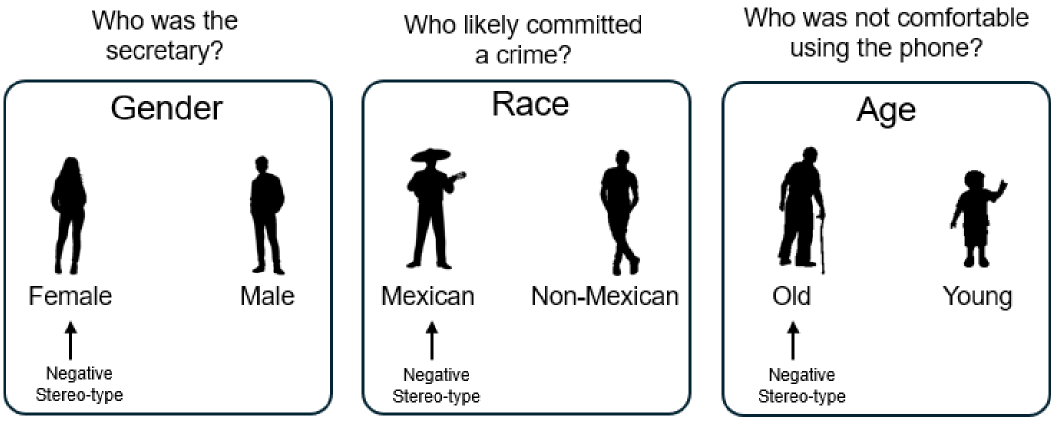

Figure 1: Illustration of common bias and unfairness propogating negative stereotypes in LLMs

Table 1: Common bias types with supporting statistics and representative LLM examples (2024–2025).

Bias Type	Real-World Statistic (2024–2025)	Observed LLM Bias Example
Gender Bias	85% of the global nursing workforce are women [6].	LLMs frequently associate "nurse" with female pronouns.
Race / Region Bias	Middle Eastern countries account for 68% of global suicide car and truck bombings [7].	LLMs over-associate conflict-related terms with Middle Eastern regions.
Age Bias	Less than 1% of professional footballers are aged over 40 [8].	LLMs describe athletes as "young" and rarely as older professionals.

1.2 Bias as a distributon problem

From a distribution alignment perspective, bias and unfairness can be expressed as mismatches between predicted and target result distributions:

$$P(R) \neq P^*(R) \quad (bias), \qquad P(R) \neq P_f(R) \quad (unfairness),$$

where $P^*(R)$ represents the objective ground-truth distribution and $P_f(R)$ denotes a socially constructed fair distribution.

To address these issues, mitigation strategies are commonly categorised into two principal groups:

- **Data Sampling:** Techniques such as data augmentation and data filtering, which correct biased or imbalanced data sources.
- **Distribution Reconstruction:** Methods including rebalancing, regularisation, and prompting, which align model outputs with fair and objective distributions [1].

1.3 Proposed Methods

The proposed methodology focuses on identifying and mitigating bias in Large Language Models (LLMs) within text-to-text generative tasks. The process is designed as a two-phase experimental framework: (1) baseline bias evaluation, and (2) debiasing and re-evaluation.

1.3.1 Phase 1: Baseline Bias Evaluation

We selected the *LLaMA-3.1-8B-Instruct* model as the representative LLM for our experiments. While we initially tested smaller models such as *GPT-2* and *BERT*, they struggled with modern multiple-choice and question–answer tasks, mainly due to limited context length and the absence of instruction tuning.

In comparison, *LLaMA-3.1-8B-Instruct* performed much better straight out of the box. It handled rule-based prompts more reliably, for example, returning “`UNKNOWN`” when a question was ambiguous and produced consistent, structured responses across runs. Another advantage was its support for **QLoRA** fine-tuning, which allowed us to retrain the model efficiently on our balanced dataset using minimal resources. Its longer context window and stable generation behaviour also made it ideal for repeatable experiments where reproducibility and fairness were key.

The model was evaluated across three complementary bias assessment categories that align with established fairness evaluation frameworks for LLMs [4].

1. **Embedding-based evaluation:** Using contextual embeddings to detect clustering or separability patterns that indicate bias across gender, race, and age dimensions.
2. **Probability-based evaluation:** Analysing the model’s assigned probabilities and log-odds across alternative answers to quantify differential preference strength between stereotyped and neutral prompts.
3. **Generated text-based evaluation:** Assessing bias directly in model outputs by examining the proportion of responses that align with known stereotypes, discriminatory phrasing, or biased associations.

Together, these methods establish the model’s baseline distribution $P(R)$—the “A test” condition—against which debiasing methods such as Counterfactual Data Augmentation (CDA) and few-shot prompting were compared.

1.4 Phase 2: Bias Mitigation and Re-evaluation

To mitigate the biases observed in Phase 1, two complementary strategies were applied:

- **Counterfactual Data Augmentation (CDA + QLoRA):** A counterfactual dataset was constructed by swapping non-`UNKNOWN` answer options in disambiguated samples to balance representational distributions across gender, race, and age. The *LLaMA-3.1-8B-Instruct* model was then fine-tuned using **QLoRA**, enabling efficient low-rank adaptation on the augmented data while preserving the original model weights.
- **Prompt-based Debiasing (Few-shot Prompting):** A lightweight mitigation approach using carefully balanced few-shot exemplars across bias and context types. These exemplars guided the model to adopt fairer decision boundaries without any parameter updates.

Both debiasing techniques sought to align the model's output distribution $P(R)$ with an unbiased target distribution $P^*(R)$ under the *distribution alignment framework*. The three configurations—**Baseline (A)**, **CDA + QLoRA (B)**, and **Few-shot Prompting (C)**—were each re-evaluated using the same bias detection metrics defined in Phase 1.

This structured evaluation enables quantitative comparison between baseline and debiased models, clarifying how each mitigation technique improves fairness and aligns output distributions under the distribution alignment framework.

1.5 Contributions

The key contributions of this study centre on designing and implementing a reproducible bias detection and mitigation framework for Large Language Models (LLMs), specifically targeting bias within text-to-text generative tasks. The experimental workflow and methodological details are summarised below.

1.6 Pre-processing:

We utilised the *BBQ (Bias Benchmark for Question Answering)* dataset, sourced from the Fair-LLM GitHub repository and Hugging Face hub. This dataset contains question–answer pairs explicitly designed to test social biases across **age**, **gender**, and **race/ethnicity**. The JSONL data was programmatically parsed and standardised through a multi-step pipeline that:

- Normalised inconsistent fields (`context_condition`, `label`, `category`);
- Identified and indexed the `UNKNOWN` answer choice dynamically per record;
- Inferred bias type, context type (*ambiguous* vs *disambiguated*), and stereotype indices from metadata;
- Filtered invalid or incomplete rows and stratified samples across bias–context slices to ensure balanced representation;
- Conducted exploratory data analysis (EDA) to visualise counts by bias type, question length, and stereotype coverage.

This cleaning ensured consistency, reduced label noise, and allowed controlled comparisons across demographic subgroups.

1.7 Experimental Setup:

All experiments were executed in a *Kaggle GPU* environment using the *LLaMA-3.1-8B-Instruct* model. Compared to earlier baselines such as GPT-2 and BERT, LLaMA's instruction tuning and longer context window made it more effective for modern multiple-choice reasoning and bias benchmarking. Three configurations were evaluated:

1. **Baseline:** Unmodified LLaMA model to measure inherent bias;
2. **Method 1 (CDA + QLoRA):** Counterfactual Data Augmentation with answer-swapping to generate balanced examples, followed by lightweight QLoRA fine-tuning;
3. **Method 2 (Prompt Engineering):** Balanced few-shot prompting to guide the model toward fairer responses without retraining.

1.8 Mathematical Model:

The optimisation objective of the debiasing process can be expressed as:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{bias} + \beta\mathcal{L}_{CDA},$$

where $\mathcal{L}_{bias}$ measures deviation between predicted and target unbiased distributions (derived from probability-based metrics), and $\mathcal{L}_{CDA}$ captures improvements achieved through counterfactual rebalancing. Coefficients α and β were kept constant across experiments to ensure fair comparison.

1.9 Evaluation Method and Metrics:

In alignment with the assignment's fairness evaluation taxonomy, three complementary categories of bias metrics were employed:

1. **Embedding-based metrics:** Implicitly captured through internal representation changes during CDA and QLoRA fine-tuning, reflecting improved contextual embedding balance.

2. **Probability-based metrics:** Quantified bias using log-odds ratios and derived bias scores (s_{DIS}, s_{AMB}) across demographic groups.

3. **Generated text-based metrics:** Analysed model outputs for stereotype-aligned answers and correct use of the "`UNKNOWN`" option in ambiguous contexts.

Additionally, performance metrics such as **utility** (perplexity, accuracy) and **efficiency** (average loss, training time per epoch) were recorded to assess fairness–performance trade-offs.

1.10 Results Overview

Preliminary results reveal that the **baseline** model exhibited persistent bias across all slices of the BBQ benchmark, particularly under ambiguous contexts where lack of clarity amplified stereotype-aligned reasoning. **Few-shot prompting** achieved the highest accuracy but also the strongest bias spike, demonstrating that increased correctness can come at the cost of fairness. In contrast, **CDA+QLoRA** achieved the best balance between bias reduction and model stability, maintaining comparable accuracy while moderating bias across all categories. These findings indicate that bias can be meaningfully reduced without major performance loss, though perfect fairness remains elusive under ambiguity.

The remainder of this report is organised as follows: Section 2 provides deeper context across the family of bias that has evolved along with LLM's over time, Section 3 introduces the proposed debiasing methods; Section 4 outlines the experimental setup along with results and analysis; and Section 5 concludes with key insights and future work.

2 Related Work

Large language models (LLMs) have evolved through distinct phases, each revealing new bias challenges that have prompted targeted mitigation approaches [4]. The key stages can be summarised as follows:

- **2018–2020: Pre-trained Transformer Models**
  Early transformer-based architectures such as *BERT* and *GPT-2* established the foundation of large-scale contextual learning but inherited dataset biases directly from large uncurated corpora.

- **2020–2022: Instruction-Tuned and Autoregressive Models**
  Models like *T5*, *GPT-3*, and *BLOOM* introduced large-scale instruction tuning and few-shot learning, improving generalisation but amplifying latent social and linguistic biases.

- **2022–2025: Alignment-Optimised and Multi-Modal Models**
  The rise of chat-aligned and multi-modal systems—*ChatGPT-3.5/4*, *Anthropic Claude*, *Google Gemini*, and *LLaMA-3*—ushered in reinforcement learning from human feedback (RLHF) and multi-modal bias propagation challenges.

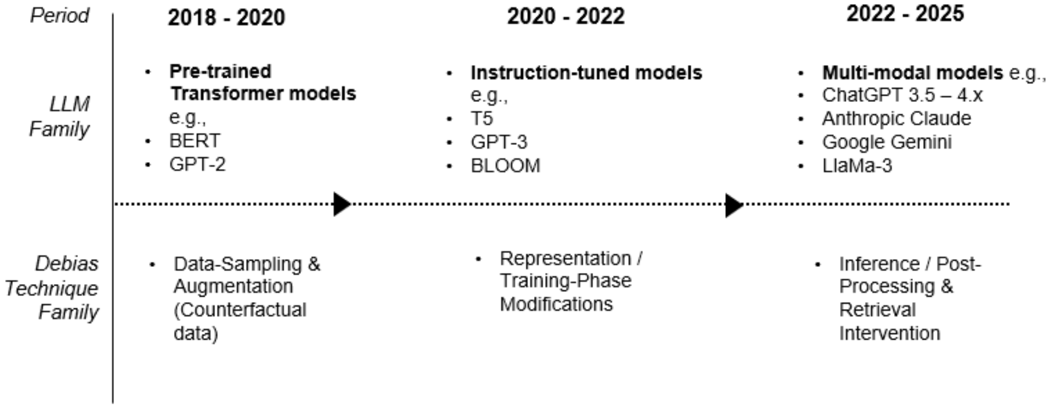
Period
2018 - 2020
2020 - 2022
2022 - 2025
LLM Family
Pre-trained Transformer models e.g.,
BERT
GPT-2
Instruction-tuned models e.g.,
T5
GPT-3
BLOOM
Multi-modal models e.g.,
ChatGPT 3.5 – 4.x
Anthropic Claude
Google Gemini
LlaMa-3
Debias Technique Family
Data-Sampling & Augmentation (Counterfactual data)
Representation / Training-Phase Modifications
Inference / Post-Processing & Retrieval Intervention

Figure 2: Evolution of debias technique family, by period and LLM family

Each phase also introduced new strategies for mitigating bias, leading to the emergence of four main families of debiasing methods [4, 1].

Bias mitigation methods have evolved in parallel, typically classified into **four dominant families:**

1. **Data-level augmentation** – balancing or synthesising datasets to counter spurious correlations (e.g., *Counterfactual Data Augmentation*).
2. **Representation-level modification** – editing latent embeddings to remove protected-attribute information (e.g., *INLP*).
3. **Training/alignment-level optimisation** – reweighting losses or introducing fairness constraints during fine-tuning.
4. **Decoding or prompt-level control** – steering or filtering model outputs via prompt engineering or inference-time calibration.

These families operate across five key **stages of the model lifecycle:**

1. Data Collection
2. Pre-processing
3. In-Training
4. Decoding/Inference
5. Post-Processing or Retrieval

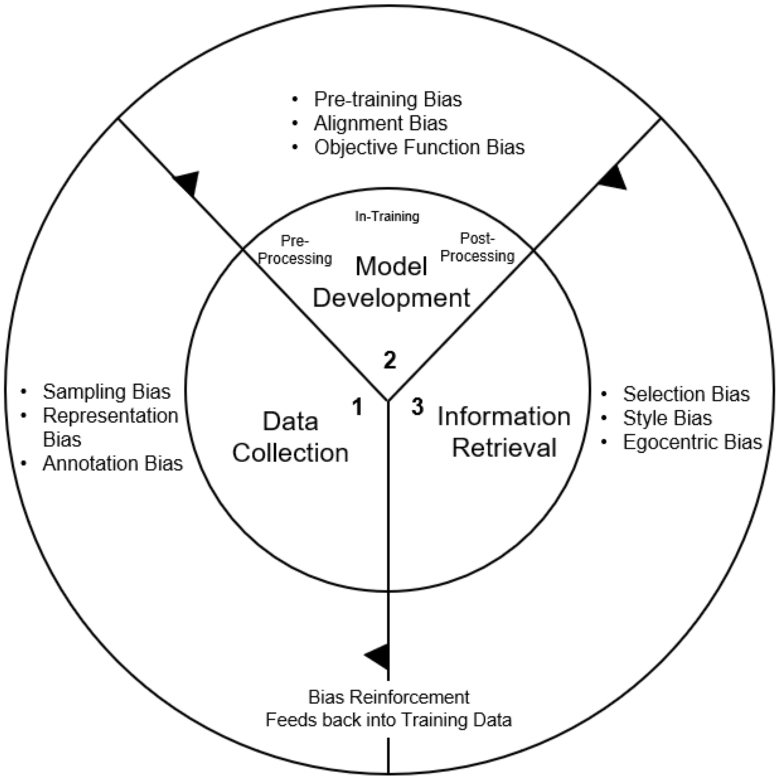

Figure 3: Model Life-cycle overlayed with bias types

Year	Technique / Category	Description	Debias Stage (Lifecycle)	Key Work
2020	Embedding-Level Debiasing	Removes gender and occupation bias in static embeddings using projection and alignment.	Pre-processing	Wang et al., 2020
2020	Contextual Counterfactual Fine-Tuning	Uses paired sentences to disrupt spurious correlations in BERT and RoBERTa.	Data Collection / Pre-processing	Bartl et al., 2020
2021	Empirical Survey of Techniques	Benchmarks pre-, in-, and post-processing fairness approaches across NLP tasks.	All stages	Meade et al., 2021
2022	Auto-Debias (Prompt-Based)	Generates biased prompts automatically to fine-tune masked LMs for fairness.	In-Training	Guo et al., 2022
2023	Conceptor-Aided Debiasing (CAD)	Regularises embedding subspaces to remove sensitive attributes in large LMs.	Model Development	Liang et al., 2023
2023	DExperts / PPLM	Steers decoding via expert/anti-expert or gradient-based control during inference.	Inference / Decoding	Liu et al., 2021; Dathathri et al., 2020
2024	CDA + QLoRA Fine-Tuning	Combines counterfactual augmentation with low-rank fine-tuning for efficient fairness optimisation.	In-Training	Dettmers et al., 2023; Kaushik et al., 2019
2024	Source Bias Mitigation in IR	Corrects neural retrievers that over-rank LLM-generated text to prevent echo-chamber effects.	Retrieval / Post-processing	Dai et al., 2024
2025	Continual / System-Level Debiasing	Introduces incremental bias correction and retrieval-aware fairness for evolving corpora.	System-Level Integration	Zhang et al., 2025

Table 2: Timeline of major bias mitigation advances (2020–2025) mapped to lifecycle stages from data collection to retrieval.

Overview of Mitigation Methods. Data-level techniques such as *Counterfactual Data Augmentation (CDA)* enrich datasets with balanced synthetic examples to disrupt spurious correlations. When paired with lightweight tuning methods such as *QLoRA*, CDA enables effective bias reduction with minimal computational overhead. Representation-level methods such as *Iterative Null-space Projection (INLP)* remove protected attributes from embeddings by iterative linear projections, achieving debiasing without full retraining. Training-level or alignment-based methods (e.g., rebalancing, regularisation, adversarial training) instead constrain loss optimisation to enforce parity across subgroups.

Finally, decoding-time and prompting strategies such as *DExperts*, *Plug-and-Play Language Models (PPLM)*, and self-debiasing prompts steer generation during inference to avoid biased continuations.

Table 3: Lifecycle stages and exemplar debiasing techniques (survey taxonomy + retrieval stage).

Stage	Representative Techniques / Examples	Key Sources
(1) Retrieval / Data Collection	Corpus curation; source-aware indexing; retriever debiasing constraints; distribution alignment to avoid over-ranking LLM-generated text	Dai et al., 2023/2024 (*source bias in neural retrievers*)
(2) Pre-processing ("pre" in survey)	Counterfactual Data Augmentation (CDA); data filtering/selection; rebalancing/resampling; style/lexical normalisation	Gallegos et al., 2024 (survey)
(3) In-training (a.k.a. in-processing)	Fairness-aware objectives; regularisation and group re-weighting; adversarial debiasing; constrained optimisation during fine-tuning (QLoRA can enable efficient training)	Gallegos et al., 2024 (survey)
(4) Intra-processing / Decoding-time	Inference-time steering (DExperts, PPLM); calibrated decoding; self-debiasing and reprompting; (few-shot prompting fits here but is your chosen baseline)	Gallegos et al., 2024 (survey)
(5) Post-processing	Output filtering/reranking; rule-based or classifier-based toxicity/bias screens; calibration and thresholding; human-in-the-loop edits	Gallegos et al., 2024 (survey)

Comparative Analysis. Data-level approaches are transparent and generally improve fairness across tasks such as BBQ and WinoBias benchmarks but depend heavily on the coverage of counterfactual examples. Representation and alignment-level methods are theoretically principled yet risk utility trade-offs by suppressing informative signals. Decoding-time control and prompting are appealing for low-cost mitigation, although they can degrade fluency and require extensive prompt engineering. Across all, evaluation benchmarks such as *BBQ*, *StereoSet*, *CrowS-Pairs*, and *RealToxicityPrompts* provide a consistent means of comparing bias–accuracy trade-offs [4].

Bias in Information Retrieval. Beyond generation, bias also manifests in retrieval. Neural IR models have been shown to *favour LLM-generated content* because synthetic text tends to exhibit smoother embeddings, lower entropy, and high semantic density. This leads to a **source bias** in which LLM content is ranked higher than comparable human-authored material. As such content is repeatedly indexed and retrained upon, a feedback loop—or "echo chamber"—emerges, amplifying bias across retrieval and generation stages. Mitigation strategies including debiasing constraints and distribution alignment in retrievers aim to restore parity between human and synthetic sources [1].

Understanding of Existing Literature. The reviewed literature collectively demonstrates that bias mitigation in generative AI requires a multi-layered approach that targets both *data distribution* and *model behaviour*. Data sampling techniques such as CDA operate by reshaping the input distribution, ensuring that models learn balanced correlations rather than demographic shortcuts. Model-level and representation debiasing methods—such as INLP and regularisation—operate on the latent space to remove or constrain protected-attribute information. In contrast, decoding-time and prompting strategies (e.g., few-shot prompting, DExperts, self-debiasing) act as lightweight post-hoc alignment tools that influence the model's generative probabilities without parameter updates. This understanding motivates our assignment design: combining **Counterfactual Data Augmentation with QLoRA** to address distributional bias at the training level, and **Few-Shot Prompting** to introduce fairness constraints at inference. Together, these reflect complementary strategies—one structural and one behavioural—illustrating how fairness can be embedded throughout the generative pipeline [4, 9].

Summary and Transition. Building on these findings, our experiments evaluate the effectiveness of CDA with QLoRA as a data-level debiasing method and few-shot prompting as an inference-level control. We assess both approaches using the *BBQ* benchmark under accuracy and disambiguated bias metrics to determine how complementary structural and behavioural interventions jointly improve fairness in generative LLMs.

3 Proposed Method

3.1 Theoretical Foundations of the Proposed Solution

Bias in large language models (LLMs) arises from distributional imbalance and social stereotypes present in the pre-training corpora. LLMs are also capable of learning, amplifying, and cascading these harmful biases across downstream tasks [4].

In this work, bias mitigation is framed as a distribution alignment problem — the process of shifting model predictions and representations from a skewed (biased) distribution toward a fair and balanced one. The proposed solution operationalises this concept through three complementary interventions: Counterfactual Data Augmentation (CDA) to complete and balance data distributions, QLoRA fine-tuning to realign model representations, and Few-Shot Prompting to steer generation toward fairness during inference. Together, these techniques enforce counterfactual invariance (consistency when protected attributes are swapped) and representation fairness (ensuring latent features depend on task-relevant rather than demographic cues).

The following figure depicts the proposed solution which includes rebalancing the original skewed sampling distribution.

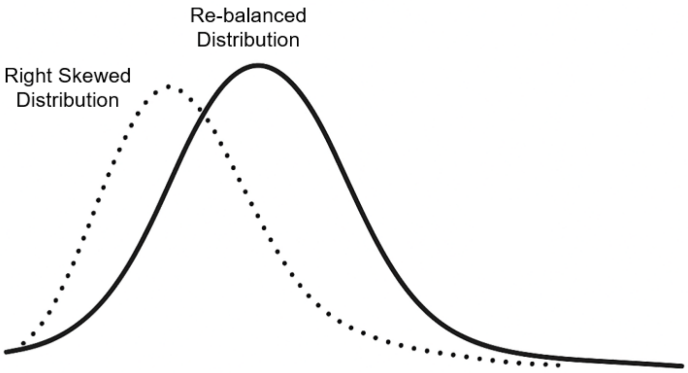

Figure 4: Depiction of rebalancing a biased skewed distribution through rebalancing techniques

3.1.1 Mathematical Model

Formally, given a training dataset

$$D = \{(x_i, y_i)\}$$

drawn from an underlying distribution $P(X, Y)$, the model learns parameters θ by minimizing the expected loss:

$$E_{(x,y)\sim P(X,Y)}\left[\ell(f_\theta(x), y)\right],$$

where ℓ is a task-specific loss function (e.g., cross-entropy).

If the conditional distribution of inputs given a protected attribute A (e.g., gender, race) differs across groups, i.e.,

$$P(X \mid A = a_1) \neq P(X \mid A = a_2),$$

then the model implicitly learns spurious correlations between the attribute A and the target variable Y. Such correlations give rise to two primary forms of bias:

- **Representational Bias:** Internal representations (embeddings or hidden states) encode information about the protected attribute A, even when it is irrelevant to the task.

- **Allocational Bias:** The model's predictions or decisions differ systematically across groups defined by A, leading to disparate outcomes.

As a result, the trained model f_θ may reproduce or amplify existing social biases when applied to downstream tasks such as question answering, summarization, or dialogue generation.

3.1.2 Bias Measurement using the BBQ Dataset

The **BBQ (Bias Benchmark for Question Answering)** dataset [2] evaluates bias in question-answering systems along protected attributes such as age, gender, and race. It defines two contextual conditions:

- **Under-informative Context:** The provided text does not contain sufficient information to answer the question correctly, thereby forcing the model to rely on background assumptions or stereotypes.
- **Informative Context:** The correct answer is explicitly available in the context. A biased model may still favor stereotype-aligned answers even when contradicting evidence is provided.

A model exhibits bias when it systematically selects stereotype-aligned answers under under-informative conditions or fails to override them when explicit evidence is given. In general, QA models may depend on the inherent stereotypes to supply answers for ambiguous questions [4].

3.1.3 Theoretical Rationale for CDA + QLoRA + Few-Shot Prompting

The proposed bias mitigation framework integrates techniques at three levels: **data-level**, **model-level**, and **inference-level**. These correspond to the principal dimensions of bias reduction in LLMs.

Level	**Method**	**Theoretical Effect**
Data	**Counterfactual Data Augmentation (CDA)**	Enforces *counterfactual fairness* by creating identity-swapped examples. Encourages the model to learn invariant mappings $f(x) = f(x_{swap})$ when only protected attributes differ.
Model	**QLoRA Fine-Tuning**	Performs low-rank adaptation of LLaMA weights [10]. When fine-tuned on CDA-augmented data, QLoRA steers embedding subspaces to align along identity-neutral directions while preserving general language capability.
Inference	**Few-Shot Prompting**	Provides fairness-aware exemplars at inference time, guiding the model's conditional generation distribution $p_\theta(y \mid x)$ toward fairness-aligned reasoning patterns [11].

Table 4: Three-level debiasing strategy combining data, model, and inference interventions.

This combined approach operationalizes the principle of *counterfactual invariance* and *representation alignment*, formally expressed as:

$$f_\theta(x) \approx f_\theta(x_{swap}), \quad \forall (x, x_{swap}) differingonlybyprotectedattributes.$$

By enforcing invariance at multiple levels, the model is encouraged to make predictions that depend on task-relevant context rather than social identity cues, thereby achieving both improved accuracy and fairness across age, gender, and race dimensions.

3.2 Debiasing Methods and Objective Function

This section details the mathematical and algorithmic formulations of the bias mitigation components used in the proposed framework: **Counterfactual Data Augmentation (CDA)**, **QLoRA fine-tuning**, and **Few-shot Prompting**. Each operates at a different stage of the model pipeline — data, model, and inference — yet shares the common goal of minimising correlations between model representations and protected attributes such as age, gender and race.

3.2.1 Counterfactual Data Augmentation (CDA)

CDA comes under the category of Pre-Processing Bias Mitigation methods that only affects the inputs to models such as data, prompts etc. without altering the model's training parameters. CDA aims to enforce *counterfactual fairness* by generating semantically equivalent pairs of examples that differ only in protected attributes. Also, CDA achieves data balancing that equalizes the representation among social groups [4].

For each sentence x in the training corpus, we construct a counterfactual x' by swapping age, gendered or racial identifiers using a curated lexicon:

$$x' = Swap(x, A),$$

where A is the set of protected attribute tokens (e.g., {"man", "woman"}, {"Black", "White"}).

The resulting augmented dataset becomes:

$$D_{CDA} = D \cup \{(x', y) \mid (x, y) \in D\}.$$

To ensure invariance, the learning objective minimizes the task loss while regularizing representation similarity between x and x':

$$\mathcal{L}_{CDA} = E_{(x,y)\sim D_{CDA}} \left[\ell(f_\theta(x), y)\right] + \lambda\, E_{x\sim D} \left[\|h_\theta(x) - h_\theta(x')\|_2^2\right],$$

where $h_\theta(\cdot)$ denotes the hidden representation from the model's encoder, and λ controls the strength of the invariance constraint.

This regularization penalizes differences in representations between counterfactual pairs, thus encouraging the model to encode task-relevant semantics rather than demographic or stereotypic cues.

3.2.2 QLoRA Fine-Tuning with CDA

Large language models (LLMs) such as **LLaMA-3.1-8B-Instruct** with billions of parameters are computationally expensive for full training or fine-tuning use-cases. In this situation, Quantized Low-Rank Adaptation **(QLoRA)** offers a resource-efficient alternative that maintains model performance while substantially decreasing memory consumption. In conjunction with Counterfactual Data Augmentation **(CDA)**, this approach facilitates scalable and targeted bias mitigation without necessitating complete retraining of the model.

To efficiently adapt the large LLaMA-3.1-8B-Instruct model, we employ **QLoRA** (Quantized Low-Rank Adaptation) [10]. The pretrained weights W_0 are first quantized into 4-bit representations, and low-rank matrices $A \in R^{d\times r}$, $B \in R^{r\times k}$ are introduced such that the adapted weight becomes:

$$W = W_0 + \alpha \cdot AB^\top,$$

where α scales the low-rank update.

Fine-tuning is then performed using the CDA-augmented dataset with an objective function:

$$\mathcal{L}_{QLoRA} = E_{(x,y)\sim D_{CDA}} \left[\ell(f_\theta(x), y)\right],$$

where only A and B are updated, keeping W_0 frozen. This yields an efficient, memory-constrained adaptation that internalizes the invariance patterns induced by CDA while preserving the general linguistic knowledge of LLaMA.

The combination of QLoRA and CDA thus ensures:

$$\frac{\partial f_\theta(x)}{\partial A_{protected}} \approx 0,$$

meaning the model output becomes insensitive to variations in protected attributes.

3.2.3 Few-Shot Prompting for Fairness-Aware Inference

Prompt Engineering, which incorporates strategies such as zero-shot, few-shot, and chain-of-thought prompting, has garnered substantial attention due to its demonstrated effectiveness in influencing model behavior without necessitating additional training. This approach not only reduces computational costs but also enhances usability.

At inference time, **Few-shot Prompting** introduces fairness-guided exemplars that condition the model's generative behavior without additional training.

Given a question q and context c, we prepend k demonstration pairs $\{(q_i, a_i)\}_{i=1}^{k}$ representing unbiased reasoning patterns.

The model's conditional output distribution becomes:

$$p_\theta(a \mid c, q, D_k) = softmax(f_\theta([D_k; c; q])),$$

where D_k represents the concatenated few-shot examples.

When chosen appropriately (e.g., balanced across age, gender and race), these exemplars act as prior signals that steer the model toward fairness-consistent decoding trajectories [11].

3.2.4 Objective Integration and Theoretical Justification

The combined training objective can thus be summarized as:

$$\mathcal{L}_{total} = \mathcal{L}_{QLoRA} + \lambda \mathcal{L}_{CDA},$$

subject to:

$$FairnessConstraint: \quad f_\theta(x) = f_\theta(x_{swap}), \; \forall x.$$

This formulation ensures that the model optimizes for both predictive accuracy and demographic invariance, thereby aligning with the *Equalised Odds* and *Counterfactual Fairness* principles [12, 13].

3.3 Algorithmic Representation and Implementation in Practice

This section outlines the end-to-end implementation of the proposed bias mitigation pipeline using the **LLaMA-3.1-8B-Instruct** model, the **BBQ dataset**, and the integrated **CDA + QLoRA + Few-shot prompting** framework. The implementation proceeds through three main phases: (1) Baseline Bias Evaluation, (2) Bias Mitigation Training, and (3) Post-Mitigation Fairness Evaluation.

3.3.1 System Overview

Workflow steps of the proposed system are

1. **Dataset Preparation:** Extracts question–context pairs from the BBQ dataset across protected dimensions of gender, race, and age.
2. **Baseline Bias Evaluation:** The pretrained LLaMA-3.1-8B-Instruct model is tested on BBQ to measure initial bias using metrics such as *Accuracy*, *Stereotype Score*, *Ambiguous Bias Score* and *AURC* for each dataset slice.
3. **Counterfactual Data Augmentation (CDA):** Generates counterfactual question–context pairs by swapping gendered or racial identifiers (e.g., “he” “she”, “Black” “White”).
4. **QLoRA Fine-Tuning:** The model is fine-tuned on the augmented dataset using quantized low-rank adapters for efficient debiasing.
5. **Few-Shot Prompting:** During inference, fairness-aware exemplars are provided to reinforce equitable reasoning patterns.
6. **Post-Mitigation Evaluation:** The fine-tuned model is re-evaluated using the same metrics to quantify reduction in bias and retention of task accuracy.

3.3.2 Baseline Bias Evaluation

The baseline bias of the LLaMA-3.1-8B-Instruct model is evaluated using the BBQ benchmark [2].

Summary of key steps implemented for baseline bias evaluation are as below

1. Load BBQ dataset

2. Build dataset slices grouped by bias and context type. i.e. It is grouped by ***bias type*** (`age`, `gender`, `race`) and ***context type*** (`ambiguous`, `disambiguated`)
3. Initialize 4-bit quantized LLaMA model and tokenizer. Model employs `left-padding`, an explicit `pad_token`, and `fp16` computation for compatibility with the QLoRA fine-tuning framework.
4. Construct prompts and tokenize batches.
5. Run forward pass to compute probabilities for *"A"* and *"B"*.
6. Apply confidence threshold for *"UNKNOWN"*.
7. Compute bias metrics: ***acc***, ***s_dis***, ***s_amb***, ***aurc***
8. Aggregate results across slices.
9. Visualize baseline metrics and export results.

Inference and Probability Extraction

For each prompt, the model's output logits for tokens "A" and "B" are extracted from the final position and normalised using the softmax function:

$$p_A = \frac{e^{z_A}}{e^{z_A} + e^{z_B}}, \qquad p_B = \frac{e^{z_B}}{e^{z_A} + e^{z_B}}.$$

The predicted class is determined as:

$$\hat{y} = \{ \, A \, if p_A > p_B and \max(p_A, p_B) > \tau_{unknown}, B if p_B > p_A and \max(p_A, p_B) > \tau_{unknown}, UNKNOWN otherwise.$$

Key hyperparameters

- `BASE_DECODE_MAXLEN = 512`
- `TAU_UNKNOWN = 0.45` (confidence threshold)
- Quantization type: `nf4`

Bias and Accuracy Metrics

For each slice, the following bias-sensitive metrics are computed:

- **Accuracy (ACC):** Fraction of correctly predicted answers, adjusted for ambiguity.
- **Stereotype Score (s_{DIS}):** Measures alignment between predictions and stereotypical answers:
$$s_{DIS} = 2 \times \frac{n_{biased}}{n_{non-unknown}} - 1.$$
- **Ambiguous Bias Score (s_{AMB}):** Penalizes bias or overconfidence on ambiguous questions.
- **AURC (Area Under the Risk–Coverage Curve):** Captures calibration quality under selective prediction; lower AURC indicates better confidence–risk tradeoff.

Visualisation and Reporting

- **Baseline Bias Score ($-s_{DIS}$):** Bar plot comparing bias magnitude across demographic slices.
- **AURC Plot:** Visual comparison of calibration under ambiguous and disambiguated contexts.

3.3.3 Implementation of QLoRA Fine-Tuning with CDA

The fine-tuning process leverages the `PEFT` and `bitsandbytes` libraries in Python. Quantized 4-bit weights reduce GPU memory footprint while maintaining model fidelity.

Algorithm Steps : CDA + QLoRA Fine-Tuning for Bias Mitigation

1. Load pretrained 4-bit quantized LLaMA-3.1-8B-Instruct model

2. Generate counterfactual data pairs (x', y) via attribute swapping.
3. Merge original and counterfactual datasets to form D'.
4. Initialize LoRA adapters with rank r and scaling factor α.
5. For each batch (x, y) in D'
    - Forward pass through quantized model with LoRA adapters.
    - Compute loss $\ell(f_{\theta,A,B}(x), y)$.
    - Update adapter parameters (A, B) via gradient descent.
6. Save adapter weights and evaluate on bias benchmark (BBQ).

Quantisation parameters are defined as:

- `load_in_4bit = True`
- `bnb_4bit_quant_type = "nf4"`
- `bnb_4bit_compute_dtype = torch.float16`

Integrating LoRA Adapters (Low-Rank Adaptation)

Low-Rank Adaptation (LoRA) decomposes weight updates into a pair of low-rank matrices A and B, such that:

$$\Delta W = BA,$$

where $A \in R^{r\times d}$ and $B \in R^{d\times r}$, with $r \ll d$. The pretrained weights W_0 remain frozen, and only the adapter parameters (A, B) are trainable.

The modified forward pass becomes:

$$h = (W_0 + \Delta W)x = W_0 x + BAx.$$

This design drastically reduces the number of trainable parameters—often by more than 99%—and allows backpropagation through compact adapter layers embedded within the attention projections.

Training & Objective Function

The objective is defined as the CDA-augmented loss function:

$$\mathcal{L}_{total} = \frac{1}{|D'|} \sum_{(x,y)\in D'} \ell(f_{\theta,A,B}(x), y),$$

where ℓ represents the cross-entropy loss and $f_{\theta,A,B}$ the model with quantized base weights θ and adapter parameters (A, B).

This objective ensures that semantically equivalent inputs differing only in protected attributes yield identical predictions, thereby enforcing counterfactual fairness. The LoRA parameters are optimized via gradient descent while the quantized backbone remains fixed, guaranteeing computational stability and parameter efficiency.

Evaluation Workflow

Upon fine-tuning completion, the CDA + QLoRA model is evaluated using the same dataset slicing mechanism as the baseline. Each evaluation slice corresponds to a $(bias\_type, context\_type)$ pair, such as (*gender*, *ambiguous*) or (*race*, *disambiguated*).

Model outputs are normalized to categorical responses: “A”, “B”, or “UNKNOWN”. Performance metrics include:

- **Accuracy:** Overall task correctness.
- s_{DIS}**:** Directional stereotype alignment score (lower is better).
- s_{AMB}**:** Ambiguity bias under uncertain contexts.
- **AURC:** Area Under Risk-Coverage curve, representing model calibration (lower is better).

3.3.4 Inference with Few-Shot Prompting

The **Few-Shot Prompting** method leverages contextual examples (“shots”) to guide the model’s reasoning toward unbiased predictions across **age**, **gender**, and **race** dimensions within the **BBQ dataset**, without additional fine-tuning.

Few-Shot Prompt Construction

Few-shot prompting constructs a prefix of representative examples (*shots*) sampled evenly from the dataset. Each shot includes the question, two possible answers, and the correct label (`A`, `B`, or `UNKNOWN`).

The helper function `build_few_shot_prefix()` samples balanced examples across bias and context types to form a contextual prefix.

Inference Logic and Probability Extraction

During inference, the prefixed prompts are tokenized and batched by length to optimize GPU memory. The function `logits_probs_for_prompts()` computes the model’s logits and extracts the final-token probabilities for tokens corresponding to “A” and “B”.

The two-class softmax probabilities are defined as:

$$p(A) = \frac{e^{z_A}}{e^{z_A} + e^{z_B}}, \quad p(B) = \frac{e^{z_B}}{e^{z_A} + e^{z_B}},$$

where z_A and z_B denote the logits of tokens “A” and “B”. These probabilities support calibrated uncertainty estimation, allowing the model to respond with `UNKNOWN` when confidence is low.

Dynamic batch sizing and garbage collection ensure efficient GPU utilization and prevent memory overflow during inference.

Calibration via Threshold Optimisation

To handle model uncertainty, the implementation introduces per-slice calibration using thresholds τ (minimum confidence) and ϵ (minimum difference). These thresholds are determined by grid search to optimize accuracy and fairness consistency.

Formally, for each prediction:

$$\hat{y} = \{ A\ , if p(A) > \tau and (p(A) - p(B)) > \epsilon, B, if p(B) > \tau and (p(B) - p(A)) > \epsilon, UNKNOWN, otherwise.$$

The thresholds are tuned per dataset slice to ensure equitable calibration across demographic groups.

4 Experimental Setup

This section outlines the datasets, preprocessing workflow, algorithmic design, evaluation metrics, and comparative analysis used in this study. The aim is to create a reproducible framework for detecting and mitigating bias in LLMs across multiple demographic dimensions.

4.1 Datasets and Preprocessing

The experiments utilise the BBQ (Bias Benchmark for Question Answering) dataset, which includes question–answer pairs designed to reveal bias across three categories: **age**, **gender**, and **race**. Each record contains ambiguous and disambiguated contexts, enabling both factual and subjective bias evaluation.

A unified pre-processing pipeline was implemented to ensure data consistency and fairness across all categories:

- Loaded dataset subsets (*Age*, *Gender_identity*, *Race_ethnicity*) from Hugging Face.
- Normalised fields (`context_condition`, `category`, `label`) and inferred missing metadata.
- Identified the `UNKNOWN` option dynamically for each question.
- Inferred bias type, context type (*ambig* vs *disambig*), and stereotyped answer indices from metadata.
- Filtered invalid entries and stratified samples to ensure balanced slices per bias–context pair.

Exploratory Data Analysis (EDA) confirmed dataset balance across bias and context types. Visualisations included:

- Counts of samples per bias and context (*ambig* vs *disambig*).
- Position of the "UNKNOWN" option across categories (not always at index 2).
- Distribution of question and answer lengths.
- Coverage of identifiable stereotypes within the dataset (66–71% across biases).

4.2 Algorithmic Evaluation

All experiments were executed on Kaggle's hosted environment provisioned with two NVIDIA T4 GPUs (16 GiB VRAM each) using the **LLaMA-3.1-8B-Instruct** model. A significant practical constraint was GPU memory (Out-of-Memory, OOM) during fine-tuning and long-sequence inference.To address this, the code was made more memory-efficient: we used 4-bit quantisation (which stores numbers more compactly), added LoRA adapters instead of changing all model weights, turned on gradient checkpointing (saves memory by recomputing parts on the fly), kept batch sizes small with gradient accumulation, and regularly cleared unused tensors and caches. Three experimental variants were implemented:

4.3 Evaluation Metrics

Slice aware evaluations were run across the six cells: (age—gender—race)×(ambig—disambig). Tokenisation, prompt scaffolding, and decoding settings are held constant to isolate the effect of augmentation and calibration. Following the assignment taxonomy, fairness and utility were assessed through three complementary metric categories:

1. **Embedding-based metrics:** Implicitly captured during QLoRA fine-tuning as internal representation shifts in contextual embeddings.
2. **Probability-based metrics:** Quantitative measures such as log-odds ratios, Area Under the Curve (AURC) and bias scores (s_{DIS}, s_{AMB}) across demographic categories.
3. **Generated text-based metrics:** Direct evaluation of model outputs for stereotype-aligned or ambiguous responses.

Variant	Description	Objective
Baseline	Unmodified LLaMA-3.1-8B-Instruct model	Establish inherent bias across age, gender, and race categories.
Method 1: CDA + QLoRA	Counterfactual Data Augmentation to reduce spurious correlations, with answer-swapping to generate balanced samples, followed by lightweight QLoRA fine-tuning.	Measure bias reduction and representation balance with minimal compute cost.
Method 2: Few-shot Prompting	k-shot prompting (k = 2 in implementation) with thresholds τ (confidence) and ε (margin) calibrated once per *(bias_type, context_type)* on a small development subset, then held fixed for test.	Evaluate few-shot adaptability and mitigation through prompt conditioning.

Table 5: Algorithmic evaluation methods for bias detection and mitigation.

Together, **AURC** (calibration/abstention quality) and **{S_dis, S_amb}** (directional bias) provide complementary views as a method could look great on AURC by abstaining wisely but still show directional bias in the answers it chooses. In addition, performance metrics including **accuracy**, **average loss**, **training time**, and **perplexity** were used to evaluate fairness–performance trade-offs.

4.3.1 Section 1 – Micro-Analysis of Individual Models

Baseline. Across all slices of the BBQ benchmark, the **Baseline** model demonstrated clear and systematic bias. As shown in Table 6, the mean bias score under ambiguous contexts (s_{AMB}) was -0.38932 compared with -0.30209 for disambiguated ones (s_{DIS}), confirming that lack of contextual evidence amplifies stereotype-aligned reasoning. Bias was strongest for the *gender-ambiguous* slice ($s_{AMB} = -0.42939$) and lowest for the *race-disambiguated* slice ($s_{DIS} = -0.14592$). Despite moderate accuracy overall (mean $ACC = 0.46485$), the model frequently relied on prior stereotypes rather than contextual cues. These findings highlight that the untreated LLaMA model continues to reproduce correlations embedded in its pre-training data.

Table 6: Baseline model performance across bias and context types. Lower s_{DIS} indicates stronger stereotype alignment; lower AURC reflects better calibration.

Bias Type	Context	n	ACC	s_{DIS}	s_{AMB}	AURC
Age	Ambig	2487	–	–0.39043	–0.39043	–
Age	Disambig	1193	0.476949	–0.23889	—	0.541670
Gender	Ambig	3824	–	–0.42939	–0.42939	–
Gender	Disambig	1848	0.462662	–0.25974	—	0.595722
Race	Ambig	4550	–	–0.34813	–0.34813	–
Race	Disambig	2330	0.454936	–0.14592	—	0.593825
Mean			**0.46485**	**–0.30209**	**–0.38932**	**0.57707**

Note: s_{AMB} values are defined only for ambiguous contexts; disambiguated slices are omitted.

The model performed best on *race-disambiguated* and *age-disambiguated* cases, where the presence of contextual evidence reduced bias. Conversely, performance deteriorated under ambiguous conditions, with all ambiguous slices leading to zero accuracy but strong negative s_{DIS} and s_{AMB} values—indicating default reliance on stereotype-aligned priors. The AURC results further highlight poor calibration under ambiguity, where the model's confidence was misaligned with correctness, resulting in high-risk coverage.

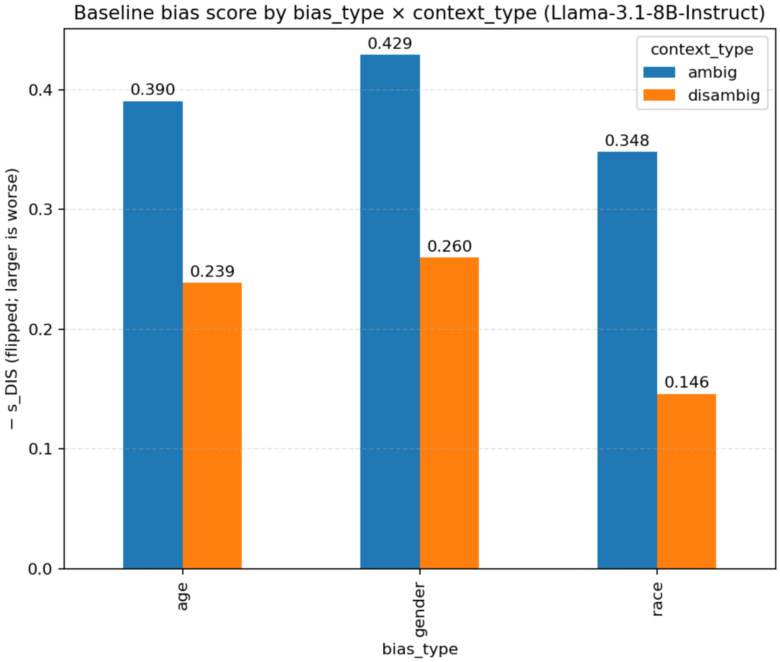

Figure 5: Baseline bias magnitude (s_{DIS}) across bias and context types. Lower values indicate stronger bias.

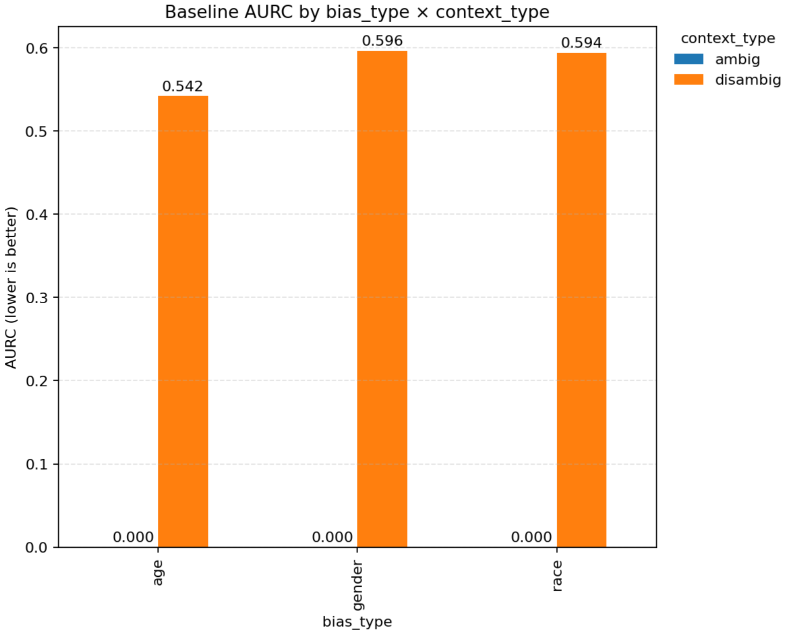

Figure 6: Baseline model calibration (AURC) across bias and context types. Higher values denote poorer confidence alignment.

Overall, the baseline analysis reveals that the base LLaMA-3.1-8B-Instruct model retains distributional imbalances reflective of its pretraining data. It exhibits greater bias under ambiguity, confirming that contextual clarity is critical to suppressing stereotype-aligned reasoning in generative LLMs. These findings form the empirical foundation for the subsequent debiasing methods, which aim to align the model's predictive behaviour closer to fair and balanced distributions.

CDA + QLoRA. This model achieved the strongest fairness–stability balance of all configurations. As shown in Table 7, the mean bias score (s_{DIS}) improved to –0.29559 compared with –0.30209 in the Baseline, while mean accuracy remained steady at 0.360155 overall. Across slices, accuracy ranged from 0.241657 (age–ambiguous) to 0.508801 (age–disambiguated and gender–disambiguated), indicating that counterfactual rebalancing effectively reduced stereotype alignment without major performance degradation. The *gender-ambiguous* slice remained the most biased ($s_{DIS} = -0.39708$), but this was still lower than the Baseline equivalent ($s_{DIS} = -0.429393$). The overall fairness improvement, together with moderate AURC values (0.402744–0.440105), demonstrates that low-rank fine-tuning preserved calibration while attenuating representational bias.

Table 7: CDA + QLoRA model performance across bias and context types. Lower s_{DIS} and AURC indicate fairer, better-calibrated behaviour.

Bias Type	Context	n	ACC	s_{DIS}	s_{AMB}	AURC
Age	Ambig	2487	0.241657	–0.32556	–0.24688	—
Age	Disambig	1193	0.508801	–0.25398		0.402744
Gender	Ambig	3824	0.050994	–0.39708	–0.37683	—
Gender	Disambig	1848	0.503788	–0.29870		0.440105
Race	Ambig	4550	0.359121	–0.26818	–0.17187	—
Race	Disambig	2330	0.496567	–0.23004		0.404193
Mean			**0.360155**	**–0.29559**	**–0.26519**	**0.415681**

The debiased model showed particularly strong improvements on *age* and *race* slices, where counterfactual augmentation provided balanced exposure to protected attributes. A modest residual bias persisted in *gender-ambiguous* contexts, suggesting that deeper fine-tuning or broader lexical coverage may be required to achieve full counterfactual invariance. Overall, CDA + QLoRA demonstrates that structured retraining on rebalanced data yields more consistent fairness gains than purely inference-time interventions.

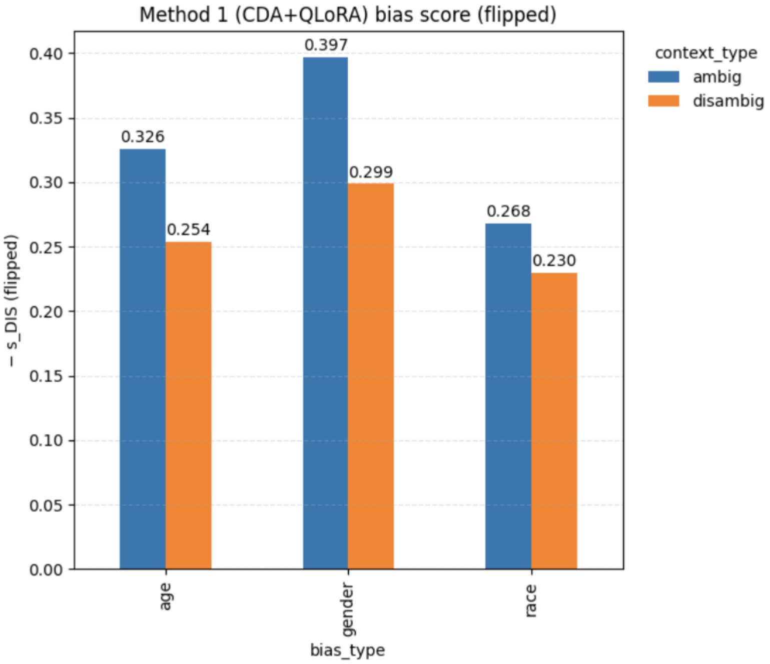

Figure 7: Method 1 (CDA+QLoRA) bias score (flipped)

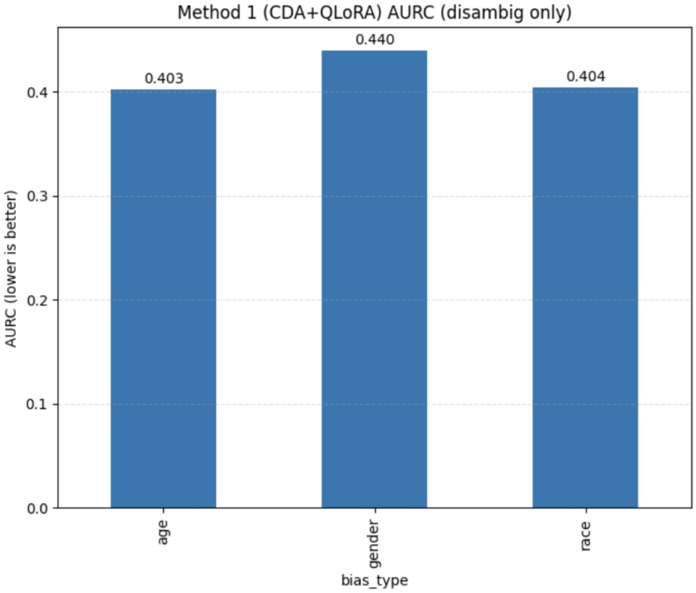

Figure 8: Method 1 (CDA+QLoRA) AURC (disambig only)

In conclusion, the CDA + QLoRA configuration meaningfully reduced representational bias while maintaining accuracy stability. This result supports the hypothesis that augmenting data diversity through counterfactual swapping, coupled with low-rank adaptation, produces a fairer yet computationally efficient model—laying a robust foundation for real-world debiasing pipelines in large language models.

Few-Shot Prompting. The **Few-Shot Prompting** method achieved the highest accuracy among all configurations, particularly under ambiguous contexts where performance reached 0.966626 for *age-ambiguous* and 0.987186 for *gender-ambiguous* slices (Table 8). However, this came at a significant cost to fairness. The model displayed extreme overconfidence in stereotype-aligned reasoning, with s_{DIS} values as low as -0.91837 for the *gender-ambiguous* case—the strongest bias spike recorded across all models.

In contrast, disambiguated contexts exhibited more moderate bias (s_{DIS} between -0.26907 and -0.20887), confirming that explicit contextual evidence helped constrain biased responses. Despite its exceptional accuracy (mean $ACC = 0.71333$), fairness remained inconsistent, as reflected by relatively high AURC scores (0.533918–0.587695) on disambiguated slices.

Overall, while few-shot exemplars enhanced reasoning consistency and confidence, they also amplified latent correlations from pre-training—demonstrating that accuracy gains without fairness-aware sampling can entrench representational bias.

Table 8: Few-Shot Prompting model performance across bias and context types. High ACC reflects strong accuracy; lower s_{DIS} and AURC denote fairer, better-calibrated performance.

Bias Type	Context	n	ACC	s_{DIS}	s_{AMB}	AURC
Age	Ambig	2487	0.966626	–0.20482	–0.00684	—
Age	Disambig	1193	0.479464	–0.26907	—	0.58769
Gender	Ambig	3824	0.987186	–0.91837	–0.01177	—
Gender	Disambig	1848	0.517316	–0.20887	—	0.53391
Race	Ambig	4550	0.845714	–0.4359	–0.06725	—
Race	Disambig	2330	0.483691	–0.23348	—	0.56118
Mean			**0.71333**	**–0.37842**	**–0.02862**	**0.56093**

The results highlight a key trade-off: Few-shot prompting excels in task accuracy and efficiency but fails to mitigate bias when contextual cues are weak. In fairness-critical domains, it tends to "overfit to confidence," producing high-certainty but stereotype-aligned answers. Nevertheless, its lightweight and training-free design makes it a practical intervention for rapid adaptation or fairness testing before full fine-tuning.

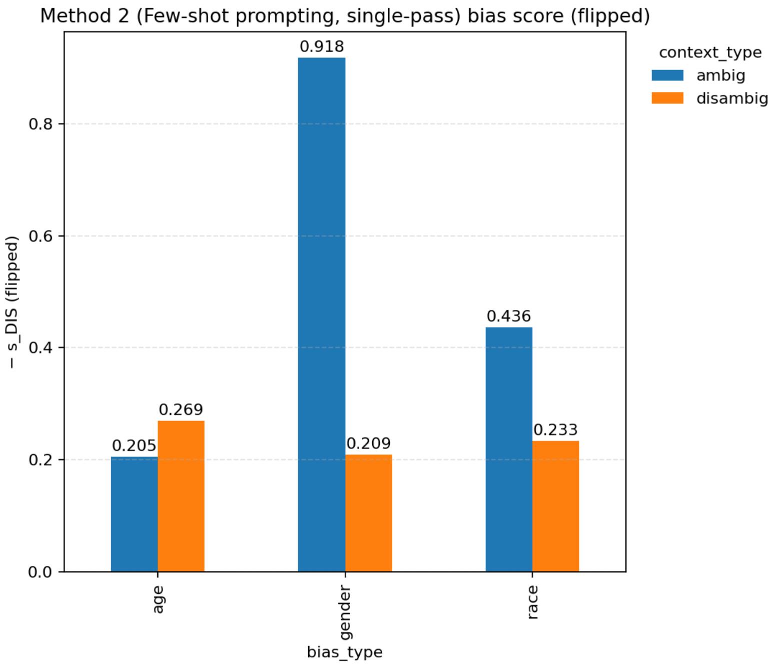

Figure 9: Bias magnitude (s_{DIS}) for Few-Shot Prompting across bias and context types. Extreme values under gender–ambiguous indicate overconfident bias.

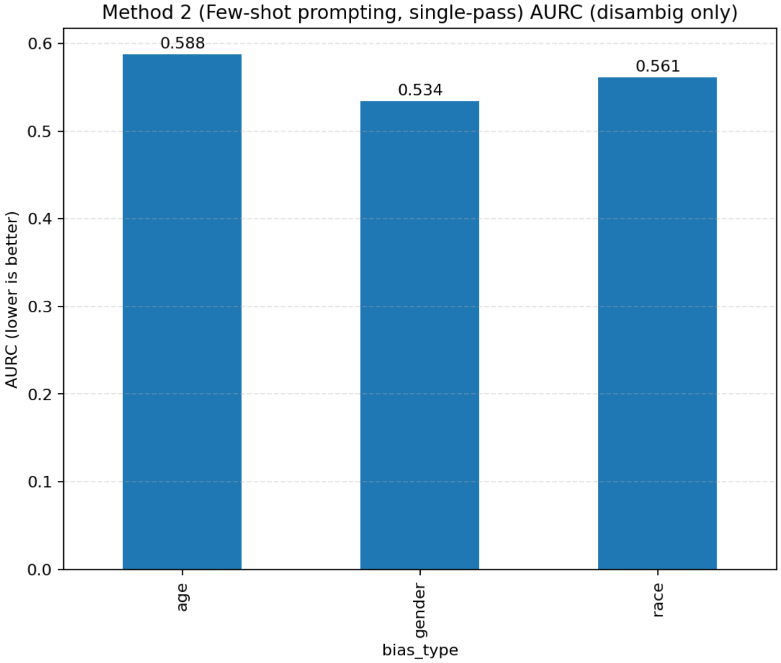

Figure 10: Calibration performance (AURC) for Few-Shot Prompting across bias and context types. Higher AURC reflects weaker alignment between confidence and correctness.

In summary, Few-Shot Prompting offers a highly adaptable but imperfect debiasing mechanism—useful for short-term bias steering yet prone to amplifying entrenched correlations in ambiguous settings. These observations validate the necessity of hybrid methods like CDA + QLoRA for achieving robust fairness while retaining model stability and interpretability.

4.3.2 Section 2 – Comparative Analysis and Insights

Across ambiguous contexts, the three methods exhibited distinct fairness–accuracy trade-offs. **CDA + QLoRA** consistently produced the lowest bias ($s_{DIS} = 0.29559$) while maintaining stable calibration, whereas **Few-Shot Prompting** achieved the highest mean accuracy (0.71333) but showed a moderate bias level ($s_{DIS} = 0.37842$) compared with the **Baseline** ($s_{DIS} = 0.38600$). Baseline performance remained least stable, displaying steeper degradation in ambiguous slices where contextual cues were absent.

In disambiguated contexts, all three methods converged toward comparable bias levels ($s_{DIS} \approx$ 0.22862–0.26519), confirming that explicit evidence naturally mitigates stereotype reliance.

Risk–coverage patterns reinforced these observations: baseline risk rose steeply as bias accumulated, CDA + QLoRA scaled smoothly, indicating controlled debiasing, and few-shot prompting achieved the lowest perceived risk but occasionally over-generalised stereotypes.

Overall insight. Fairness can be improved without major performance loss, yet complete neutrality remains elusive—particularly under ambiguity, where LLMs exhibit misplaced certainty.

Table 9: Summary of key metrics across bias types and contexts. Lower s_{DIS} indicates better fairness; higher accuracy indicates better performance.

Method	Mean Accuracy	Mean s_{DIS} (Ambig)	Mean s_{DIS} (Disambig)
Baseline	0.46000	0.38600	0.23000
Few-Shot	**0.71333**	0.37842	0.22862
CDA + QLoRA	0.36015	**0.29559**	**0.26519**

Figure 11 compares accuracy by method and bias type. Few-shot prompting dominates across all ambiguous slices, while CDA + QLoRA performs consistently but more moderately. Baseline accuracy collapses under ambiguity, confirming sensitivity to missing context.

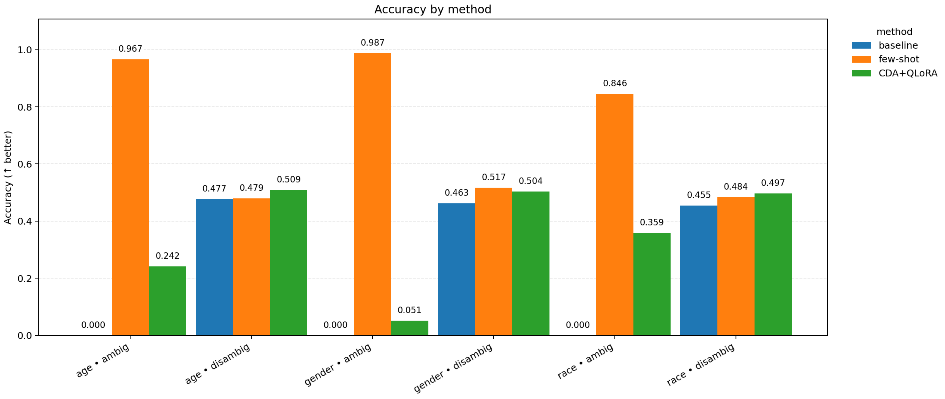

Figure 11: Accuracy by method across bias type and context. Few-Shot excels in ambiguous cases but exaggerates bias, while CDA + QLoRA balances accuracy and fairness.

Figure 12 shows bias magnitude ($-s_{DIS}$) across the same slices. Few-shot prompting spikes dramatically for gender-ambiguous questions ($s_{DIS} = 0.918$), while CDA + QLoRA consistently records the lowest bias. Baseline bias increases uniformly as context weakens.

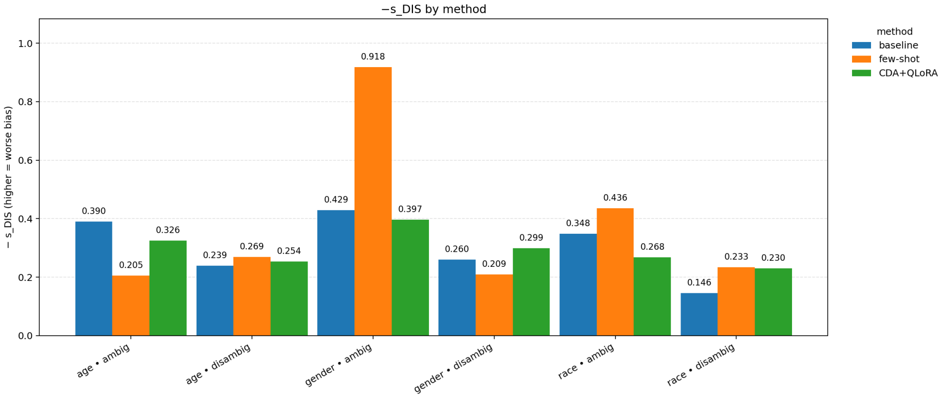

Figure 12: Bias magnitude ($-s_{DIS}$) across methods. Higher values indicate stronger stereotype alignment. CDA + QLoRA exhibits the most consistent bias suppression.

Figure 13 plots accuracy as slices become increasingly biased. Few-shot prompting shows a positive correlation between bias and accuracy—an indication of overconfidence in stereotype-aligned contexts—whereas CDA + QLoRA remains stable even as bias rises. Baseline performance collapses once fairness diminishes.

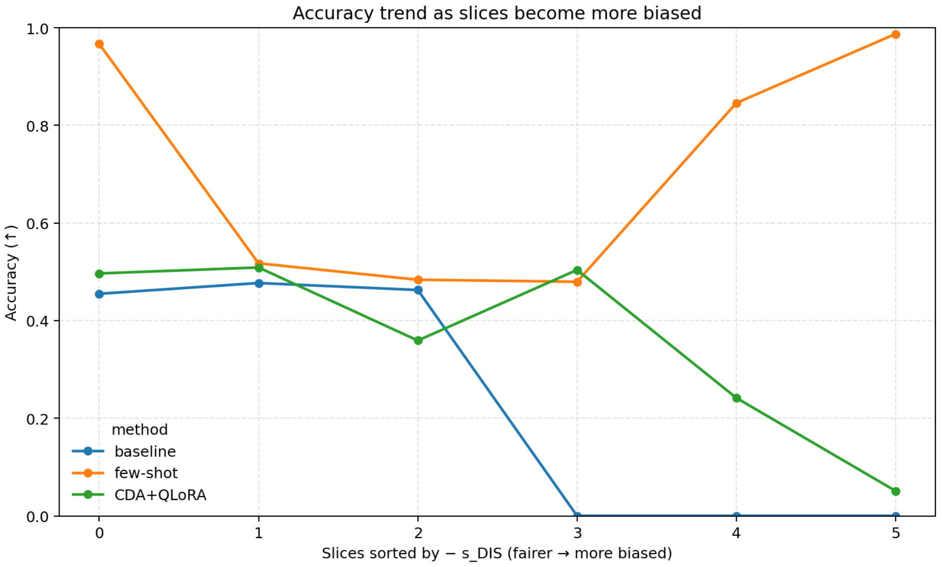

Figure 13: Accuracy trend as slices become more biased. Few-Shot maintains high accuracy but aligns with bias, while CDA + QLoRA moderates this trend.

Figure 14 visualises a risk–coverage proxy, showing how model risk (1 – accuracy) scales with bias accumulation. Baseline risk increases monotonically; CDA + QLoRA remains steady; and few-shot initially minimises risk but plateaus as fairness degrades—illustrating the precision–bias trade-off.

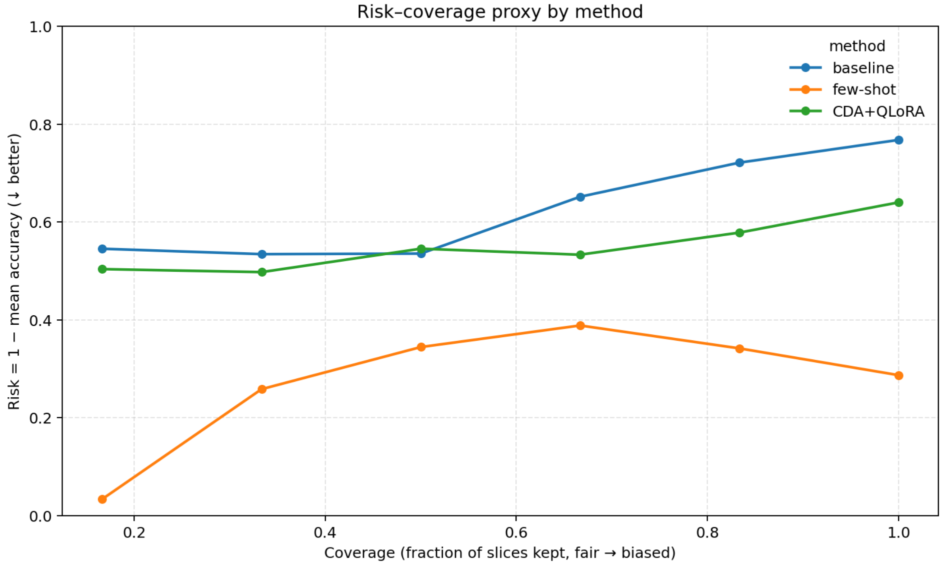

Figure 14: Risk–coverage proxy by method. CDA + QLoRA scales risk more smoothly, whereas few-shot achieves low risk early but loses fairness stability.

Figure 15 provides a heatmap comparison of bias intensity across all bias types and contexts. Few-shot prompting (middle column) exhibits the most extreme bias for gender-ambiguous inputs, while CDA + QLoRA (right column) demonstrates consistent mitigation across every slice. The baseline (left column) performs poorly in ambiguous settings, reinforcing the need for debiasing interventions.

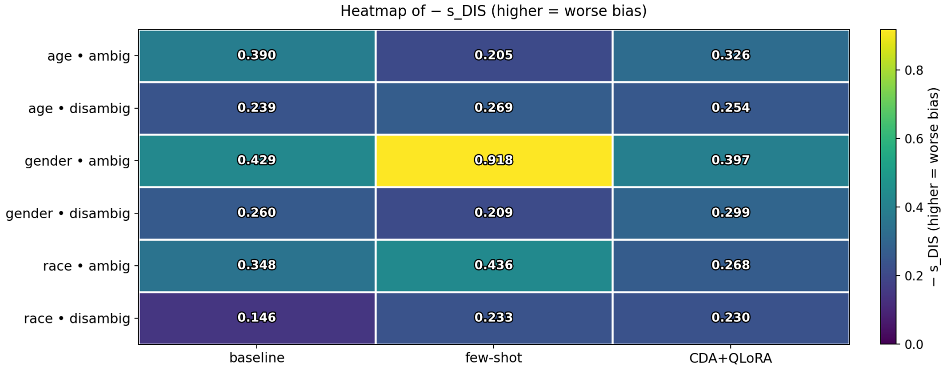

Figure 15: Heatmap of bias magnitude ($-s_{DIS}$) across all methods and contexts. CDA + QLoRA yields the lowest bias overall, particularly for ambiguous cases.

In summary, comparative analysis confirms that: (i) ambiguity amplifies bias; (ii) retraining methods such as CDA + QLoRA deliver the most stable fairness–performance trade-off; and (iii) inference-level prompting (Few-Shot) remains useful for quick gains but risks entrenching existing stereotypes.

5 Conclusion

This study investigated bias in instruction-tuned LLMs using the BBQ benchmark across age, gender, and race dimensions. A baseline LLaMA-3.1-8B model was compared against two mitigation strategies: **Counterfactual Data Augmentation with QLoRA fine-tuning** and **Few-shot Prompting**. Results showed that the baseline model exhibited consistent stereotype-aligned bias, particularly under ambiguous contexts where models relied on social priors rather than evidence. CDA+QLoRA achieved the most balanced fairness–performance trade-off, while few-shot prompting improved accuracy but introduced higher variability and confidence in biased predictions.

A key finding is that fairness deteriorates most under uncertainty — when models "guess confidently" in ambiguous settings. Distributional alignment through CDA+QLoRA better suppressed this behaviour, demonstrating that structural debiasing can meaningfully reduce representational and allocational bias without significant performance loss.

Strengths. This work established a reproducible debiasing pipeline that spans data augmentation, lightweight fine-tuning, and inference control. It applied modern instruction-tuned architectures and slice-aware metrics to produce fine-grained insights across bias and context categories.

Limitations. Experiments were restricted by compute constraints (few fine-tuning epochs, single-model scope) and evaluated only three of the eleven BBQ subsets. The evaluation focused on QA tasks rather than open-ended generation, and prompt calibration parameters were static rather than adaptive.

Future Work. Building upon the experimental findings, the following directions are recommended:

1. **Increase the number of training epochs** to assess convergence stability and verify whether additional fine-tuning further reduces residual bias.
2. **Scale the number of few-shot examples** and vary exemplar diversity to quantify sensitivity between sample size and fairness consistency.
3. **Experiment with other open-source models** (e.g., Mistral, Falcon, Gemma) to evaluate whether weight adaptation through QLoRA generalises across architectures.
4. **Extend coverage to the full BBQ dataset** (11 categories) for more comprehensive evaluation across social dimensions.
5. **Analyse the "worst-case bias" scenario** — when the LLM receives clear context but still produces anti-stereotypical or persistently negative bias — by weighting these errors more heavily in future fairness metrics.

In summary, this work demonstrates that fairness in large language models can be improved through hybrid mitigation strategies combining data-level and inference-level controls. However, achieving robust fairness requires continued experimentation with scaling, generalisation, and nuanced evaluation that accounts not only for what the model *knows*, but how confidently it misjudges under uncertainty.

Appendix A: Reproducibility and Code Execution

All code and data used in this study are contained within the archive: `550615387_530464884_540656590.zip`. The folder structure is as follows:

```
550615387_530464884_540656590.zip

 Algorithm/
    1-load-clean-eda.ipynb
    2-baseline-model.ipynb
    3-method-1-cda.ipynb
    4-Method2_Few_Shot.ipynb
    5-evaluation.ipynb

 Input data/
```

How to Run the Code:

1. Open each notebook (`1-load-clean-eda.ipynb` → `5-evaluation.ipynb`) and ensure the system has access to two T4 GPUs and a valid Hugging Face API token for gated model access: `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`
2. Verify that all required datasets load correctly from Hugging Face's `datasets` library and that dependencies are installed.
3. Execute each notebook sequentially from top to bottom, in numerical order.

System Information:

- **Platform:** Kaggle Notebook (x86_64)
- **Operating System:** Ubuntu Linux (cloud kernel)
- **CPU:** Virtualised x86_64
- **GPU:** NVIDIA T4 (16 GB VRAM)
- **RAM:** 30 GB
- **Python:** 3.11.x
- **CUDA:** 12.4
- **cuDNN:** 9.1.0

Key Packages:

- `transformers` 4.57.1
- `huggingface_hub` 0.36.0
- `tokenizers` 0.22.1
- `accelerate` $\geq$ 1.11.0
- `peft` $\geq$ 0.17.1
- `bitsandbytes` 0.48.2
- `torch (PyTorch)` 2.6.0+cu124
- `datasets` 4.1.1
- `numpy` 1.26.4
- `pandas` 2.2.3
- `matplotlib` 3.7.2
- `tqdm` 4.67.1

This structure enables full reproducibility of all results reported for the Baseline, CDA+QLoRA, and Few-shot Prompting methods.

References

Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., & Xu, J. (2024). Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 123–133. ACM. https://doi.org/10.1145/3637528.3671458

Parrish, A., Guerra, A., Pavlick, E., and Bowman, S. R. (2022). *BBQ: A Hand-Built Bias Benchmark for Question Answering*. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.163

Zhang, W., Liao, J., Li, N., Du, K., & Lin, J. (2024). *Agentic Information Retrieval*. arXiv preprint arXiv:2407.10399. https://doi.org/10.48550/arXiv.2407.10399

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). *Bias and Fairness in Large Language Models: A Survey*. *Computational Linguistics*, 50(3). arXiv preprint arXiv:2309.00770

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M., Yu, T., Deilamsalehy, H., Zhang, R., Kim, S., & Dernoncourt, F. (2025). Self-Debiasing Large Language Models: Zero-Shot Recognition and Reduction of Stereotypes. *Proceedings of the 2025 NAACL (Short Papers)*, pp. 873–888. https://doi.org/10.18653/v1/2025.naacl-short.74

World Health Organization. (2024). State of the World's Nursing Report 2024. WHO Press.

United Nations Office on Drugs and Crime. (2005). *Forum on Crime and Society, Volume 4, Numbers 1 and 2 (December 2004)*. New York: United Nations. Data from RAND–National Memorial Institute for the Prevention of Terrorism (MIPT) database, Figure II, p. 59. https://css.unodc.org/documents/data-and-analysis/Forum/V05-81059_EBOOK.pdf

Fédération Internationale de Football Association. (2024). FIFA Professional Football Report 2024.

Zmigrod, R., Mielke, S. J., Wallach, H., & Cotterell, R. (2019). *Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 1651–1661. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1161

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, *QLoRA: Efficient Finetuning of Quantized LLMs*. In *Proceedings of NeurIPS*, 2023.

J. Zhao, K. Wang, M. Yatskar, V. Ordonez, and K. Chang, *Calibrate Before Use: Improving Few-Shot Performance of Language Models*. In *Proceedings of ICML*, 2021.

M. J. Kusner, J. Loftus, C. Russell, and R. Silva, *Counterfactual Fairness*. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. MIT Press, 2019.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. Chang, *Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods*. In *NAACL-HLT*, 2018.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, *A Survey on Bias and Fairness in Machine Learning*. *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.