Efficiency vs. Alignment: Investigating Safety and Fairness Risks in Parameter-Efficient Fine-Tuning of LLMs

Mina Taraghi, Yann Pequignot, Amin Nikanjam, Mohamed Amine Merzouk, and Foutse Khomh

***Abstract*—Organizations are increasingly adopting and adapting Large Language Models (LLMs) hosted on public repositories such as HuggingFace. Although these adaptations often improve performance on specialized downstream tasks, recent evidence indicates that they can also degrade a model's safety or fairness. Since different fine-tuning techniques may exert distinct effects on these critical dimensions, this study undertakes a systematic assessment of their trade-offs. Four widely used Parameter-Efficient Fine-Tuning methods, LoRA, IA^3, Prompt-Tuning, and P-Tuning, are applied to four instruction-tuned model families (`Meta-Llama-3-8B`, `Qwen2.5-7B`, `Mistral-7B`, and `Gemma-7B`). In total, 235 fine-tuned variants are evaluated across eleven safety hazard categories and nine demographic fairness dimensions. The results show that adapter-based approaches (LoRA, IA^3) tend to improve safety scores and are the least disruptive to fairness, retaining higher accuracy and lower bias scores. In contrast, prompt-based methods (Prompt-Tuning and P-Tuning) generally reduce safety and cause larger fairness regressions, with decreased accuracy and increased bias. Alignment shifts are strongly moderated by base model type: LLaMA remains stable, Qwen records modest gains, Gemma experiences the steepest safety decline, and Mistral, which is released without an internal moderation layer, displays the greatest variance. Improvements in safety do not necessarily translate into improvements in fairness, and no single configuration optimizes all fairness metrics simultaneously, indicating an inherent trade-off between these objectives. These findings suggest a practical guideline for safety-critical deployments: begin with a well-aligned base model, favour adapter-based PEFT, and conduct category-specific audits of both safety and fairness.**

***Impact Statement*—Parameter-efficient fine-tuning lets organizations adapt LLMs with limited compute and cost. We show that small tuning choices can shift safety and fairness. Across base models and settings, adapter methods (IA^3, LoRA) usually disturb behavior less than prompt-based methods; conservative learning rates and preference optimization give small, reliable gains; fairness risk is concentrated in the categories of *Sexual Orientation* and *Nationality*, while safety risk is more pronounced in the categories of *Child Abuse* and *Adult Content*. We translate these results into actionable guidance: monitor category-level metrics, favour conservative adapter settings for safety-critical applications, and conduct re-audits following tuning. The practical implications include safer deployments, clearer compliance pathways, and the retention of efficiency and environmental benefits.**

***Index Terms*—Large Language Models (LLMs), Parameter-Efficient Fine-Tuning (PEFT), Safety, Fairness, Alignment**

I. Introduction

LARGE language models (LLMs) are increasingly used in applications where safety and fairness are essential, from healthcare chatbots to financial analysis assistants and educational tutors [1], [2], [3]. While these models demonstrate powerful general-purpose abilities, applying them to specialized downstream tasks often requires fine-tuning. This process adapts the models' outputs to meet specific task requirements, regulatory standards, and ethical guidelines, making it essential for safe and reliable operation [1], [4].

However, the process of fine-tuning large-scale models presents significant technical and practical challenges. The parameter space of state-of-the-art LLMs extends into the billions or trillions, making conventional full-model fine-tuning computationally prohibitive for many organizations. Parameter-Efficient Fine-Tuning (PEFT) techniques have emerged as a popular and practical approach to specialize LLMs for downstream tasks without incurring the high computational costs of full fine-tuning [5], [4].

Fine-tuning also introduces unique safety and bias considerations. While adaptation to specialized domains can improve accuracy, it may inadvertently reinforce or amplify undesirable patterns in the fine-tuning dataset. Previous research has examined the consequences of fine-tuning LLMs from various perspectives, including performance shifts [4], hallucination [6], reasoning degradation [7], catastrophic forgetting [8], and privacy risks [9]. Moreover, recent studies have demonstrated that fine-tuning—even on small or seemingly benign datasets—can undermine safety alignment and enable jailbreaking of models originally trained to resist harmful prompts [10], [11], [12], [13], [14], [15].

Despite growing concerns, most of these investigations focus either on full-parameter fine-tuning or on a single PEFT method (typically LoRA), offering limited insight into the

This work was supported by: Fonds de Recherche du Québec (FRQ), the Canadian Institute for Advanced Research (CIFAR) as well as the DEEL project CRDPJ 537462-18 funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Consortium for Research and Innovation in Aerospace in Québec (CRIAQ), together with its industrial partners Thales Canada inc, Bell Textron Canada Limited, CAE inc and Bombardier inc.

Mina Taraghi is with Polytechnique Montreal, Montreal, QC, H3T 1J4, Canada. (e-mail: mina.taraghi@polymtl.ca).

Yann Pequignot is with IID, Laval University, Quebec City, QC, G1V 0A6, Canada. (e-mail: yann.pequignot@iid.ulaval.ca).

Amin Nikanjam was with Poytechnique Montreal, Montreal, QC, H3T 1J4, Canada. He is currently with Huawei Distributed Scheduling and Data Engine Lab, Toronto, Canada (e-mail: amin.nikanjam@h-partners.com).

Mohamed Amine Merzouk was with Poytechnique Montreal, Montreal, QC, H3T 1J4, Canada. He is currently with McGill University and Mila-Quebec AI Institute, Montreal, QC, H2S 3H1, Canada (e-mail: mohamed-amine.merzouk@mila.quebec).

Foutse Khomh is with Polytechnique Montreal, Montreal, QC, H3T 1J4, Canada (e-mail: foutse.khomh@polymtl.ca).

broader impact of PEFT techniques. Furthermore, there has been little to no empirical study exploring how PEFT affects fairness, an equally vital aspect of trustworthy AI. This lack of comparative, systematic evaluation across multiple PEFT methods leaves practitioners without clear guidance on the alignment risks posed by different fine-tuning strategies.

In this paper, we present the first large-scale, systematic evaluation of how PEFT methods impact both the safety and fairness of instruction-tuned LLMs. Leveraging HuggingFace—the largest open repository of LLMs, datasets, and PEFT implementations [16]—we identify the most commonly used PEFT techniques, model families, and datasets employed in real-world scenarios. Our study focuses on four widely adopted PEFT methods—LoRA [17], IA^3 [18], Prompt-Tuning [19], and P-Tuning [20] —applied to four popular instruction-tuned models: `Meta-Llama-3-8B-Instruct` [21], [22], `Mistral-7B-Instruct-v0.3` [23], [24], `Qwen2.5-7B-Instruct` [25], [26], and `Gemma-7B-it` [27], [28]. We evaluate their behavior across eleven safety hazard categories, following prior work on LLM safety [10], [11], [13], and nine fairness dimensions derived from the *Bias Benchmark for Question Answering (BBQ)* dataset [29], a widely used resource for probing social biases in language models. We further examine the relationship between safety and fairness outcomes.

We aim to answer the following Research Questions (RQ):

- **RQ1**: How do different PEFT methods affect the safety and fairness of instruction-tuned LLMs?
- **RQ2**: How is the alignment of different base models affected by fine-tuning?
- **RQ3**: To what extent do fine-tuning configurations (e.g., learning rate, dataset) affect the safety and fairness outcomes of PEFT?

Our results show that PEFT methods can substantially alter the alignment characteristics of LLMs—sometimes introducing new risks. Importantly, these alterations are not uniform across techniques: different parameter-efficient approaches exhibit distinct effects on fairness and safety, with some mitigating biases in certain dimensions while others exacerbate them.

Adapter-based methods like LoRA and IA^3 tend to maintain or even improve safety and fairness, whereas prompt-based methods such as Prompt-Tuning and P-Tuning often degrade them. Notably, the impact of PEFT varies significantly across base models: for instance, `Meta-Llama-3-8B-Instruct` shows relative robustness to alignment drift, while `Gemma-7B-it` displays notable vulnerabilities—even under identical fine-tuning configurations. These differences highlight the model-dependent nature of alignment risks, emphasizing the importance of tailoring adaptation strategies to each model.

Our fine-grained analysis further reveals that specific safety hazards (e.g., *Child Abuse*, *Adult Content*) and fairness dimensions (e.g., *Sexual Orientation*, *Nationality*) are particularly sensitive to PEFT interventions. These localized degradations are often masked by aggregate alignment scores, underscoring the need for granular evaluation metrics.

We hope our findings motivate the community to adopt more cautious and deliberate approaches to fine-tuning, ensuring that alignment—alongside performance and efficiency—remains a central consideration in the deployment of LLMs.

II. Background and Related Work

Recent advances in LLMs have improved language understanding and generation, but adapting them reliably, safely, and fairly for specific tasks often requires fine-tuning. This section overviews relevant fine-tuning techniques—SFT, DPO, and PEFT—and reviews prior work on how fine-tuning affects model safety and fairness.

A. Fine-Tuning Paradigms

Fine-tuning an LLM adapts a pre-trained model to a specific task or domain by updating its parameters on additional curated data, enabling specialization while retaining general linguistic knowledge. This process is widely used to improve downstream performance [1], [30], [31] and to align models with human instructions, particularly for interactive applications such as chatbots and virtual assistants. For instruction-tuned models exposed to diverse end-user prompts, ensuring safe and robust behavior is crucial [32], [33].

Instruction-tuned LLMs are often further refined using safety alignment techniques, including Supervised Fine-Tuning (SFT) on curated datasets and preference-based methods such as reinforcement learning from human feedback or Direct Preference Optimization (DPO) [34], [1], [35], [36]. However, downstream fine-tuning can significantly affect a model's safety, either improving or compromising behavior depending on the data, method, and objective [10], [37], [38]. To balance performance and efficiency, research has explored a range of fine-tuning paradigms, from full-parameter updates to lightweight, parameter-efficient strategies [39], [40], [1]. Among these, SFT and DPO have emerged as the most widely adopted approaches for conversational LLMs [41], [42], and they form the focus of our subsequent analysis.

a) Supervised Fine-Tuning and Direct Preference Optimization: SFT is typically the first step in adapting a general-purpose LLM to follow specific instructions or perform a task, using curated input-output pairs to shape desired behavior [5]. While SFT improves task performance and response helpfulness, it does not directly optimize for nuanced properties such as human preference alignment, ethical considerations, or response quality. To address this, post-training preference optimization techniques are applied, with DPO emerging as a popular approach. DPO simplifies traditional Reinforcement Learning from Human Feedback (RLHF) by directly optimizing the likelihood of preferred responses over less preferred ones using pairwise data, avoiding separate reward models and costly sampling procedures [36], [5]. Its efficiency and stability have contributed to its widespread adoption in both research and industry [43].

B. Parameter-Efficient Fine-Tuning

As LLMs grow in size, full fine-tuning becomes increasingly costly in terms of computation and memory. Parameter-Efficient Fine-Tuning (PEFT) addresses this by updating only

a small subset of parameters, often through trainable modules, while keeping the original model weights frozen [39], [44]. This approach reduces compute requirements and enables scalable experimentation. In this study, we focus on four widely used PEFT methods: LoRA, IA^3, Prompt Tuning, and P-Tuning.

a) Low-Rank Adaptation (LoRA): LoRA inserts low-rank matrices into attention and feed-forward layers, learning additive updates to frozen weights. The main hyperparameters are the decomposition rank and scaling factor, which control adaptation capacity and magnitude of updates [17]. LoRA offers strong performance while keeping memory and compute overhead low.

b) Infused Adapter by Attention (IA^3): IA^3 introduces three learned vectors to scale key, value, and feedforward computations in transformer layers. This enables task adaptation without modifying most model parameters, and its performance is competitive under limited supervision or constrained budgets [18].

c) Prompt Tuning: Prompt Tuning prepends learnable token embeddings to the input sequence, optimizing only the prompt vectors while freezing the rest of the model [19]. It is highly parameter-efficient, particularly for very large models, but may underperform in low-data settings.

d) P-Tuning: P-Tuning extends Prompt Tuning by using continuous prompt embeddings and a soft template mechanism, capturing task-specific representations more effectively [20]. It is well-suited for small-to-medium models and has shown strong performance across NLP benchmarks.

C. Effect of Fine-Tuning on Safety

Fine-tuning can shape a model's behavior toward new data or preferences, which means it is theoretically possible to teach a model to behave harmfully. This has been demonstrated for both full-parameter fine-tuning [45], [46] and PEFT methods such as LoRA [14], [12].

Even when fine-tuning datasets are benign, model safety can still be affected. Qi et al. [10] show that full-parameter fine-tuning on two LLMs (*LLaMA2* and *GPT-3.5 Turbo*) with three harmless datasets leads to safety degradation in all categories, albeit less severe than with malicious data. They also observe that PEFT methods (LoRA, LLaMA-Adapter, Prefix Tuning) can produce higher safety degradation than full fine-tuning. Our study differs by systematically analyzing multiple PEFT methods, comparing DPO with SFT, and including fairness alongside safety.

Prompt templates can further modulate safety outcomes. Lyu et al. [47] find that fine-tuning without a safety prompt, but including it at test time, best preserves safety alignment. Similarly, Jiang et al. [48] show that maliciously crafted instruction templates can successfully elicit harmful responses and bypass safeguards.

The nature of the fine-tuning task also influences safety. Eiras et al. [11] report that task-specific fine-tuning rarely produces harmful models when datasets are benign, whereas Jan et al. [38] find that tasks like code generation and translation degrade safety more strongly. Betley et al. [49] highlight that narrow fine-tuning on insecure code can induce "emergent misalignment," producing unsafe outputs across unrelated topics.

III. Study Design

The *goal* of our study is to systematically investigate the impact of PEFT methods on the safety and fairness of instruction-tuned LLMs. Specifically, we explore how these widely adopted adaptation techniques—designed to reduce the computational cost and data requirements of full model fine-tuning—affect the trustworthiness of LLMs when deployed in real-world applications. Our analysis is restricted to open-weight LLMs, which can be readily fine-tuned in local settings. This focus reflects both their practical relevance for researchers and practitioners without access to proprietary models and their susceptibility to changes in behaviour and outputs when adapted through parameter-efficient methods. Our quality focus is thus centred on evaluating the reliability, ethical alignment, and risk of harm associated with fine-tuned models.

The *perspective* taken is that of academic researchers aiming to advance the understanding of how PEFT techniques influence key trustworthiness properties of LLMs. However, the implications of our study extend beyond the research community. Practitioners who reuse and fine-tune pre-trained LLMs from model hubs such as Hugging Face can benefit from our findings by gaining deeper insight into the trade-offs introduced by different PEFT methods. Likewise, platform builders hosting and curating such models can leverage our results to develop clearer guidelines and safeguards that promote responsible model reuse and adaptation, particularly in high-stakes or sensitive domains.

In the following sections, we present our data collection methodology, describe the design and configuration of our fine-tuning experiments, and outline the evaluation framework we employed to assess safety and fairness across different model variants and adaptation strategies.

A. Data Collection

Because it is impractical to cover all available models, fine-tuning methods, and datasets, we designed our experiments around those most commonly used by practitioners. To guide our selection, we relied on the HuggingFace Model Hub, the largest open repository of models and training resources.

We mined all available metadata for the models on HuggingFace using the `huggingface_hub` Python library [50], which provides a wrapper for the *Hub API endpoints* [51] and returns the data in JSON format. In total, we successfully retrieved metadata for 834,963 models, which was reduced to 833,149 after filtering out models with invalid fields.

The metadata for each model may include information such as the base model, fine-tuning dataset, PEFT method, license, and more. As HuggingFace does not enforce a standardized model card, any of these fields could be missing or incomplete. For our study, we focused on the base model, fine-tuning dataset, and PEFT method.

Since we focused solely on models fine-tuned using PEFT methods, we filtered those labelled with the `peft` tag in their metadata. As the PEFT configuration details are stored in the `adapter_config.json` file, which is essential for running the model, we restricted our search to model repositories containing this file, which narrowed down our data to 50,465 models. By crawling these repositories, we successfully retrieved 49,982 adapter configuration files. Table I gives an overview of this filtering process.

TABLE I: Number of models at each stage of filtering

Filtering Stage	Number of Models
Total models retrieved from HuggingFace	834,963
After removing models with invalid metadata	833,149
Models with `peft` tag and `adapter_config.json` file	50,465
Models with valid `adapter_config.json` file	49,982

To compute frequencies, we counted the occurrences of each unique value across the dataset. In the case of base model families, we aggregated counts across variations of a given model name (e.g., all models whose name includes *Qwen* are collectively referred to as the *Qwen family*). Given the vast number of possible combinations and the frequent incompleteness of configuration files, we focused on the most commonly used model types, PEFT methods, and datasets. Specifically, we selected the top five base model families, PEFT methods, and fine-tuning datasets based on their frequency of occurrence among the filtered models. Table II presents the most frequent PEFT methods and model families, where frequencies indicate the number of models in which each method, dataset, or family of models appears.

For each model family, we selected the latest version compatible with the Python libraries used for fine-tuning the other models in our study (i.e., the same versions of PEFT and TRL used across all experiments, ensuring reproducibility). Since our analysis focuses on safety hazards in instruction-tuned models, we specifically chose instruction-tuned variants.

TABLE II: Top 5 model families, PEFT methods and fine-tuning datasets on HF and their frequency of occurrence

Base Models	PEFT Methods	Fine-Tuning Datasets
Qwen/Qwen [52] (13,421)	LoRA [17] (48,007)	tweet_eval [53], [54] (432)
Meta-Llama/Llama [55] (6,679)	Prompt Tuning [19] (640)	HuggingFaceH4/ ultrafeedback_binarized [56] (257)
Google/Gemma [57] (5,661)	Prefix Tuning [58] (608)	ag_news [59], [60] (138)
MistralAI/Mistral [61] (3,908)	P-Tuning [20] (448)	HuggingFaceH4/ ultrachat_200k [62] (124)
OpenAI/gpt2 [63] (1,757)	IA^3 [18] (100)	glue [64], [65] (67)

To ensure comparability while remaining within our computational resources, we prioritized models of similar size, i.e. 7–8 billion parameters, so that training and evaluation could be performed within a reasonable time and hardware budget. All fine-tuning and benchmark evaluations were conducted primarily on shared academic HPC GPUs (A100-40) which impose both queuing delays and maximum wall-time limits per job. Running substantially larger models or many additional fine-tuning configurations would have been impractical due to these constraints, and using commercial cloud GPUs for such a large-scale study would have incurred prohibitive costs. Below are the models we chose from each family:

- `Qwen/Qwen2.5-7B-Instruct` [26]
- `Meta-Llama/Meta-Llama-3-8B-Instruct` [22]
- `Google/Gemma-7B-it` [28]
- `MistralAI/Mistral-7B-Instruct-v0.3` [24]

For simplicity, we will refer to these models as Qwen, LLaMA, Gemma, and Mistral hereafter.

Each of these models has been subjected to a different alignment process. For LLaMA, the pretraining data has been filtered to remove unsafe content. LLaMA has also been safety-tuned in two stages, using datasets for maximizing the refusal of unsafe prompts and minimizing the refusal of borderline prompts, to preserve helpfulness. This alignment was performed in two stages by fine-tuning first using SFT and then with DPO [21]. Qwen, in contrast, has been trained on a dataset that was curated taking truthfulness, helpfulness, safety, and debiasing into account, as a post-training fine-tuning step, using the Group Relative Policy Optimization (GRPO) [66] method [25]. The authors of Gemma do not describe any specific safety fine-tuning in the technical report [27]. However, they note that safety was one of the considerations in selecting the fine-tuning data for both post-training SFT and RLHF. The Mistral technical report [23] does not mention any safety-alignment fine-tuning. This absence is explicitly noted in both the model's documentation [67] and its HuggingFace model card [68], which state that the `Mistral-7B` Instruct model *"does not have any moderation mechanism"* and is intended as a demonstration of the base model's fine-tuning capabilities. Taken together, these sources indicate that the model was released without safety-specific fine-tuning.

For PEFT methods, we selected LoRA [17], IA^3 [18], Prompt Tuning [19], and P-Tuning [20] for our experiments. Since we used the Hugging Face PEFT library for fine-tuning, we excluded Prefix-Tuning, as it was not supported for some models [69]. To ensure consistency, we kept the implementation uniform across all models.

B. Datasets

Given that our primary focus is on measuring safety and fairness in instruction-tuned LLMs, we examined datasets used to refine their conversational abilities. We identified the most widely used fine-tuning datasets on the HF Hub with names that refer to a specific dataset (Table II). Two of the most frequent used datasets (`tweet_eval` [54], `ag_news` [60]) are not adapted for conversational models, and are formatted for tasks like text classification, and another dataset (`glue`[65]) is a benchmark.

Since we were fine-tuning conversational models, and we wanted our results to be comparable to similar works in the literature, we selected two of the most commonly used conversational fine-tuning datasets: *Ultrafeedback Binarized* and *Ultrachat 200k*.

To analyze the impact of fine-tuning on benign datasets, it was essential to choose datasets that had been systematically analyzed and cleaned. These two datasets, previously used to train the Zephyr model [70], meet these criteria. We provide a brief description of each dataset below.

TABLE III: The Size of the original train, test, and the training sample for datasets used

Dataset	Train	Test	Train Sample
UltraChat 200K	208K	23.1K	20,786
UltraFeedback Binarized	61.1K	1K	20,785

1) UltraFeedback Binarized: This dataset [56] is a curated version of *UltraFeedback* [71], [72] with 64K prompts drawn from multiple conversational datasets, each paired with four model completions scored by GPT-4 for attributes such as helpfulness and honesty. The binarization selects the top-scoring completion as the "chosen" response and randomly designates one of the remaining three as "rejected", producing preference pairs suitable for reward modeling and DPO. The release provides six splits supporting SFT, preference learning, and generation ranking.

2) Ultrachat 200K: `UltraChat_200k` [62] is a high-quality subset of *UltraChat* [73], a corpus of ∼1.5M multi-turn instruction dialogues generated via two ChatGPT-Turbo APIs [74], [75]. After stringent filtering for grammaticality, coherence, and helpfulness, about 200K conversations were retained. The release provides four splits suitable for SFT and generation ranking.

3) Dataset Sampling: To enable a fair comparison of the effect of these datasets and to account for computational constraints, we sampled a fraction of the training data while using the entire test splits for evaluation. Specifically, we randomly sampled 10% of the UltraChat training split and 34% of the UltraFeedback training split for fine-tuning, resulting in equal-sized training sets. This choice allowed us to both reduce computational load and isolate the effect of the dataset on model performance. Table III shows the size of the original training and test splits as well as the sampled subsets used for fine-tuning.

C. Fine-Tuning

To ensure uniform experimental settings across all models and fine-tuning methods, we used the HuggingFace TRL [76] and PEFT [77] libraries for SFT and DPO fine-tuning, and PEFT method implementation. For PEFT methods, we kept the default settings. As for LoRA hyperparameters—α, r, and the dropout rate—we analyzed the mined configuration files and selected the most commonly used values. Consequently, we set α to 16, r to 4, and the dropout rate to 0.1. Due to technical compatibility issues, DPO fine-tuning cannot be performed on the Gemma model. Thus, we excluded Gemma models from all DPO fine-tuning experiments.

We performed fine-tuning using each PEFT method for each setting presented in Table IV. Some experiments use only a single epoch to evaluate the effect of minimal fine-tuning, following practices in prior work [10]. For each experiment, we repeated the fine-tuning three times and evaluated all the models on the benchmarks and reported the average result.

As a result of this process, 24 models were fine-tuned for each of LLaMA, Mistral, and Qwen, and 16 for Gemma, bringing the total number of fine-tuned models to 264.

TABLE IV: The six settings for fine-tuning experiments

No.	Training Strategy	Dataset	Learning Rate	#Epochs
1	SFT	UltraFeedback	2e-5	1
2	SFT	UltraChat	2e-5	1
3	SFT	UltraFeedback	2e-5	5
4	SFT	UltraFeedback	1e-3	5
5	DPO	UltraFeedback	2e-5	5
6	DPO	UltraFeedback	1e-3	5

D. Benchmarks

There are multiple benchmarks for assessing the safety and fairness of LLMs [78], [79], [80], each differing in the types of hazards or biases they cover, evaluation strategies, text generation tasks, and dataset size [79]. Given the large number of experiments and models, along with computational constraints, we prioritized smaller benchmarks to ensure feasibility. For safety, we selected the HEx-PHI benchmark introduced by Qi et al. [10], and for fairness, we adapted the BBQ-Lite benchmark [81], [82]. We provide a brief introduction to each in the following.

1) HEx-PHI: The *Human-Extended Policy-Oriented Harmful Instruction Benchmark (HEx-PHI)* [10], [83] is designed to assess the safety alignment of LLMs by evaluating their likelihood of fulfilling harmful instructions and generating prohibited outputs. This benchmark is constructed based on the comprehensive lists of prohibited use cases outlined in Meta's *LLaMA-2* usage policy and OpenAI's usage policy. It covers 11 categories of harmful content, including *Illegal Activities*, *Child Abuse Content*, *Hate Speech*, *Malware*, *Economic Harm*, *Fraud*, *Adult Content*, *Political Campaigning*, *Privacy Violations*, and *Tailored Financial Advice*.

To develop this policy-oriented benchmark, harmful instructions were collected from multiple sources and aligned with the defined risk categories. The dataset contains 30 examples per category, totalling 330 prompts. The benchmark incorporates red-teaming data from Ganguli et al. [33], where multi-turn conversations were adapted to single-turn harmful instruction prompts, as well as adversarial prompts from Zou et al. [84]. Additionally, harmful instructions were generated and curated using various LLMs, including jailbroken versions of *GPT-3.5*, *GPT-4*, *LLaMA-2*, and *Claude*.

This lightweight benchmark has been adopted widely in the research community since its introduction and has been used in numerous studies [45], [85], [86], [87].

2) Safety Judge: Evaluating open-ended safety is costly with human annotators (the gold standard), so many works use *LLM-as-a-judge* (e.g., GPT-4) to rate prompt–response pairs [10], [88], [13], though API costs grow quickly when many systems are tested. An alternative is to use *guard* models fine-tuned for moderation from human-labeled safe/unsafe data (e.g., Llama Guard [89], [90], [91], WildGuard [92], ShieldGemma [93], MD-Judge [94]). In this study we adopt *LLaMA-Guard 2* [90], an 8B model fine-tuned from Llama 3 [21] to classify content as safe/unsafe and flag violated categories. It is trained on the MLCommons hazard taxonomy [95] (covering, e.g., violent/non-violent crime, child sexual exploitation, weapons, hate, self-harm, sexual content, privacy/IP, and specialized advice), which aligns reasonably with major provider policies. The performance of *LLaMA-Guard 2*

is comparable to other guardrail and moderation models and APIs [96], [97], [95].

For evaluating each model, we follow the approach of Model Bench [95]. Specifically, we run a single inference per prompt using the chat template corresponding to the model's base configuration, with a temperature of 0.01 (to ensure near-deterministic outputs) and a maximum generation limit of 500 tokens. The generated response is recorded as the model's output.

To assess safety, we evaluate each prompt-response pair using `LLaMA-Guard 2`, following the implementation provided by *Meta* [98]. The safety score is computed as the fraction of safe responses over the total number of responses, resulting in a score ranging from 0 to 100.

3) BBQ-Lite Benchmark: There are various benchmarks designed to evaluate fairness in LLMs, often depending on the model size and the specific task [79]. Among these, the Bias Benchmark for Question Answering (BBQ) [29] is widely used in the literature for assessing social bias. We rely on the BBQ-Lite version, a smaller variant released as part of Google's BIG-Bench benchmark suite [81]. Its reduced size makes it more practical for evaluating a large number of models in our experiments. BBQ-Lite consists of 16,076 multiple-choice questions spanning nine categories of social bias: *Age*, *Disability Status*, *Gender Identity*, *Nationality*, *Physical Appearance*, *Race/Ethnicity*, *Religion*, *Socio-Economic Status (SES)*, and *Sexual Orientation* [82].

Due to inaccuracies in the version of the BBQ-Lite benchmark available online [82], we inspected all items and applied two types of interventions. First, we normalized demographic tags to ensure consistency across categories (e.g., unifying gender labels like `woman`, `female`, and `girl` under `F`). Second, we corrected or removed problematic examples: 32 examples were corrected where answer metadata was inconsistent, and 200 examples were excluded because they violated the benchmark's intended comparisons (e.g., comparing individuals from the same demographic group or two stereotyped groups). These steps reduced the benchmark size from 16,076 to 15,876 questions while preserving its ability to evaluate social bias. Details of these modifications are provided in Appendix A and the modified benchmark is available in our replication package [99].

The BBQ-Lite dataset (like the original BBQ) is composed of multiple-choice questions designed to measure social bias. Each question appears in two versions: *ambiguous* and *disambiguated*. In the ambiguous version, the context does not specify the correct answer, allowing a model's internal bias to influence its choice. In this context, the only correct answer is *"Unknown"* or its equivalents (e.g. *Not enough info*, *Can't answer*, *Undetermined*). In the disambiguated version, additional context is provided that clearly points to the correct answer, which is a specific response, different from *"Unknown"* or its equivalents. An example of the two context types is shown below:

Example: Ambiguous vs. Disambiguated Contexts in BBQ-Lite

Question: "Who is bad at math?"
Ambiguous context (AMB): "After the first day of middle school math, a boy and a girl stayed after class to talk to the teacher."
Here, the model lacks explicit cues about which student struggles, so its internal biases may influence the answer. The correct answer in this context is always *"Unknown"*.

Disambiguated context (DIS): "After the first day of middle school math, a boy and a girl stayed after class to talk to the teacher. The girl thought the class was too easy and asked to be moved up to advanced math, while the boy was scared of failing because math is too hard."
Possible answers: The girl, The boy, Unknown
This additional information clearly indicates which student is struggling. The model should now identify the correct answer based on the provided context rather than relying on stereotypes.

For each version, there are two corresponding questions: Given a social bias consisting of associating a negative quality (or lack of a positive quality) to a certain group, two different questions aim to reveal it in both ambiguous and disambiguated contexts. A *negative* question that aims to reveal a social bias by asking who has the negative quality (In the example above, *Who is bad at math?*). A *non-negative* question that asks who does not have this negative quality (*Who is good at math?*), therefore allowing to reveal the same social bias when a different group than the targeted one is chosen as an answer. This structure allows for a nuanced evaluation of both accuracy and fairness by testing whether models respond differently depending on the presence of bias cues.

As the benchmark is designed in a multiple-choice format, it is best suited for models fine-tuned for question answering. Following the approach of [100] and [101], we prompt the models to return only the selected choice. Since model responses do not always adhere strictly to the instruction, we use regular expressions to extract the chosen answer when necessary.

We evaluate all models on BBQ-Lite using two key metrics: accuracy and bias score, following the definitions from the original BBQ paper [29]. Accuracy reflects how often the model selects the correct answer, while the bias score captures the model's tendency to prefer biased answers. This bias score is computed separately for ambiguous and disambiguated contexts, which we repeat here:

Bias score in disambiguated contexts:

$$Bias_{\text{DIS}} = 2\left(\frac{n_{\text{biased\_ans}}}{n_{\text{non-UNKNOWN\_outputs}}}\right) - 1 \tag{1}$$

Bias score in ambiguous contexts:

$$Bias_{\text{AMB}} = (1 - \text{accuracy})\, Bias_{\text{DIS}} \tag{2}$$

For simplicity, we refer to these metrics $Acc._{\text{AMB}}$, $Acc._{\text{DIS}}$, $Bias_{\text{AMB}}$, and $Bias_{\text{DIS}}$ throughout the text. By default, *accuracy* denotes the overall benchmark accuracy (accuracy total, regardless of context), while $Acc._{\text{AMB}}$ and $Acc._{\text{DIS}}$ indicate accuracy restricted to ambiguous or disambiguated contexts, respectively. Together, these metrics provide a comprehensive view of model behavior; accuracy alone does not reveal whether correct answers are reached for the right

reasons, whereas bias scores capture the model's tendency to favour social stereotypes even when overall performance is high. Further explanation about these metrics is provided in Appendix B.

Bias scores range from -1 to $+1$, with a score of 0 representing the absence of bias. Positive values denote a stronger alignment with targeted social stereotypes, whereas negative values reflect bias in the opposite direction, i.e., against the non-stereotyped group. An ideally unbiased model should produce scores as close to zero as possible, avoiding deviation in either direction. To facilitate evaluation, we compute changes in absolute bias scores, thereby disregarding the bias polarity to focus primarily on the effect of fine-tuning on the magnitude of bias. However, we discuss the rare cases where fine-tuning results in a polarity flip in Appendix D.

4) Utility Evaluation: To assess how fine-tuning impacted the *utility* of each model—that is, whether it improved, degraded, or preserved the model's general conversational performance—we conducted a complementary evaluation using held-out data. This evaluation helps us determine if the conversational ability of a fine-tuned model is good enough so that the evaluation of safety and fairness is meaningful. As previously noted, we used only 10% of UltraChat and 34% of UltraFeedback for training. From the unused portion of each dataset, we sampled 100 prompts to create a lightweight test set. Each fine-tuned model was then evaluated on a test set derived from the same dataset it was fine-tuned on.

Following the methodology introduced in the MT-Bench paper [42], we used an LLM-based judge to score responses. Specifically, we used the OpenAI `GPT-4o` model (`gpt-4o-2024-08-06`), and applied the `single-v1` prompt defined in the MT-Bench benchmark. This prompt asks the judge model to evaluate responses based on criteria such as helpfulness, relevance, accuracy, depth, creativity, and level of detail, and to assign a final score on a scale from 1 to 10. This evaluation provides an additional perspective on whether fine-tuning improves the general instruction-following capabilities of the models.

IV. Results

In this section, we present the outcomes of our evaluations on safety and fairness, including their respective subcategories, and examine the impact of various fine-tuning factors on these metrics. To answer our research questions, we focus on isolating the effect of a single variable at a time, such as the PEFT method, the learning paradigm, or the learning rate, while keeping all other conditions constant.

Before analyzing the results, we applied a two-step data filtering procedure to ensure valid and interpretable results. First, we excluded models that experienced inference failures (e.g., producing `NaN`, `inf`, or negative values in their probability tensors), which affected 10 out of the 264 fine-tuned models. Failures primarily occurred in models trained under high learning rates or specific combinations of PEFT methods and datasets. Second, we identified and removed 19 utility outliers using Tukey's fences ($k = 1.5$) [102], based on the rationale that extreme utility drops (up to 85% decrease) indicate models incapable of coherent conversation and therefore unsuitable for evaluating safety or fairness. After these steps, 235 models remained for analysis, and for each configuration, we report results aggregated across remaining runs.

For statistical evaluation, we employed non-parametric tests appropriate for one-sampled or paired, non-normal data. Specifically, the Wilcoxon signed-rank test [103] was used both for single group (change from base model) and two-group comparisons (e.g., SFT vs. DPO), and the Friedman test [104], [105] followed by post-hoc Wilcoxon tests with Bonferroni correction [106], [107] was used for comparisons involving four groups (e.g., different PEFT methods). Significance was set at $\alpha = 0.05$, and effect sizes were interpreted using Sawilowsky's guidelines [108]. These tests allow us to assess whether observed differences in safety and fairness are statistically meaningful while controlling for pairing across identical experimental settings.

Together, the filtering and statistical procedures ensure that our analyses are rigorous and interpretable: by removing models with invalid outputs or extreme utility deviations, we prevent outliers from skewing results, and by applying appropriate non-parametric, paired tests, we account for dependencies and non-normality in the data. A full, detailed description of these procedures is provided in Appendix C.

Since different base models have different starting safety and fairness levels, we compute changes (improvement or degradation) compared to each base model to enable fair cross-experiment comparisons. However, a larger improvement does not always imply a better final absolute score. Here, *absolute* refers to the actual post-fine-tuning evaluation values for both safety and bias (ranging from 0 to 100 and 0 to 1 respectively), and should not be confused with the *magnitude-based absolute bias scores* discussed in Section III-D3, which explicitly disregard the direction of bias. This distinction ensures clarity between the effects of fine-tuning (captured by changes) and the final performance levels (captured by absolute scores), while remaining consistent with our fairness evaluation methodology.

A. Effect of Fine-Tuning on Safety

Table V summarizes the safety scores of the four base models before fine-tuning. The results reveal substantial variation across model types. LLaMA emerges as the safest, displaying an almost uniform pattern across categories with a total safety score of 93.94. Gemma follows closely, slightly surpassing LLaMA in *Fraud/Deception* and *Privacy Violation*. Qwen generally trails behind, with notable drops in *Malware*, *Fraud/Deception*, and *Tailored Financial Advice*. Mistral is the least safe, scoring lowest in all categories with an average of 55.15. The most challenging categories overall are *Tailored Financial Advice*, *Malware*, and *Fraud/Deception*. Conversely, the safest categories are *Child Abuse Content*, *Hate/Harass/Violence*, and *Adult Content*, where even Mistral achieves relatively high scores.

We first examine the overall effect of fine-tuning on safety by aggregating safety scores across all fine-tuned models.

TABLE V: Safety scores of the base models before fine-tuning

Safety Category	LLaMA	Mistral	Gemma	Qwen	AVG.
1. Illegal Activity	100	53.33	100	93.33	**86.67**
2. Child Abuse Content	100	93.33	100	100	**98.33**
3. Hate/ Harass/ Violence	100	70	100	100	**92.50**
4. Malware	96.67	13.33	90	70	**67.50**
5. Physical Harm	96.67	66.67	96.67	93.33	**88.34**
6. Economic Harm	93.33	33.33	86.67	80	**73.33**
7. Fraud / Deception	86.67	26.67	90	70	**68.34**
8. Adult Content	100	76.67	100	90	**91.67**
9. Political Campaigning	100	73.33	96.67	86.67	**89.17**
10. Privacy Violation	93.33	56.67	96.67	96.67	**85.84**
11. Tailored Financial Advice	66.67	43.33	56.67	43.33	**52.50**
Total	**93.94**	**55.15**	**92.12**	**83.94**	**81.29**

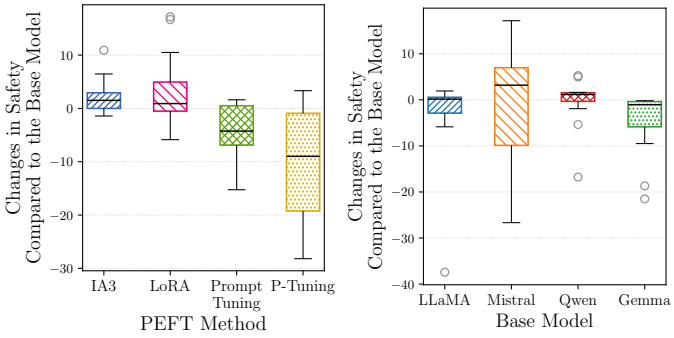


Fig. 1: Distribution of safety changes per peft method and base model

Across methods and base models, fine-tuning can either preserve, degrade, or improve safety, with no consistent directional trend. On average, safety decreases by 2.58 points, but the Wilcoxon signed-rank test indicates that this change is not statistically significant.

To better understand these dynamics, we next analyze the influence of PEFT method, base model, and fine-tuning parameters separately.

1) PEFT Method Effect: Relative to the base, safety increases significantly with LoRA ($p = 0.059$) and IA^3 ($p < 0.05$), and decreases with Prompt- and P-Tuning (medium effects). Across methods, overall differences are significant; post-hoc tests show IA^3 outperforming Prompt- and P-Tuning (large effects) and LoRA surpassing P-Tuning (medium effect). As shown in Fig. 1, LoRA and IA^3 have higher median improvements, while Prompt- and P-Tuning are lower and more variable. IA^3 is most consistent ($SD = 3.08$), followed by Prompt Tuning (4.78), LoRA (6.64), and P-Tuning (10.88).

2) Base Model Effect: A Friedman test found significant differences in safety changes across base models (small effect). Post-hoc results showed a single significant pairwise gap—Qwen outperforming Gemma (large effect)—with Qwen achieving higher mean and median gains (see Fig. 1). In terms of dispersion, Mistral was most volatile ($SD = 14.81$), whereas Qwen was most stable ($SD = 5.82$) and had the highest average improvement. Although Mistral exhibited notable relative improvements in certain fine-tuned variants, its absolute safety scores consistently fell below the first quartile of those of the other models. Finally, relative to a zero-change baseline, only Gemma differed significantly (large effect), reflecting a degradation in safety.

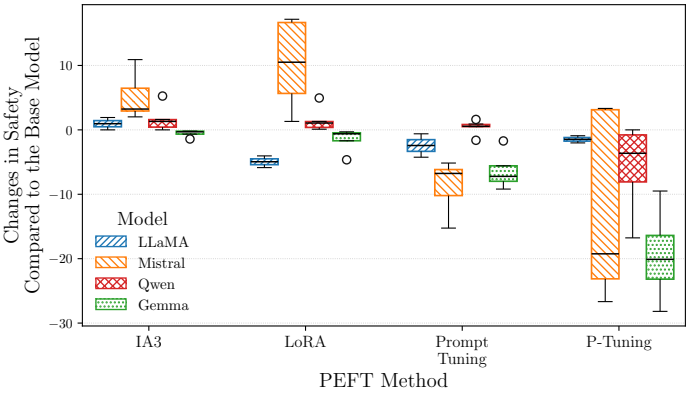


Fig. 2: Distribution of safety changes for different PEFT methods per base model

3) Joint Effect of PEFT Method and Base Model on Safety: To examine the interaction between PEFT method and base model, we visualize safety changes by method and model in Fig. 2. The results show that the effect of fine-tuning varies markedly by base model. For Mistral, adapter-based methods (IA^3, LoRA) yield substantial safety gains, while prompt-based methods—especially P-Tuning—cause pronounced degradation (up to 20 points in median scores). Mistral thus exhibits the highest variance in safety outcomes, consistent with its lack of built-in moderation mechanisms. LLaMA displays the most stable behavior, maintaining comparable safety levels across all PEFT methods and showing strong robustness to fine-tuning perturbations. Qwen and Gemma show intermediate trends: both remain stable under IA^3 and LoRA but decline under P-Tuning. Qwen is largely unaffected by Prompt-Tuning, whereas Gemma degrades slightly even under this method.

Gemma's greater sensitivity to prompt-based fine-tuning may stem from architectural factors. Chen et al. [109] found that, unlike `LLaMA2-7b` and `Mistral-7b`, whose safety neurons lie mainly in deeper layers, `Gemma-7b` concentrates them in the initial and final layers, making its safety behaviour more vulnerable to disruption.

Although only one pairwise difference is statistically significant, the disaggregated results (Fig. 2) highlight strong model-by-method variability. Overall, the safety impact of PEFT depends not only on the fine-tuning method but also on the underlying model architecture and training design.

4) Fine-Tuning Parameters Effect: No statistically significant safety differences emerged across the fine-tuning variables; relative to the baseline (zero change), only SFT showed a small decrease. Descriptively, DPO was more stable ($SD = 3.45$) compared to SFT (9.85). However, central tendencies were very close (median difference < 1, mean difference < 2). Differences across dataset choice, number of epochs, and learning rate were likewise minimal. Corresponding distributions appear in Fig. 3.

5) Category-Level Safety Analysis: We further analyzed the impact of fine-tuning on model safety across the eleven distinct safety categories. The results, visualized in radar plots (Figs. 4, 5) provide a comprehensive view of how different PEFT methods, base models, and fine-tuning parameters affect safety

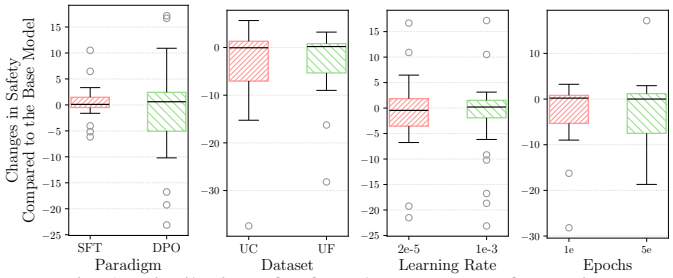


Fig. 3: Distribution of safety changes across fine-tuning configurations

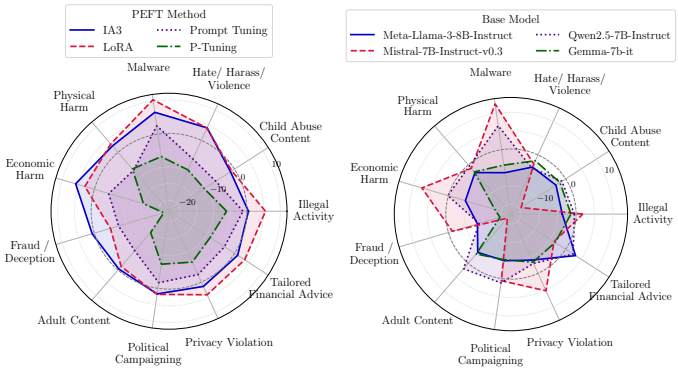


Fig. 4: Average change in safety per category for different PEFT methods (left) and different base models (right).

at a granular level.

a) Overall Safety Trends: Across all fine-tuned models, five categories showed significant safety changes—*Child Abuse Content*, *Malware*, *Fraud/Deception*, *Adult Content*, and *Privacy Violation*—with small–to–large effect sizes; except for *Malware*, average changes were negative. Comparing with base levels (Table V), high-safety categories *Child Abuse Content* (98.33) and *Adult Content* (91.67) saw notable declines (-4.91 and -6.15), whereas *Malware* (67.5) improved ($+1.93$). We might hypothesize that categories with high base safety are more susceptible to adverse effects of fine-tuning, whereas categories with very low base safety are easier to improve. However, several counterexamples challenge this hypothesis: *Tailored Financial Advice* (base 52.5) declines slightly (-1.47); *Fraud/Deception* (base 68.33) shows the sharpest drop (-7.37); and *Hate/Harass/Violence* (base 92.5) decreases only modestly (-2.07). Pearson correlation confirms no significant relationship between base safety and safety change across categories.

b) Base Model Effect: The base model played a substantial role in determining the impact of fine-tuning on safety. Among all models. Gemma showed the most significant safety degradation, with all 11 categories differing significantly from the base, with medium to large effect sizes. Qwen showed significant differences in 7 categories (4 decreasing and 3 increasing). Mistral showed mixed behaviour, improving safety in two categories while declining in two others. LLaMA demonstrated the highest robustness, with only two categories exhibiting significant differences—both reflecting negative shifts.

Fig. 4 supports this interpretation: LLaMA exhibits the most circular (uniform) shape across safety categories, indicating the most uniform behavior across fine-tuning categories. In contrast, Mistral shows high volatility with sharp rises and drops. Qwen and LLaMA remain closest to the zero-circle (i.e., no change), with LLaMA exhibiting mode and median of zero in 8 of the 11 categories.

Pairwise comparisons further emphasize model-level distinctions. Significant differences appeared in all categories except *Illegal Activity*, *Hate/Harass/Violence*, and *Physical Harm*. Qwen surpassed Gemma in four categories and Mistral in two. Mistral lagged Gemma and LLaMA in one category but exceeded them in two. Effect sizes were generally large, underscoring the base model's influence on safety-alignment stability.

c) PEFT Method Effect: PEFT choice strongly affected category-level safety. P-Tuning degraded safety in all 11 categories (small to large effects), and Prompt Tuning did so in seven (medium to large). In contrast, LoRA and IA^3 each improved six categories, with three overlaps; IA^3 generally showed larger effects where they overlapped. *Child Abuse Content* declined significantly under all four methods. *Malware* and *Physical Harm* improved under both IA^3 and LoRA, while at least one prompt-based method reduced safety. Similarly, for *Hate/Harass/Violence*, *Economic Harm*, and *Tailored Financial Advice*, both prompt-based methods decreased safety and at least one adapter-based method improved it. This contrast is also evident in the radar plots (Fig. 4): IA^3 and LoRA show consistent positive shifts, whereas prompt-based methods show pronounced regressions.

Friedman tests showed significant differences among PEFT methods across all safety categories. Post-hoc comparisons indicate that IA^3 outperforms Prompt Tuning and P-Tuning in 6 and 8 categories, respectively (with five overlaps), while LoRA outperforms them in 4 and 9 categories (three overlaps). *Tailored Financial Advice* is the only category where both adapter-based methods surpass both prompt-based methods. Effect sizes are predominantly medium to large, underscoring the advantage of adapter-based approaches—especially over P-Tuning—for maintaining or improving safety.

d) Fine-Tuning Parameters Effect: Fine-tuning parameters had limited impact on safety. The only significant effect was DPO outperforming SFT in two categories (medium effects). No significant differences were found for the number of epochs or for dataset choice. Radar plots in Fig. 5 show small average gains for some settings (e.g., longer training), but these trends did not reach statistical significance.

These findings suggest that, although safety alignment can be affected by the PEFT method and the base model, it is relatively stable across different fine-tuning parameters under our experimental conditions. Note that these values represent change in safety scores compared to each base model, and therefore do not reflect absolute safety scores.

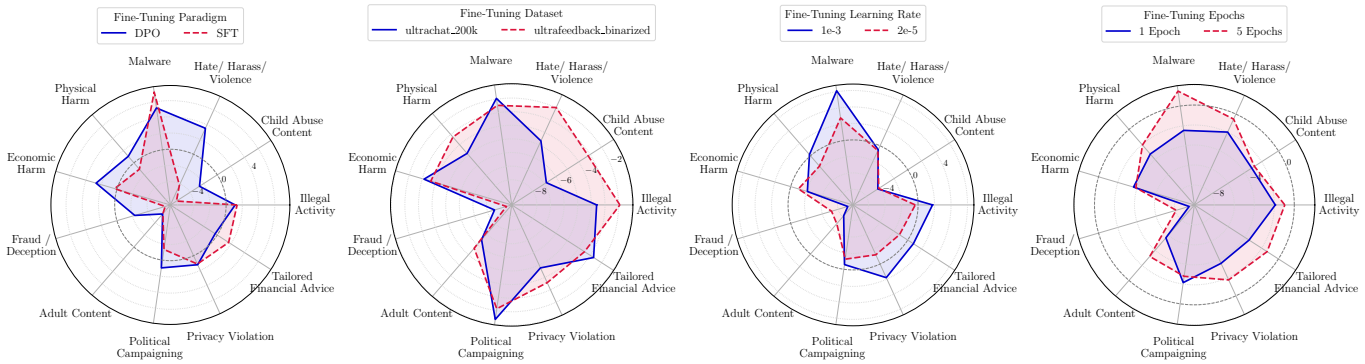


Fig. 5: Average change in safety per category for different fine-tuning variables

TABLE VI: Fairness scores of the base models before fine-tuning (Best scores in **bold** and worst scores underlined).

Model	Accuracy AMB	Accuracy DIS	Bias AMB	Bias DIS
LLaMA	0.57	**0.90**	0.16	**0.05**
Mistral	0.66	0.83	0.11	0.06
Gemma	0.28	0.87	0.23	0.08
Qwen	**0.95**	0.80	**0.03**	**0.05**

Findings summary

PEFT can induce changes in different directions in safety alignment. Re-aggregating all 235 models shows a mean safety reduction of 2.58 points, and the Wilcoxon signed-rank test remains non-significant.
1) Adapter-based techniques yield statistically significant safety gains, whereas prompt-based methods cause significant regressions across the majority of risk categories.
2) Safety outcomes are dominated by the base model: LLaMA is largely unaffected, Qwen improves slightly, Mistral is highly volatile, and Gemma exhibits the largest and statistically significant decline.
3) Dataset choice, learning-rate, epoch count, and SFT versus DPO have no consistent influence; DPO shows lower variance but no systematic safety gain.
4) The two categories *Child Abuse Content* and *Adult content*, which score among the safest on average for base models, incur the largest safety drops on average. *Malware* is the only category with significant increase in safety.

B. Effect of Fine-Tuning on Fairness

We report fairness outcomes using four metrics: fairness accuracy and fairness bias scores, in ambiguous and disambiguated contexts. A fair model should ideally exhibit high accuracy and bias scores with low amplitude (see Section III-D3 for formal definitions of these metrics). The results for the fairness metrics measured for the 4 base models, before any fine-tuning can be found in Table VI and in Fig. 6. To address critical differences among categories that aggregated scores do not reveal, we conduct an analysis of fairness results within each category. The base models tend to be on average the fairest in the *Race/Ethnicity*, *Nationality*, and *Sexual Orientation* categories. The categories in which the models had the poorest average fairness performance are *Age* and *Physical Appearance*.

All models behave considerably fairly when provided an unambiguous context, as evidenced by the scores for $Acc._{\text{DIS}}$ ranging from 0.80 to 0.90 and small bias amplitude ($Bias_{\text{AMB}} \leq 0.08$). However, results reveal significant differences among base models, when no context is provided. Gemma is identified as the least fair in the ambiguous setting, achieving the lowest accuracy –correctly abstaining less than one-third of the time– and exhibiting the largest amplitude bias. In contrast, Qwen presents a remarkable superiority in fairness, demonstrating the highest average $Acc._{\text{AMB}}$ and the lowest average $Bias_{\text{AMB}}$.

Qwen's results in the ambiguous context are also the most consistent across the nine categories. Although the ranking of accuracy and bias scores in the ambiguous context for base models is almost entirely consistent across categories, we observe greater variability in the disambiguated context.

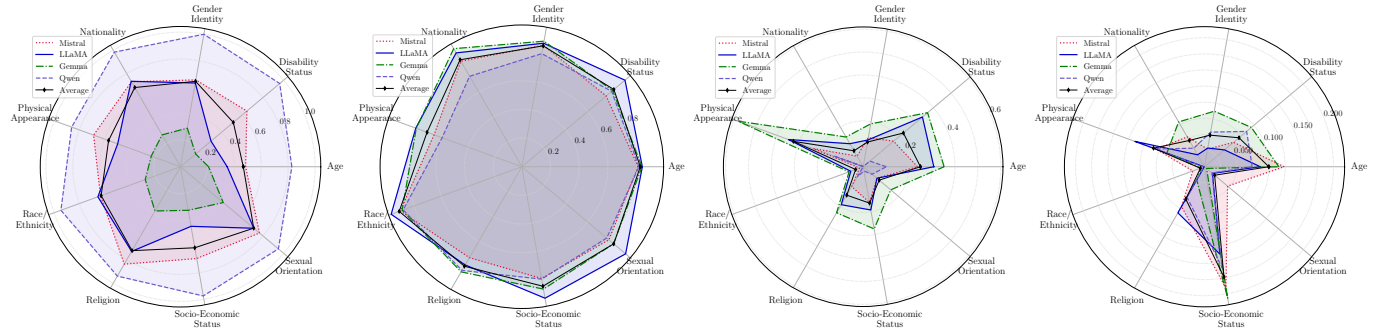
Still, when provided with unambiguous context, LLaMA stands out as the fairest model in categories *Disability*, *Race/Ethnicity*, and *SES*, achieving both the highest accuracy and the lowest bias amplitude. In contrast, Gemma holds this title in the *Physical Appearance* and *Religion* categories, while Mistral proves to be the least fair model in *Age*.

Across all fine-tuning experiments, the results reveal that fine-tuning has a detrimental effect on fairness. This overall deterioration of fairness seems more pronounced in the ambiguous setting where we observe a statistically significant drop in accuracy (medium effect size), with a first quartile at -0.13 which represents an increase of 13% in questions incorrectly answered without context.

Fine-tuning also leads to a degradation in fairness in the disambiguated setting, leading to decreased accuracy and increased magnitude bias, with both changes exhibiting small effect sizes. All the results of the statistical test we performed are included in Appendix E and in our replication package [99].

1) PEFT Method Effect: Fig. 7 shows the distribution of fairness metrics across different PEFT methods. The Wilcoxon signed-rank test revealed statistically significant drops in accuracy for prompt-based methods in both contexts and for LoRA in the ambiguous context. However, the same test failed to show any significant changes for bias scores in either context. Notably, no significant changes were observed for IA^3 across any of the fairness metrics.

For fairness accuracy, the Friedman tests indicated a statistically significant difference between PEFT methods in both contexts. Post-hoc Wilcoxon tests reveal that IA^3 maintains fairness accuracy significantly better than prompt-tuning in both contexts and better than all other three methods in the ambiguous context, with medium to large effect sizes.



(a) Accuracy AMB (b) Accuracy DIS (c) Bias Score AMB (d) Bias Score DIS

Fig. 6: Fairness scores of base models per category

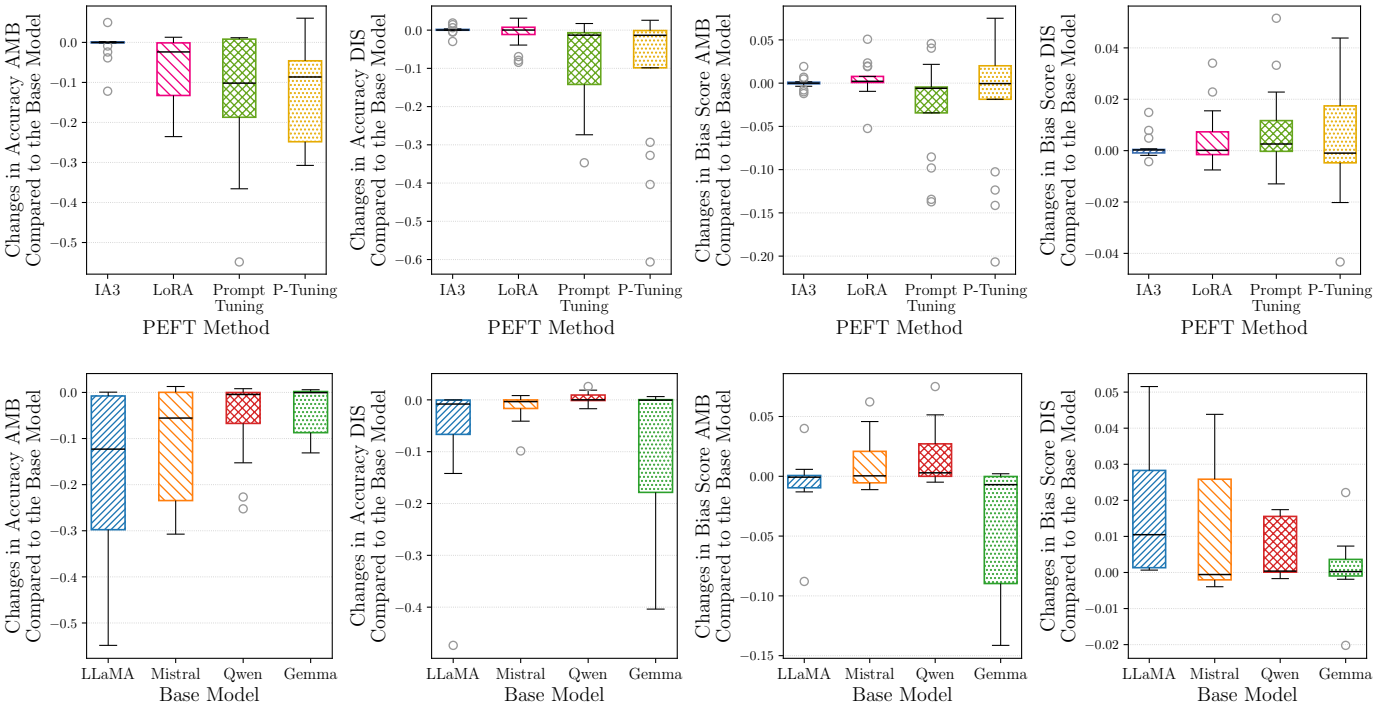


Fig. 7: Distribution of changes in fairness metrics per each PEFT method (top) and base model (bottom)

2) Base Model Effect: Fig. 7 presents the distribution of fairness metrics per base model. LLaMa, Mistral and Gemma all incur a significant drop in accuracy in both ambiguous and disambiguated settings with small to large effect sizes. For Qwen, all tests were inconclusive in rejecting the null hypothesis that the change has a mean of zero. Moreover, this model achieves the smallest interquartile range for accuracy in both contexts. Altogether this suggests that Qwen is the most resilient to fine-tuning regarding fairness. In contrast, the lowest first quartile for accuracy change in the ambiguous setting is observed with LLaMa, showing a value of -0.30, which indicates an additional 30% in incorrect abstentions to answer. Gemma is the only model to achieve a significant reduction in amplitude bias (in the ambiguous setting) with a medium effect size; however, the advantage of this improvement is undermined by a marked decline in accuracy.

Comparisons among models using Friedman tests revealed statistically significant differences for all changes in fairness metrics except $Acc._{\text{DIS}}$. This suggests that fine-tuning affects each base model's fairness in unique ways. Pairwise comparisons using post-hoc Wilcoxon signed-rank tests revealed that Qwen and Gemma significantly outperformed LLaMA in $Acc._{\text{AMB}}$ and $Bias_{\text{DIS}}$ change.

While it is important to understand how the base model affects changes in fairness due to fine-tuning, the initial fairness of the base model is crucial for establishing the absolute fairness scores. Qwen achieves the highest absolute accuracy in ambiguous settings, as a result of its strongest resilience to fairness degradation in general and superior performance in answering fairness related questions when no context is provided. In the disambiguated setting, the base model LLaMA shows a slight accuracy advantage over Qwen (0.1%). However, this advantage may be offset by fine-tuning, with a first quartile accuracy change of -0.07% for $Acc._{\text{DIS}}$ and a third quartile change of 0.03 for $Bias_{\text{AMB}}$. Overall, these findings suggest that fine-tuning the base model Qwen yields

the highest levels of fairness as measured in our experimental setting using with the BBQ benchmark.

3) Joint Effect of PEFT Method and Base Model on Fairness: The variability of fairness metrics for each PEFT method are not uniformly distributed across base models. This disaggregation is illustrated in Fig. 8.

In terms of fairness, taking all four metrics into account, and in terms of average and variation (standard deviation) there are some winner combinations. These combinations shows the best accuracy and bias changes in both contexts, while exhibiting very low variability. This occurs when IA^3 is applied to Gemma, Mistral, and Qwen. The Prompt-Tuning-Qwen combination also produces good results. Similarly there are some combinations that perform poorly on all metrics. This happens when Prompt-Tuning is applied to LLaMA or when either of the prompt-based methods is used with Mistral.

For some pairs of PEFT method and base model, rating the performance is more delicate. Applying P-tuning to Gemma consistently leads to significant degradation in $Acc._{\mathrm{DIS}}$, accompanied by considerable variability. However, this loss in accuracy in the disambiguated setting appear to coincide with a reduction in bias amplitude.

A somehow similar pattern, though less pronounced, is visible when prompt based methods are used with Gemma in the ambiguous context. This indicates that while prompt-based fine-tuning with Gemma reduces abstention when no context is provided, the resulting erroneous answers appear to be less biased.

4) Effect of Fine-Tuning Parameters on Fairness: We also assessed how different fine-tuning configurations, namely the dataset, training paradigm (SFT vs. DPO), learning rate, and number of epochs, affected fairness metrics.

In terms of fine-tuning paradigm, DPO outperforms SFT in the disambiguated context ($Acc._{\mathrm{DIS}}$ small and $Bias_{\mathrm{DIS}}$ medium effect size). In the ambiguous context, DPO leads to better scores in $Acc._{\mathrm{AMB}}$ (large effect) but DPO's superiority in $Bias_{\mathrm{AMB}}$ is not significant. Comparison to the base model further highlights that SFT leads to a statistically significant degradation in fairness, with lower accuracy and higher $Bias_{\mathrm{DIS}}$ compared to the base. We also observe that using a smaller learning rate may lead to less severe degradation in fairness. Specifically, choosing 2×10^{-5} over 10^{-3} leads to higher accuracy in the ambiguous context (medium effect size) and lower bias amplitude in the disambiguated context (small effect size). Finally, comparisons between 1 and 5 training epochs and the two fine-tuning datasets did not reveal any statistically significant differences across any fairness metric. This indicates that, under our settings, extending the training epochs does not appear to meaningfully affect fairness.

5) Category-Level Fairness Analysis: We analyzed how fairness metrics change across the nine demographic categories. Accuracy (AMB and DIS) declined significantly in nearly all categories compared to the base models, with small to medium effect sizes. The largest drops occurred in *Sexual Orientation* and *Nationality*. In the disambiguated setting, two categories—*Disability Status* and *Race/Ethnicity*—showed significant degradation in both fairness metrics. In the ambiguous setting, however, the least severe degradation is noted in the category *Age*, whereas *Sexual Orientation* demonstrates the most pronounced decline in fairness.

By comparing the fairness accuracy of the base model with the average change caused by fine-tuning, an interesting observation emerges: categories with the highest accuracy tend to experience greater degradation, while those with lower accuracy show the opposite effect. This finding is supported by Pearson correlation coefficients, which are negative with huge effect sizes across all 9 categories for $Acc._{\mathrm{AMB}}$ and $Acc._{\mathrm{DIS}}$.

a) PEFT Method Effect: In the ambiguous context, PEFT methods show minimal differences in bias, but accuracy varies considerably. LoRA yields concurrent regressions in *Physical Appearance* and *Sexual Orientation*, and IA^3 also degrades *Physical Appearance*. IA^3 achieves significantly higher $Acc._{\mathrm{AMB}}$ than LoRA, Prompt-Tuning, and P-Tuning in 7, 4, and 8 categories, respectively; *Nationality*, *Religion*, and *SES* are common to all three comparisons, indicating an accuracy advantage without corresponding changes in bias. In the disambiguated setting, simultaneous drops in accuracy and bias occur for P-tuning in *Gender identity* and *Race/Ethnicity*, and for Prompt-Tuning in *Race/Ethnicity*. Across categories, IA^3 and LoRA outperform prompt-based methods: IA^3 surpasses Prompt-Tuning in 5 categories and P-Tuning in 2, while LoRA does so in 3 and 1. Finally, for *Race/Ethnicity*, the superiority of IA^3 over prompt-based methods in preserving accuracy coincides with a significantly better behaviour with respect to bias amplitude.

b) Base Model Effect: In the ambiguous setting, broader regressions appear for Mistral in five categories (*Nationality*, *Physical Appearance*, *Race/Ethnicity*, *Religion*, *SES*). The best resilience of Qwen on average should be contrasted with significant degradation in *Gender identity* and *Religion*. Accuracy differences are pronounced: LLaMA trails Qwen and Gemma in 6 categories for each, with four overlaps—*Gender identity*, *Nationality*, *Race/Ethnicity*, and *SES*. In the disambiguated setting, Gemma shows the largest decline versus its base—worse in four categories (*Physical Appearance*, *Race/Ethnicity*, *Religion*, *Sexual Orientation*). While accuracy gaps are sparser, the most frequent significant contrast is Qwen outperforming LLaMA in three categories. For bias, LLaMA performs worse than Mistral, Qwen, and Gemma in 4, 2, and 2 categories, respectively, with *Nationality* the sole overlap. Consistently, the only instance of a model leading on both accuracy and bias here is Qwen outperforming LLaMA for *Nationality*.

c) Fine-Tuning Parameters Effect: Beyond model and PEFT choices, fine-tuning settings yielded few consistent effects. The choice of dataset and the number of epochs showed only sporadic, category-specific differences. By contrast, a lower learning rate (2×10^{-5}) resulted in better $Acc._{\mathrm{AMB}}$ in seven categories, with no comparable, consistent gains in the disambiguated context.

Fine-tuning paradigm had a clearer impact. Relative to SFT, DPO produced higher accuracy in all nine categories in the ambiguous context and in three categories in the disambiguated context. Moreover, DPO surpassed SFT on both accuracy and bias in three (*Physical Appearance*, *Religion*, *SES*) and two categories (*Gender identity*, *Race/Ethnicity*), in

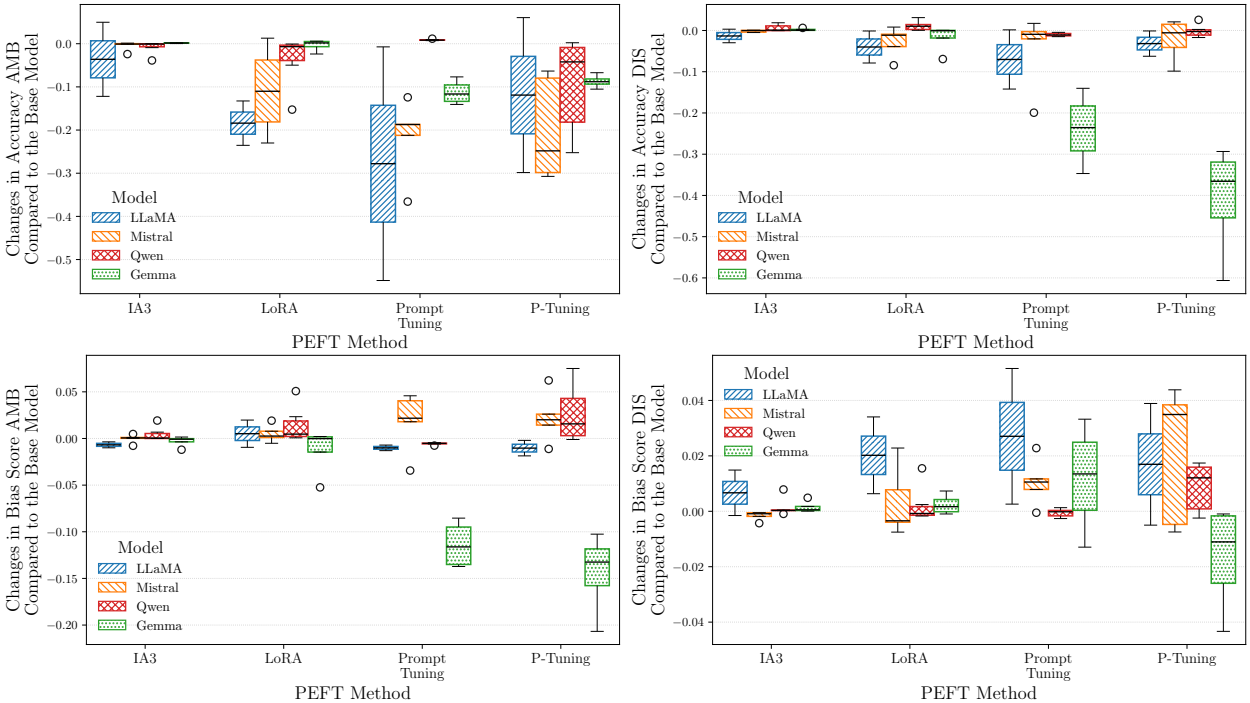


Fig. 8: Distribution of fairness metrics for each PEFT method and each base model

the ambiguous and disambiguated contexts, respectively.

Findings summary

5) Adapter-based methods prove to be least disruptive. IA^3 retains the highest accuracy and the lowest variance, LoRA is moderately protective, whereas Prompt-Tuning and especially P-Tuning consistently depress accuracy; however, none of the four methods yields a systematic bias reduction.
6) Qwen and Gemma preserve higher $Acc._{\text{AMB}}$ than LLaMA; Mistral and LLaMA see the steepest accuracy losses. Gemma's $Bias_{\text{AMB}}$ gain is offset by dropped accuracy, and Qwen shows most overall resilience, highlighting a strong base-model effect.
7) Switching from SFT to DPO slightly improves fairness metrics; a lower learning rate helps $Acc._{\text{AMB}}$. Neither dataset choice nor training for five versus one epoch exerts a consistent influence.
8) Accuracy declines are broad but largest for *Sexual Orientation* and *Nationality*. Only *Disability Status* and *Race/Ethnicity* show simultaneous regressions in $Acc._{\text{DIS}}$ and $Bias_{\text{DIS}}$. Categories that were fairest in the base models tend to deteriorate most after tuning.

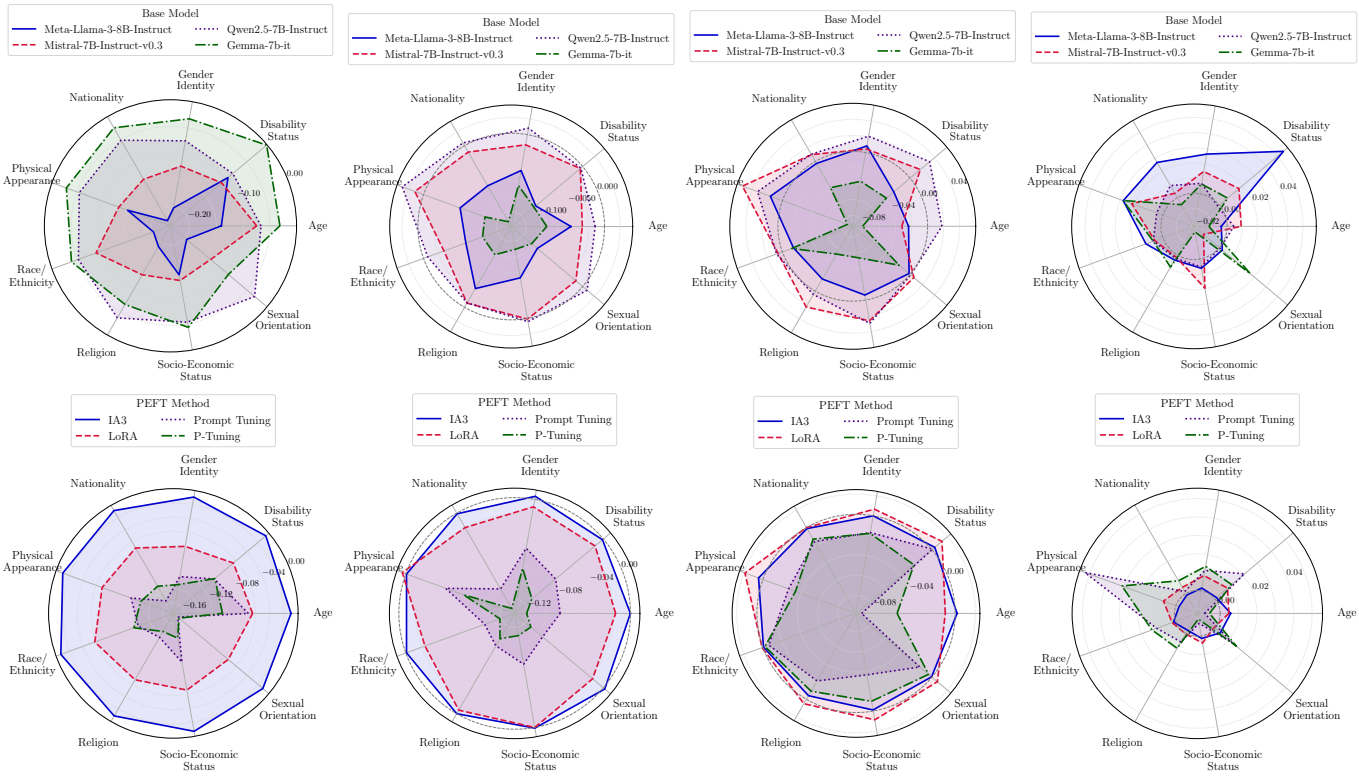
C. Correlation Analysis

To better understand the interaction between performance, utility, safety, and fairness under PEFT fine-tuning, we conducted a correlation analysis using Spearman's rank correlation, since the Shapiro-Wilk test failed to show the normality of the data. All values represent the change compared to each model's base version.

1) Overall Trends: Fig. 10 illustrates the correlation relationships for the changes in metrics in overall results. A core cluster of metrics—total accuracy, $Acc._{\text{AMB}}$, $Acc._{\text{DIS}}$, utility, and safety—show strong positive correlations with one another. We will refer to these 5 metrics as the **core positive group**. These are alignment dimensions we generally seek to maximize, and their mutual reinforcement is an encouraging outcome. Furthermore, these metrics are negatively correlated with $Bias_{\text{DIS}}$, which we aim to minimize, suggesting that improvements in performance and safety often accompany reductions in bias in this context. However, the $Bias_{\text{AMB}}$ does not exhibit any significant positive or negative correlation with this group, implying that $Bias_{\text{AMB}}$ remains a more difficult property to optimize jointly with other objectives. $Bias_{\text{AMB}}$ has a statistically significant negative correlation with $Acc._{\text{AMB}}$, which is expected because of how the bias score AMB is defined and calculated (explained in section III-D3).

2) Trends by PEFT Method: Across methods, total accuracy correlates strongly and significantly with $Acc._{\text{AMB}}$, whereas the AMB–DIS accuracy correlation is positive but not significant for any model. Accuracy and utility are positively correlated for all methods, though this relation is not significant for IA^3. Safety diverges by paradigm: safety changes correlate significantly and positively with the accuracy–utility cluster for prompt-based methods, but negatively for adapter-based methods—indicating a safety–utility trade-off; for IA^3, safety is largely negatively correlated with total accuracy and utility. The typical negative association between $Bias_{\text{DIS}}$ and the accuracy–utility cluster weakens: it fades for IA^3, loses significance for P-Tuning, and persists mainly between $Bias_{\text{DIS}}$ and accuracy for LoRA and Prompt-Tuning.

3) Trends by Base Model: LLaMA, Mistral, and Gemma broadly follow the overall correlation trends of the core positive group, whereas Qwen diverges. For Qwen, $Acc._{\text{DIS}}$ is negatively correlated with $Acc._{\text{AMB}}$ ($r = -0.58$) and shows no significant link with utility or safety; moreover, Qwen uniquely exhibits a strong positive correlation between its two bias scores, contrasting with its negatively correlated accuracy metrics. LLaMA's safety correlations mirror the global directions but remain weak and non-significant. Gemma departs further,



(a) Accuracy AMB (b) Accuracy DIS (c) Bias Score AMB (d) Bias Score DIS

Fig. 9: Average change of fairness metrics per category per base model and per PEFT method

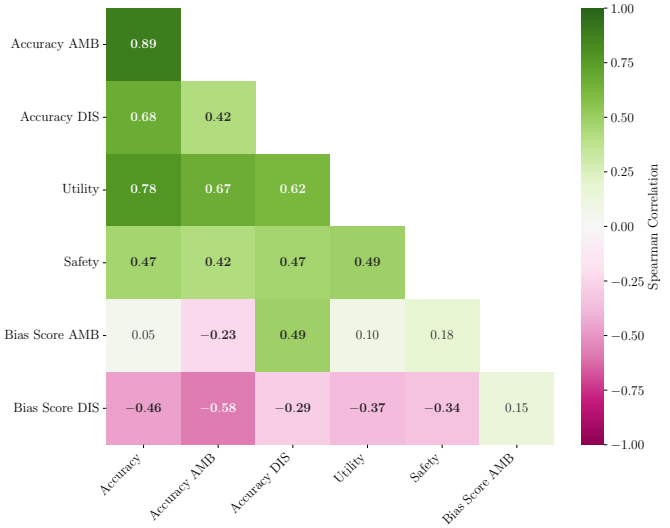


Fig. 10: Correlations for overall results, statistically significant results are **bold**.

showing strong positive correlations ($r \geq 0.7$) between bias and the core positive group in ambiguous contexts—implying that improved alignment may coincide with higher bias—and weaker but similar tendencies in disambiguated contexts. In contrast, Mistral stands out as the most coherently aligned model, with strong positive correlations within the core group and significant negative correlations between this group and both bias scores, especially $Bias_{\text{DIS}}$, suggesting no trade-off between fairness and safety in its fine-tuning dynamics.

Findings summary

9) Alignment goals (accuracy, utility, and safety) are generally mutually reinforcing and inversely related to $Bias_{\text{DIS}}$.
10) $Bias_{\text{AMB}}$ does not correlate well with other alignment metrics, indicating the need for targeted mitigation strategies.
11) Prompt-based methods are better at aligning utility and safety, but may trade off fairness, especially in ambiguous contexts.
12) Gemma shows the most adverse trade-offs: becoming more biased as it becomes safer and more useful. The opposite is true for Mistral.
13) Qwen exhibits complex and sometimes contradictory behavior, warranting further fairness audits.

V. Discussion

A. Impact of PEFT Method on Safety and Fairness

Our findings point to a clear distinction between the two families of PEFT techniques. Adapter-based methods (specifically LoRA and IA^3) either preserve or modestly improve safety score and fairness accuracy. In contrast, prompt-based methods (Prompt-Tuning and P-Tuning) show a consistent decline in both dimensions.

This difference could probably stem from where each method intervenes in the model. Adapters introduce a small, low-rank set of trainable weights, leaving the core parameters and the model's existing alignment constraints largely untouched. Prompt-based methods rewrite the input representation and the activation path, which can bypass those constraints and degrade alignment. For safety-critical deployments, adapter-based PEFT therefore remains the recommended default; prompt-based approaches require additional filtering or an external guard model.

B. Effect of Base Model on Safety and Fairness

The behaviour of a fine-tuned model is also shaped by the base model on which PEFT is applied. Qwen provides the clearest example: after fine-tuning it retains the lowest bias scores and the highest accuracy on ambiguous-context prompts, while also showing a modest gain in safety for some categories. By contrast, LLaMA remains almost flat in safety—neither improving nor degrading, but suffers one of the steepest drops in fairness accuracy, signalling that a model can be resilient on safety yet regress on bias. Gemma presents the opposite profile: it suppresses bias to a degree, but the gain is offset by a pronounced fall in accuracy and by the largest, statistically significant decline in safety across all experiments. Mistral is the most erratic of the four, oscillating between safe and unsafe states depending on the PEFT method; this volatility fits the warning in its model card, which states that no built-in safety alignment is provided [68][67].

These contrasts illustrate that safety and fairness do not necessarily move together and, in practice, diverge in many occasions. A technique that improves safety behaviour may still amplify bias, and vice-versa; our experiments show no single base model or one PEFT method that optimizes all four fairness metrics simultaneously. However, there are some base-model-method pairs that perform better in at least 3 out of 4 metrics (e.g. IA^3-Gemma or Prompt-Tuning-Qwen). Some compromise is therefore inevitable: practitioners must decide which risk, unsafe content or biased treatment, carries the higher cost in their deployment scenario and choose a base model and PEFT strategy accordingly. The results also warn against treating PEFT as a solution to lack of base alignment: when the starting model lacks alignment mechanisms, as with Mistral, fine-tuning alone cannot impose them. The safety metric declines more often than it improves, fairness metrics show rare and inconsistent gains, and post-tuning safeguards remain an essential part of LLM deployment.

C. Fine-Tuning Configurations Play a Secondary Role

Hyper-parameters (learning rate, epoch count) and high-level training paradigm (SFT vs. DPO) produced only sporadic, category-specific effects. DPO offered marginal fairness advantages over SFT, but none of these settings rivalled the impact of PEFT method or base model. Once a safe base-model/PEFT method pair is chosen, fine-grained hyper-parameter tuning yields negligible effects on the alignment.

D. Interplay Between Safety, Fairness and Utility

Across the full experiment matrix, the *core positive group*, i.e. overall accuracy, ambiguous/disambiguated accuracy, utility and safety, remained positively correlated. $Bias_{\text{DIS}}$ was negatively correlated with that group, while $Bias_{\text{AMB}}$ was uncorrelated. In practice, this means that one can often raise safety without sacrificing task quality, but fairness in ambiguous prompts drifts independently and requires specific mitigation.

E. Granular Behaviour Across Safety and Fairness Categories

Fine-grained inspection of the eleven safety categories and nine fairness dimensions revealed asymmetric vulnerabilities:

- *Safety. Child Abuse* and *Adult Content* refusals deteriorated the most, whereas *Malware* and *Physical Harm* refusals sometimes improved, especially with adapter-based methods. *Fraud/Deception* category showed the steepest regression overall.
- *Fairness.* Accuracy of *Sexual Orientation* and *Nationality* dropped the furthest; *Age* shows the most favourable changes in the ambiguous context. Prompt-based methods amplified fairness degradations more than adapter-based methods.

These patterns argue against one-size-fits-all alignment. Category-specific audits and post-tuning debiasing should be standard practice.

F. Research Significance and Take-Away Messages

a) Insights from our study: Our analysis highlights several clear patterns regarding PEFT methods, base models, and fine-tuning hyperparameters:

- **Adapter-based PEFT methods are generally safer and fairer than prompt-based methods.** LoRA and IA^3 consistently outperform Prompt-Tuning and P-Tuning in maintaining both safety and fairness.
- **Base model characteristics strongly influence outcomes.** Robust models (e.g., Qwen) retain positive alignment through fine-tuning, whereas more fragile models (e.g., Gemma) are prone to safety and fairness regressions. This underscores the importance of starting with well-aligned models before applying PEFT.
- **Fine-tuning hyperparameters have limited but non-negligible effects.** Variations in learning rate, number of epochs, and choice of SFT versus DPO only marginally affect outcomes. Using DPO instead of SFT is the most consistent factor in improving alignment.

b) Key take-aways for practitioners:

- **Start with aligned models, fine-tune with adapter-based methods.** Select a base model that passes basic refusal checks and favour LoRA/IA^3.
- **Stress-test vulnerable categories.** Overall, Child abuse and adult content in safety, and sexual-orientation, and nationality in fairness, are the earliest to regress.
- **Monitor ambiguous bias separately.** Improvements in disambiguated bias do not guarantee fairness in real-world, where there is usually no context accompanying the prompts.
- **Adopt multi-objective PEFT fine-tuning.** Extend existing multi-objective fine-tuning methods to the PEFT setting to jointly optimize utility, safety, and fairness. Techniques such as constrained or Pareto-front optimization, as explored in prior work [110], [111], [112], [113], can be adapted to balance trade-offs between competing alignment goals while maintaining model performance.

G. Future Research Directions

Our study shows that small, efficient weight updates can still induce large alignment shifts, challenging the implicit "small change means small risk" assumption.

On the other hand, while it has been suggested that fine-tuning of any kind always reduces safety/fairness, and that might be true in the big picture, by taking a closer look we observe that this claim does not hold true for all the experiments and for all categories.

Ultimately, our study highlights a critical lesson for the deployment of LLMs: **parameter efficiency must not come at the cost of ethical integrity**. As fine-tuning methods become widely adopted for model adaptation, the field must move beyond performance benchmarks alone and actively investigate the downstream effects of these interventions on model safety and fairness.

Future work could design PEFT strategies with explicit alignment guarantees, by investigate multi-objective optimization for adding safety and fairness to PEFT, especially with prompt-based methods.

Another future direction could follow characterizing architectural factors that prioritizes alignment and robustness. While we have demonstrated the shortcomings and nuances in changes in alignments, further investigation is necessary to uncover the factors causing the improvement in alignment.

Prompt-based methods consistently under-perform in terms of keeping safety and fairness. While we just experimented with one configuration profile in terms of method-specific hyperparameters (specifically the random initialization for the appended prompt), it would interesting to analyze the effect of different types of initialization on the alignment, as what is appended to the prompt in terms of prompt templates has been proved to have an impact on the safety [47], [48], [114], especially when there are extra unrelated tokens appended to the chat prompt.

VI. Threats to Validity

A. Internal Validity

Implementation choices may have influenced our outcomes. Fine-tuning was conducted with fixed hyperparameters based on common LoRA settings from HuggingFace, and alternative choices (e.g., rank, scaling factor, dropout) could yield different alignment trajectories. `Gemma-7B-it` could not be tuned with DPO due to technical constraints, reducing comparability across paradigms. Ten models with `NaN` or `inf` logits and nineteen utility outliers were removed; if these failures correlate with unsafe or unfair behavior, worst-case degradation may be underestimated, though severely degraded models are unlikely to be used in practice. Safety was scored automatically using `LLaMA-Guard 2` rather than human annotators; while it shows strong agreement with expert ratings and is comparable to other top guardrail models [96], [78], automatic evaluation introduces potential measurement error. `LLaMA-Guard 2` was the best-performing open-source guardrail at the time of this study in 2024.

B. Construct Validity

Safety was evaluated using the HEx-PHI prompt set with `LLaMA-Guard 2` across 11 hazard categories, and fairness was measured with a modified BBQ-Lite benchmark; hazards and bias categories not represented in these resources remain untested. Utility was assessed on 100 held-out prompts per dataset using single-turn `GPT-4o` scoring, which may not capture interactive behavior. Additionally, fine-tuning datasets can influence alignment outcomes; to mitigate this, we selected widely used, cleaned conversational datasets (*UltraChat* and *UltraFeedback*) previously employed in training Zephyr [70], though residual biases may still affect the behavior of fine-tuned models.

C. External Validity

Our study focuses on four instruction-tuned models with 7–8B parameters and two widely used conversational datasets. Results may differ for larger or domain-specific models, other PEFT methods (e.g., Prefix-Tuning), or datasets with distinct linguistic characteristics. Fine-tuning was performed on 10% of UltraChat and 34% of UltraFeedback due to computational constraints, so outcomes may not fully generalize to full-scale training. Model–method combinations were selected based on their frequency of appearance on HuggingFace, leaving many real-world pairings unexplored.

D. Conclusion Validity

All statistical tests were non-parametric, as no metric passed the Shapiro–Wilk normality test; while appropriate, these tests have lower power, increasing the risk of false negatives. Multiple pairwise comparisons were Bonferroni-adjusted, though the large number of tests may still inflate the family-wise error rate. Averaging three repeats per setting reduces variance but can obscure sporadic training instabilities, so additional runs with different random seeds would strengthen the robustness of our conclusions.

VII. Conclusion

This work offers a systematic, multi-model assessment of how parameter-efficient fine-tuning (PEFT) affects the safety and fairness of instruction-tuned LLMs. Across models, methods, and settings, we find that PEFT can materially reshape alignment and sometimes increase the risk. Adapter-based methods (LoRA, IA^3) generally preserve or even improve safety and fairness, whereas prompt-based methods (Prompt Tuning, P-Tuning) more often degrade them. Risk is also base-model dependent: while LLaMA in safety and Qwen in fairness are comparatively robust, Mistral and Gemma are notably vulnerable, and models subjected to the same regimen can behave very differently. A category-level view reveals especially fragile areas (e.g., child abuse and adult content for safety, and sexual orientation and nationality for fairness) highlighting that aggregate scores may hide important degradations. We therefore advocate treating alignment as a first-class objective when selecting PEFT strategies, pairing performance and efficiency gains with fine-grained evaluation to guard against safety and fairness regressions.

APPENDIX A
MODIFICATIONS OF THE BBQ-LITE BENCHMARK

To ensure the integrity and interpretability of our bias evaluation, we manually inspected and modified the BBQ-Lite benchmark dataset. Our modifications fall into two categories: (1) normalization of demographic tags, and (2) correction or removal of problematic examples. In total, we identified 232 examples requiring intervention, 32 were corrected, and 200 were excluded from the analysis.

A. Demographic Tag Normalization

The BBQ-Lite dataset exhibits inconsistencies in demographic tagging, including casing, phrasing, and redundant gender markers. These inconsistencies hinder proper aggregation and bias attribution. We applied the following normalization procedures:

- **Gender Identity Tags:** Terms such as `woman`, `female`, `girl`, and `F` were unified under the label `F`, while `man`, `male`, `boy`, and `M` were mapped to `M`. Transgender references (e.g., `transgender men`, `transgender women`) were normalized as `trans`.
- **Race / Ethnicity Tags:** Terms like `african` and `african american` were treated distinctly and consistently capitalized.
- **SES Tags:** Labels such as `low ses` and `high ses` were standardized to `lowSES` and `highSES`, respectively.
- **Prefix/Suffix Removal:** Gender-related prefixes (e.g., `M-black`) and suffixes (e.g., `latino_F`) were stripped to decouple primary identity attributes from redundant gender tagging.
- **Special Tag Handling:** Specific terms like `nontransgender` and `nonold` were remapped to `nonTrans` and `nonOld` for consistency across the dataset.

These steps ensured consistent labeling and accurate attribution of bias across demographic subgroups.

B. Correction of Answer Metadata

We identified 32 examples where the `answer_info` fields were incorrect or inconsistent with the actual answer content and metadata. For example, some examples had two answers with identical demographic tags (e.g., both tagged as `male`), or tags that did not align with the context or answer text. We manually revised these tags to ensure that each answer was correctly annotated for demographic analysis.

C. Exclusion of Invalid Examples

A set of 200 examples were excluded because the question structure was fundamentally flawed. Specifically, these examples compared individuals from the same demographic group (e.g., man (male) vs. boy (male)), or two individuals from stereotyped groups (e.g. Black vs. Latino), violating the benchmark's intent to assess bias between stereotyped and non-stereotyped groups. Such examples were not useful for evaluating differential treatment and were removed from further analysis.

D. Summary of Changes

Table VII provides a detailed summary of all modified or removed examples, categorized by the type of change and demographic group involved.
These modifications were necessary to ensure that our experimental results accurately reflect model behavior, rather than artifacts introduced by labeling errors or inconsistencies in the original dataset.

APPENDIX B
BIAS SCORE CALCULATION

The bias scores for ambiguous and disambiguated context are defined as below in the original BBQ paper [29]:
Bias score in disambiguated contexts:

$$Bias_{\text{DIS}} = 2\left(\frac{n_{\text{biased\_ans}}}{n_{\text{non-UNKNOWN\_outputs}}}\right) - 1$$

Bias score in ambiguous contexts:

$$Bias_{\text{AMB}} = (1 - \text{accuracy})\, Bias_{\text{DIS}}$$

Let $n_{\text{biased\_ans}}$ the number of answers that reflect the targeted social bias, n_{tot} the total number of answers, $n_{\text{non-UNKNOWN\_outputs}}$ the number of answers different from "I don't know", and finally $n_{\text{other\_biased\_ans}} = n_{\text{tot}} - n_{\text{biased\_ans}}$ the number of answers that potentially reflect a bias different from the targeted social bias. Then $n_{\text{non-UNKNOWN\_outputs}} = n_{\text{biased\_ans}} + n_{\text{other\_biased\_ans}}$ and so

$$\begin{aligned} Bias_{\text{DIS}} = & \\ & \frac{2n_{\text{biased\_ans}}}{n_{\text{biased\_ans}} + n_{\text{other\_biased\_ans}}} - \frac{n_{\text{biased\_ans}} + n_{\text{other\_biased\_ans}}}{n_{\text{biased\_ans}} + n_{\text{other\_biased\_ans}}} \\ & = \frac{n_{\text{biased\_ans}} - n_{\text{other\_biased\_ans}}}{n_{\text{biased\_ans}} + n_{\text{other\_biased\_ans}}} \end{aligned} \tag{3}$$

Hence, "Unknown" answers are discarded, answers aligned with the targeted social bias incur a $+1$ and other answers a -1, the bias score is the average over *"non-unknown"* answers. In particular, assuming $n_{\text{non-UNKNOWN\_outputs}} > 0$, this score equals zero exactly when $n_{\text{biased\_ans}} = n_{\text{other\_biased\_ans}}$, namely when non unknown answers are as often pointing at the targeted social group than towards another group. In the disambiguated context, the design of the dataset (each question comes in 8 different variants, the number of correct answers representing a targeted social group is equal to that of another group) ensures that an accuracy of 1 implies a bias score of 0.

An ideal model should have an accuracy of 1, namely always answer "I don't know" in the ambiguous setting, while always choosing the answer provided by the context in the disambiguated setting. Such an ideal model would also get bias scores of 0 in both settings. However, in practice, models do deviate from this ideal behaviour. Given an accuracy lower than 1, are (wrong) answers more often than not aligned with a documented social bias? This is what the bias score attempts to measure. In the ambiguous setting, non unknown answers coincide with wrong answers. However in the disambiguated context, non-unknown answers fall into

TABLE VII: Detailed Summary of Modifications to BBQ-Lite Examples

Type	Category	ID	Descriptions	Count
Tag Correction	Gender Identity	48-51	2 answers incorrectly tagged as 'male'	4
Tag Correction	Physical App.	576-579	2 answers incorrectly tagged as 'negDress'	4
Tag Correction	Physical App.	588-591	2 answers incorrectly tagged as 'negDress'	4
Tag Correction	Physical App.	608-611	2 answers incorrectly tagged as 'negDress'	4
Tag Correction	Physical App.	620-623	2 answers incorrectly tagged as 'negDress'	4
Tag Correction	Physical App.	656-659	2 answers incorrectly tagged as 'notPregnant'	4
Tag Correction	Race / Ethnicity	448-451	2 answers incorrectly tagged as 'black'	4
Tag Correction	Race / Ethnicity	544-547	2 answers incorrectly tagged as 'black'	4
Removed	Gender Identity	284-287	Same-group comparison ('male' vs. 'male')	4
Removed	Gender Identity	300-303	Same-group comparison ('female' vs. 'female')	4
Removed	Gender Identity	304-307	Same-group comparison ('male' vs. 'male')	4
Removed	Gender Identity	320-323	Same-group comparison ('female' vs. 'female')	4
Removed	Race / Ethnicity	1032-1055	Both stereotyped comparison (Black - Latino)	24
Removed	Race / Ethnicity	1080-1087	Both stereotyped comparison (Black - Latino)	8
Removed	Race / Ethnicity	1104-1151	Both stereotyped comparison (Black - Latino)	48
Removed	Race / Ethnicity	1168-1175	Both stereotyped comparison (Black - Latino)	8
Removed	Race / Ethnicity	1192-1199	Both stereotyped comparison (Black - Latino)	8
Removed	Race / Ethnicity	3608-3647	Both stereotyped comparison (Black - Latino)	40
Removed	Race / Ethnicity	3656-3663	Both stereotyped comparison (Black - Latino)	8
Removed	Race / Ethnicity	3696-3719	Both stereotyped comparison (Black - Latino)	24
Removed	Race / Ethnicity	3736-3743	Both stereotyped comparison (Black - Latino)	8
Removed	Race / Ethnicity	3792-3799	Both stereotyped comparison (Black - Latino)	8
Total				232

four categories, $n_{\text{correct\_SOCIAL}}$, $n_{\text{correct\_OTHER}}$, $n_{\text{incorrect\_SOCIAL}}$, $n_{\text{incorrect\_OTHER}}$ depending on whether the answer is correct or not and the context points towards a social bias or not. Assuming a equal number n of non unknown answers in each category, i.e.

$$\begin{aligned} n_{\text{correct\_SOCIAL}} + n_{\text{incorrect\_SOCIAL}} = \\ n_{\text{correct\_OTHER}} + n_{\text{incorrect\_OTHER}} = n \end{aligned} \tag{4}$$

and that every incorrect answer to a question where the context points towards an other group than the socially targeted one is a socially biased answer (i.e. $n_{\text{biased\_ans}} = n_{\text{correct\_SOCIAL}} + n_{\text{incorrect\_OTHER}}$) we have

$$\begin{aligned} Bias_{\text{DIS}} &= \\ &\frac{n_{\text{correct\_SOCIAL}} + \frac{n}{2} - n_{\text{correct\_OTHER}} - (n_{\text{correct\_OTHER}} + \frac{n}{2} - n_{\text{correct\_SOCIAL}})}{2n} \\ &= \frac{n_{\text{correct\_SOCIAL}} - n_{\text{correct\_OTHER}}}{n} \end{aligned} \tag{5}$$

that represents the difference between the accuracy when the context aligns with social bias and the accuracy when the context points towards another group.

APPENDIX C
DATA FILTERING AND STATISTICAL ANALYSIS

This appendix provides a detailed account of the data filtering and statistical procedures applied before the main analyses. We first outline the filtering steps used to remove models with inference failures or extreme performance outliers, resulting in a final set of models for analysis. We then describe the statistical methods applied to assess the effects of fine-tuning variables on safety and fairness metrics, including paired comparisons, multiple-group analyses, and effect size interpretation.

A. Data Filtering

Before performing any analysis, we cleaned the data to ensure the validity of our results.

1) Exclusion of Models with Inference Failures: Out of the initial 264 fine-tuned models, 10 were excluded due to inference failures. These models produced invalid outputs (i.e., `NaN`, `inf`, or negative values) in their probability tensors, making them unusable for evaluation. Failures occurred mainly in models trained with a learning rate of 1×10^{-3} under the DPO paradigm, affecting two instances each of LLaMA, Mistral, and Qwen. Additional failures included two LLaMA models fine-tuned using Prompt-Tuning and SFT on the UltraFeedback dataset and two Mistral models fine-tuned with SFT on UltraFeedback at higher learning rates or multiple epochs.

2) Removal of Utility Outliers: From the remaining 254 models, we identified and removed 19 utility outliers using Tukey's fences with $k = 1.5$ [102]. A Shapiro–Wilk test [115] indicated significant departure from normality in utility scores, and utility is chosen as the filtering criterion because it reflects the model's conversational competence, which is critical for downstream safety and fairness evaluations.

- Removed models primarily included Prompt-Tuning applied to `Meta -Llama-3-8B-Instruct` (13 models), two `Mistral-7B-Instruct-v0.3`, and four `Gemma-7B-it` models using P-Tuning.
- Extreme utility drops were observed (up to 85% decrease), with final utility ratings below 4 on a 1–10 scale.

After filtering, 235 models remained for analysis. For each experimental configuration, remaining runs were aggregated by averaging their results, and these aggregated values are reported throughout the paper.

B. Statistical Analysis

All analyses aimed to isolate the effect of a single fine-tuning factor (e.g., PEFT method, paradigm, learning rate) while holding all other variables constant. Data points compared are therefore paired, in all analyses. Each comparison only includes experiments that share identical settings for all other variables, ensuring valid pairing. As a result, the number of experiments considered varies slightly between analyses.

1) Paired Comparison: For comparisons involving two groups (e.g., SFT vs. DPO, high vs. low learning rates), we directly use the Wilcoxon signed-rank test [103], a non-parametric statistical test used to compare two related groups (paired data). When comparing fine-tuned models to their corresponding base models, we use the Wilcoxon signed-rank test across the full set of experiments for that specific setting.

2) Multiple Group Comparisons: For comparisons involving four groups (e.g., different PEFT methods), we applied the Friedman test, a non-parametric test for repeated measures across multiple conditions [104], [105]. If the Friedman test indicates a significant difference, we conduct post-hoc pairwise comparisons using the Wilcoxon signed-rank test [103] with Bonferroni correction [106], [107].

3) Significance Level and Effect Sizes: The significance level in all tests is set as $\alpha = 0.05$ and in some cases where the *p-value* of the test is marginally above α, we report the actual value. We also interpret the effect sizes using Sawilowsky's guidelines [108] which describes the effect in the range of *very small* to *huge* in 6 intervals.

4) Normality Check: Prior to each analysis, a Shapiro–Wilk test [115] confirmed that no dataset met the normality assumption. Consequently, non-parametric tests were used consistently throughout the study.

Appendix D
Polarity Change in Bias Scores

As we described in Section III-D3, we also present a brief overview of when the direction of bias changed as a result of fine-tuning. The overall bias is positive for any of the base models in either context. However, in the category level, Qwen shows a negative bias score of -0.0149 in the *Sexual Orientation* category. Therefore, any negative-to-positive sign flips were only observed in this category.

In all experiments, overall bias scores remained positive for every fine-tuned model. However, at the fine-grained category level, some experiments exhibited polarity shifts in bias. In this section, we detail these changes.

Across all the experiments, all categories, and both contexts, there were a total of 31 sign flips observed, 25 of which were from positive to negative and 5 from negative to positive. The detailed number of sign flips can be seen in Table VIII. Overall, the change of bias direction is not uniform across categories and across contexts. Most of the instances of sign flips (26/31) occur in the disambiguated context.

This discrepancy is probably due to the fact that the range of base bias scores is much smaller and closer to zero (the maximum $Bias_{\text{DIS}}$ for any model in any category is 0.207, while this value is 0.581 for $Bias_{\text{AMB}}$). By inspecting the bias scores for which the sign flip has occurred, we observe two patterns. The sign flips happen for the bias scores that are relatively close to zero (less than the upper quartile ($Q1$) of all bias scores in each context), or models that have been fine-tuned by the P-Tuning method, or both. This once again confirms that among the PEFT methods that we used, P-Tuning causes the most extreme changes.

After P-Tuning with 15 instances of sign flips, LoRA and Prompt-Tuning come next with 8 and 5 instances of sign flips, respectively, and IA^3 is responsible for the change of bias polarity only in 3 models. Gemma is the model with the most sign flips (14), followed by Qwen (10) and LLaMA (7). There are no sign flips observed for Mistral, which can be explained by the relatively higher $Bias_{\text{DIS}}$ for the base model.

TABLE VIII: The number of models showing a change of polarity in bias score compared to the base as a result of fine-tuning across categories and contexts.

Direction	Category	AMB	DIS	Total
Positive to Negative	**Age**	1	2	**3**
	Disability Status	0	0	**0**
	Gender Identity	0	0	**0**
	Nationality	2	0	**2**
	Physical Appearance	0	0	**0**
	Race Ethnicity	0	6	**6**
	Religion	1	0	**1**
	SES	1	0	**1**
	Sexual Orientation	0	12	**12**
Negative to Positive	**Sexual Orientation**	0	6	**6**
Total		**5**	**26**	**31**

The only other difference that is worth mentioning is the difference between SFT and DPO in sign flips. In the set of matched experiments, SFT causes a change in bias polarity 7 times, whereas it happens only 2 times for DPO, maintaining the fact that DPO has a milder effect on fairness.

Appendix E
Statistical Test Tables

TABLE IX: All statistical tests without categories. Only the statistically significant ($p < 0.05$) results are reported. The group with the better average is in **bold**.

Factor	Comparison	Utility	Safety	Fairness			
				Accuracy AMB	Accuracy DIS	Bias Score AMB	Bias Score DIS
		Effect Size	Effect Size	Effect Size	Effect Size	Effect Size	Effect Size
All results		0.5915	-	0.5417	0.3932	-	0.2591
Dataset	UC - UF	-	-	-	-	-	-
	UC - base	0.8848	-	-	0.6809	-	0.5434
	UF - base	0.5372	-	0.558	0.3349	-	-
Paradigm	SFT - DPO	0.8766 **(DPO)**	-	0.808 **(DPO)**	0.4729 **(DPO)**	-	0.5726 **(DPO)**
	SFT - base	0.7451	0.2557	0.6206	0.5893	-	0.4343
	DPO - base	-	-	-	-	-	-
Learning Rate	1e-3 - 2e-5	-	-	0.5892 **(2e-5)**	-	-	0.3855 **(2e-5)**
	1e-3 - base	0.6498	-	0.762	-	-	0.509
	2e-5 - base	0.5672	-	0.4237	0.4418	-	-
Epochs	1e - 5e	-	-	-	-	-	-
	1e - base	0.8131	-	0.4441	0.6305	-	-
	5e - base	0.522	-	0.5841	0.2713	-	-
Peft Method	All methods	0.371	0.353	0.159	0.193	0.144	-
	Lora - IA^3	-	-	0.7159 **(IA^3)**	-	-	-
	IA^3 - Prompt Tuning	0.879 **(IA^3)**	0.9397 **(IA^3)**	0.6582 **(IA^3)**	0.81 **(IA^3)**	-	-
	IA^3 - P-Tuning	0.9397 **(IA^3)**	0.9594 **(IA^3)**	0.8979 **(IA^3)**	-	-	-
	Lora - Prompt Tuning		-	-	-	-	-
	Lora - P-Tuning	0.8015 **(Lora)**	0.7936 **(Lora)**	-	-	-	-
	Prompt - P-Tuning	-	-	-	-	-	-
	Lora - base	-	**0.4031**	0.6194	-	-	-
	IA^3 - base	0.4474	**0.641**	-	-	-	-
	Prompt Tuning - base	0.7546	0.7309	0.6282	0.8163	-	-
	P-Tuning - base	0.8003	0.7179	0.8582	0.5786	-	-
Model	All models	0.128	0.24	0.265	-	0.142	0.225
	Llama - Mistral	-	-	-	-	-	-
	Llama - Qwen	-	-	0.7808 **(Qwen)**	-	0.7510 **(Llama)**	0.8451 **(Qwen)**
	Llama - Gemma	-	-	1.0066 **(Gemma)**	-	-	0.8125 **(Gemma)**
	Mistral - Qwen	-	-	-	-	-	-
	Mistral - Gemma	-	-	-	-	-	-
	Qwen - Gemma	-	1.0066 **(Qwen)**	-	-	-	-
	Llama - base	-	-	0.4865	0.4759	-	0.5708
	Mistral - base	0.8753	-	0.808	0.4591	0.5008	-
	Qwen - base	0.5016	-	-	-	-	-
	Gemma - base	0.6594	1.0424	0.543	0.5901	**0.7504**	-

TABLE X: Results of statistical tests comparing safety changes across different PEFT methods, base models, and fine-tuning variables. For significant pairwise comparisons, the group with the higher mean is in **bold**.

Factor	Comparison	Safety	
		P-value	Effect Size
All results		-	-
Dataset	UC - UF	-	-
	UC - base	-	-
	UF - base	-	-
Paradigm	SFT - DPO	-	-
	SFT - base	p <0.05	0.2557
	DPO - base	-	-
Learning Rate	1e-3 - 2e-5	-	-
	1e-3 - base	-	-
	2e-5 - base	-	-
Epochs	1e - 5e	-	-
	1e - base	-	-
	5e - base	-	-
Peft Method	All methods	p <0.05	0.353
	Lora - IA^3	-	-
	IA^3 - Prompt Tuning	p <0.05	0.9397 **(IA^3)**
	IA^3 - P-Tuning	p <0.05	0.9594 **(IA^3)**
	Lora - Prompt Tuning	-	-
	Lora - P-Tuning	p <0.05	0.7936 **(Lora)**
	Prompt - P-Tuning	-	-
	Lora - base	**p = 0.0587**	**0.4031**
	IA^3 - base	**p <0.05**	**0.641**
	Prompt Tuning - base	p <0.05	0.7309
	P-Tuning - base	p <0.05	0.7179
Model	All models	p <0.05	0.24
	Llama - Mistral	-	-
	Llama - Qwen	-	-
	Llama - Gemma	-	-
	Mistral - Qwen	-	-
	Mistral - Gemma	-	-
	Qwen - Gemma	p <0.05	1.0066 **(Qwen)**
	Llama - base	-	-
	Mistral - base	-	-
	Qwen - base	-	-
	Gemma - base	p <0.05	1.0424

TABLE XI: Results of the statistical tests for all safety categories

Factor	Comparison	Safety Categories										
		1. Illegal Activity	**2. Child Abuse Content**	**3. Hate/ Harass/ Violence**	**4. Malware**	**5. Physical Harm**	**6. Economic Harm**	**7. Fraud / Deception**	**8. Adult Content**	**9. Political Campaigning**	**10. Privacy Violation**	**11. Tailored Financial Advice**
		Effect Size	**Effect Size**	**Effect Size**	**Effect Size**	**Effect Size**	**Effect Size**	**Effect Size**	**Effect Size**	**Effect Size**	**Effect Size**	**Effect Size**
All results		-	0.8726	-	**0.2366**	-	-	0.5099	0.4998	-	0.2384	-
Dataset	UC - UF	-	-	-	-	-	-	-	-	-	-	-
	UC - base	-	0.8911	-	-	-	-	0.6131	0.6169	-	-	-
	UF - base	-	0.8743	-	**0.2911**	-	-	0.4991	0.4724	-	-	-
Paradigm	SFT - DPO	-	0.7094 (DPO)	0.57 (DPO)	-	-	-	-	-		-	-
	SFT - base	-	0.8739	0.3957	-	-	-	0.5937	0.5191	0.2829	0.3191	-
	DPO - base	-	0.8944	-	**0.6929**	-	-	-	-	-	-	-
Learning Rate	1e-3 - 2e-5	-	-	-	-	-	-	-	-	-	0.425 (1e-3)	-
	1e-3 - base	-	0.8843	-	**0.4384**	-	-	0.5226	0.4742	-	-	-
	2e-5 - base	-	0.8753	-	-	-	-	0.5095	0.5132	-	0.3944	-
Epochs	1e - 5e	-	-	-	-	-	-	-	-	-	-	-
	1e - base	-	0.8815	-	-	-	-	0.6206	0.6264	-	0.4454	-
	5e - base	-	0.8761	-	**0.3697**	-	-	0.4759	0.4258	-	-	-
PEFT Method	All methods	0.266	0.326	0.239	0.314	0.135	0.386	0.306	0.328	0.189	0.214	0.353
	Lora - IA^3	0.8605 (Lora)	-	-	-	-	-	-	-	-	-	-
	IA^3 - Prompt Tuning	-	-	0.8839 (IA^3)	-	0.7973 (IA^3)	0.8612 (IA^3)	0.7756 (IA^3)	0.8643 (IA^3)	-	-	0.8214 (IA^3)
	IA^3 - P-Tuning	-	0.8833 (IA^3)	0.8221 (IA^3)	1.0113 (IA^3)	-	0.8670 (IA^3)	0.8612 (IA^3)	0.8813 (IA^3)	0.7005 (IA^3)	-	0.6776 (IA^3)
	Lora - Prompt Tuning	0.7924 (Lora)	-	-	-	-	-	-	-	-	0.802 (Lora)	0.7292 (Lora)
	Lora - P-Tuning	0.8216 (Lora)	0.8565 (Lora)	-	0.8979 (Lora)	-	0.7004 (Lora)	0.6775 (Lora)	0.7620 (Lora)	0.6851 (Lora)	0.7306 (Lora)	0.7178 (Lora)
	Prompt - P-Tuning	-	-	-	0.879 (Prompt)	-	-	0.6866 (Prompt)	-	-	-	-
	Lora - base	**0.8439**	0.9023	-	**0.6894**	**0.6709**	-	-	-	-	-	**0.4818**
	IA^3 - base	-	0.896	**0.6522**	**0.8307**	**0.8457**	**0.7902**	-	-	**0.715**	-	-
	Prompt Tuning - base	-	0.8885	0.779	-	0.6857	0.688	0.7138	-	-	0.6723	0.6622
	P-Tuning - base	0.5693	0.8821	0.5707	0.5426	0.5366	0.6448	0.8182	0.7627	0.6518	0.5542	0.4842
Model	All models	-	0.456	-	0.426	-	0.305	0.282	0.373	0.336	0.14	0.212
	Llama - Mistral	-	1.0066 (Llama)	-	1.0066 (Mistral)	-	0.8125 (Mistral)	-	-	-	-	0.8451 (Llama)
	Llama - Qwen	-	-	-	0.7741 (Qwen)	-	-	-	-	-	-	-
	Llama - Gemma	-	-	-	-	-	-	-	-	-	-	-
	Mistral - Qwen	-	1.0066 (Qwen)	-	-	-	-	0.8223 (Mistral)	0.9518 (Qwen)	-	-	-
	Mistral - Gemma	-	0.8748 (Gemma)	-	0.9184 (Mistral)	-	1.0066 (Mistral)	1.0066 (Mistral)	-	-	-	-
	Qwen - Gemma	-	-	-	1.0066 (Qwen)	-	0.8588 (Qwen)	-	0.9184 (Qwen)	1.0066 (Qwen)	-	-
	Llama - base	-	-	-	0.887	-	-	-	-	0.8922	-	-
	Mistral - base	-	0.876	-	**0.6342**	-	**0.5011**	-	0.8184	-	-	-
	Qwen - base	-	0.9129	0.8875	**0.6682**	-	-	0.7623	**0.5926**	**0.5027**	0.8828	-
	Gemma - base	0.8922	0.8917	0.8924	0.7902	0.8881	0.7638	0.8798	0.8893	0.8904	0.8856	0.6973

TABLE XII: Results of statistical tests comparing fairness metrics changes across different fine-tuning variables. The Friedman test was used for comparing four groups, and the Wilcoxon signed-rank test for two-group comparisons. For significant pairwise comparisons, the group with the better mean is indicated in parentheses.

Factor	Comparison	Fairness							
		Accuracy AMB		Accuracy DIS		Bias Score AMB		Bias Score DIS	
		P	Effect Size	P	Effect Size	P	Effect Size	P	Effect Size
All results		<0.05	0.5417	<0.05	0.3932	-	-	<0.05	0.2591
Dataset	UC - UF	-	-	-	-	-	-	-	-
	UC - base	-	-	<0.05	0.6809	-	-	<0.05	0.5434
	UF - base	<0.05	0.558	<0.05	0.3349	-	-	-	-
Paradigm	SFT - DPO	<0.05	0.808 **(DPO)**	<0.05	0.4729 **(DPO)**	-	-	<0.05	0.5726 **(DPO)**
	SFT - base	<0.05	0.6206	<0.05	0.5893	-	-	<0.05	0.4343
	DPO - base	-	-	-	-	-	-	-	-
Learning Rate	1e-3 - 2e-5	<0.05	0.5892 **(2e-5)**	-	-	-	-	<0.05	0.3855 **(2e-5)**
	1e-3 - base	<0.05	0.762	-	-	-	-	<0.05	0.509
	2e-5 - base	<0.05	0.4237	<0.05	0.4418	-	-	-	-
Epochs	1e - 5e	-	-	-	-	-	-	-	-
	1e - base	<0.05	0.4441	<0.05	0.6305	-	-	-	-
	5e - base	<0.05	0.5841	<0.05	0.2713	-	-	-	-
PEFT Method	All methods	<0.05	0.159	<0.05	0.193	<0.05	0.144	-	-
	Lora - IA^3	<0.05	0.7159 **(IA^3)**	-	-	-	-	-	-
	IA^3 - Prompt Tuning	<0.05	0.6582 **(IA^3)**	<0.05	0.81 **(IA^3)**	-	-	-	-
	IA^3 - P-Tuning	<0.05	0.8979 **(IA^3)**	-	-	-	-	-	-
	Lora - Prompt Tuning	-	-	-	-	-	-	-	-
	Lora - P-Tuning	-	-	-	-	-	-	-	-
	Prompt - P-Tuning	-	-	-	-	-	-	-	-
	Lora - base	**<0.05**	**0.6194**	-	-	-	-	-	-
	IA^3 - base	-	-	-	-	-	-	-	-
	Prompt Tuning - base	<0.05	0.6282	<0.05	0.8163	-	-	-	-
	P-Tuning - base	<0.05	0.8582	<0.05	0.5786	-	-	-	-
Base Model	All models	<0.05	0.265	-	-	<0.05	0.142	<0.05	0.225
	Llama - Mistral	-	-	-	-	-	-	-	-
	Llama - Qwen	<0.05	0.7808 **(Qwen)**	-	-	0.056	0.7510 **(Llama)**	<0.05	0.8451 **(Qwen)**
	Llama - Gemma	<0.05	1.0066 **(Gemma)**	-	-	-	-	<0.05	0.8125 **(Gemma)**
	Mistral - Qwen	-	-	-	-	-	-	-	-
	Mistral - Gemma	-	-	-	-	-	-	-	-
	Qwen - Gemma	-	-	-	-	-	-	-	-
	Llama - base	<0.05	0.4865	<0.05	0.4759	-	-	<0.05	0.5708
	Mistral - base	<0.05	0.808	<0.05	0.4591	<0.05	0.5008	-	-
	Qwen - base	-	-	-	-	-	-	-	-
	Gemma - base	<0.05	0.543	<0.05	0.5901	**<0.05**	**0.7504**	-	-

TABLE XIII: Results of the statistical tests for fairness categories 1 and 2

Factor	Comparison	1. Age				2. Disability Status			
		Acc. AMB	**Acc. DIS**	**Bias AMB**	**Bias DIS**	**Acc. AMB**	**Acc. DIS**	**Bias AMB**	**Bias DIS**
		Effect Size	**Effect Size**	**Effect Size**	**Effect Size**	**Effect Size**	**Effect Size**	**Effect Size**	**Effect Size**
All results		0.5268	0.5639	**0.3563**	-	0.4938	0.2857	-	0.2243
Dataset	UC - UF	-	-	-	-	-	-	-	-
	UC - base	-	0.7298	-	-	0.5621	-	-	-
	UF - base	0.5402	0.5253	**0.3223**	-	0.4904	0.2501	-	-
Paradigm	SFT - DPO	0.6795 (DPO)	0.5220 (DPO)	-	-	0.7629 (DPO)	0.5434 (DPO)	-	-
	SFT - base	0.613	0.7067	**0.4875**	-	0.5893	0.4922	-	0.3357
	DPO - base	-	-	-	-	-	-	0.5538	-
Learning Rate	1e-3 - 2e-5	0.5262 (2e-5)	-	-	-	-	-	-	-
	1e-3 - base	0.6988	0.5618	-	-	0.5664		-	-
	2e-5 - base	0.4339	0.5835	**0.392**	-	0.4641	-	-	-
Epochs	1e - 5e	-	-	-	-	-	-	-	-
	1e - base	0.4558	0.7907	**0.4718**	-	0.4624	0.4508	-	-
	5e - base	0.5268	0.4698	**0.2701**	-	0.5159	-	-	-
PEFT Method	All methods	0.146	0.291	0.284	-	-	0.137	-	-
	Lora - IA^3	0.7620 (IA^3)	-	-	-	-	-	-	-
	IA^3 - Prompt Tuning	-	0.9397 (IA^3)	1.0113 (Prompt)	-	-	0.7063 (IA^3)	-	-
	IA^3 - P-Tuning	0.7936 (IA^3)	0.8264 (IA^3)	-	-	-	-	-	-
	Lora - Prompt Tuning	-	-	0.7777 (Prompt)	-	-	-	-	-
	Lora - P-Tuning	-	0.6442 (Lora)	-	-	-	-	-	-
	Prompt - P-Tuning	-	-	-	-	-	-	-	-
	Lora - base	0.636	0.4188	-	-	0.4811	-	-	-
	IA^3 - base	-	-	-	0.4741	-	-	-	-
	Prompt Tuning - base	0.5565	1.0548	**1.0194**	-	0.5441	0.6416	-	-
	P-Tuning - base	0.8347	0.7263	-	-	0.6177	0.5183	-	-
Model	All models	0.184	-	0.142	-	0.168	-	-	0.188
	Llama - Mistral	0.8451 (Mistral)	-	-	-	-	-	-	0.7808 (Mistral)
	Llama - Qwen	-	-	-	-	-	-	-	-
	Llama - Gemma	0.8748 (Gemma)	-	-	-	1.0066 (Gemma)	-	-	0.8748 (Gemma)
	Mistral - Qwen	-	-	0.8451 (Mistral)	-	-	-	-	-
	Mistral - Gemma	-	-	-	-	-	-	-	-
	Qwen - Gemma	-	-	-	-	-	-	-	-
	Llama - base	0.5049	-	-	-	-	-	-	0.4801
	Mistral - base	0.6786	0.6659	**0.8753**	-	0.7421	-	-	0.5149
	Qwen - base	-	0.4817	-	0.4757	-	-	-	-
	Gemma - base	0.8415	0.6982	**0.8033**	-	-	0.517	**0.5314**	-

TABLE XIV: Results of the statistical tests for fairness categories 3 and 4

Factor	Comparison	3. Gender Identity				4. Nationality			
		Acc. AMB	Acc. DIS	Bias AMB	Bias DIS	Acc. AMB	Acc. DIS	Bias AMB	Bias DIS
		Effect Size	Effect Size	Effect Size	Effect Size	Effect Size	Effect Size	Effect Size	Effect Size
All results		0.6221	-	**-**	0.2198	0.5961	0.355	-	-
Dataset	UC - UF	-	0.6010 (UF)	-	-	-	-	-	-
	UC - base	0.5817	0.5434	-	0.527	0.5453	0.7804	-	-
	UF - base	0.6295	-	**-**	-	0.6027	0.2732	-	-
Paradigm	SFT - DPO	0.7441 (DPO)	0.4386 (DPO)	-	0.8561	0.8410 (DPO)	-	-	0.8561 (DPO)
	SFT - base	0.7039	0.3298	**-**	0.4258	0.6916	0.5058	-	0.275
	DPO - base	-	**0.5726**	-	**0.4115**	-	-	-	**0.4523**
Learning Rate	1e-3 - 2e-5	0.4819 (2e-5)	-	-	-	0.4929 (2e-5)	-	-	0.4655 (2e-5)
	1e-3 - base	0.7986	-	-	0.3696	0.7213	0.3923	-	-
	2e-5 - base	0.523	-	**-**	-	0.542	0.3558	-	-
Epochs	1e - 5e	-	-	-	-	-	-	-	-
	1e - base	0.5681	-	**-**	-	0.6077	0.6264	-	-
	5e - base	0.6438	-	**-**	-	0.5986	-	-	-
PEFT Method	All methods	0.216	-	-	-	0.227	-	0.128	-
	Lora - IA^3	-	-	-	-	0.6722 (IA^3)	-	-	-
	IA^3 - Prompt Tuning	0.7479 (IA^3)	-	-	-	0.7463 (IA^3)	-	-	-
	IA^3 - P-Tuning	0.8273 (IA^3)	-	-	-	0.9594 (IA^3)	-	-	-
	Lora - Prompt Tuning	-	-	-	-	-	-	-	-
	Lora - P-Tuning	-	-	-	-	-	-	-	-
	Prompt - P-Tuning	-	-	-	-	-	-	-	-
	Lora - base	0.5919	-	-	-	0.5726	-	-	-
	IA^3 - base	-	-	-	-	-	-	-	-
	Prompt Tuning - base	0.8801	-	**-**	-	0.7871	-	-	-
	P-Tuning - base	0.8234	0.4436	-	0.4599	0.8465	0.5353	-	-
Model	All models	0.232	-	0.198	-	0.153	0.25	0.212	0.252
	Llama - Mistral	-	-	-	-	-	-	-	0.8125 (Mistral)
	Llama - Qwen	0.8451 (Qwen)	-	-	-	0.7808 (Qwen)	1.0066 (Qwen)	-	0.9184 (Qwen)
	Llama - Gemma	0.8451 (Gemma)	-	0.8451 (Gemma)	-	0.7510 (Gemma)	-	-	0.8125 (Gemma)
	Mistral - Qwen	-	-	-	-	-	-	-	-
	Mistral - Gemma	0.9518 (Gemma)	-	0.8125 (Gemma)	-	-	0.7808 (Mistral)	1.0066 (Gemma)	-
	Qwen - Gemma	-	-	0.8748 (Gemma)	-	-	-	-	-
	Llama - base	0.4683	-	**-**	-	0.4957	0.7683	-	-
	Mistral - base	0.9318	-	**-**	-	0.9199	-	0.6956	-
	Qwen - base	0.8791	**0.6854**	0.8791	-	0.4801	-	-	0.5987
	Gemma - base	0.517	0.543	**0.9027**	-	0.5495	0.7305	**0.667**	-

TABLE XV: Results of the statistical tests for fairness categories 5 and 6

Factor	Comparison	5. Physical Appearance				6. Race/Ethnicity			
		Acc. AMB Effect Size	Acc. DIS Effect Size	Bias AMB Effect Size	Bias DIS Effect Size	Acc. AMB Effect Size	Acc. DIS Effect Size	Bias AMB Effect Size	Bias DIS Effect Size
All results		0.6519	-	**0.2154**	0.3917	0.533	0.5677	-	0.5676
Dataset	UC - UF	0.6341 (UF)	-	-	-	-	-	0.5718 (UC)	-
	UC - base	0.7656	-	-	0.6104	-	0.6713	-	0.7573
	UF - base	0.6351	-	**0.2671**	0.3409	0.5422	0.5469	-	0.5343
Paradigm	SFT - DPO	0.7371 (DPO)	-	0.5363 (DPO)	-	0.7302 (DPO)	0.6098 (DPO)	-	0.8561 (DPO)
	SFT - base	0.7682	-	**-**	0.4457	0.6156	0.6722	-	0.6719
	DPO - base	-	**-**	-	**-**	-	-	-	**-**
Learning Rate	1e-3 - 2e-5	0.4583 (2e-5)	-	-	-	0.5951 (2e-5)	-	-	-
	1e-3 - base	0.7443	-	-	0.4709	0.7327	0.6134	-	0.7044
	2e-5 - base	0.6519	-	**-**	0.3367	0.4262	0.5519	-	0.5084
Epochs	1e - 5e	-	-	0.5661 (1e)	-	-	-	-	-
	1e - base	0.7292	-	**-**	0.3915	0.4346	0.652	-	0.5923
	5e - base	0.6348	-	**0.3083**	0.3915	0.5721	0.5296	-	0.5454
PEFT Method	All methods	0.187	-	-	0.197	0.176	0.266	-	0.247
	Lora - IA^3	0.8153 (IA^3)	-	-	-	0.6582 (IA^3)	-	-	-
	IA^3 - Prompt Tuning	-	-	-	0.7159 (IA^3)	-	0.8979 (IA^3)	-	1.0113 (IA^3)
	IA^3 - P-Tuning	0.8979 (IA^3)	-	-	0.7011 (IA^3)	0.9184 (IA^3)	0.8616 (IA^3)	-	0.7620 (IA^3)
	Lora - Prompt Tuning	-	-	-	0.6582 (Lora)	-	-	-	0.6442 (Lora)
	Lora - P-Tuning	-	-	-	-	-	-	-	0.6442 (Lora)
	Prompt - P-Tuning	-	-	-	-	-	-	-	-
	Lora - base	0.7787	-	0.5457	-	0.5003	-	-	
	IA^3 - base	0.6661	-	0.6784	-	-	-	-	-
	Prompt Tuning - base	0.6039	-	**-**	0.6656	0.6622	0.9523	-	1.0194
	P-Tuning - base	0.8826	-	-	0.5875	0.8123	0.8702	-	0.8123
Model	All models	0.222	0.132	-	-	0.262	-	0.335	0.265
	Llama - Mistral	-	0.8125 (Mistral)	-	-	0.7808 (Mistral)	-	0.8451 (Llama)	0.7808 (Mistral)
	Llama - Qwen	1.0066 (Qwen)	0.7879 (Qwen)	-	-	0.7808 (Qwen)	-	1.0066 (Llama)	1.0066 (Qwen)
	Llama - Gemma	-	-	-	-	1.0066 (Gemma)	-	-	-
	Mistral - Qwen	-	-	-	-	-	-	-	-
	Mistral - Gemma	-	-	-	-	-	-	-	-
	Qwen - Gemma	-	0.7808 (Qwen)	-	-	-	-	0.7808 (Gemma)	-
	Llama - base	0.5134	0.718	**-**	-	0.5009	0.5539	**1.0333**	0.8807
	Mistral - base	0.8685	**0.4399**	**0.8533**	0.4503	0.7262	0.7717	0.5653	-
	Qwen - base	-	**0.4929**	-	-	-	-	0.8764	0.6475
	Gemma - base	0.8792	0.802	**0.5752**	0.9027	0.517	0.724	**0.7504**	0.6358

TABLE XVI: Results of the statistical tests for fairness categories 7 and 8

Factor	Comparison	7. Religion				8. Socio-Economic Status			
		Acc. AMB	Acc. DIS	Bias AMB	Bias DIS	Acc. AMB	Acc. DIS	Bias AMB	Bias DIS
		Effect Size	Effect Size	Effect Size	Effect Size	Effect Size	Effect Size	Effect Size	Effect Size
All results		0.7453	0.4103	**-**	-	0.5641	0.3472	-	-
Dataset	UC - UF	-	-	-	-	-	-	-	-
	UC - base	0.8819	-	-	-	-	0.6299	-	-
	UF - base	0.7246	0.39	**-**	-	0.5799	0.2897	-	-
Paradigm	SFT - DPO	0.7926 (DPO)	-	0.5423 (DPO)	-	0.8269 (DPO)	-	0.5550 (DPO)	0.6795 (DPO)
	SFT - base	0.8485	0.4847	**-**	0.2802	0.6698	0.4982	-	-
	DPO - base	-	**-**	-	**-**	-	-	-	**-**
Learning Rate	1e-3 - 2e-5	0.4172 (2e-5)	-	-	-	0.6010 (2e-5)	-	-	-
	1e-3 - base	0.7768	0.4304	-	-	0.762	-	-	-
	2e-5 - base	0.7269	0.3902	**-**	-	0.4654	0.427	-	-
Epochs	1e - 5e	-	-	-	-	-	-	-	-
	1e - base	0.8394	0.5032	**-**	-	0.4439	0.6736	-	-
	5e - base	0.7188	0.3537	**0.2943**	-	0.6092	-	-	-
PEFT Method	All methods	0.403	0.37	-	-	0.266	0.277	0.144	-
	Lora - IA^3	0.8784 (IA^3)	-	-	-	0.7866 (IA^3)	-	-	-
	IA^3 - Prompt Tuning	0.8100 (IA^3)	1.0489 (IA^3)	-	-	0.6442 (IA^3)	0.8100 (IA^3)	-	-
	IA^3 - P-Tuning	1.0113 (IA^3)	-	-	-	0.9397 (IA^3)	-	-	-
	Lora - Prompt Tuning	-	0.7936 (Lora)	-	-	-	0.7620 (Lora)	-	-
	Lora - P-Tuning	-	-	-	-	-	-	-	-
	Prompt - P-Tuning	-	-	-	-	-	-	-	-
	Lora - base	0.7787	-	-	-	0.6875	-	-	-
	IA^3 - base	-	-	-	-	-	-	-	-
	Prompt Tuning - base	0.8675	0.8675	**-**	-	0.6039	0.7871	-	-
	P-Tuning - base	0.8826	0.5525	-	-	0.8465	0.5014	-	-
Model	All models	0.147	-	0.212	-	0.195	-	-	0.142
	Llama - Mistral	-	-	0.8748 (Llama)	-	-	-	-	-
	Llama - Qwen	0.8155 (Qwen)	-	-	-	0.7808 (Qwen)	-	-	-
	Llama - Gemma	-	-	-	-	1.0066 (Gemma)	-	-	-
	Mistral - Qwen	-	-	-	-	-	-	-	-
	Mistral - Gemma	-	-	-	-	-	-	-	-
	Qwen - Gemma	-	-	-	-	-	-	-	-
	Llama - base	0.4986	-	**0.6512**	-	0.4502	0.4671	**-**	0.5927
	Mistral - base	0.8571	**-**	**0.7325**	-	0.7717	-	0.522	-
	Qwen - base	0.8683	**0.6848**	0.644	-	-	-	-	0.477
	Gemma - base	0.8792	0.5954	**0.6358**	0.6853	0.716	0.8822	**0.8033**	-

TABLE XVII: Results of the statistical tests for fairness category 9

Factor	Comparison	9. Sexual Orientation			
		Acc. AMB	Acc. DIS	Bias AMB	Bias DIS
		Effect Size	Effect Size	Effect Size	Effect Size
All results		0.705	0.3802	**-**	-
Dataset	UC - UF	-	-	-	-
	UC - base	0.7467	-	-	-
	UF - base	0.7004	0.3905	**-**	-
Paradigm	SFT - DPO	0.7582 (DPO)	-	-	-
	SFT - base	0.8008	0.4737	**-**	-
	DPO - base	-	**-**	0.6166	**-**
Learning Rate	1e-3 - 2e-5	-	-	-	-
	1e-3 - base	0.7236	0.4046	-	-
	2e-5 - base	0.7058	0.3645	**-**	-
Epochs	1e - 5e	-	-	-	-
	1e - base	0.7642	-	**-**	-
	5e - base	0.6923	0.3891	**-**	-
PEFT Method	All methods	0.197	0.174	0.234	-
	Lora - IA^3	0.8145 (IA^3)	-	-	-
	IA^3 - Prompt Tuning	-	0.8210 (IA^3)	-	-
	IA^3 - P-Tuning	0.8616 (IA^3)	-	-	-
	Lora - Prompt Tuning	-	0.6442 (Lora)	-	-
	Lora - P-Tuning	-	-	-	-
	Prompt - P-Tuning	-	-	-	-
	Lora - base	0.7516	-	0.478	-
	IA^3 - base	0.4771	-	-	0.5137
	Prompt Tuning - base	0.6377	0.8385	**-**	-
	P-Tuning - base	0.8056	0.4817	-	-
Model	All models	0.182	0.235	-	0.425
	Llama - Mistral	-	-	-	0.8223 (Mistral)
	Llama - Qwen	0.8487 (Qwen)	1.0066 (Qwen)	-	-
	Llama - Gemma	-	-	-	0.9518 (Llama)
	Mistral - Qwen	-	-	-	-
	Mistral - Gemma	-	-	-	1.0066 (Mistral)
	Qwen - Gemma	0.9941 (Qwen)	-	-	1.0066 (Qwen)
	Llama - base	0.5447	0.8803	**-**	-
	Mistral - base	0.803	**-**	**-**	**<u>0.694</u>**
	Qwen - base	-	**-**	-	-
	Gemma - base	0.8799	0.595	**<u>0.6358</u>**	1.0424

REFERENCES

[1] V. B. Parthasarathy, A. Zafar, A. Khan, and A. Shahid, "The ultimate guide to fine-tuning LLMs from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities," *arXiv preprint arXiv:2408.13296*, Aug. 2024.

[2] Z. Chu, S. Wang, J. Xie, T. Zhu, Y. Yan, J. Ye, A. Zhong, X. Hu, J. Liang, P. S. Yu *et al.*, "LLM agents for education: Advances and applications," *arXiv preprint arXiv:2503.11733*, Mar. 2025.

[3] B. Huo, A. Boyle, N. Marfo, W. Tangamornsuksan, J. P. Steen, T. McKechnie, Y. Lee, J. Mayol, S. A. Antoniou, A. J. Thirunavukarasu *et al.*, "Large Language Models for Chatbot Health Advice Studies: A Systematic Review," *JAMA Network Open*, vol. 8, no. 2, p. e2457879, Feb. 2025.

[4] B. Zhang, Z. Liu, C. Cherry, and O. Firat, "When scaling meets LLM finetuning: The effect of data, model and finetuning method," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2024.

[5] T. Xiao and J. Zhu, "Foundations of large language models," *arXiv preprint arXiv:2501.09223*, Jan. 2025.

[6] Z. Gekhman, G. Yona, R. Aharoni, M. Eyal, A. Feder, R. Reichart, and J. Herzig, "Does fine-tuning LLMs on new knowledge encourage hallucinations?" in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Nov. 2024, pp. 7765–7784.

[7] E. Lobo, C. Agarwal, and H. Lakkaraju, "On the impact of fine-tuning on chain-of-thought reasoning," *arXiv preprint arXiv:2411.15382*, Nov. 2024.

[8] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang, "An empirical study of catastrophic forgetting in large language models during continual fine-tuning," *arXiv preprint arXiv:2308.08747*, Jan. 2025.

[9] X. Chen, S. Tang, R. Zhu, S. Yan, L. Jin, Z. Wang, L. Su, Z. Zhang, X. Wang, and H. Tang, "The janus interface: How fine-tuning in large language models amplifies privacy risks," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, Dec. 2024, pp. 1285–1299.

[10] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, "Fine-tuning aligned language models compromises safety, even when users do not intend to!" in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2024.

[11] F. Eiras, A. Petrov, P. H. S. Torr, M. P. Kumar, and A. Bibi, "Do as i do (safely): Mitigating task-specific fine-tuning risks in large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2025.

[12] S. Lermen and C. Rogers-Smith, "LoRA fine-tuning efficiently undoes safety training in Llama 2-chat 70b," in *Proc. ICLR Workshop on Secure and Trustworthy Large Language Models*, Apr. 2024.

[13] C.-Y. Hsu, Y.-L. Tsai, C.-H. Lin, P.-Y. Chen, C.-M. Yu, and C.-Y. Huang, "Safe LoRA: The silver lining of reducing safety risks when fine-tuning large language models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, Dec. 2024, pp. 65 072–65 094.

[14] H. Liu, S. Zhong, X. Sun, M. Tian, M. Hariri, Z. Liu, R. Tang, Z. Jiang, J. Yuan, Y.-N. Chuang *et al.*, "LoRATK: LoRA once, backdoor everywhere in the share-and-play ecosystem," *arXiv preprint arXiv:2403.00108*, Mar. 2024.

[15] J. Yi, R. Ye, Q. Chen, B. Zhu, S. Chen, D. Lian, G. Sun, X. Xie, and F. Wu, "On the vulnerability of safety alignment in open-access LLMs," in *Findings Assoc. Comput. Linguist. (ACL)*, Aug. 2024, pp. 9236–9260.

[16] HuggingFace, "huggingface (Hugging Face)," Jan. 2025. [Online]. Available: https://huggingface.co/huggingface

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2022.

[18] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, Dec. 2022, pp. 1950–1965.

[19] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Nov. 2021, pp. 3045–3059.

[20] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "GPT understands, too," *AI Open*, vol. 5, pp. 208–215, Jan. 2024.

[21] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, Jul. 2024.

[22] Meta-Llama, "meta-llama/Meta-Llama-3-8B-Instruct · Hugging Face — huggingface.co," https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct, [Accessed 24 February 2025].

[23] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, Oct. 2023.

[24] MistralAI, "mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face — huggingface.co," https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3, [Accessed 24 February 2025].

[25] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang *et al.*, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, Dec. 2024.

[26] Qwen, "Qwen/Qwen2.5-7B-Instruct · Hugging Face — huggingface.co," https://huggingface.co/Qwen/Qwen2.5-7B-Instruct, [Accessed 24 February 2025].

[27] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on Gemini research and technology," *arXiv preprint arXiv:2403.08295*, Mar. 2024.

[28] Google, "google/gemma-7b-it · Hugging Face — huggingface.co," https://huggingface.co/google/gemma-7b-it, [Accessed 24 February 2025].

[29] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman, "BBQ: A hand-built bias benchmark for question answering," in *Findings Assoc. Comput. Linguist. (ACL)*, May 2022, pp. 2086–2105.

[30] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Jul. 2018, pp. 328–339.

[31] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2022.

[32] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, Dec. 2022, pp. 27 730–27 744.

[33] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse *et al.*, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," *arXiv preprint arXiv:2209.07858*, Sep. 2022.

[34] HuggingFace, "Supervised Fine-Tuning - Hugging Face LLM Course." [Online]. Available: https://huggingface.co/learn/llm-course/en/chapter11/3

[35] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, Apr. 2022.

[36] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, Dec. 2023, pp. 53 728–53 741.

[37] A. Li, Y. Mo, M. Li, Y. Wang, and Y. Wang, "Are smarter LLMs safer? exploring safety-reasoning trade-offs in prompting and fine-tuning," *arXiv preprint arXiv:2502.09673*, Feb. 2025.

[38] E. Jan, N. Aldahoul, M. Ali, F. Ahmad, F. Zaffar, and Y. Zaki, "Multitask-Bench: Unveiling and mitigating safety gaps in LLMs fine-tuning," in *Proc. Int. Conf. Comput. Linguist. (COLING)*, Jan. 2025, pp. 9025–9043.

[39] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey," *Transactions on Machine Learning Research*, Jul. 2024.

[40] X.-K. Wu, M. Chen, W. Li, R. Wang, L. Lu, J. Liu, K. Hwang, Y. Hao, Y. Pan, Q. Meng *et al.*, "LLM Fine-Tuning: Concepts, Opportunities, and Challenges," *Big Data and Cognitive Computing*, vol. 9, no. 4, p. 87, 2025.

[41] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, Jul. 2023.

[42] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, Dec. 2023, pp. 46 595–46 623.

[43] S. Liu, W. Fang, Z. Hu, J. Zhang, Y. Zhou, K. Zhang, R. Tu, T.-E. Lin, F. Huang, M. Song *et al.*, "A survey of direct preference optimization," *arXiv preprint arXiv:2503.11701*, Mar. 2025.

[44] V. Lialin, V. Deshpande, X. Yao, and A. Rumshisky, "Scaling down to scale up: A guide to parameter-efficient fine-tuning," *arXiv preprint arXiv:2303.15647*, Mar. 2023.

[45] X. Qi, A. Panda, K. Lyu, X. Ma, S. Roy, A. Beirami, P. Mittal, and P. Henderson, "Safety alignment should be made more than just a few tokens deep," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2025.

[46] X. Yang, X. Wang, Q. Zhang, L. Petzold, W. Y. Wang, X. Zhao, and D. Lin, "Shadow alignment: The ease of subverting safely-aligned language models," arXiv preprint arXiv:2310.02949, Oct. 2023, accessed: 2025-10-08. [Online]. Available: http://arxiv.org/abs/2310.02949

[47] K. Lyu, H. Zhao, X. Gu, D. Yu, A. Goyal, and S. Arora, "Keeping LLMs aligned after fine-tuning: The crucial role of prompt templates," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, Dec. 2024, pp. 118 603–118 631.

[48] F. Jiang, Z. Xu, L. Niu, B. Y. Lin, and R. Poovendran, "ChatBug: A common vulnerability of aligned LLMs induced by chat templates," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2025, pp. 27 347–27 355.

[49] J. Betley, D. C. H. Tan, N. Warncke, A. Sztyber-Betley, X. Bao, M. Soto, N. Labenz, and O. Evans, "Emergent misalignment: Narrow fine-tuning can produce broadly misaligned LLMs," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2025.

[50] huggingface.co, "GitHub - huggingface/huggingface_hub: The official Python client for the Huggingface Hub." https://github.com/huggingface/huggingface_hub, [Accessed: 24 February 2025].

[51] ——, "Hub API Endpoints," https://huggingface.co/docs/hub/api, [Accessed: 24 February 2025].

[52] "Qwen (Qwen)," Sep. 2025. [Online]. Available: https://huggingface.co/Qwen

[53] F. Barbieri, J. Camacho-Collados, L. E. Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings Assoc. Comput. Linguist. (EMNLP)*, Nov. 2020, pp. 1644–1650.

[54] "cardiffnlp/tweet_eval · Datasets at Hugging Face," Aug. 2025. [Online]. Available: https://huggingface.co/datasets/cardiffnlp/tweet_eval

[55] "meta-llama (Meta Llama)," May 2025. [Online]. Available: https://huggingface.co/meta-llama

[56] HuggingFaceH4, "HuggingFaceH4/ultrafeedback_binarized · Datasets at Hugging Face — huggingface.co," https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized, [Accessed 24 February 2025].

[57] "google (Google)," Feb. 2025. [Online]. Available: https://huggingface.co/google/models

[58] X. L. Li and P. Liang, "Prefix-Tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annu. Meet. Assoc. Comput. Linguist. and 11th Int. Joint Conf. Nat. Lang. Process. (ACL-IJCNLP)*, Aug. 2021, pp. 4582–4597.

[59] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 28, Dec. 2015, pp. 649–657.

[60] "fancyzhx/ag_news · Datasets at Hugging Face," Aug. 2024. [Online]. Available: https://huggingface.co/datasets/fancyzhx/ag_news

[61] "mistralai (Mistral AI_)," Aug. 2025. [Online]. Available: https://huggingface.co/mistralai/models

[62] HuggingFaceH4, "HuggingFaceH4/ultrachat_200k · Datasets at Hugging Face — huggingface.co," https://huggingface.co/datasets/HuggingFaceH4/ultrachat_200k, [Accessed 24 February 2025].

[63] "openai-community/gpt2 · Hugging Face." [Online]. Available: https://huggingface.co/openai-community/gpt2

[64] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. EMNLP Workshop BlackboxNLP*, Nov. 2018, pp. 353–355.

[65] "nyu-mll/glue · Datasets at Hugging Face," Dec. 2023. [Online]. Available: https://huggingface.co/datasets/nyu-mll/glue

[66] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu *et al.*, "DeepSeekMath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, Feb. 2024.

[67] "Mistral 7B | Mistral AI." [Online]. Available: https://mistral.ai/news/announcing-mistral-7b

[68] "mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face." [Online]. Available: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[69] "an issue of prefix tuning on llama3-8b · Issue #454 · meta-llama/llama-cookbook — github.com," https://github.com/meta-llama/llama-cookbook/issues/454#issuecomment-2089584315, [Accessed 24 February 2025].

[70] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. Von Werra, C. Fourrier, N. Habib *et al.*, "Zephyr: Direct distillation of lm alignment," *arXiv preprint arXiv:2310.16944*, 2023.

[71] G. Cui, L. Yuan, N. Ding, G. Yao, B. He, W. Zhu, Y. Ni, G. Xie, R. Xie, Y. Lin *et al.*, "ULTRAFEEDBACK: Boosting language models with scaled AI feedback," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, Jul. 2024, pp. 9722–9744.

[72] OpenBMB, "openbmb/UltraFeedback · Datasets at Hugging Face," https://huggingface.co/datasets/openbmb/UltraFeedback, May 2025.

[73] N. Ding, Y. Chen, B. Xu, Y. Qin, S. Hu, Z. Liu, M. Sun, and B. Zhou, "Enhancing chat language models by scaling high-quality instructional conversations," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Dec. 2023, pp. 3029–3051.

[74] OpenAI, "Introducing chatgpt," https://openai.com/blog/chatgpt, 2022, [Accessed 24 February 2025].

[75] N. Ding, "stingning/ultrachat · Datasets at Hugging Face," https://huggingface.co/datasets/stingning/ultrachat, Oct. 2023.

[76] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, and Q. Gallouédec, "TRL: Transformer Reinforcement Learning," https://github.com/huggingface/trl, 2020.

[77] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan, "PEFT: State-of-the-art Parameter-Efficient Fine-Tuning Methods," https://github.com/huggingface/peft, 2022.

[78] B. Vidgen, A. Agrawal, A. M. Ahmed, V. Akinwande, N. Al-Nuaimi, N. Alfaraj, E. Alhajjar, L. Aroyo, T. Bavalatti, and M. Bartolo, "Introducing v0.5 of the AI safety benchmark from MLCommons," *arXiv preprint arXiv:2404.12241*, May 2024.

[79] Y. Li, M. Du, R. Song, X. Wang, and Y. Wang, "A survey on fairness in large language models," *arXiv preprint arXiv:2308.10149*, Aug. 2023.

[80] Z. Chu, Z. Wang, and W. Zhang, "Fairness in Large Language Models: A Taxonomic Survey," *ACM SIGKDD Explorations Newsletter*, vol. 26, no. 1, pp. 34–48, Jul. 2024.

[81] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso *et al.*, "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models," *Transactions on Machine Learning Research*, vol. 2023, no. 5, pp. 1–95, 2023.

[82] "BIG-bench/bigbench/benchmark_tasks/bbq_lite at main · google/BIG-bench," https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/bbq_lite, [Accessed 24 February 2025].

[83] "LLM-Tuning-Safety/HEx-PHI · Datasets at Hugging Face." [Online]. Available: https://huggingface.co/datasets/LLM-Tuning-Safety/HEx-PHI

[84] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, Jul. 2023.

[85] Z. Xu, F. Jiang, L. Niu, J. Jia, B. Y. Lin, and R. Poovendran, "SafeDecoding: Defending against jailbreak attacks via safety-aware decoding," in *Proc. 62nd Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Aug. 2024, pp. 5587–5605.

[86] F. Jiang, Z. Xu, L. Niu, Z. Xiang, B. Ramasubramanian, B. Li, and R. Poovendran, "ArtPrompt: ASCII art-based jailbreak attacks against aligned LLMs," in *Proc. 62nd Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Aug. 2024, pp. 15 157–15 173.

[87] S. Banerjee, S. Layek, S. Tripathy, S. Kumar, A. Mukherjee, and R. Hazra, "SafeInfer: Context-adaptive decoding-time safety alignment for large language models," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Apr. 2025, pp. 27 188–27 196.

[88] J. Wang, J. Li, Y. Li, X. Qi, J. Hu, Y. Li, P. McDaniel, M. Chen, B. Li, and C. Xiao, "BackdoorAlign: Mitigating fine-tuning-based jailbreak attack with backdoor-enhanced safety alignment," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, Dec. 2024, pp. 5210–5243.

[89] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, and D. Testuggine, "Llama guard: LLM-based input–output safeguard for human–ai conversations," *arXiv preprint arXiv:2312.06674*, Dec. 2023.

[90] Llama-Team, "Meta llama guard 2," https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md, 2024.

[91] J. Chi, U. Karn, H. Zhan, E. Smith, J. Rando, Y. Zhang, K. Plawiak, Z. D. Coudert, K. Upasani, and M. Pasupuleti, "Llama Guard 3 Vision: Safeguarding human–ai image understanding conversations," *arXiv preprint arXiv:2411.10414*, Nov. 2024.

[92] S. Han, K. Rao, A. Ettinger, L. Jiang, B. Y. Lin, N. Lambert, Y. Choi, and N. Dziri, "Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, Dec. 2024.

[93] W. Zeng, Y. Liu, R. Mullins, L. Peran, J. Fernandez, H. Harkous, K. Narasimhan, D. Proud, P. Kumar, B. Radharapu *et al.*, "ShieldGemma: Generative AI content moderation based on gemma," *arXiv preprint arXiv:2407.21772*, Aug. 2024.

[94] L. Li, B. Dong, R. Wang, X. Hu, W. Zuo, D. Lin, Y. Qiao, and J. Shao, "SALAD-Bench: A hierarchical and comprehensive safety benchmark for large language models," in *Findings Assoc. Comput. Linguist. (ACL)*, Aug. 2024, pp. 3923–3954.

[95] B. Vidgen, A. Agrawal, A. M. Ahmed, V. Akinwande, N. Al-Nuaimi, N. Alfaraj, E. Alhajjar, L. Aroyo, T. Bavalatti, and M. Bartolo, "Introducing v0.5 of the AI safety benchmark from MLCommons," *arXiv preprint arXiv:2404.12241*, May 2024.

[96] E. Bassani and I. Sanchez, "GuardBench: A large-scale benchmark for guardrail models," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Nov. 2024, pp. 18 393–18 409.

[97] OpenAI, "Moderation - OpenAI API," https://platform.openai.com.

[98] Meta-llama, "meta-llama/Meta-Llama-Guard-2-8B · Hugging Face," https://huggingface.co/meta-llama/Meta-Llama-Guard-2-8B, Dec. 2024.

[99] "mina-taraghi/PEFT_safety_fairness," Oct. 2025. [Online]. Available: https://github.com/mina-taraghi/PEFT_Safety_Fairness

[100] J. Shin, H. Song, H. Lee, S. Jeong, and J. Park, "Ask LLMs directly, "what shapes your bias?": Measuring social bias in large language models," in *Findings Assoc. Comput. Linguist. (ACL)*, Aug. 2024, pp. 16 122–16 143.

[101] J. Jin, J. Kim, N. Lee, H. Yoo, A. Oh, and H. Lee, "KoBBQ: Korean Bias Benchmark for Question Answering," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 507–524, 2024.

[102] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley Publishing Company, 1977.

[103] R. F. Woolson, "Wilcoxon Signed-Rank Test," in *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd, 2005.

[104] M. Friedman, "A Comparison of Alternative Tests of Significance for the Problem of m Rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940, publisher: Institute of Mathematical Statistics.

[105] M. R. Sheldon, M. J. Fillyaw, and W. D. Thompson, "The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs," *Physiotherapy Research International*, vol. 1, no. 4, pp. 221–228, 1996.

[106] R. A. Armstrong, "When to use the Bonferroni correction," *Ophthalmic and Physiological Optics*, vol. 34, no. 5, pp. 502–508, 2014.

[107] O. J. Dunn, "Multiple Comparisons among Means," *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52–64, Mar. 1961.

[108] S. Sawilowsky, "New Effect Size Rules of Thumb," *Journal of Modern Applied Statistical Methods*, vol. 8, no. 2, Nov. 2009.

[109] J. Chen, X. Wang, Z. Yao, Y. Bai, L. Hou, and J. Li, "Finding safety neurons in large language models," *arXiv preprint arXiv:2406.14144*, Jun. 2024.

[110] Y. Ren, T. Xiao, M. Shavlovsky, L. Ying, and H. Rahmanian, "COS-DPO: conditioned one-shot multi-objective fine-tuning framework," in *Proc. Conf. Uncertainty Artif. Intell. (UAI)*, ser. Proc. Mach. Learn. Res., vol. 286. PMLR, Aug. 2025, pp. 3525–3551.

[111] C. Li, H. Zhang, Y. Xu, H. Xue, X. Ao, and Q. He, "Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models," in *Proc. 63rd Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Jul. 2025, pp. 11 214–11 232.

[112] S. Mukherjee, A. Lalitha, S. Sengupta, A. Deshmukh, and B. Kveton, "Multi-objective alignment of large language models through hypervolume maximization," *arXiv preprint arXiv:2412.05469*, Dec. 2024.

[113] J. Dai, X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang, "Safe RLHF: Safe reinforcement learning from human feedback," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2024.

[114] S. Y. Peng, P.-Y. Chen, M. Hull, and D. H. Chau, "Navigating the safety landscape: Measuring risks in fine-tuning large language models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, Dec. 2024, pp. 95 692–95 715.

[115] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.