Efficiency at What Cost? Safety and Fairness in Parameter-Efficient Fine-Tuning of LLMs

Anonymous Authors

***Abstract*—Organizations are increasingly adopting and adapting large language models 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, 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 large language models 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; and risk concentrates in age and race/ethnicity. We translate these results into guidance: track category-level metrics, prefer conservative adapter settings for safety-critical use, and re-audit after tuning. The impact is practical: safer deployments, clearer compliance paths, and efficiency and environmental benefits retained.**

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

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.

Manuscript submitted October 10, 2025.

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 I 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 [1]:
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

$$\begin{aligned} Bias_{\text{DIS}} = & \\ & \frac{2n_{\text{biased\_ans}}}{n_{\text{biased\_ans}} + n_{\text{other\_biased\_ans}}} - \frac{n_{\text{biased\_ans}} + n_{\text{other\_biased\_ans}}}{n_{\text{biased\_ans}} + n_{\text{other\_biased\_ans}}} \\ & = \frac{n_{\text{biased\_ans}} - n_{\text{other\_biased\_ans}}}{n_{\text{biased\_ans}} + n_{\text{other\_biased\_ans}}} \end{aligned} \tag{1}$$

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 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.

$$\begin{aligned} n_{\text{correct\_SOCIAL}} + n_{\text{incorrect\_SOCIAL}} = \\ n_{\text{correct\_OTHER}} + n_{\text{incorrect\_OTHER}} = n \end{aligned} \tag{2}$$

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

$$\begin{aligned} Bias_{\text{DIS}} = & \\ & \frac{n_{\text{correct\_SOCIAL}} + \frac{n}{2} - n_{\text{correct\_OTHER}} - (n_{\text{correct\_OTHER}} + \frac{n}{2} - n_{\text{correct\_SOCIAL}})}{2n} \\ & = \frac{n_{\text{correct\_SOCIAL}} - n_{\text{correct\_OTHER}}}{n} \end{aligned} \tag{3}$$

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 $k = 1.5$ [2]. A Shapiro–Wilk test [3] 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.

TABLE I: 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

- 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 [4], 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 [5], [6]. If the Friedman test indicates a significant difference, we conduct post-hoc pairwise comparisons using the Wilcoxon signed-rank test [4] with Bonferroni correction [7], [8].

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 [9] 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 [3] confirmed that no dataset met the normality assumption. Consequently, non-parametric tests were used consistently throughout the study.

APPENDIX D
STATISTICAL TEST TABLES

REFERENCES

[1] 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.

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

[3] 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.

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

[5] 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.

[6] 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.

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

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

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

TABLE II: 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 III: 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 IV: 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 V: 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 VI: 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 VII: 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 VIII: Results of the statistical tests for fairness categories 5 and 6

Factor	Comparison	5. Physical Appearance				6. Race/Ethnicity			
		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		0.6519	-	**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 IX: 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 X: 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	**-**	**-**	**0.694**
	Qwen - base	-	**-**	-	-
	Gemma - base	0.8799	0.595	**0.6358**	1.0424