ON FAIRNESS OF TASK ARITHMETIC: THE ROLE OF TASK VECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Model editing techniques, particularly task arithmetic with task vectors, offer an efficient alternative to full fine-tuning by enabling direct parameter updates through simple arithmetic operations. While this approach promises substantial computational savings, its impact on fairness has remained largely unexplored—despite growing concern over biased outcomes in high-stakes applications such as hate speech detection. In this work, we present the first systematic study of fairness in task arithmetic, benchmarking it against full fine-tuning (FFT) and Low-Rank Adaptation (LoRA). We evaluate across multiple language models and datasets using standard group fairness metrics, including Demographic Parity and Equalized Odds. Our analysis shows that task vectors can be tuned to achieve competitive accuracy while reducing disparities, and that merging subgroup-specific task vectors provides a practical mechanism for steering fairness outcomes. We further provide a theoretical bound linking task-vector scaling to fairness metrics, offering insight into the observed trade-offs. Together, these findings establish task arithmetic not only as a cost-efficient editing method but also as a fairness-aware alternative to existing adaptation techniques, laying the groundwork for responsible deployment of large language models. Our code is available at `https://anonymous.4open.science/status/fairness_task_vector-4F2F`

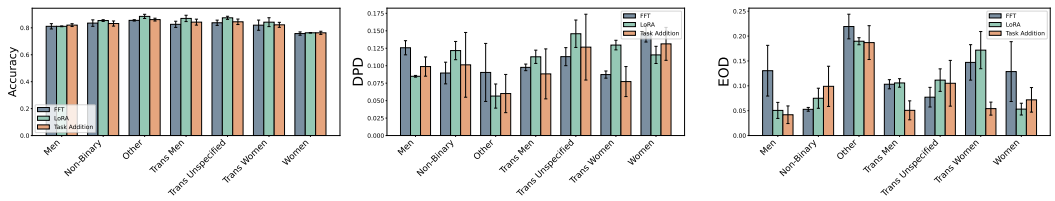
1 INTRODUCTION

As large language models (LLMs) are deployed across increasingly diverse applications, efficient techniques for adapting them to specific tasks have become essential. While model distillation and compact architectures reduce computational demands (Sanh et al., 2019b; Jiao et al., 2020; Turc et al., 2020; Abdin et al., 2024), task-specific fine-tuning (FFT) remains resource-intensive. This has motivated parameter-efficient fine-tuning (PEFT) methods such as adapters and Low-Rank Adaptation (LoRA) (Houlsby et al., 2019; Hu et al., 2022; Ben Zaken et al., 2022; Dettmers et al., 2023), which update only a small fraction of parameters.

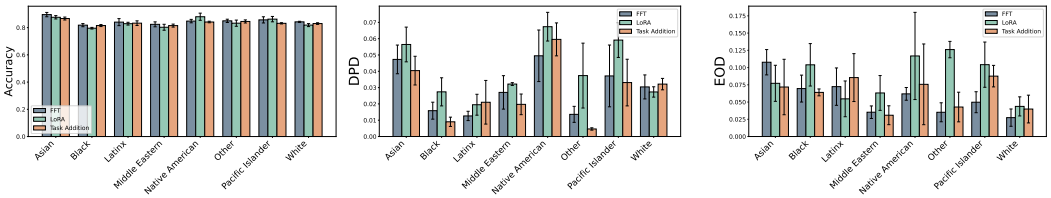
LoRA exemplifies this trade-off: it preserves most of the pretrained weights while reducing training costs. However, PEFT methods do not resolve deeper concerns. In high-stakes domains with imbalanced data—such as toxicity or hate-speech detection—they can maintain or even amplify biases (Ding et al., 2024b; Sap et al., 2019), raising concerns about fairness.

A promising alternative is task arithmetic with task vectors (Ilharco et al., 2023; Zhang et al., 2024; Yoshida et al., 2025; Yoshikawa et al., 2025). A task vector is defined as the difference between a fine-tuned model and its base counterpart. By adding, subtracting, or scaling such vectors, one can directly edit model behavior without gradient-based retraining. This approach offers (i) computational efficiency, (ii) fine-grained control over transferred capabilities, and (iii) enhanced interpretability when task vectors are associated with specific subgroups (Cerrato et al., 2025). Yet its fairness implications remain poorly understood. For example, enhancing performance on one demographic subgroup may inadvertently degrade outcomes for another, and the trade-offs with standard metrics such as Demographic Parity (DPD) or Equalized Odds (EOD) remain unclear.

To address this gap, we conduct the first systematic study of fairness in task arithmetic. We compare task-vector editing against both FFT and LoRA, and we further investigate whether injecting



(a) Gender-based demographic subgroups



(b) Race-based demographic subgroups

Figure 1: LoRA and FFT vs. Task addition with the optimal coefficient for the training accuracy ($\lambda = 0.8$ for gender setting and $\lambda = 0.5$ for race setting) on group-wise accuracy, demographic parity difference (DPD, lower is fairer), and equalized odds difference (EOD, lower is fairer). Error bars denote the standard error across three seeds. Columns: group-wise accuracy, DPD, EOD. No consistent pattern emerges that task addition necessarily degrades subgroup fairness relative to LoRA or FFT subgroups show improvements or comparable results under task addition, while others show small declines.

subgroup-specific task vectors into an FFT model provides additional control over fairness outcomes. Our experiments focus on hate-speech detection with LLaMA-7B (Touvron et al., 2023), and we replicate on Civil Comments (Borkan et al., 2019) with DistilBERT and Qwen2.5-0.5B Qwen Team (2025), observing consistent fairness–utility trade-offs.

Our contributions are as follows:

- **Comprehensive evaluation**: We compare FFT, LoRA, task-vector editing, and a hybrid approach that injects task vectors into FFT, analyzing their impact on fairness metrics and predictive performance (Figure 1).
- **Fairness through scaling**: We show that adjusting task-vector coefficients can substantially improve fairness while maintaining accuracy (Figure 2).
- **Subgroup-sensitive editing**: We demonstrate that merging task vectors from underrepresented subgroups allows targeted fairness adjustments with negligible accuracy loss (Figures 3a, 3b, 4a).
- **Theoretical grounding**: We derive an upper bound linking task-vector scaling to DPD, providing a principled explanation for the observed fairness–accuracy trade-offs (Section 5.1 and Appendix B).

Through this analysis, we establish task arithmetic as not only a cost-efficient model editing technique but also a fairness-aware alternative to existing adaptation methods. Our findings lay the groundwork for extending task-vector approaches toward fair and responsible deployment of LLMs.

2 Preliminaries

In this section, we first provide an overview of the fundamental concept of task vectors and the procedure known as task arithmetic, which applies these vectors to edit model behavior. We then introduce methods for merging multiple task vectors into a single model.

Task arithmetic. A task vector is defined as the difference in model parameters between a fine-tuned model on a given task and the original base model. Formally, if θ_{base} are the pre-trained

weights and θ_{task} are the weights after fine-tuning on a task, then the task vector is: $\Delta\theta = \theta_{\text{task}} - \theta_{\text{base}}$ (Ilharco et al., 2023).

This vector represents a direction in weight space such that moving the base model's weights by $\Delta\theta$ steers the model to perform well on that task. In other words, adding $\Delta\theta$ to θ_{base} yields a model with improved performance on the target task, without any additional training. Once computed, task vectors can be manipulated through simple arithmetic operations to edit model behavior directly in weight space (Ilharco et al., 2023; Ortiz-Jimenez et al., 2024). Key operations include:

> **Addition:** Given two task vectors $\Delta\theta_A$ and $\Delta\theta_B$ (for tasks A and B), their sum can be applied to the base model ($\theta_{\text{base}} + \Delta\theta_A + \Delta\theta_B$) to produce a model that exhibits improved performance on both tasks A and B (Ilharco et al., 2023). This task addition effectively combines knowledge from multiple tasks into one model.
>
> **Negation:** Using the negative of a task vector, $-\Delta\theta$, one can subtract a task's influence. For example, applying $\theta_{\text{base}} - \Delta\theta_A$ (or equivalently $\theta_{\text{base}} + (-\Delta\theta_A)$) yields a model with reduced performance on task A—effectively unlearning or forgetting it—while preserving other behaviors (Ilharco et al., 2023). This is useful for removing undesirable skills or biases.
>
> **Scalar scaling:** Multiplying a task vector by a scalar λ adjusts the strength of the edit. For example, using $\theta_{\text{base}} + \lambda\Delta\theta_A$ allows partial ($0 < \lambda < 1$) or amplified ($\lambda > 1$) application of a task's effect. This scaling provides fine-grained control over how strongly the task knowledge is injected into the model.

Merging task vectors. Since task vectors reside in a common weight space, they can be merged by simple addition with tunable scaling. Formally, given a base model θ_0 and task vectors $\Delta\theta_i$, one can construct a merged model as:

$$\theta_{\text{merged}} = \theta_0 + \sum_i \lambda_i \, \Delta\theta_i \,, \tag{1}$$

where each coefficient λ_i controls the influence of task i. Varying λ_i thus directly modulates how strongly the i-th task's knowledge is injected, allowing fine-grained blending of capabilities. Indeed, adding multiple task vectors with $\lambda_i = 1$ endows a model with all those capabilities simultaneously (Ilharco et al., 2023). Optimizing the λ_i values (i.e., learning an anisotropic scaling for each vector) further improves the composition by balancing contributions and reducing interference between tasks (Zhang et al., 2024).

3 Related Work

Task arithmetic: efficiency and interpretability. Task vectors offer a computationally efficient framework for editing and analyzing model behavior. Once a task vector is computed—namely, the weight difference between a base model and its fine-tuned variant (Ilharco et al., 2023; Zhang et al., 2024; Yoshida et al., 2025)—no additional training data or retraining is required to transfer or remove task-specific capabilities. By treating each fine-tuning update as a direction in weight space, practitioners can combine or negate these updates through simple addition or subtraction (Ilharco et al., 2023). This modularity not only reduces computational overhead but also enhances interpretability by isolating the contribution of each task.

Beyond modularity, task arithmetic can reveal valuable information about how and where a model adapts to new tasks. Li et al. (2024) show a near-linear relationship between data size and the norm of a task vector, suggesting that over-represented tasks can dominate weight space shifts in multi-task settings. In addition, the orientation of task vectors can indicate synergies or conflicts among tasks (Li et al., 2025), and decomposing these vectors by layer can pinpoint which parts of the model are most affected (Zhang et al., 2024; Gargiulo et al., 2025). Hence, task vectors offer a promising lens for diagnosing training dynamics and identifying potential biases.

Group fairness metrics in binary text classification. There are two canonical criteria that capture complementary harms and enable comparability with prior PEFT–fairness work: Demographic Parity Difference (DPD) for allocation disparity (gaps in positive selection rates across groups) and

Equalized Odds Difference (EOD) for error-rate disparity (gaps in TPR and FPR) (Hardt et al., 2016; Feldman et al., 2015; Kennedy et al., 2020a). Both are standard in auditing toolkits and empirical studies (Bellamy et al., 2018; Fairlearn contributors, 2025) and are the prevailing baseline in the literature we build upon (Fraenkel, 2020; Pitoura, 2019; Quan et al., 2023). For score-based classifiers with group-agnostic thresholds and (approximate) calibration, many group fairness desiderata reduce to constraints on (i) selection rates and (ii) class-conditional error rates (Hardt et al., 2016; Kleinberg et al., 2017). DPD targets (i); EOD targets (ii). Impossibility results imply that, when base rates differ, one cannot satisfy calibration, selection parity, and error parity simultaneously; reporting DPD and EOD therefore exposes the relevant trade-off frontier without auxiliary counterfactual assumptions (Kleinberg et al., 2017). Accuracy-parity is not, by itself, a principled fairness guarantee; it's generally tracked for monitoring (Barocas et al., 2023). Formal definitions in Appx. A; implementation in §4.1.

FFT and LoRA under fairness constraints. Parameter-efficient methods such as LoRA (Hu et al., 2022) address computational bottlenecks by training only a small set of parameters, yet they do not inherently solve fairness issues. In some cases, LoRA yields comparable subgroup performance to full fine-tuning (Ding et al., 2024b), while in others, it fails to mitigate toxic behaviors or biases (Das et al., 2024). The variance in outcomes depends on factors like the rank of the LoRA matrices, the base model's quality, and the distribution of training data (Das et al., 2024).

Merging tasks and fairness composition. Despite the potential efficiency gains and interpretability offered by task arithmetic, the merging of task vectors for multiple groups can trigger new challenges. For instance, simply summing vectors may lead to "negative transfer," where updates beneficial to one subgroup degrade performance for another (Ding et al., 2024a; Yu et al., 2020). In highly imbalanced settings, merging models through supervised fine-tuning can also disproportionately favor majority groups while disadvantaging minorities (Cross et al., 2024).

Additionally, prior work shows that fairness guarantees often do not compose: even if individual components satisfy group or individual fairness in isolation, composing them can break those guarantees (Dwork & Ilvento, 2018). This motivates our focus on post-hoc task-arithmetic edits: adding or scaling subgroup task vectors can be viewed as composing behaviors, and interactions among subgroup-specific task vectors can produce unpredictable shifts in metrics like Demographic Parity and Equalized Odds (Gohar et al., 2023). Consequently, identifying effective ways to adjust task vectors—such as through scalar scaling—remains a key step toward fairness-aware model editing. This work aims to fill that gap by systematically evaluating how these operations influence both fairness and overall model accuracy.

In parallel, multi-task fairness methods such as Multi-Task-Aware Fairness (Wang et al., 2021), Learning-to-Teach Fairness-Aware MTL (Roy & Ntoutsi, 2022), and FairBranch (Roy et al., 2024) manage fairness–accuracy trade-offs during training. Our study complements these by asking: when we edit models *after training* via task vectors, can simple controls (e.g., λ-scaling) recover fairer behavior without retraining?

4 Experimental Setup

4.1 Configuration

Building on the experimental framework established by Ding et al. (2024b), we adopted their evaluation and experimental procedure to assess the fairness implications of LoRA in comparison to FFT. In our work, we extend this analysis by focusing on how task arithmetic compares to both LoRA and FFT in terms of fairness and performance. The detailed experimental setup is provided in Appendix C.

Gender Subgroups		**Race Subgroups**	
Men	817	Asian	311
Non-binary	114	Black	1,007
Trans men	178	Latinx	368
Trans unspecified	173	Native American	153
Trans women	148	Middle Eastern	493
Women	2,057	Pacific Islander	138
Other	59	White	580
		Other	302
Total	**3,546**	**Total**	**3,352**

Table 1: Berkeley D-Lab Hate Speech data statistics in the gender and race subgroups.

Datasets. We use a modified version of the *Berkeley D-Lab Hate Speech* dataset originally introduced by Kennedy et al. (2020a) and adapted by Ding et al. (2024b), the research we are building upon. Our dataset contains a total of 6,898 tweet-sized

text snippets annotated for hate speech and categorized by sensitive attributes: *Race* and *Gender*, each further divided into fine-grained subgroups (e.g., *Women*, *Non-binary*, *Men* within *Gender*) as shown in Table 1. We frame hate speech detection as a binary classification task: given a text snippet, the model predicts whether it constitutes hate speech (e.g., hatespeech in the Gender subset may target Non-binary or Trans Women). Each example includes both the hate speech label and one or more protected attribute annotations (e.g., *gender* = woman, *race* = Asian). These are used to assess subgroup-level performance and fairness metrics. This setting supports rigorous fairness analysis due to its rich attribute annotations and real-world relevance (Kennedy et al., 2020a). To test generalization beyond hate speech, we apply our methods to the *Civil Comments* dataset (Borkan et al., 2019), a large-scale toxicity corpus with sensitive-attribute labels. We treat toxicity as binary with a 0.5 threshold; comments above this are positive "flagged". Fairness is evaluated across Gender and Race subgroups.

Evaluation metrics and fairness scope. Since we cast hate-speech and toxicity detection as binary classification, for each protected attribute (e.g., Gender, Race/Ethnicity), we compute subgroup-resolved metrics: *DPD* measures selection-rate disparity as the maximum absolute gap in flag rates across subgroups. *EOD* measures error-rate disparity by requiring both true-positive and false-positive rates to be comparable. *Accuracy-parity gap* is the maximum absolute difference in accuracy across subgroup pairs and serves as a stability indicator. We report per-subgroup values along with macro-averages and worst-group results. These choices mirror established practice and enable direct comparison to prior PEFT–fairness evaluations discussed in §3. Formal definitions and computation details appear in Appendix A.

4.2 Protocol

We evaluate our methods on a main generative base model, LLaMA2-7B (Touvron et al., 2023)[1], and two compact baselines for CivilComments toxicity, DistilBERT (Sanh et al., 2019a)[2] and Qwen2.5-0.5B (Qwen Team, 2025)[3]. Our fairness evaluations focus on two sensitive attributes: gender and race across both datasets, computing subgroup accuracy, DPD, and EOD. These selections span distinct architectures (decoder-only vs. encoder-only), parameter scales, tasks, and label taxonomies.

For FFT, the pretrained model was fine-tuned on the combined training data from all subgroups of the target attribute (gender or race). Evaluation was then performed on the test data from each corresponding subgroup, enabling fine-grained assessment of both performance and fairness. For LoRA, we followed the same training and evaluation procedure as FFT. The rank of LoRA's adaptation modules was set to 8, following Ding et al. (2024b).

For task arithmetic, we applied a compositional fine-tuning approach. The training data was partitioned by subgroup (gender or race), and FFT was applied separately to each subgroup's data to produce fine-tuned models θ_i. From these, we computed task vectors $\Delta\theta_i$ relative to the base model. These vectors were then merged using the approach described in Eq. (1), with a single, uniform scaling coefficient λ applied to all vectors. λ served as the sole hyperparameter in the merging process and was tuned on the training data. The evaluation metrics were computed in the same manner as for FFT and LoRA.

Task vector coefficient adjustment. Building on the task vector merging framework introduced in Eq. (1), we further explore the impact of the scaling coefficient λ on fairness outcomes. Specifically, we vary the uniform task vector coefficient λ across a broad range (from 0.0 to 1.0 with 0.1 intervals) and evaluate how this adjustment influences subgroup-level fairness metrics, including accuracy, DPD, and EOD.

Impact of worst-performing subgroup task vectors on fairness and performance. To investigate whether incorporating task vectors from underperforming subgroups can improve fairness

[1]LLaMA 2 is licensed under the LLAMA 2 Community License, Copyright (c) Meta Platforms, Inc. All Rights Reserved. See `https://ai.meta.com/llama/license`.

[2]DistilBERT is released under the Apache 2.0 License. See `https://github.com/huggingface/transformers/blob/main/LICENSE`.

[3]See the Qwen2.5-0.5B model card and license details at `https://huggingface.co/Qwen/Qwen2.5-0.5B`.

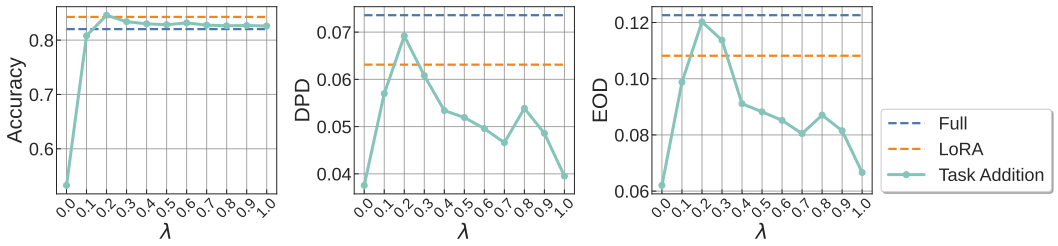


Figure 2: Varying the task arithmetic coefficient λ and comparing against FFT (purple dashed) and LoRA (orange dashed) for macro-averaged accuracy (left), demographic parity difference (DPD, center), and equalized odds difference (EOD, right) on the **gender** subset. Higher accuracy is better; lower DPD/EOD indicate improved group fairness. For $\lambda \gtrsim 0.3$, task addition maintains *competitive accuracy* while *typically lowering* DPD/EOD relative to both baselines.

without sacrificing overall performance, we first identified the lowest-performing subgroups within each attribute based on the average of DPD and EOD under the FFT setting. We excluded the "others" group from this analysis as it does not reflect the characteristics of any specific subgroup. This selection was informed by both our experimental results and those reported in Ding et al. (2024b), which showed consistent patterns. For gender, the worst-performing subgroups were men and women; for race, they were Asian and Native American. We constructed a new model variant by injecting a worst-performing subgroup task vector worst-performing subgroup task vector into the base fine-tuned model:

$$\theta_{\text{new}} = \theta_{\text{SFT}} + \lambda(\theta_{\text{worst-performing subgroup}} - \theta_0),$$

where λ controls the strength of the task vector injection. We varied λ from 0.0 to 1.0 at 0.2 intervals to analyze the effect of this targeted addition on subgroup fairness metrics and overall accuracy.

5 Results

5.1 Theoretical intuition.

We complement our empirical findings with an analytical upper bound that links task-vector scaling to fairness metrics.

Theorem (informal). *Consider the merged model $\theta(\lambda) = \theta_0 + \sum_g \lambda\, \Delta\theta_g$, where $\Delta\theta_g$ denotes the task vector for subgroup g. Then the demographic parity difference (DPD) satisfies*

$$\text{DPD}(\theta(\lambda)) \;\le\; 2L \sum_g \big|\lambda - 1\big|\, \|\Delta\theta_g\|_2, \quad \textit{for a Lipschitz constant } L.$$

Intuitively, deviations of the scaling coefficient λ from the balanced setting ($\lambda = 1$) enlarge disparities in proportion to the norms of subgroup task vectors. This explains why fairness disparities shrink as $\lambda \to 1$, consistent with the empirical trends observed in Figure 2. A full derivation and tighter constants are provided in Appendix B.

5.2 Empirical results overview.

Figures 1a and 1b compare FFT, LoRA, and task addition across gender and race subgroups for hate speech detection on LLaMA-2. For task addition, we selected $\lambda = 0.8$ for gender, $\lambda = 0.5$ for race, as it achieved the highest average training accuracy across three random seeds within the tested range $\lambda \in [0.0, 1.0]$. These visualizations provide a direct comparison of subgroup-wise model behavior. From the subgroup-level bar plots in Figure 1, we observe that accuracy remains consistently high and comparable across all three adaptation methods, regardless of subgroup. On *Civil Comments*, on both DistilBERT and Qwen-2.5, Task Addition reduces group disparities while keeping accuracy competitive. (see Appendix. E and Table 4 for full CIs/results).

We also observe that, relative to FFT, task addition improves fairness in five of seven gender subgroups and in three of eight race subgroups, with no single method dominating across all groups.

The effect in fairness being subgroup-dependent, motivates treating λ as a deliberate tuning knob and inspecting subgroup behavior explicitly. As shown in Appendix B.2, theoretically, task addition realizes a group-weighted ERM in the linearized model. Concretely, $\theta(\lambda) = \theta_0 + \sum_g \lambda_g \Delta\theta_g$ coincides with the one-step minimizer of a first-order surrogate where subgroup g is re-weighted by λ_g. This explains the smooth fairness–utility frontier traced by sweeping λ, and Theorem 5.1 predicts larger parity swings for groups with larger $\|\Delta\theta_g\|_2$. The observed curves in Fig. 2 align with those predictions without further assumptions.

Taken together, the empirical trends and their first-order mechanism align with prior literature: our macro-averaged accuracy, DPD, and EOD findings for FFT and LoRA are consistent with (Ding et al., 2024b). Moreover, the reductions perspective of Agarwal et al. (2018) and the equalized-odds criterion of Hardt et al. (2016) anticipate precisely the trade-off behavior we document, reinforcing the robustness of our evaluation and interpretation.

5.3 Controlling accuracy and fairness metrics through lambda.

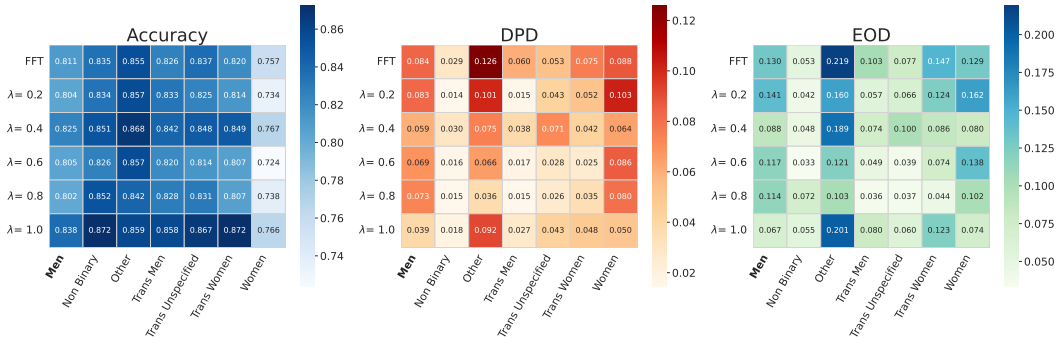
Figure 2 illustrates the overall performance of FFT, LoRA, and task arithmetic as the scaling coefficients for task addition vary from 0.0 to 1.0. We observe how varying the task-arithmetic coefficient λ impacts macro-averaged accuracy (left), demographic parity difference (DPD, center), and equalized odds difference (EOD, right) on a gender subset of the data. As λ increases from 0.0 to 0.2, we observe a peak in accuracy, but this configuration yields higher DPD and EOD, indicating reduced fairness. Beyond $\lambda = 0.3$, accuracy remains competitive compared to FFT and LoRA, while both DPD and EOD progressively decline, suggesting that fairness improves without severely compromising performance. Notably, these task addition curves stay consistently lower than FFT and LoRA in terms of DPD and EOD at higher λ values. Overall, this ablation could indicate that tuning λ provides a practical mechanism for balancing accuracy and fairness objectives, offering guidelines for practitioners who wish to fine-tune fairness outcomes while maintaining strong predictive performance.

5.4 Subgroup-targeted vectors: gains with trade-offs

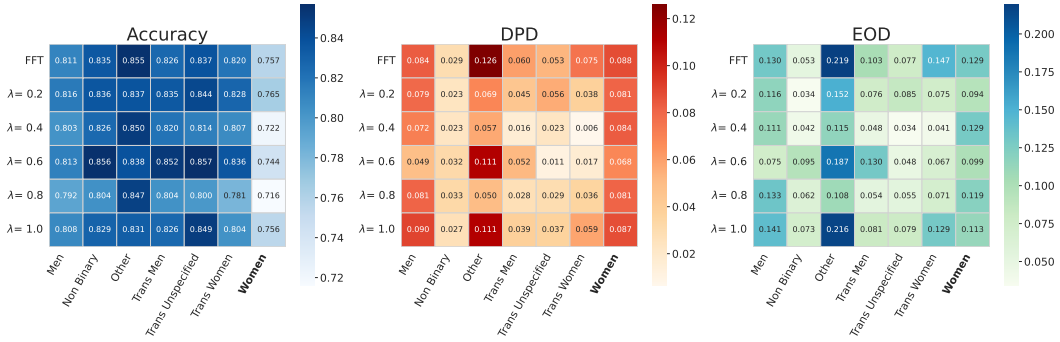
To further analyze the effects of subgroup-specific task composition, Figure 3a–3b illustrate heatmaps where the y-axis lists each method or configuration under evaluation: FFT as baseline, followed by task arithmetic with varying scaling coefficients (0.0 to 1.0 with 0.2 intervals). The x-axis represents the subgroups— (e.g., Women, Trans, etc. for Gender). Each cell shows the corresponding performance metric (e.g., macro-averaged accuracy, DPD, or EOD for a given method on a specific subgroup. For these experiments, we added the task vector of the worst-performing subgroups (Women and Men for the gender dataset subset, and Asian, and Native American for the race dataset subset) to the FFT model, as explained earlier.

We generally observe that increasing the scaling coefficient λ tends to improve overall accuracy, consistent with the trends observed in Figure 2. However, effects are not uniform across all subgroups. In the gender-based plots, for example, the Asian subgroup consistently achieves the highest accuracy and lowest DPD/EOD—highlighting a recurring tradeoff where performance gains for one group may exacerbate disparities for others. When the Women task vector is added (Figure 3b), accuracy improves for the Trans Women subgroups. However, fairness metrics for subgroups such as Men tend to worsen as the scaling coefficient λ increases.

In Figure 3a, injecting the Men task vector improves performance for some subgroups, yet Women consistently show lower accuracy and do not see consistent fairness improvements at higher λ. Some groups (e.g., Other, Trans Men, Trans Women) begin with relatively poor fairness under FFT and show partial improvements with task vector addition. Still, these improvements are not universal—for example, the Other subgroup often retains high EOD values regardless of λ. Likewise, Native American accuracy remains mostly unchanged across λ, while fairness metrics can deteriorate when injecting task vectors for other groups. To visualize these results in more detail, Figure 4a shows macro-averaged accuracy, DPD, and EOD for the Men task vector added to the FFT model. The plots illustrate how varying the scaling coefficient λ impacts overall performance and fairness, highlighting the effects of subgroup-specific task injection. We can observe in Figure 4a that injecting the Men task vector into the FFT model results in a slight accuracy gain and a clear monotonic



(a) When **Men** task vector added to the FFT model on the **gender** subset.



(b) When **Women** task vector added to the FFT model on the **gender** subset.

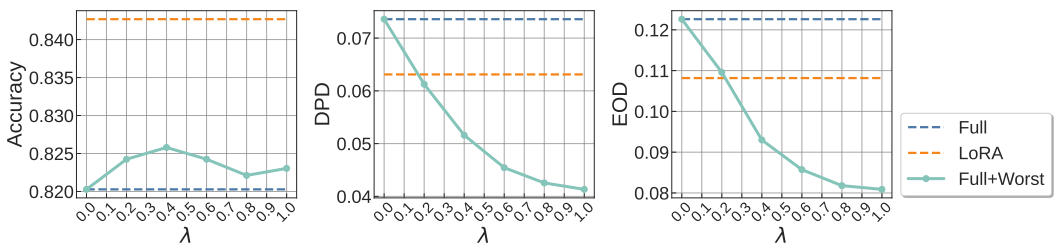
Figure 3: Heatmaps of Accuracy (left), DPD (center), and EOD (right) for gender (top) and race (bottom) subgroups under the baseline FFT model ($\lambda = 0.0$) and with increasing λ values from 0.2 to 1.0 in 0.2 increments. The task vector for Men was added on the gender subset (top), and the task vector for Women was added on the gender subset (bottom). Darker cells indicate higher values on each metric's scale; for DPD/EOD, lower values are better.

decrease in both DPD and EOD as λ increases—indicating a favorable and consistent improvement in fairness on the gender subset.

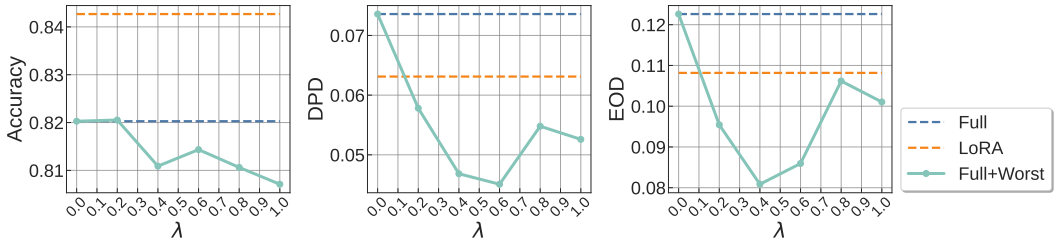
However, Figure 4b and the additional plots in Figures 10 and 11 in Appendix D.2 show more varied patterns as seen on Figures 3a and 3b. When injecting the Native American task vector (Figure 11), accuracy remains stable while fairness seems to decrease (increased DPD and EOD). Asian (Figure 10) shows the same behavior as injecting the Men task vector (Figure 4a), positive increase of fairness metrics as λ increases. These results show that injecting task vectors shifts fairness and performance in a group-specific manner, tracing a clear fairness–utility frontier. This heterogeneity is expected: per §5.2 and Theorem 5.1, sensitivity scales with $|\Delta\theta_g|_2$. Practically, task-vector merging thus offers a *subgroup-conditioned* control knob: identifying which $\Delta\theta_g$ help or hurt which groups provides a new actionable design consideration that SFT/LoRA do not expose, and that hasn't been explored in previous task arithmetic literature.

6 Conclusion and Limitations

Conclusion. In this study, we investigated the impact of a task arithmetic approach using task vectors on fairness, in comparison to conventional FFT and LoRA methods. We conducted detailed experiments to assess how the task addition affects prediction accuracy and fairness metrics, including the DPD and EOD across various subgroups. The results indicate that, with appropriate settings of the scalar coefficient λ, the task arithmetic method can improve DPD and EOD without significantly compromising overall model accuracy. Notably, using low to moderate values of the task vector coefficient effectively reduced prediction bias in minority groups compared to FFT and



(a) When **Men** task vector added to the FFT model on the **gender** subset.



(b) When **Women** task vector added to the FFT model on the **gender** subset.

Figure 4: Impact of injecting both the **Men** and **Women** subgroup task vectors into the FFT model on the gender data subset. The plot illustrates how scaling coefficient λ reduces DPD and EOD, outperforming the baseline FFT (blue dashed) and LoRA (orange dashed), with negligible impact on macro-averaged accuracy.

LoRA.We observe this pattern across two datasets (hate speech, toxicity) and three model families (LLaMA-2, DistilBERT, Qwen-2.5).

Furthermore, the task arithmetic framework allows for subgroup-specific evaluation and adjustment of model updates, enhancing interpretability—a key advantage of this method in the context of fairness. This interpretability facilitates the mitigation of excessive bias or adverse effects on particular groups, ultimately enabling more balanced model training.

Limitations. Despite these promising results, several challenges remain. The effectiveness of task arithmetic depends on dataset characteristics and subgroup distributions, necessitating further investigation into its generalizability across different tasks and domains. Moreover, future work should explore algorithms for automatically optimizing the scalar coefficient λ and for balancing trade-offs among multiple subgroups.

In summary, our study demonstrates that task arithmetic using task vectors offers a promising approach for controlling model fairness. Further experimental validation, application to diverse tasks, and developing trade-off optimization methods are essential for improving fairness in broader and more realistic deployment scenarios.

Reproducibility statement. We provide code [4], configs, and scripts to reproduce all experiments, including data preprocessing, training, and evaluation. All datasets and base models used are open-source/publicly available; we include scripts to fetch the exact versions. Exact hyperparameters, model identifiers, and implementation details are documented in the appendix, along with seeds and hardware/software specs. Results are reported over multiple runs, and we provide instructions to regenerate all figures and tables from logged outputs.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A

[4] Our code is available at `https://anonymous.4open.science/status/fairness_task_vector-4F2F`

highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 10–15 Jul 2018.

Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 120–129. PMLR, 09–15 Jun 2019.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint*, 2018. URL https://arxiv.org/abs/1810.01943.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1–9, 2022.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Civil comments: A new public text data set of annotated online comments. In *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom (NLP4IF): Censorship, Disinformation, and Propaganda*, 2019. URL https://arxiv.org/abs/1903.04561.

Mattia Cerrato, Marius Köppel, Alexander Segner, and Stefan Kramer. Fair interpretable learning via correction vectors. *arXiv preprint*, 2025.

James I. Cross, Wei Chuangpasomporn, and John A. Omoronyia. Bias in medical ai: Implications for clinical decision-making. *PLOS Digital Health*, 1(1):e0000561, Nov 2024. doi: 10.1371/journal.pdig.0000561. Published on 7 Nov 2024.

Saswat Das, Marco Romanelli, Cuong Tran, Zarreen Reza, Bhavya Kailkhura, and Ferdinando Fioretto. Low-rank finetuning for llms: A fairness perspective. *arXiv preprint*, 2024.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chuntao Ding, Zhichao Lu, Shanguang Wang, Ran Cheng, and Vishnu N. Boddeti. Mitigating task interference in multi-task learning via explicit task routing with non-learnable primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024a.

Zhoujie Ding, Ken Liu, Pura Peetathawatchai, Berivan Isik, and Sanmi Koyejo. On fairness of low-rank adaptation of large models. *arXiv preprint*, July 2024b. URL https://colmweb.org/. Published: 10 Jul 2024, Last Modified: 25 Aug 2024.

Cynthia Dwork and Christina Ilvento. Group fairness under composition. In *FAT/ML Workshop*, 2018. URL https://www.fatml.org/media/documents/group_fairness_under_composition.pdf. Workshop paper.

Fairlearn contributors. *Reductions — Fairlearn 0.13.0.dev0 documentation*, 2025. URL https://fairlearn.org/main/user_guide/mitigation/reductions.html. Accessed: March 16, 2025.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 259–268. ACM, 2015. doi: 10.1145/2783258.2783311. URL https://dl.acm.org/doi/10.1145/2783258.2783311.

Aaron Fraenkel. *Fairness and Algorithmic Decision Making*. UC San Diego, 2020. URL https://afraenkel.github.io/fairness-book/intro.html. Lecture Notes for UCSD course DSC 167.

Antonio Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. Task singular vectors: Reducing task interference in model merging. *arXiv preprint arXiv:2412.00831*, 2025. URL https://arxiv.org/abs/2412.00831. Version 3, 3 Jan 2025.

Usman Gohar, Nuno Ribeiro, and Harichandra Ramadurgam. Towards understanding fairness and its composition in ensemble machine learning. *arXiv preprint arXiv:2102.96452*, 2023. URL https://arxiv.org/pdf/2212.04593. Version 3, 25 Mar 2023.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pp. 3323–3331, 2016. URL https://arxiv.org/abs/1610.02413.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mohammad Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shawn Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. Published as a conference paper at ICLR 2023.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4163–4174, 2020.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. –, 2020a.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*, 2020b.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2017.43.

Hongkang Li, Yinhua Zhang, Shuai Zhang, Pin-Yu Chen, Sijia Liu, and Meng Wang. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/pdf?id=vRvVVb0NAz. Published as conference paper at ICLR 2025.

Mingxin Li, Zhijie Nie, Yanzhao Zhang, Dingkun Long, Richong Zhang, and Pengjun Xie. Improving general text embedding model: Tackling task conflict and data imbalance through model merging. *arXiv preprint arXiv:2410.15035v1*, 2024. URL https://arxiv.org/abs/2410.15035. 19 Oct 2024.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. *arXiv preprint*, 2024.

Evaggelia Pitoura. Towards diversity-aware, fair, and unbiased data management. In *ISIP 2019*, Heraklion, Greece, May 9 2019.

Tangkun Quan, Fei Zhu, Quan Liu, and Fanzhang Li. Learning fair representations for accuracy parity. *Engineering Applications of Artificial Intelligence*, 119:105819, 2023. doi: 10.1016/j.engappai.2023.105819.

Qwen Team. Qwen2.5-0.5b. https://huggingface.co/Qwen/Qwen2.5-0.5B, 2025. Model card; accessed 2025-09-19.

Arjun Roy and Eirini Ntoutsi. Learning to teach fairness-aware deep multi-task learning. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2022*, volume 13715 of *Lecture Notes in Computer Science*, pp. 710–726. Springer, 2022. doi: 10.1007/978-3-031-26387-3_43. URL https://arxiv.org/abs/2206.08403.

Arjun Roy, Christos Koutlis, Symeon Papadopoulos, and Eirini Ntoutsi. Fairbranch: Mitigating bias transfer in fair multi-task learning. In *Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024. doi: 10.1109/IJCNN60899.2024.10651221. URL https://arxiv.org/abs/2310.13746. Also available as arXiv:2310.13746.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2) at NeurIPS 2019*, 2019a. URL https://arxiv.org/abs/1910.01108.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019b.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Catalin Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3592–3601, 2020.

Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H. Chi. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1748–1757. ACM, 2021. doi: 10.1145/3447548.3467326. URL https://arxiv.org/abs/2106.02705.

Kotaro Yoshida, Yuji Naraki, Takafumi Horie, Ryosuke Yamaki, Ryotaro Shimizu, Yuki Saito, Julian McAuley, and Hiroki Naganuma. Mastering task arithmetic: τ jp as a key indicator for weight disentanglement. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1VwWi6zbxs.

Yuya Yoshikawa, Ryotaro Shimizu, Takahiro Kawashima, and Yuki Saito. Transferring visual explainability of self-explaining models through task arithmetic. *arXiv preprint arXiv:2507.04380*, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/file/3fe78a8acf5fda99de95303940a2420c-Paper.pdf.

Frederic Z. Zhang, Paul Albert, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. Knowledge composition using task vectors with learned anisotropic scaling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/. Main Conference Track.

APPENDIX

A FAIRNESS METRICS

A.1 DEMOGRAPHIC PARITY DIFFERENCE (DPD) (AGARWAL ET AL., 2018; 2019)

DPD measures how varied the model's rate of positive predictions is across attributes. This metric is calculated as follows:

$$M_{\text{DPD}} = \Big|\Pr[f(X) = 1 \mid A = 1] - \Pr[f(X) = 1 \mid A = 0]\Big|,$$

where A is the sensitive attributes, $f(X)$ is the prediction from the models, and X is the feature vector. The larger the DPD, the greater the difference in prediction outcomes across attributes, indicating greater unfairness in the model predictions.

A.2 EQUALIZED ODDS DIFFERENCE (EOD) (DING ET AL., 2024B)

EOD is a metric that measures whether the model exhibits similar predictive performance in terms of true and false positives, regardless of the attribute.

$$M_{\text{eod}} = \max\left\{M_{\text{TP}}, M_{\text{FP}}\right\}. \tag{2}$$

Here, letting Y denote the true label, M_{TP} and M_{FP} are defined as follows:

$$M_{\text{TP}} = \Big|\Pr[f(X) = 1 \mid Y = 1, A = 1] - \Pr[f(X) = 1 \mid Y = 1, A = 0]\Big|,$$

$$M_{\text{FP}} = \Big|\Pr[f(X) = 1 \mid Y = 0, A = 1] - \Pr[f(X) = 1 \mid Y = 0, A = 0]\Big|.$$

A.3 ACCURACY PARITY

Accuracy parity refers to the expectation that a classifier achieves comparable accuracy across different sensitive attribute groups. Formally, accuracy parity is satisfied when the probability of correct classification is equal across groups, i.e.,

$$\mathbb{E}(Y = \hat{Y} \mid S = 0) = \mathbb{E}(Y = \hat{Y} \mid S = 1), \tag{3}$$

This notion of fairness ensures that all subgroups receive equally reliable predictions, and is particularly relevant in applications where consistent model performance across demographics is critical. Unlike statistical parity or equal opportunity, accuracy parity focuses on equal overall correctness rather than specific error types or outcome rates (Quan et al., 2023).

We observed **high degree of accuracy parity** in both gender and race settings, as the accuracy differences between subgroups are negligible, indicating that the model performs consistently across all groups.

B DPD UPPER BOUND AND OPTIMAL TASK–VECTOR SCALING

B.1 NOTATION AND ASSUMPTIONS

A1 Smooth predictions. Soft scores p_θ satisfy $|p_\theta(x) - p_{\theta'}(x)| \le L \left\|\theta - \theta'\right\|_2 \; \forall x$.

A2 Task vectors. For each group $g \in \{1, \dots, G\}$, $\Delta\theta_g := \theta_0^{(g)} - \theta_0$ is obtained with the *same* learning rate and schedule.

A3 Scaling coefficients. Coefficients obey $\sum_{g=1}^{G} \lambda_g = G$.

A4 Symmetric data-generating process. The joint distribution satisfies $\mathcal{D} = \bigcup_g \mathcal{D}_g$ where all $\mathcal{D}_g$ share the same conditional distribution except for the sensitive attribute label.

The merged model is

$$\theta(\boldsymbol{\lambda}) \;=\; \theta_0 + \sum_{g=1}^{G} \lambda_g \, \Delta\theta_g.$$

Demographic Parity Difference (DPD) reads

$$\mathrm{DPD}(\theta) \;=\; \Big|\mathbb{E}_{\mathcal{D}_1}[p_\theta] - \mathbb{E}_{\mathcal{D}_0}[p_\theta]\Big|.$$

B.2 Task Addition and Weighted ERM

Lemma 1 (First-order link)**.** *Let $\ell(\theta; x)$ be the training loss. For any non-negative $\{\lambda_g\}$,*

$$\theta(\boldsymbol{\lambda}) \;\approx\; \arg\min_{\theta} \sum_{g=1}^{G} \lambda_g \, \mathbb{E}_{x \sim \mathcal{D}_g}\big[\ell(\theta_0; x) + \nabla_\theta \ell(\theta_0; x)^\top (\theta - \theta_0)\big].$$

That is, task addition gives the first-order *solution of a group-weighted ERM.*

Proof. Insert the linear Taylor expansion of ℓ at θ_0 and minimise the resulting quadratic form; the solution is exactly $\theta(\boldsymbol{\lambda})$. □

Implication. Deviation $|\lambda_g - 1|$ alters the group weights and therefore *directly pushes DPD upward*, as made explicit in Proposition 1 below.

B.3 DPD Upper Bound

Proposition 1 (DPD bound)**.** *Under Assumptions **A1–A4**,*

$$\mathrm{DPD}\big(\theta(\boldsymbol{\lambda})\big) \;\le\; 2L \sum_{g=1}^{G} |\lambda_g - 1| \, \|\Delta\theta_g\|_2.$$

Proof. Define $\bar{\theta} := \theta_0 + \frac{1}{G} \sum_g \Delta\theta_g$. Assumption **A4** gives $\mathrm{DPD}(\bar{\theta}) = 0$. Put $f(x) := p_{\theta(\boldsymbol{\lambda})}(x) - p_{\bar{\theta}}(x)$. Then $\mathrm{DPD}(\theta(\boldsymbol{\lambda})) = |\mathbb{E}_{\mathcal{D}_1}[f] - \mathbb{E}_{\mathcal{D}_0}[f]|$. Triangle and Jensen yield $\le 2L \, \|\theta(\boldsymbol{\lambda}) - \bar{\theta}\|_2$. Finally, $\theta(\boldsymbol{\lambda}) - \bar{\theta} = \sum_g (\lambda_g - 1)\Delta\theta_g$ and the triangle inequality give the stated bound. □

C Experimental details

C.1 Computational Resources and Software Environment

Hardware and Software: All experiments presented in this study were performed using computational resources equipped with two NVIDIA H100 GPUs. The experiments leveraged a GPU environment consisting of CUDA 12.1.0, cuDNN 9.0.0, and NCCL 2.20.5.

The experiments were conducted using Python 3.9.18, incorporating several essential Python libraries specifically optimized for deep learning tasks. The primary libraries included PyTorch (version 2.6.0), transformers (version 4.49.0), tokenizers (version 0.21.1), DeepSpeed (version 0.16.4), and Accelerate (version 1.5.2).

The training experiments utilized the DeepSpeed framework with the following key configurations: a gradient accumulation step of 4, optimizer offloaded to the CPU, zero redundancy optimizer at stage 2 (ZeRO-2), and mixed precision training employing FP16 and BF16 for enhanced performance and memory efficiency. All experiments were conducted with a total computational cost of approximately 30 GPU-hours.

Protocol: We fine-tuned models based on the Llama-7B (Touvron et al., 2023) architecture obtained via HuggingFace repositories. Each model was trained for 4 epochs, employing a cosine learning rate scheduler with a learning rate of 1×10^{-5}, a warm-up ratio of 0.01, and a weight decay of 0.001. Training utilized a per-device batch size of 2, with an effective batch size of 16 achieved through gradient accumulation. Reproducibility was ensured by setting a random seed of 13, 14, 15 across all experiments.

For Qwen2.5 experiments, models were trained for 2 epochs using a learning rate of 2×10^{-5}, a batch size of 16, and a sample fraction of 25% of the Civil Comments dataset. DistilBERT experiments utilized 2 epochs with a learning rate of 1×10^{-5}, a batch size of 16, and the full dataset (100% sample fraction). Both architectures employed a weight decay of 0.01 and evaluation/save strategies set to"epoch" with early stopping enabled.

For Low-Rank Adaptation (LoRA) experiments were conducted with a rank (lora_r) of 8, scaling factor (lora_alpha) of 16, and no dropout.

C.2 Dataset

We use the Berkeley D-Lab hatespeech detection dataset (Kennedy et al., 2020b) [5] for our experiments.

The dataset is divided into subgroups based on the following attributes: *Race or Ethnicity*, *Religion*, *National Origin or Citizenship Status*, *Gender Identity*, *Sexual Orientation*, *Age*, and *Disability Status*. In our study, we use some of these subgroups to evaluate fairness.

Following Das et al. (2024), we binarize the hate speech score associated with each review using a threshold of 0.5 to determine whether the review constitutes hate speech. When multiple annotations exist for the same instance, we obtain one human annotation to avoid duplication.

D Additional Results

Here, we present results focusing on diverse subgroups, which we could not include in the main paper due to space constraints.

D.1 Comparison of FFT, LoRA, and Task Arithmetic

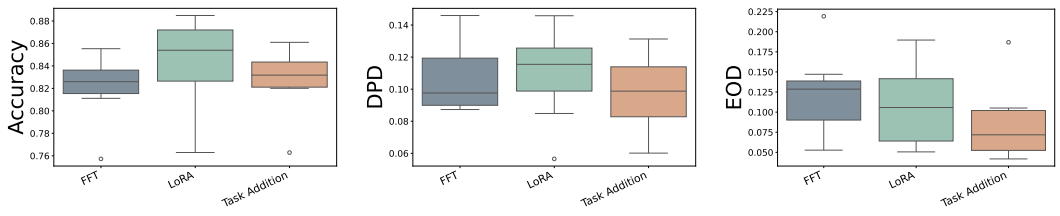


Figure 5: Boxplots of group-wise accuracy, demographic parity difference (DPD), and equalized odds difference (EOD) for —FFT, LoRA, and task addition with coefficient ($\lambda = 0.8$) —evaluated on the **gender** subset of the data. Higher accuracy is desirable, whereas lower DPD and EOD values indicate improved fairness. Boxplots show medians, interquartile ranges, and variability (with standard error across three seeds). While accuracy is similar across methods, Task Addition generally yields lower DPD and EOD medians than FFT and LoRA, suggesting a better balance between performance and fairness, though overlapping distributions imply these differences are not uniformly significant.

[5] https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech

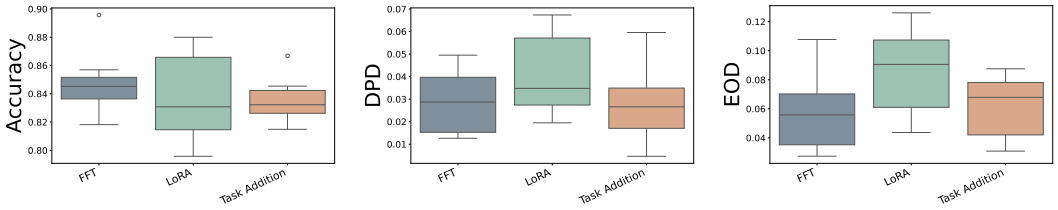


Figure 6: Boxplots of group-wise accuracy, demographic parity difference (DPD), and equalized odds difference (EOD) for —FFT, LoRA, and Task Addition with optimal coefficient ($\lambda = 0.5$) —evaluated on the **race** subset of the data. Higher accuracy is desirable, whereas lower DPD and EOD values indicate improved fairness. Boxplots show medians, interquartile ranges, and variability (with standard error across three seeds).

Figure 7 illustrates the overall performance of FFT, LoRA, and task arithmetic as the scaling for task arithmetic vary from 0.0 to 1.0. Trends observed reinforced results on the gender subset on Figure 2. Overall, λ provides a practical mechanism for balancing accuracy and fairness objectives, and similarly there is a peak at $\lambda = 0.2$ for highest accuracy, and higher DPD and EOD (less fairness).

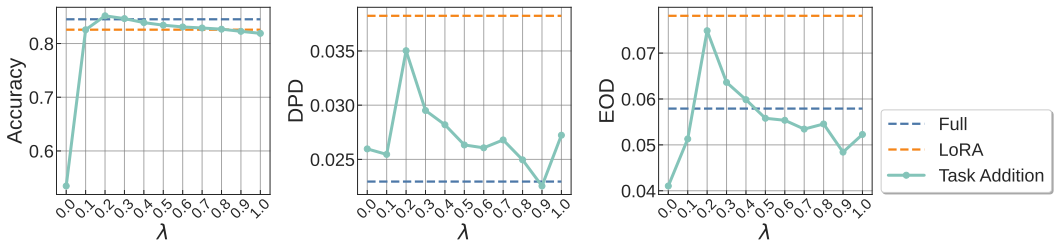


Figure 7: On a **race-focused** subset, we vary task arithmetic's coefficient λ and compare it against FFT (purple dashed) and LoRA (orange dashed). The plots show group-wise accuracy (left), demographic parity difference (DPD, center), and equalized odds difference (EOD, right). Higher accuracy is better, while lower DPD and EOD indicate improved fairness. As λ changes, task arithmetic remains competitive in accuracy and can reduce fairness gaps relative to the baselines.

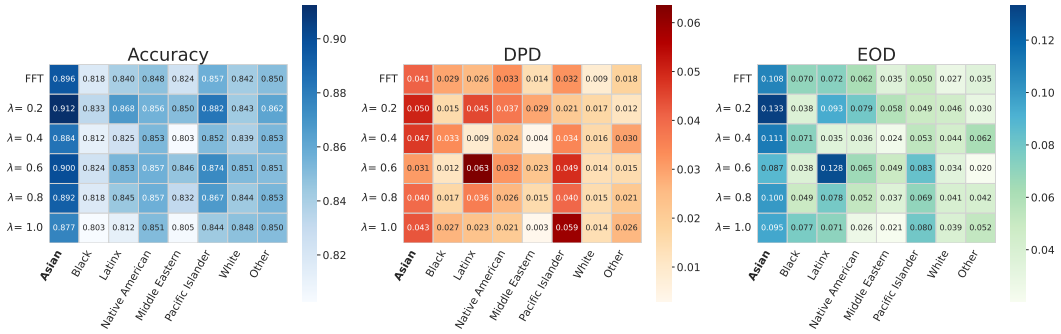


Figure 8: The task vector corresponding to **Asian** was added to the FFT model on the race data subset. Heatmap of Accuracy (left), DPD (center), and EOD (right) under the baseline (FFT) and increasing λ values (0.2 to 1.0). Darker cells indicate higher values in each metric's scale; for DPD/EOD, lower is better.

D.2 Subgroup-Specific Task Addition to FFT

We include additional heatmaps that visualize subgroup-wise performance across FFT and varying scaling coefficients for the FFT model injected with a worst-performing subgroup. These supplementary plots, which follow the same setup described earlier, are consistent with the trends observed in Figures 3a–3b.

In both gender and race subgroup experiments, increasing the scaling coefficient λ generally leads to improved macro-averaged accuracy. However, its impact on fairness metrics—DPD and EOD—is less predictable and varies across subgroups. For instance, some subgroups benefit from improved fairness as their corresponding task vectors are added, while others experience increased disparity, even if accuracy remains stable or improves.

This nuanced behavior reflects a broader pattern: gains in performance for certain subgroups can sometimes come at the expense of fairness for others. Injecting task vectors from worst-performing subgroups does not consistently reduce disparities and, in some cases, can amplify them.

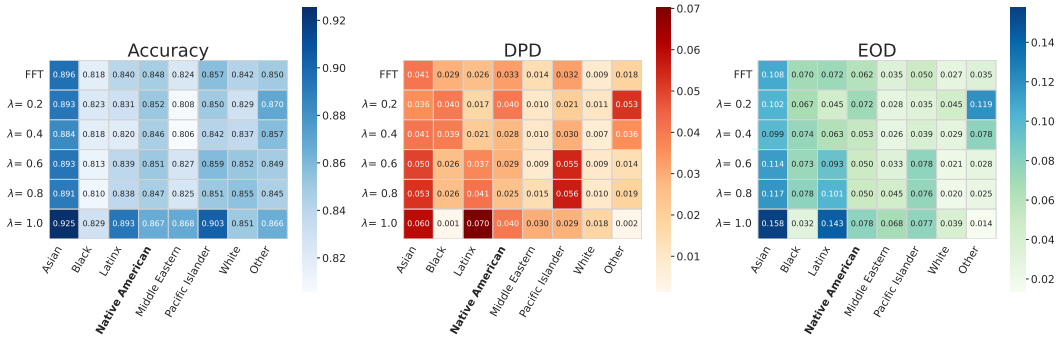


Figure 9: The task vector corresponding to **Native American** was added to the FFT model on the race data subset. Heatmap of Accuracy (left), DPD (center), and EOD (right) under the baseline (FFT) and increasing λ values (0.2 to 1.0). Darker cells indicate higher values in each metric's scale; for DPD/EOD, lower is better.

Figures 11–4b present additional results for the Full+Worst configuration, in which task vectors from the worst-performing subgroups (Native American, Asian, Men, and Women) are added to the FFT model. These plots show macro-averaged accuracy, DPD, and EOD as a function of the scaling coefficient λ.

Across these figures, we observe mixed effects: while accuracy generally remains stable or improves slightly, fairness outcomes vary by subgroup. In Figure 11, DPD and EOD worsen despite minimal accuracy changes. Meanwhile, Figure 4b reveals stable performance with minor fairness

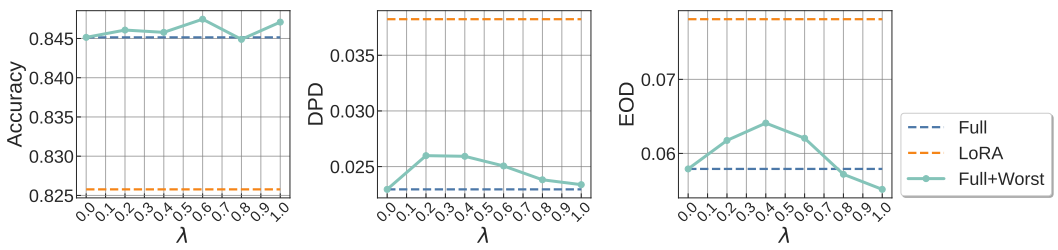


Figure 10: Effect of adding the **Asian** task vector to the FFT model on the **race** subset. Accuracy keeps competitive with increasing λ, and both DPD and EOD decrease consistently.

improvements, though gains are not consistent across metrics. These results further emphasize that task vector injection alone does not ensure universal fairness improvements and often introduces subgroup-specific trade-offs.

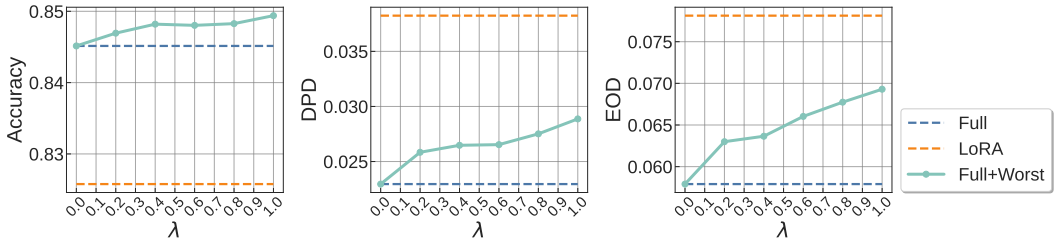


Figure 11: Results of injecting the **Native American** task vector into the FFT model. Accuracy shows minimal change across λ, while DPD and EOD increase (worsen fairness).

E Additional Experiments on Civil Comments

Protocol & uncertainty. Unless noted, we follow the LLaMA-2 setup (Section 4.2): SFT and LoRA (r=8) to obtain subgroup-specific models, compute task vectors w.r.t. the pretrained base, and merge with a uniform scalar λ. We sweep λ on the validation split (maximize overall accuracy) and evaluate on the test split. Uncertainty is 95% stratified bootstrap over the test set (2,000 resamples, preserving group $\times$ label frequencies). When multiple seeds are used, we pool predictions before resampling. For accuracy, we additionally report Wilson CIs when relevant.

At a glance. On Civil Comments with DistilBERT (67M), task addition maintains accuracy within $\sim$0.6–1.1pp of SFT/LoRA while reducing fairness gaps: for *gender*, DPD drops by $\approx$41–54% and EOD by $\approx$34–47%; for *race*, DPD drops by $\approx$41–58% and EOD by $\approx$58–73% (midpoint comparisons). These patterns align with LLaMA-2 on the Berkeley D-Lab dataset (Table 4). As a complementary cross-architecture check, Qwen-2.5-0.5B on *gender* exhibits the same qualitative λ-controlled trade-off, improving substantially over LoRA with competitive accuracy.

E.1 Civil Comments — Gender

Notes. Relative to LoRA, Qwen-2.5-0.5B task addition halves DPD/EOD ($\sim$54–56%) while regaining $\sim$3.3pp accuracy; relative to SFT, accuracy is lower and fairness is mixed (DPD comparable; EOD higher). DistilBERT shows consistent reductions in DPD/EOD with $\lesssim$1pp accuracy cost.

E.2 Civil Comments — Race

Discussion. Together with LLaMA-2 on Berkeley D-Lab (Table 4), these experiments indicate that the λ-controlled fairness–utility trade-off extends across architectures and datasets: task addition typically preserves accuracy within $\sim$1pp while materially reducing worst-case DPD/EOD.

Table 2: **Civil Comments (Gender).** Headline metrics (Accuracy ↑, worst-case DPD ↓, worst-case EOD ↓). Entries are 95% CIs from stratified bootstrap; point estimates marked with † will be replaced by CIs computed using the same protocol.

Model/Method	Accuracy	Worst-DPD	Worst-EOD
DistilBERT SFT	0.9457–0.9476	0.0887–0.1101	0.6157–0.6433
DistilBERT LoRA	0.9447–0.9453	0.0735–0.0812	0.5024–0.5084
DistilBERT Task Addition	**0.9395**†	**0.0454**†	**0.3358**†
Qwen-2.5-0.5B SFT[1]	0.884–0.886	0.093–0.119	0.060–0.084
Qwen-2.5-0.5B LoRA[1]	0.774–0.790	0.210–0.251	0.232–0.362
Qwen-2.5-0.5B Task Addition[1]	**0.810–0.820**	**0.100–0.103**	**0.130–0.143**

† Point estimates; CIs to be computed with the same bootstrap.

Table 3: **Civil Comments (Race).** Headline metrics (Accuracy ↑, worst-case DPD ↓, worst-case EOD ↓). Models evaluated for this attribute are shown. CIs are 95% stratified bootstrap; † indicates point estimates to be replaced by CIs.

Model/Method	Accuracy	Worst-DPD	Worst-EOD
DistilBERT SFT	0.9467–0.9473	0.0987–0.0995	0.2568–0.3544
DistilBERT LoRA	0.9446–0.9453	0.1360–0.1425	0.4649–0.4895
DistilBERT Task Addition	**0.9362**†	**0.0580**†	**0.1289**†

F USE OF LARGE LANGUAGE MODELS (LLMS)

Scope of assistance. For polishing grammar, wording, concision, and transitions in the abstract, introduction, and discussion. Light edits on figure/table captions and section headings. And style normalization, enforcing consistent terminology and tense across sections. No ideas, claims, analyses, datasets, model architectures, experiments, or results originated from an LLM.

Models and interface. Edits were produced with state-of-the-art LLMs (e.g., ChatGPT/GPT-class models) via a standard chat interface. To preserve anonymity, no identifying information (author names, affiliations, or URLs) was included in prompts. For data privacy, no proprietary data, code, or non-public results were provided. We avoided uploading full drafts and removed any metadata that could compromise double-blind review.

Prompts and examples. Typical prompts included: *"Please copyedit the following paragraph for clarity and brevity without changing technical meaning."* and *"Standardize terminology (task vectors, task arithmetic) and flag any ambiguous phrasing."* The models were instructed not to add facts or alter technical content.

Model	**Race (95% CI)**	Accuracy	Worst DPD	Worst EOD
LLaMA2-7B	SFT	0.7901–0.9039	0.0000–0.0345	0.0000–0.0730
	LoRA	0.7599–0.9143	0.0000–0.0459	0.0000–0.1087
	Task addition	**0.7972–0.8724**	**0.0000–0.0265**	**0.0000–0.1308**
DistilBERT	SFT	0.9467–0.9473	0.0987–0.0995	0.2568–0.3544
	LoRA	0.9446–0.9453	0.1360–0.1425	0.4649–0.4895
	Task addition	**0.9362**	**0.0580**	**0.1289**
Model	**Gender (95% CI)**	Accuracy	Worst DPD	Worst EOD
LLaMA2-7B	SFT	0.7914–0.8491	0.0621–0.1125	0.0000–0.1794
	LoRA	0.8031–0.8823	0.0535–0.0596	0.0105–0.0906
	Task addition	**0.8031–0.8823**	**0.0259–0.0943**	**0.0000–0.0858**
DistilBERT	SFT	0.9457–0.9476	0.0887–0.1101	0.6157–0.6433
	LoRA	0.9447–0.9453	0.0735–0.0812	0.5024–0.5084
	Task addition	**0.9395**	**0.0454**	**0.3358**
Qwen-2.5-0.5B[1]	SFT	0.884–0.886	0.093–0.119	0.060–0.084
Qwen-2.5-0.5B	LoRA	0.774–0.790	0.210–0.251	0.232–0.362
Qwen-2.5-0.5B	**Task addition**	**0.810–0.820**	**0.100–0.103**	**0.130–0.143**

Table 4: 95% confidence intervals. Models evaluated for each attribute are shown: LLaMA2-7B on Berkeley D-Lab; DistilBERT and Qwen-2.5-0.5B on Civil Comments (Qwen-2.5 for gender). Task addition maintains accuracy while showing competitive or improved fairness compared to SFT and LoRA.