On Fairness of Task Arithmetic: The Role of Task Vectors

Laura Gomezjurado Gonzalez[1♯] Hiroki Naganuma[2,3♯] Kotaro Yoshida[4♯]
Yuji Naraki[5] Takafumi Horie[6] Ryotaro Shimizu[7]

[1]Stanford University [2]Mila [3]Univ. de Montréal [4]Inst. of Science Tokyo
[5]Independent [6]Kyoto Univ. [7]ZOZO Research

♯Equal contribution

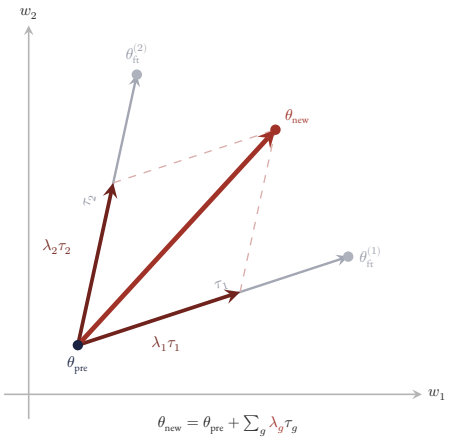
ICLR 2026

Motivation

- **Task arithmetic** enables efficient model editing via parameter-space operations — without retraining.
- Yet its **impact on fairness** remains largely unexplored.
- We present the **first systematic study** of how task vectors influence demographic biases.
- Benchmarked against FFT and LoRA on high-stakes tasks.

Research Question

Can we control subgroup disparities *directly in weight space* by manipulating task vectors, without retraining?



Methodology & Framework

Task Vector:

$$\tau = \theta_{\text{ft}} - \theta_{\text{pre}}$$

Fairness-Aware Merging:

$$\theta_{\text{new}} = \theta_{\text{pre}} + \lambda \cdot \tau$$

Models & Data:

- LLaMA-2-7B — hate speech (D-Lab)
- DistilBERT, Qwen2.5-0.5B — toxicity
- ViT-Base/16 — age (UTKFace)

Fairness Metrics:

- **Accuracy** — overall task performance
- **DPD** (Demographic Parity Difference): measures prediction rate disparity across subgroups
- **EOD** (Equalized Odds Difference): measures TPR/FPR disparity across subgroups

Training Methods Compared:

- Full Fine-Tuning (FFT)
- LoRA
- Task Addition (ours)

Theoretical Insight

Both DPD and EOD share a common upper-bound term $U(\lambda)$:

$$U(\lambda) = 2L \sum_g |\lambda_g - 1| \, \|\Delta\theta_g\|_2$$

DPD Bound

$$\mathrm{DPD}(\theta(\lambda)) \;\leq\; U(\lambda)$$

EOD Bound

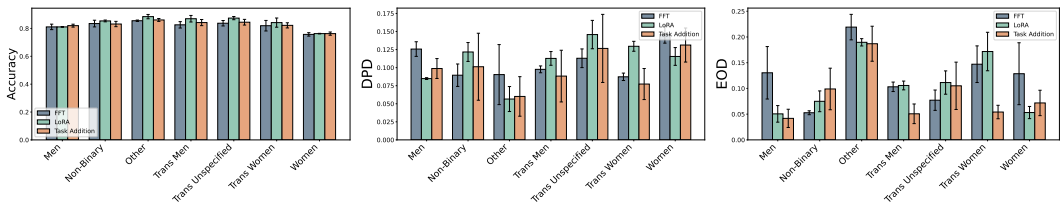
$$\mathrm{EOD}(\theta(\lambda)) \;\leq\; \mathrm{EOD}(\bar{\theta}) + 4\sqrt{(B_0 + B_1)\, U(\lambda)}$$

- B_0, B_1 are class-specific density constants.
- Both bounds decrease as $\lambda_g \to 1$,
  confirming the **Fairness–Utility frontier** is navigable by tuning λ.
- Theory provides guarantees that task vector scaling *monotonically controls* bias metrics.

Performance Comparison

Task Addition vs. FFT and LoRA on LLaMA-2-7B (mean ± SE, 5 runs).

A. Gender Subgroups

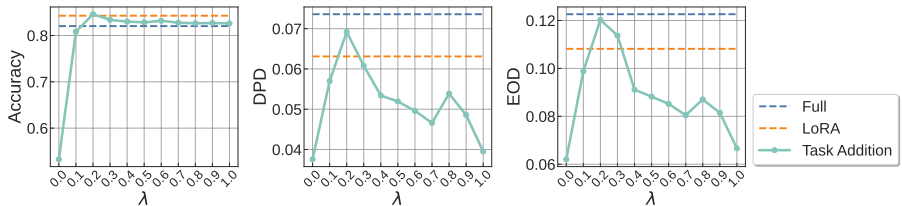


Key Result 1

Task Addition achieves **comparable accuracy** to FFT/LoRA while consistently **reducing bias** (lower DPD/EOD).

Tuning Fairness via λ

We sweep λ from 0.1 to 1.2 and plot accuracy, DPD, and EOD for gender-stratified hate speech. Dashed lines: FFT and LoRA baselines.



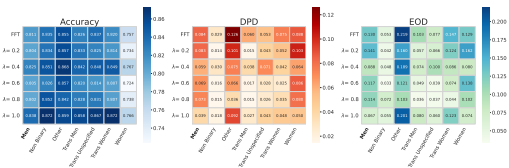
Key Result 2

At $\lambda \in [0.5, 0.8]$, accuracy stabilizes while **fairness improves significantly**. The coefficient λ acts as a "control knob" for fairness — unavailable in standard fine-tuning.

Subgroup Injection & Post-hoc Addition to SFT

Subgroup-Specific Injection

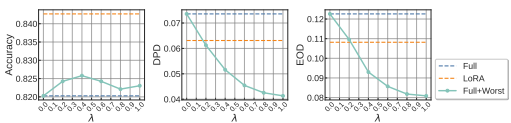
We merge task vectors from individual subgroups (e.g., "Men") into the base model.



Subgroup vectors enable **targeted fairness adjustments**; the impact is heterogeneous, highlighting careful vector selection.

Post-hoc Addition to SFT

We inject subgroup task vectors into a fine-tuned (SFT) model to refine fairness *post-hoc*.



Post-hoc injection provides an additional lever for fairness correction **without retraining**.

Conclusions

1. **First Systematic Study** — Task arithmetic influences model fairness across NLP and vision tasks.
2. **Controllability** — λ navigates the Accuracy–Fairness trade-off without retraining.
3. **Interpretability** — Subgroup task vectors trace demographic adaptation in weight space.
4. **Theory** — Upper bounds on DPD and EOD link λ to fairness metrics with provable guarantees.

Thank you!