Federated Learning with Ethical Constraints via Subspace Projection

Samir Poudel

March 15, 2026

1 Problem Statement

When language models fine-tune on user data, they absorb biases present in that data. This creates a fundamental challenge: how do you personalize models to individual users without amplifying harmful patterns like demographic stereotypes or toxic language? Data filtering blocks legitimate personalization, and post-hoc bias detection happens too late because the model already learned the bias during training. We need a method that prevents bias during training, not after.

This problem is critical in federated learning settings where you cannot inspect client data directly. Edge AI deployments for mobile keyboards, email assistants, and healthcare applications require personalized models that cannot tolerate amplified user biases.

2 Proposed Solution

I will adapt Subspace Projection Aggregation (SPA) to act as a bias filter during federated learning. The core idea: identify which model parameters control ethical behavior, then lock those parameters during personalization.

2.1 Technical Approach

Step 1: Identify the Safety Subspace

I will fine-tune Llama-3-70B on Anthropic's HH-RLHF dataset [1], which contains 170k human preference comparisons for helpful and harmless responses. I use LoRA (Low-Rank Adaptation) via the HuggingFace PEFT library with rank $r = 16$ and $\alpha = 32$, targeting the query and value projection matrices (`q_proj`, `v_proj`) in each transformer layer, reducing trainable parameters to roughly 0.1% of the full model. During fine-tuning, I will track LoRA adapter changes using singular value decomposition (SVD). Parameters that change most significantly form the "safety subspace."

Step 2: Constrained Local Training

I will simulate 10-15 federated clients using the Flower framework [2]. Each client fine-tunes on biased datasets (Wiki-Bias and BOLD [3]) using LoRA with the same rank and target modules as Step 1. The key constraint: local LoRA updates cannot modify safety subspace parameters. Mathematically, if W_{safety} represents the safety subspace projection matrix, local gradient updates ∇_{local} are projected to the orthogonal complement: $\nabla_{\text{constrained}} = (I - W_{\text{safety}} W_{\text{safety}}^T) \nabla_{\text{local}}$.

Step 3: Server-Side Aggregation with DPO

I will aggregate client updates using SPA, which handles heterogeneous model updates through subspace projection, then apply Direct Preference Optimization (DPO) [6] at the server to maintain global preference alignment without reinforcement learning.

Step 4: Fairness Evaluation I will measure bias using HELM benchmark fairness metrics [5], specifically demographic parity and equalized odds on the BOLD dataset, comparing four methods: standard FedAvg, unconstrained LoRA fine-tuning, full fine-tuning on biased data, and the proposed SPA with safety subspace constraints.

3 Expected Outcomes

My method should achieve two goals:

1. Match personalization quality: Task-specific accuracy within 2% of unconstrained methods
2. Reduce bias: Demographic parity gap reduced by 40-50% compared to FedAvg baseline

4 Implementation Details

Hardware: MTSU Hamilton cluster, dual NVIDIA RTX PRO 6000 Blackwell GPUs (192GB HBM3e each).

Software: HuggingFace PEFT (LoRA), Flower 1.12+ (federated orchestration), Llama-3-70B, PyTorch 2.0+ with DeepSpeed, HELM benchmark suite.

Datasets:

- HH-RLHF: 170k preference pairs ($\sim$2GB)
- Wiki-Bias + BOLD: 83k samples combined ($<$500MB)
- Total training data: $<$3GB (fits in GPU memory)

Timeline (8 weeks):

- Weeks 1–2: Environment setup and safety subspace identification
- Weeks 3–4: Constrained federated training pipeline
- Weeks 5–6: Experiments (10–15 clients, 5 federated rounds)
- Week 7: HELM evaluation and statistical analysis
- Week 8: Final report and documentation

5 Significance and Novelty

Federated learning deployments cannot inspect client data for bias. This project provides mathematical guarantees that personalization preserves safety constraints.

Novel Contribution: Existing work uses subspace methods for efficiency [7] or privacy [4], and fair federated learning addresses bias through client selection and adaptive aggregation. No prior work uses subspace constraints to lock ethical parameters during local LoRA fine-tuning. This project combines federated learning, preference alignment via DPO, and fairness in NLP, falling under "Novel methods" per the course guidelines.

References

[1] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[2] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Titouan Parcollet, Pedro PB de Gusmão, and Nicholas D Lane. Flower: A friendly federated learning framework. *arXiv preprint arXiv:2007.14390*, 2020.

[3] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872, 2021.

[4] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021.

[5] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[6] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[7] Weiran Sun, Zhenyu Jiang, Kyeongbo Choi, Myeongho Kang, and John Paisley. Subspace learning for effective meta-learning. *arXiv preprint arXiv:2209.12581*, 2022.