SI252 Course Project Proposal: The Alchemy of Federated PEFT: Turning Configuration Exploration into Training Efficiency Gold

Linjie Gu(2023233134), PengCheng Wang(2024233198), Shuang Chen(2024233246)

***Abstract*—Fine-tuning pre-trained foundation models (PTMs) has revolutionized deep learning by enabling state-of-the-art performance across diverse tasks with minimal adaptation [1]. In federated learning (FL), leveraging PTMs can alleviate data scarcity and accelerate convergence through effective knowledge transfer [2]. However, conventional FL approaches typically commit to a fixed PTM and a predetermined fine-tuning strategy, which forces the system to "work hard" along a single suboptimal trajectory—a phenomenon we refer to as the *pre-trained model trap*. This trap manifests when over-provisioned models incur excessive communication overhead, while under-capacitated models fail to manage data heterogeneity. Moreover, the privacy constraints and resource variability inherent in FL further complicate optimal model selection and fine-tuning configuration. To address these challenges, we propose VISTA (Visionary Intelligent Selection for Trap Avoidance), an online framework that concurrently explores multiple candidate PTMs and parameter-efficient fine-tuning (PEFT) configurations. VISTA employs a bandit learning strategy to eliminate suboptimal candidates with confidence and utilizes Lyapunov optimization to ensure that the additional computation and memory switching costs remain within acceptable bounds. Comprehensive experiments demonstrate that VISTA significantly reduces communication overhead and accelerates convergence compared to conventional fixed-strategy baselines.**

***Index Terms*—Federated learning, Pre-trained models, Model selection, Fine-tuning, Resource-aware optimization.**

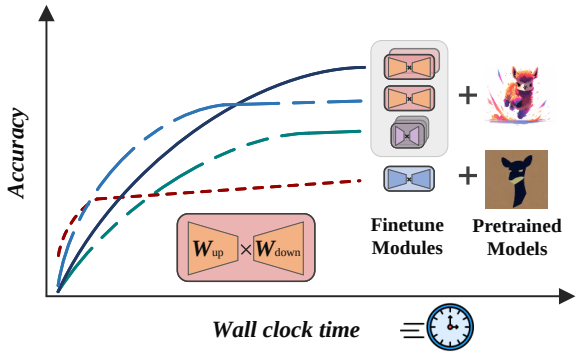

Fig. 1: Performance Comparison with different pretrained model selection and configuration

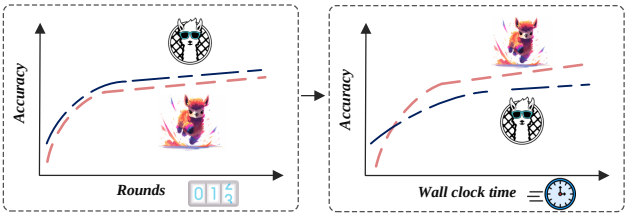

Fig. 2: Pretrained Model Entrapment

I. Introduction

Federated Learning (FL) has emerged as the de facto paradigm for training machine learning models on distributed private data—a capability increasingly vital in the era of large foundation models. While pretrained large foundation models like Llama and Qwen demonstrate unprecedented capabilities, their development relies on exhaustive public internet data, their performance improvement in vertical domain still relies on finetuning via high quality professional data. This leaves high-quality personal data, often trapped in isolated edge devices due to privacy regulations (e.g., GDPR) and sector-specific silos (e.g., healthcare records), as the next frontier for model improvement. FL enables collaborative learning while preserving data locality, making it theoretically ideal for unlocking this value.

The emergence of large foundation models like Llama and DeepSeek has redefined machine learning capabilities through exhaustive pretraining on web-scale public data. These models demonstrate unprecedented performance in general-domain tasks, yet their adaptation to vertical domains—medical diagnosis, legal analysis, personalized education—requires high-quality professional data that often resides in edge devices. This specialized data encodes domain expertise and user preferences critical for model alignment, but privacy regulations (e.g., GDPR) and institutional policies prevent its centralized collection. Federated Learning (FL) addresses this fundamental tension by enabling collaborative model refinement while preserving data locality, positioning itself as the essential paradigm for adapting foundation models to specialized domains.

Although promising the future of federated tuning on PTMs, the fundamental challenge of training efficiency can't

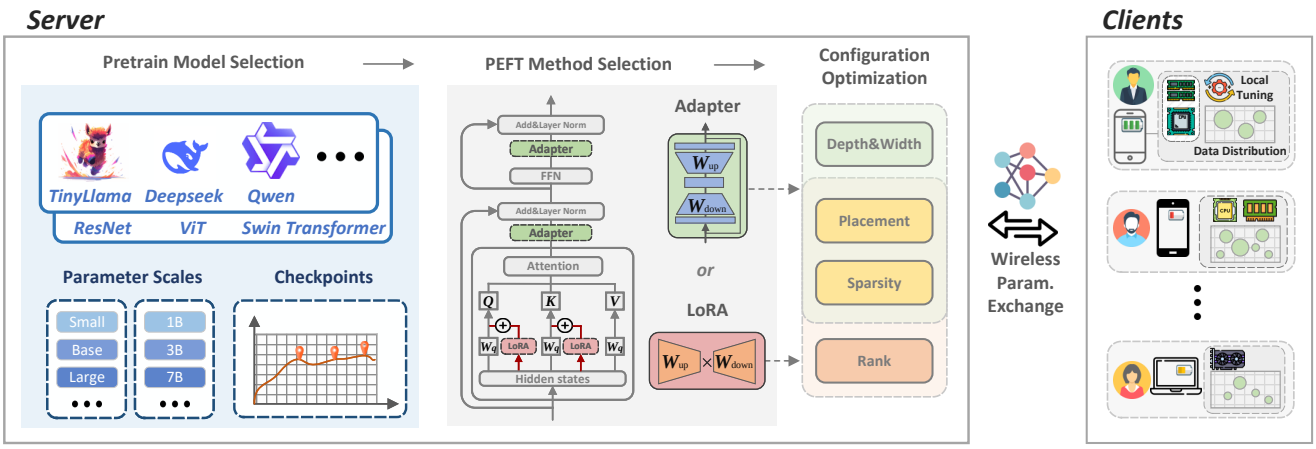

Fig. 3: Problem overview

be ignored, especially when trying to deploy computation intensive LM training paradaigm on edge devices. Existing efficiency optimization approaches bifurcate into two lineages: 1) Training from scratch via parameter pruning [3] or heterogeneous architecture design [4], and 2) Fine-tuning PTMs using parameter-efficient methods (PEFT) like adapters [5] or LoRA [6]. While these methods reduce communication overhead, they share a critical blind spot: *the assumption of compatibility between PEFT-configuration and the FL task*. Centralized learning avoids this pitfall through iterative PTM experimentation enabled by full data access and scaling laws [7]. FL's defining constraints—distributed data ownership, privacy requirements, and device heterogeneity—force irreversible upfront commitments to suboptimal configurations, manifesting in three systemic failure modes:

- **Structural Capacity Mismatch:** Static model architectures struggle to adapt to evolving FL environments. Over-provisioned PTMs like ViT-Huge incur prohibitive communication costs (63.7MB/round vs. ViT-Base's 12.4MB) for marginal accuracy gains (81.7% vs. 78.3% on CIFAR-100 [8]), while under-capacitated models like ViT-Tiny plateau at 68.2% accuracy under data heterogeneity.
- **Irreversible Domain Drift:** Pretraining-target domain discrepancies compound during FL training. ImageNet-based PTMs exhibit 19-34% accuracy gaps in medical imaging tasks compared to domain-adapted counterparts [9], yet FL's privacy constraints prevent the data inspection needed to identify optimal PTMs.
- **Rigid Configuration Lock-In:** While recent PEFT advancements for FL [5] reduce communication overhead, they focus on optimizing hyperparameters within fixed method classes. Empirical studies show that PEFT method selection impacts final accuracy more significantly than hyperparameter tuning—for instance, proper adapter configuration can outperform optimal LoRA ranks by 12.3% accuracy on clinical text tasks [10]. However, FL systems lack mechanisms to revise initial PEFT choices post-deployment.

These recurring failures emerge from a common root cause: FL currently lacks the ability to adjust PTM or PEFT configurations *after* deployment. Centralized workflows can fix early misalignments by adjusting architectures mid-training, but privacy-preserving aggregation and distributed ownership hinder such iterative experimentation in FL. Once a suboptimal setup is chosen, it can derail the training trajectory, as illustrated in Figure 3, where inaccurate LoRA choices cause irreversible performance degradation after a few dozen communication rounds.

This situation raises a fundamental question:

> *How can FL systems dynamically co-adapt PTM selection and PEFT configuration to balance accuracy, efficiency, and operational constraints in unknown distributed environments?*

Answering this question necessitates online configuration exploration, but introduces two critical challenges: 1) managing the exponential search space of PTM-PEFT combinations, and 2) controlling the cumulative costs of exploration measured through both temporal overhead (extended time-to-accuracy) and device resource consumption.

Our framework, **VISTA**, resolves this through three synergistic innovations:

- **Concurrent Partial Learning Based Configuration Exploration:** We design a parallel partial Learning exploration framework that trains multiple PTM-PEFT configurations concurrently through partitioned client groups on sampled local datasets, for reducing the exploration cost of evaluation the tuning procedure of multiple PTMs. Via the almost "free lunch" extra exploration trails, both in training time and resource budget consumption perspective, more local training trails can be conducted on candidate PTMs, which further helps the comparison of training dynamic between them and eliminate those less effective options.
- **Constrained Dueling Bandit Based Client sampling:** To overcome the natural data heterogeneous challenge of clients' data distribution and relieve of impact of data

quality difference on Model tuning dynamic evaluation, we proposed a constrained dueling bandit based client sampling method, which dueling bandit learning was utilized for estimate the client's local data quality from training feedbacks obtained from tuning on different PTMs. Lyapunov optimization technology was utilized for resource budget control. Further, the estimated client data quality score was used by the PTM selection decision, for maintaining the evaluation fairness between PTMs, which is guaranteed by Lyapunov optimization as well. Theoretical analysis will show that our method can achieve $\mathcal{O}(\sqrt{T})$ regret with zero constraint violation after $\mathcal{O}(\log T)$ rounds under a given Federated PTM selection policy π.

- **Cross-Architecture Knowledge Transfer:** To further reduce the cost of extra exploration introduced by concurrently tuning of multiple PTMs rather than straightly tuning on a intuitively selected one, we developed a knowledge distillation method for extracting knowledge from those eliminated PTMs, for recycling the effort devoted on their tuning in some level.

Experimental validation across

II. Related Work

Relevant studies in federated learning and pre-trained model adaptation can be categorized into three central areas:

Federated Learning with Pre-trained Models. Initial investigations [1], [2] revealed that using pre-trained models as initialization in FL cuts down on training time and enhances model accuracy. Later works [11], [12] explored broader aspects such as the gap between federated and centralized fine-tuning. To address communication bottlenecks, researchers proposed parameter-efficient updates via methods like LoRA or adapter modules [10], [13], [14]. Nonetheless, most solutions fix the foundation model architecture and rarely consider systematically exploring alternatives.

Optimal Fine-tuning Configuration Search. Recent FL related studies [5], [15] used reinforcement learning or adaptive heuristics to discover optimal fine-tuning configurations, such as the best adapter layers or hyperparameters for each client. In parallel, other work on sparsity and heterogeneous LoRA [16], [17] underscores the need for configuration awareness. Still, many existing approaches begin with a singular pre-trained backbone, possibly hampering flexibility if the initial choice is poor.

Model Selection for ML Training and Serving. Methods aimed at measuring the transferability of pre-trained models [9], [18]–[21] offer insights on how well a specific model may adapt to a downstream task. Although these techniques are useful, they often ignore key FL factors such as communication costs, client resource heterogeneity, and limited direct data visibility. Recent theoretical work [22] underscores the need for adaptive model selection strategies in collaborative contexts, hinting at the benefits of our exploration-based approach. Some previous works about model serving system with heterogeneous candidate models have realized the importance of trading off between models running cost and capability, and make attempts on selection or dynamic scheduling [23]–[25]. Also, some recent work turn their focus on PTM included serving systems, and emphasize of adaptively scheduling PEFT parameters for better service quality-latency(service cost) trade-off [26], [27]. However, few previous works make efforts on the online PEFT method choosing, although related works have made trail on adaptive pruning or sparsity. We argue that some performance gap can't be narrowed via extra trail(larger parameter size/ longer training) under incorrect method choice, which indicates " correct choice is more important than hard working".

III. Background and Motivation

Our work focuses on avoiding the pitfalls of prematurely committing to an unsuitable pre-trained model or fine-tuning strategy. In this section, we detail the primary challenges encountered in designing an online selection framework for FL and discuss how we addresses them.

A. Preliminary Experiments of FedPEFT

TODO

B. Efficient Exploration under Resource Budget Control

Federated learning environments impose strict limitations on computation, memory, and communication bandwidth. A naive strategy that adheres to a single configuration can be highly suboptimal if that choice underperforms. Key issues include:

1) **Candidate Pool Sizing:** Determining the optimal number of foundation models and tuning methods to explore is critical. While a larger candidate pool increases the likelihood of selecting an optimal configuration, it also demands additional computational and memory resources.
2) **Balancing Exploration and Exploitation:** It is essential to decide when to prune underperforming candidates and reallocate resources to the more promising options. VISTA employs a bandit learning strategy that enables effective candidate elimination with sufficient statistical confidence.
3) **Adapting to Heterogeneous Data:** FL systems are characterized by diverse client data distributions and varying device capabilities. A robust exploration process must dynamically adjust to these differences to ensure optimal performance across all clients.

IV. Method

A. System Modeling

Federated Learning Setup. We consider an *asynchronous* FL system composed of a central server and a set of clients $\mathcal{C} := \{1, 2, \ldots, N\}$. The server explores a library of candidate pre-trained models (PTMs) $\mathcal{M} := \{1, 2, \ldots, M\}$. Each model $i \in \mathcal{M}$ is described by an architecture tag A_i (e.g. ViT/Bert/Llama) and parameters θ_i. Fine-tuning is conducted through a parameter-efficient configuration $d \in \mathcal{D}$ such as LoRA rank or adapter width.

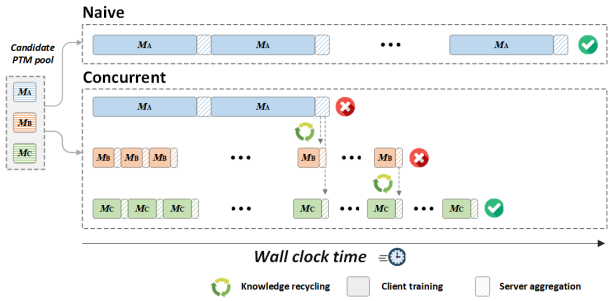

Fig. 4: A demonstration of concurrent training with three candidate PTMs

Decision Epoch & Space. Whenever *any* client group completes its local update, a decision epoch $t \in \mathbb{N}$ is triggered. At epoch$t$ the server jointly selects

1) a set of PTM–PEFT pairs to train next, $\mathcal{X}_t \subseteq \mathcal{M} \times \mathcal{D}$;
2) one *or several* client subsets of size K_t, $\mathcal{S}_t \subseteq \mathcal{C}_t$, $|\mathcal{S}_t| = K_t$, where $\mathcal{C}_t \subseteq \mathcal{C}$ denotes the currently online clients; and
3) a local mini-batch sampling rate $\alpha_t \in (0, 1]$.

We write the joint action as $\Psi_t \triangleq (\mathcal{X}_t, \mathcal{S}_t, \alpha_t, K_t)$.

Round Latency & Budget Consumption. For client j tuning PTM i with configuration d under sampling rate α, we split its latency into

$$\begin{aligned} \mu^j_{\mathrm{cp}}(i,\alpha) &: \text{ GPU/CPU compute time,} \\ \mu^j_{\mathrm{io}}(i) &: \text{ parameter or dataset (un)loading,} \\ \mu^j_{\mathrm{cm}}(i,d) &: \text{ network communication.} \end{aligned}$$

The *round-critical latency* is thus

$$\tilde{\mu}(i,d,\alpha,\mathcal{S}_t) := \max_{j\in\mathcal{S}_t}\big[\mu^j_{\mathrm{cp}} + \mu^j_{\mathrm{io}} + \mu^j_{\mathrm{cm}}\big].$$

Client j's instantaneous resource-consumption rate is

$$\rho_j(i,d,\alpha) := \frac{\eta_j(i,d,\alpha)}{\tilde{\mu}(i,d,\alpha,\mathcal{S}_t)},$$

where $\eta_j(\cdot)$ denotes energy or memory actually spent in that local round.

Let $m_\dagger$ (with fixed PEFT $\hat{d}$ and sampling $\hat{\alpha}$) be the *naïve baseline* FL would employ without exploration. Its long-term reference rate is

$$\bar{\rho}^\dagger_j := \frac{K}{N}\,\mathbb{E}\big[\rho_j(m_\dagger,\hat{d},\hat{\alpha})\big],$$

which naturally scales the single-client rate by its average participation frequency K/N.

Training Gain. For PTM i in epoch t we denote the single-round improvement (by accuracy, loss decrease, or any monotone utility) as

$$\Delta F_{i,t}(\mathcal{S}_t,\Psi_t) := F(\theta_{i,t-1}) - F\big(\theta_{i,t}\,|\,\mathcal{S}_t,\Psi_t\big),$$

where $F(\theta)$ denotes the model performance metric, *e.g.* loss and test accuracy on global test set, and we let $\mathbb{I}_t(i) = \mathbf{1}\{i \in \mathcal{X}_t\}$.

B. Sequential (Concurrency=1) Optimization

In the most basic setting we allow only $|\mathcal{X}_t| = 1$ and enforce a constant client group size $|\mathcal{S}_t| = K$. The long-term throughput maximisation reads

$$\textbf{P0:} \quad \max_{\{\Psi_t,\mathbb{I}_t,i\}} \mathbb{E}\left[\sum_{t=1}^{\infty}\sum_{i=1}^{M}\frac{\mathbb{I}_t(i)\,\Delta F_{i,t}(\mathcal{S}_t,\Psi_t)}{\tilde{\mu}(i,d_t,\alpha_t,\mathcal{S}_t)}\right], \tag{1a}$$

$$\text{s.t.} \quad |\mathcal{S}_t| = K,\ \forall t, \tag{1b}$$

$$\mathbb{E}\left[\frac{\sum_t \mathbb{I}(j\in\mathcal{S}_t)\,\eta_j(i_t,d_t,\alpha_t)}{\sum_t \tilde{\mu}(i_t,d_t,\alpha_t,\mathcal{S}_t)}\right] \le \bar{\rho}^\dagger_j,\ \forall j, \tag{1c}$$

$$\mathbb{E}\left[\frac{\sum_t \mathbb{I}_t(i)\,\frac{1}{K}\sum_{k\in\mathcal{S}_t} u_k}{\sum_t \mathbb{I}_t(i)}\right] \ge \bar{u},\ \forall i. \tag{1d}$$

where u_i represents the data quality of client$i$ and $\bar{u}$ represents the averaged data quality among the whole client set $\{N\}$. Constraint (1c) caps each client's *average* consumption below the baseline, while (1d) preserves data-quality fairness among PTMs.

Problem Decomposition. Under the mild assumption that $\Delta F_{i,t} \uparrow u_k$ (higher-quality data yields larger utility improvement), **P0** decomposes into

1) *Configuration selection*

$$\max_{\{\mathbb{I}_t,i\}} \mathbb{E}\left[\sum_t\sum_{i=1}^{M}\frac{\mathbb{I}_t(i)\,\Delta F_{i,t}}{\tilde{\mu}(i,d,\alpha,\mathcal{S})}\right]$$

s.t. quality fairness constraint (1d).

2) *Client sampling*

$$\max_{\{\mathcal{S}_t\}} \mathbb{E}\left[\sum_t\sum_{j\in\mathcal{S}_t} u_j\right]$$

s.t. $|\mathcal{S}_t| = K$, power constraint (1c).

Bandit Learning Based Client Scoring. Because true u_k is unknown, we estimate it *online* by comparing gradient feedback. For each sampled client $k \in \mathcal{S}_t$ we observe gradient v_k. For every pair $(k,\ell) \subset \mathcal{S}_t$ we construct a *leave-two-out* proxy of the global gradient

$$\tilde{G}_{-k\ell} := \frac{1}{K-2}\sum_{p\in\mathcal{S}_t\setminus\{k,\ell\}} v_p, \quad Z_{k\ell} := \langle v_k - v_\ell, \tilde{G}_{-k\ell}\rangle.$$

Maintaining pairwise win counts $w_{k,\ell}$, we build an upper-confidence estimate

$$\hat{r}_{k,\ell}(t) := \left[\frac{w_{k,\ell}}{w_{k,\ell}+w_{\ell,k}} + \sqrt{\frac{\beta\log t}{w_{k,\ell}+w_{\ell,k}}}\right]^+,$$

and compute the *Borda score* $\hat{u}_k(t) = \frac{1}{K-1}\sum_{\ell\neq k}\hat{r}_{k,\ell}(t)$.

Why Borda? Pairwise gradients share no absolute scale across PTMs; Borda aggregation only uses *relative* wins, yielding a robust total order that is transitive and less sensitive to noisy magnitudes. We prove in Appendix A that under a mild condition the Borda winner coincides with the optimal client in terms of global-gradient similarity.

Lyapunov Optimization for Budget Control. To enforce the *long-term* constraint (1c) without knowing future arrivals, we

follow stochastic network control and attach a virtual queue to each client:

$$Q_j(t+1) = \big[Q_j(t)+\eta_j(i_t,d_t,\alpha_t)-\bar{\rho}_j^{\dagger}\,\tilde{\mu}(i_t,d_t,\alpha_t,\mathcal{S}_t)+\varepsilon_t\big]^+.$$

where ε_t is a relaxation factor for zero constraint violation in constant level rounds. Minimising the one-step Lyapunov drift plus a weighted performance penalty leads to the per-round decision

$$\max_{|\mathcal{S}_t|=K} \sum_{j\in\mathcal{S}_t} \big[V_t\,\hat{u}_j(t) - Q_j(t)\big],$$

which is solved by greedily selecting the top-K clients with the highest score $V_t\hat{u}_j - Q_j$. *Why Lyapunov?* It converts long-horizon power-budget constraints into queue stability, guaranteeing (i) $\mathcal{O}(\sqrt{T})$ regret and (ii) zero constraint violation after $\mathcal{O}(\log T)$ rounds, yet keeps per-epoch optimisation lightweight.

C. Extension to Concurrent Partial Learning Based Exploration

When multiple PTM–PEFT pairs are tuned in parallel ($|\mathcal{X}_t| > 1$), all objective terms and constraints in (1) extend by summing over concurrent pairs and by accounting for parameter switch-over overhead and shared memory pressure. A full treatment is left to the final report.

REFERENCES

[1] J. Nguyen, J. Wang, K. Malik, M. Sanjabi, and M. Rabbat, "Where to begin? on the impact of pre-training and initialization in federated learning," *arXiv preprint arXiv:2206.15387*, 2022.

[2] H.-Y. Chen, C.-H. Tu, Z. Li, H.-W. Shen, and W.-L. Chao, "On the importance and applicability of pre-training for federated learning," *arXiv preprint arXiv:2206.11488*, 2022.

[3] M. Kim, S. Yu, S. Kim, and S.-M. Moon, "Depthfl: Depthwise federated learning for heterogeneous clients," in *The Eleventh International Conference on Learning Representations*, 2023.

[4] R. Liu, F. Wu, C. Wu, Y. Wang, L. Lyu, H. Chen, and X. Xie, "No one left behind: Inclusive federated learning over heterogeneous devices," ser. KDD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3398–3406. [Online]. Available: https://doi.org/10.1145/3534678.3539086

[5] D. Cai, Y. Wu, S. Wang, F. X. Lin, and M. Xu, "Efficient federated learning for modern nlp," in *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, 2023, pp. 1–16.

[6] F. Wu, J. Hu, G. Min, and S. Wang, "Adaptive rank allocation for federated parameter-efficient fine-tuning of language models," *arXiv preprint arXiv:2501.14406*, 2025.

[7] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[8] C. Xie, D.-A. Huang, W. Chu, D. Xu, C. Xiao, B. Li, and A. Anandkumar, "Perada: Parameter-efficient federated learning personalization with generalization guarantees," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 838–23 848.

[9] Y. Zhang, T. Huang, Y. Ding, D. Zhan, and H. Ye, "Model spider: Learning to rank pre-trained models efficiently," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[10] J. Bai, D. Chen, B. Qian, L. Yao, and Y. Li, "Federated fine-tuning of large language models under heterogeneous tasks and client resources," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[11] B. Wang, Y. J. Zhang, Y. Cao, B. Li, H. B. McMahan, S. Oh, Z. Xu, and M. Zaheer, "Can public large language models help private cross-device federated learning?" *arXiv preprint arXiv:2305.12132*, 2023.

[12] A. Agarwal, M. Rezagholizadeh, and P. Parthasarathi, "Practical takes on federated learning with pretrained language models," in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 454–471.

[13] Z. Wang, Y. Zhou, Y. Shi, K. Letaief, *et al.*, "Federated fine-tuning for pre-trained foundation models over wireless networks," *arXiv preprint arXiv:2407.02924*, 2024.

[14] Z. Peng, X. Fan, Y. Chen, Z. Wang, S. Pan, C. Wen, R. Zhang, and C. Wang, "Fedpft: Federated proxy fine-tuning of foundation models," *arXiv preprint arXiv:2404.11536*, 2024.

[15] T. Qi, Y. Zhan, P. Li, and Y. Xia, "Tomtit: Hierarchical federated fine-tuning of giant models based on autonomous synchronization," in *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*. IEEE, 2024, pp. 1910–1919.

[16] K. Kuo, A. Raje, K. Rajesh, and V. Smith, "Sparsity for communication-efficient lora," in *5th Workshop on practical ML for limited/low resource settings*, 2024.

[17] Y. J. Cho, L. Liu, Z. Xu, A. Fahrezi, and G. Joshi, "Heterogeneous lora for federated fine-tuning of on-device foundation models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 12 903–12 913.

[18] N. Ding, X. Chen, T. Levinboim, S. Changpinyo, and R. Soricut, "Pactran: Pac-bayesian metrics for estimating the transferability of pretrained models to classification tasks," in *European Conference on Computer Vision (ECCV)*. Springer, 2022.

[19] M. Gholami, M. Akbari, X. Wang, B. Kamranian, and Y. Zhang, "Etran: Energy-based transferability estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 18 613–18 622.

[20] X. Li, Z. Hu, Y. Ge, Y. Shan, and L.-Y. Duan, "Exploring model transferability through the lens of potential energy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 5429–5438.

[21] F. Meng, W. Shao, Z. Peng, C. Jiang, K. Zhang, Y. Qiao, and P. Luo, "Foundation model is efficient multimodal multitask model selector," in *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[22] J. Li, Z. Wu, Z. Xu, and I. King, "On the necessity of collaboration for online model selection with decentralized data," in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

[23] D. Crankshaw, X. Wang, G. Zhou, M. J. Franklin, J. E. Gonzalez, and I. Stoica, "Clipper: A {Low-Latency} online prediction serving system," in *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, 2017, pp. 613–627.

[24] Y. Wang, K. Chen, H. Tan, and K. Guo, "Tabi: An efficient multi-level inference system for large language models," in *Proceedings of the Eighteenth European Conference on Computer Systems*, 2023, pp. 233–248.

[25] Z. Li, L. Zheng, Y. Zhong, V. Liu, Y. Sheng, X. Jin, Y. Huang, Z. Chen, H. Zhang, J. E. Gonzalez, *et al.*, "{AlpaServe}: Statistical multiplexing with model parallelism for deep learning serving," in *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, 2023, pp. 663–679.

[26] Z. Zhou, X. Wei, J. Zhang, and G. Sun, "{PetS}: A unified framework for {Parameter-Efficient} transformers serving," in *2022 USENIX Annual Technical Conference (USENIX ATC 22)*, 2022, pp. 489–504.

[27] B. Wu, R. Zhu, Z. Zhang, P. Sun, X. Liu, and X. Jin, "{dLoRA}: Dynamically orchestrating requests and adapters for {LoRA}{LLM} serving," in *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, 2024, pp. 911–927.

APPENDIX

A. Consistency between Borda winner and client with best data quality

Problem setup

- n clients with local gradients $v_1,\dots,v_n \in \mathbb{R}^d$.
- True global gradient is

$$G \;=\; \frac{1}{n}\sum_{k=1}^{n} v_k.$$

- Projection score (true utility) $s_i = \langle v_i, G\rangle$. The unique maximiser is denoted $w^\star = \arg\max_i s_i$.

Then in a "leave-two-out duel", we Fix a batch $S \subset [n]$ of size K $(3 \le K \le n)$ and a pair $(i,j) \subset S$. The remaining $K-2$ gradients average to

$$\widehat{G}_{-ij} \;=\; \frac{1}{K-2} \sum_{k \in S \setminus \{i,j\}} v_k.$$

Client i beats j iff $\langle v_i, \widehat{G}_{-ij} \rangle > \langle v_j, \widehat{G}_{-ij} \rangle$.

Every unordered pair is duelled once; the *Borda score* $B_i = \#\{j \neq i : i \text{ beats } j\}$; the *Borda winner* $b^\star = \arg\max_i B_i$.

For any batch containing i, j:

$$\langle v_i - v_j,\, \widehat{G}_{-ij} \rangle \;=\; \frac{1}{K-2} \sum_{k \in S \setminus \{i,j\}} \langle v_i - v_j,\, v_k \rangle. \tag{1}$$

Taking expectation over the random batch choices gives

$$\mathbb{E}[\text{RHS}] \;=\; \frac{K-2}{n-2} \Big[n\big(s_i - s_j\big) - \big(\|v_i\|^2 - \|v_j\|^2\big) \Big]. \tag{2}$$

Hence the *sign* (and thus the expected duel outcome) is determined solely by

$$x_i := n\, s_i - \|v_i\|^2. \tag{3}$$

Indeed, i beats j in expectation iff $x_i > x_j$.

Lemma 1 (Borda monotonicity): If $x_a \ge x_b$ then $B_a \ge B_b$; if $x_a > x_b$ then $B_a > B_b$.

Theorem 1 (Necessary & sufficient): Let $\Delta_j = s_{w^\star} - s_j > 0,\ \delta_j = \|v_{w^\star}\|^2 - \|v_j\|^2$. Then

$$b^\star = w^\star \iff n\,\Delta_j > \delta_j \quad (\forall j \neq w^\star). \tag{$*$}$$

Proof: Since $x_{w^\star} - x_j = n\Delta_j - \delta_j$. Condition $(*)$ is exactly $x_{w^\star} > x_j$ for each rival. Lemma 1 then yields $B_{w^\star} > B_j\ \forall j \neq w^\star \iff b^\star = w^\star$. ∎

$(*)$ reveals that the projection gap must dominate the norm gap. If all $\|v_i\|$ are equal, the right side is 0 and winners coincide automatically. To make sure that the Borda winner is always consistent with the optimal client, we should make sure that $(*)$ always holds. Here we propose a *gradient clipping* strategy for consistency enforcement:

Clip every gradient at radius τ: $\widetilde{v}_i = \min\{1, \tau/\|v_i\|\}\, v_i$. After clipping $\tilde{\delta}_j := \|\widetilde{v}_{w^\star}\|^2 - \|\widetilde{v}_j\|^2 \in [-\tau^2, \tau^2]$.

Corollary 1 (Deterministic clip radius): Let $\Delta_{\min} = \min_{j \neq w^\star} \Delta_j$. If $\tau < \sqrt{n\,\Delta_{\min}}$ then $(*)$ holds for clipped gradients, hence $b^\star = w^\star$ deterministically.

Another possible solution is to use "gradient unit direction dueling": Define unit directions $u_i = v_i/\|v_i\|$ $(i = 1, \dots, n)$ and $G^{(u)} = \frac{1}{n}\sum_k u_k$. Duels and Borda scores computed with $\{u_i\}$ depend on

$$\hat{x}_i = n\langle u_i, G^{(u)} \rangle - 1,$$

so the norm term vanishes. Thus

Theorem 2 (Direction-only consistency): If $w^\star = \arg\max_i \langle u_i, G^{(u)} \rangle$ is unique (*no equality of directional cosines*), then with unit-direction duelling

$$b^\star = w^\star \quad \text{(always)}.$$

Note that such method is at the expense of losing gradient norm information, which may be a important characteristic for data quality evaluation sometimes, and make the dueling more sensitive to noise.

B. Interpretation for the advantage of "leave 2 out" over "full batch mean" gradient approximation

$$\begin{aligned}
&S:\ |S| = K \ge 3, \qquad R := S \setminus \{i,j\},\ |R| = K-2,\\
&v_k = v_k^\star + \varepsilon_k, \quad \varepsilon_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma),\ \Sigma \succ 0,\\
&w^\star := v_i^\star - v_j^\star, \quad w := w^\star + (\varepsilon_i - \varepsilon_j),\\
&G^\star := \mathbb{E} v_k^\star, \quad \Delta := \langle w^\star, G^\star \rangle \neq 0,\\
&\sigma_{ij}^2 := w^\top \Sigma w = w^{\star\top} \Sigma w^\star.
\end{aligned}$$

$$\bar{G} = \frac{1}{K}\sum_{p \in S} v_p, \qquad \hat{G}_{-ij} = \frac{1}{K-2}\sum_{k \in R} v_k.$$

$$Z_{ij}^{\mathrm{A}} := \langle w, \bar{G} \rangle, \quad Z_{ij}^{\mathrm{B}} := \langle w, \hat{G}_{-ij} \rangle, \quad e_{\mathrm{A}} := Z_{ij}^{\mathrm{A}} - \Delta,\ e_{\mathrm{B}} := Z_{ij}^{\mathrm{B}} - \Delta.$$

Using $\varepsilon_i \perp \varepsilon_k\ (k \neq i)$ and $\mathbb{E}[\varepsilon_i \varepsilon_i^\top] = \Sigma$,

$$\mathrm{Cov}(\varepsilon_i, Z_{ij}^{\mathrm{A}}) = \Sigma G^\star + \frac{1}{K}\Sigma\big(w^\star + I\big) \neq 0, \quad \mathrm{Cov}(\varepsilon_i, Z_{ij}^{\mathrm{B}}) = 0.$$

Thus FM's score "moves" whenever its own noise moves; L2O is fully decorrelated.

Write $\Phi_k := w^\top \varepsilon_k \sim \mathcal{N}(0, \sigma_{ij}^2)$.

$$e_{\mathrm{A}} = \frac{1}{K}(\Phi_i - \Phi_j) + \frac{1}{K}\sum_{k \in R} \Phi_k + \langle \varepsilon_i - \varepsilon_j, G^\star \rangle + \frac{1}{K}(\varepsilon_i^\top \varepsilon_i - \varepsilon_j^\top \varepsilon_j),$$

$$e_{\mathrm{B}} = \frac{1}{K-2}\sum_{k \in R} \Phi_k.$$

Conditional distribution given ε_i. Because $(\Phi_k)_{k \in R}$ are independent of $\varepsilon_i, \varepsilon_j$,

$$e_{\mathrm{B}} \mid \varepsilon_i \sim \mathcal{N}\big(0,\, \sigma_{ij}^2/(K-2)\big).$$

For e_{A} define the *random shift*

$$m(\varepsilon_i) := \langle \varepsilon_i, G^\star \rangle + \frac{\varepsilon_i^\top \varepsilon_i}{K}.$$

Then

$$e_{\mathrm{A}} \mid \varepsilon_i \sim \mathcal{N}\Big(\frac{1}{K}\Phi_i + m(\varepsilon_i),\, \sigma_{ij}^2/(K-2)\Big).$$

Hence FM has non-zero conditional mean that depends on ε_i; L2O's conditional mean is zero.

Then fix $\Delta > 0$ (the $\Delta < 0$ case is symmetric). Define events

$$E_{\mathrm{A}} := \big\{\Delta + e_{\mathrm{A}} < 0\big\}, \qquad E_{\mathrm{B}} := \big\{\Delta + e_{\mathrm{B}} < 0\big\}.$$

L2O (no shift). With variance $s_{\mathrm{B}}^2 = \sigma_{ij}^2/(K-2)$,

$$\Pr(E_{\mathrm{B}}) = \Phi\Big(-\frac{\Delta}{s_{\mathrm{B}}}\Big), \qquad \Phi(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2} dt.$$

FM (random shift). Conditioning on ε_i,

$$\Pr(E_{\mathrm{A}} \mid \varepsilon_i) = \Phi\Big(-\frac{\Delta + m(\varepsilon_i) + \frac{1}{K}\Phi_i}{s_{\mathrm{B}}}\Big).$$

Take expectation over ε_i. Because $\varepsilon_i \to \Phi\big(-(\Delta + x)/s_{\mathrm{B}}\big)$ is convex in the shift x (the second derivative of Φ is positive on $(-\infty, 0)$), Jensen's inequality gives

$$\mathbb{E}_{\varepsilon_i}\big[\,\Phi\big(-(\Delta + m(\varepsilon_i) + \tfrac{1}{K}\Phi_i)/s_{\mathrm{B}}\big)\big] \;\geq\; \Phi\Big(-\frac{\Delta}{s_{\mathrm{B}}}\Big).$$

But the left side is exactly $\Pr(E_{\mathrm{A}})$ and the right side $\Pr(E_{\mathrm{B}})$, hence

$$\Pr(E_{\mathrm{A}}) \;\geq\; \Pr(E_{\mathrm{B}}),$$

which means that the FM mis-classification probability is no smaller.

Remark 1: The inequality is *strict* unless $m(\varepsilon_i) + \frac{1}{K}\Phi_i \equiv 0$ almost surely—which cannot happen when $\Sigma \succ 0$.