

Università degli Studi di Napoli
Federico II

FedLoRAX: Trustworthy Heterogenous LoRA aggregation for Federated LLM finetuning

Makhmoor Fiza

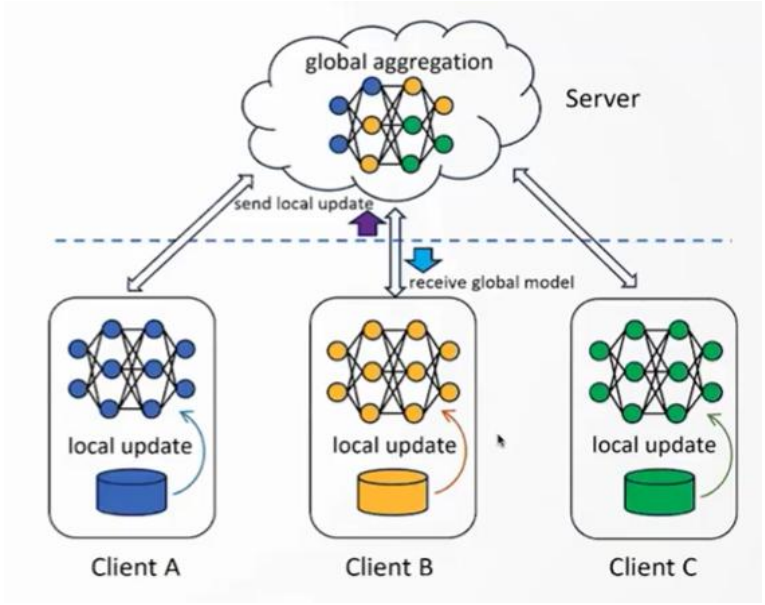
23-07-2026

Agenda

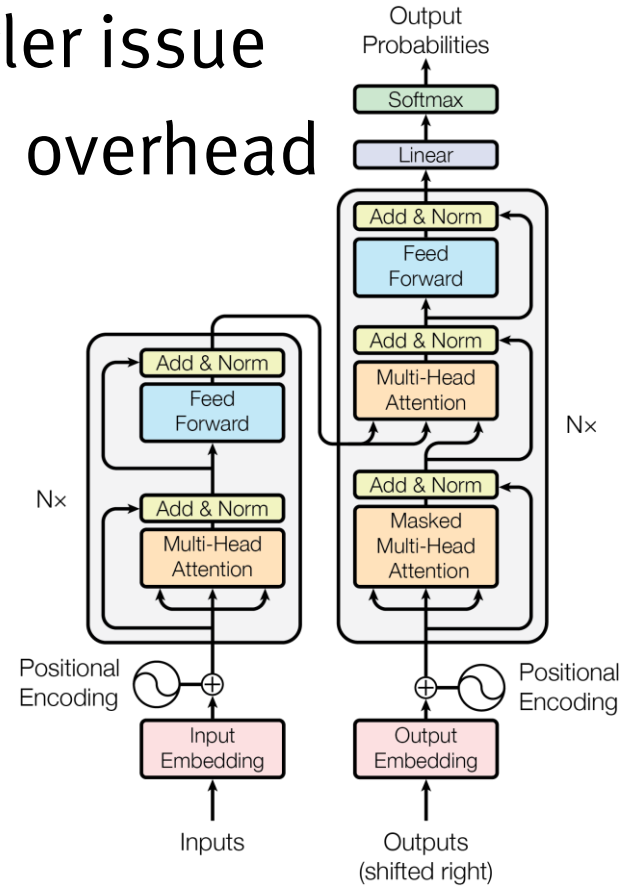
- Introduction
- Research Question
- Vanila aggregation
- Framework proposed
- Framework Proposed
- Experimentation
  - Results
- Future Work

Background

- Fixed model shared by all the parties
- Real world settings: varying resources , straggler issue
- Increasing model size leads to communication overhead
- Blackbox



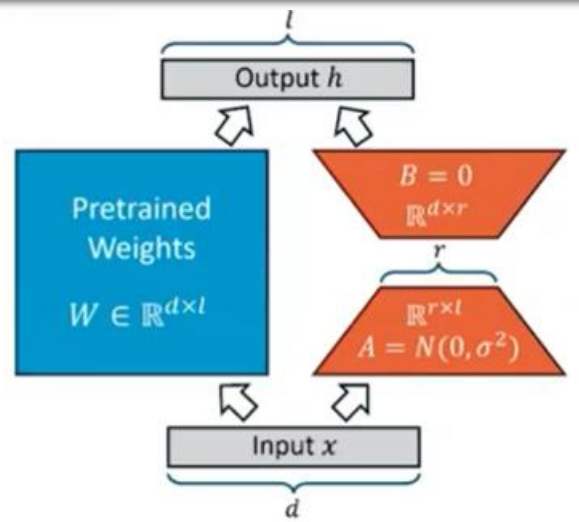


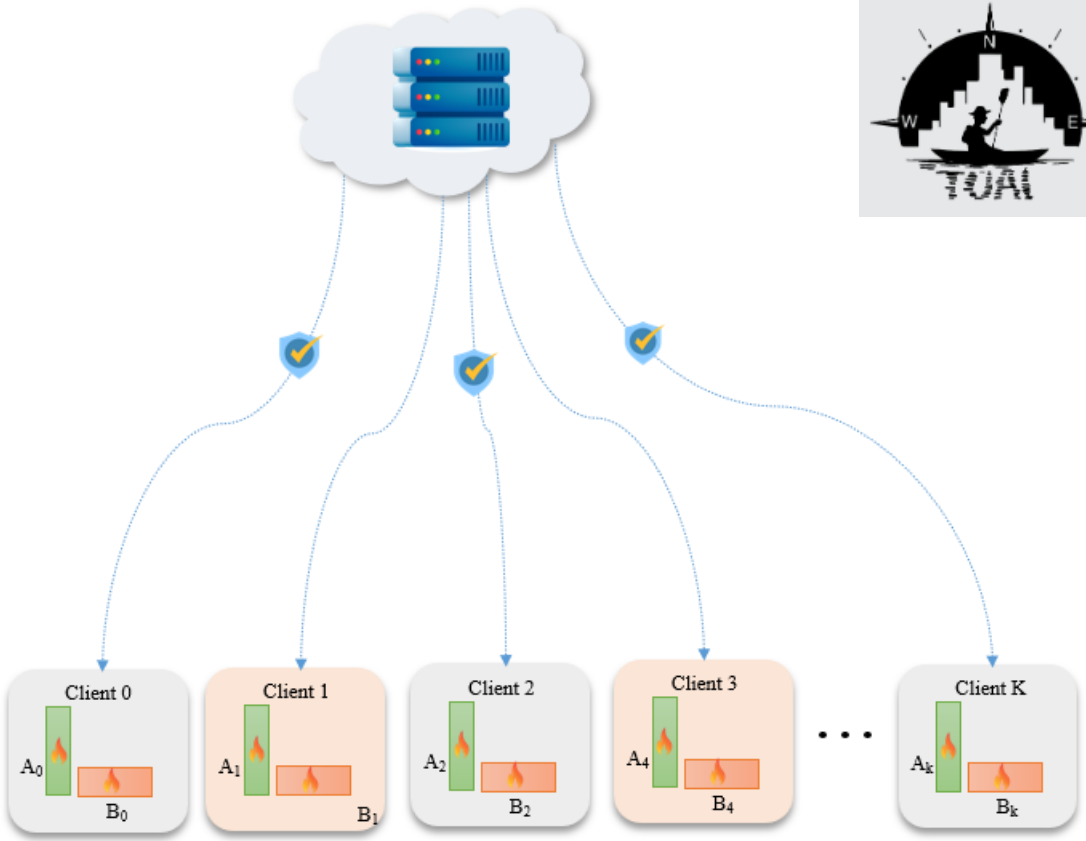


Preliminaries

- **Low rank Adaptation (LoRA)** (Hu et al. *ICLR 2022)*
  - A parameter-efficient finetuning method that only LoRA parameters are updated with the original model parameters remain frozen.

$$W = W_0 + \Delta W = W_0 + B \cdot A$$





$$\left(\sum_{k=1}^{K} \mathbf{A}_k\right) \times \left(\sum_{k=1}^{K} \mathbf{B}_k\right) \neq \sum_{k=1}^{K} (\mathbf{B}_k \mathbf{A}_k).$$

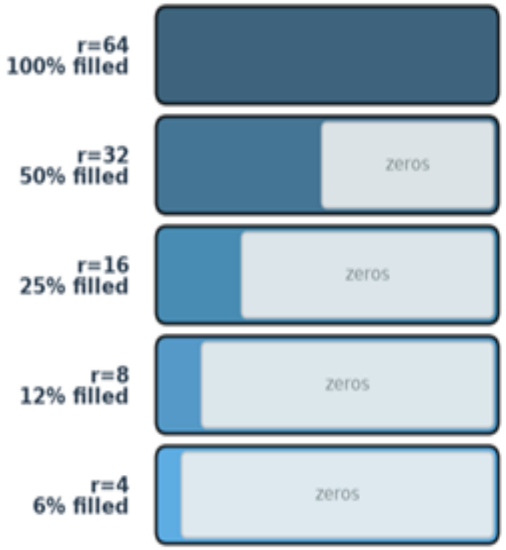
1:How to aggregate models with different structures ?

2: Can the client updates be distinguised on the basis of trust?

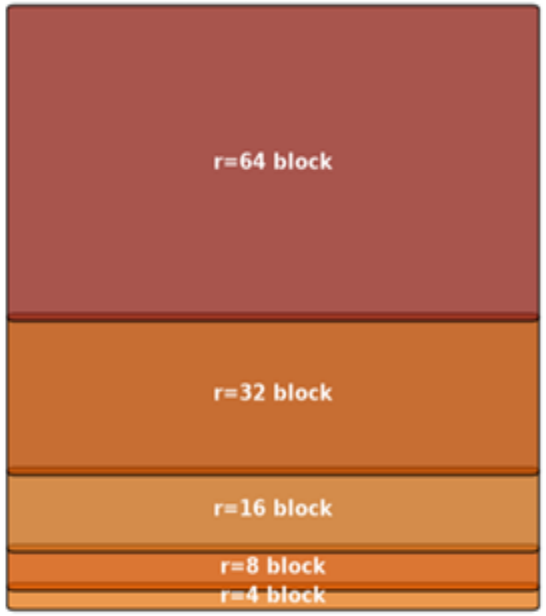
3: Can we detect and exclude anomalous or malicious clients?

Vanila Aggregation for Hetero LoRA

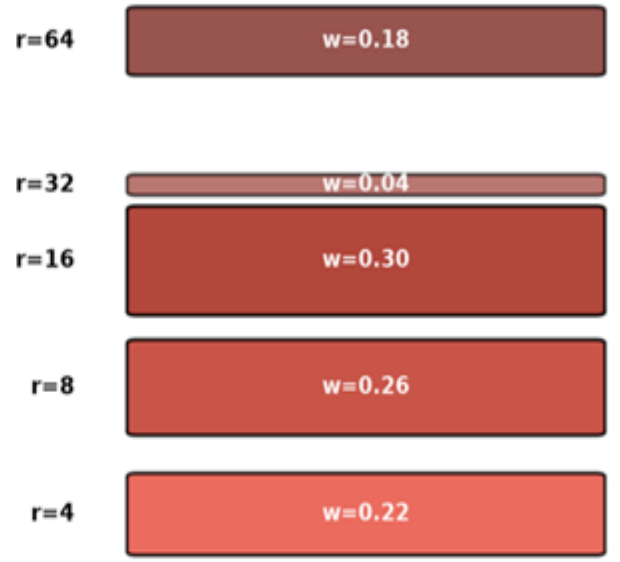
Y. J. Cho et al., "Heterogeneous LoRA for Federated Fine-tuning of On-Device Foundation Models," in *EMNLP*, 2024



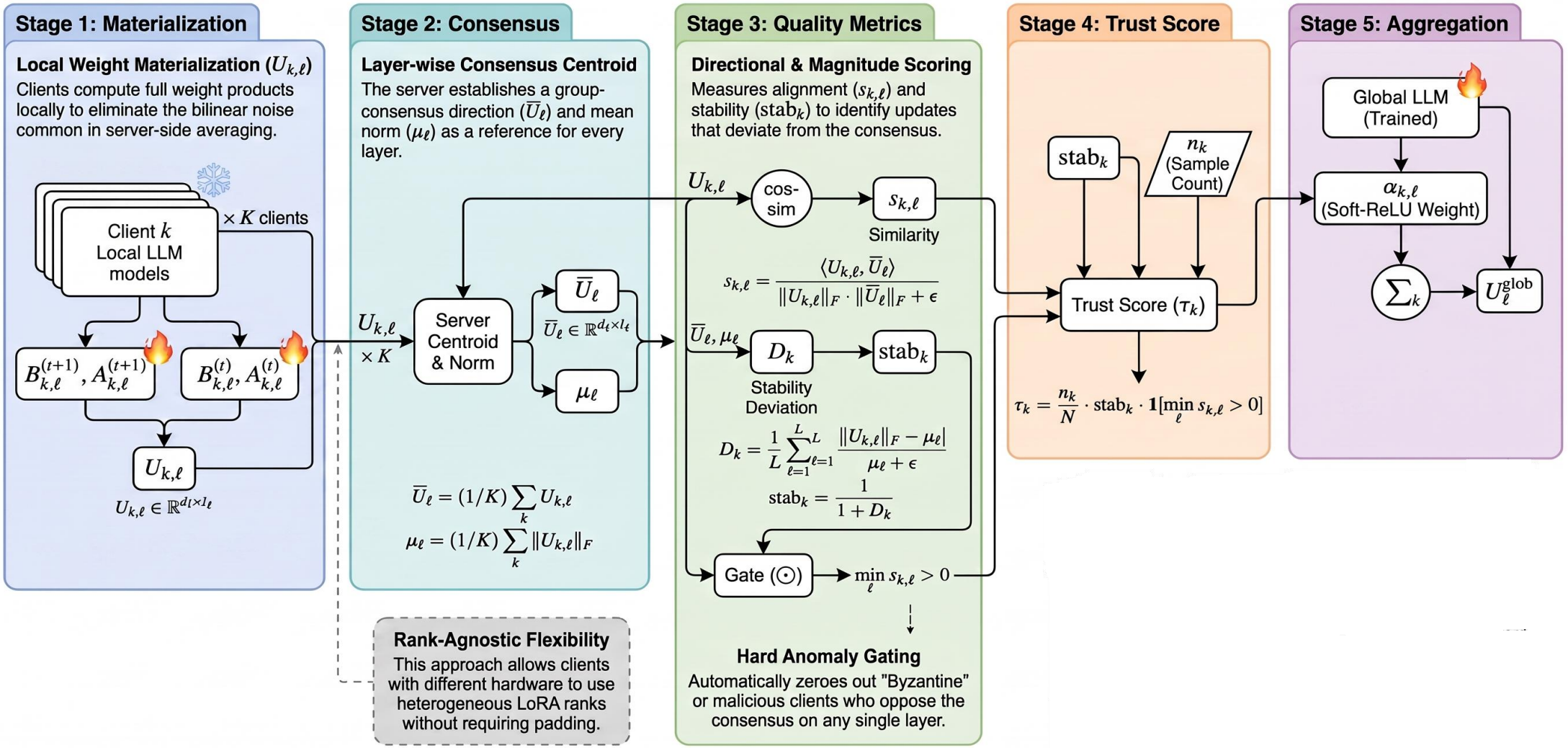
Wang et al. (NeurIPS 2025), "FLoRA: Federated Fine-Tuning Large Language Models with Heterogeneous Low-Rank Adaptations"



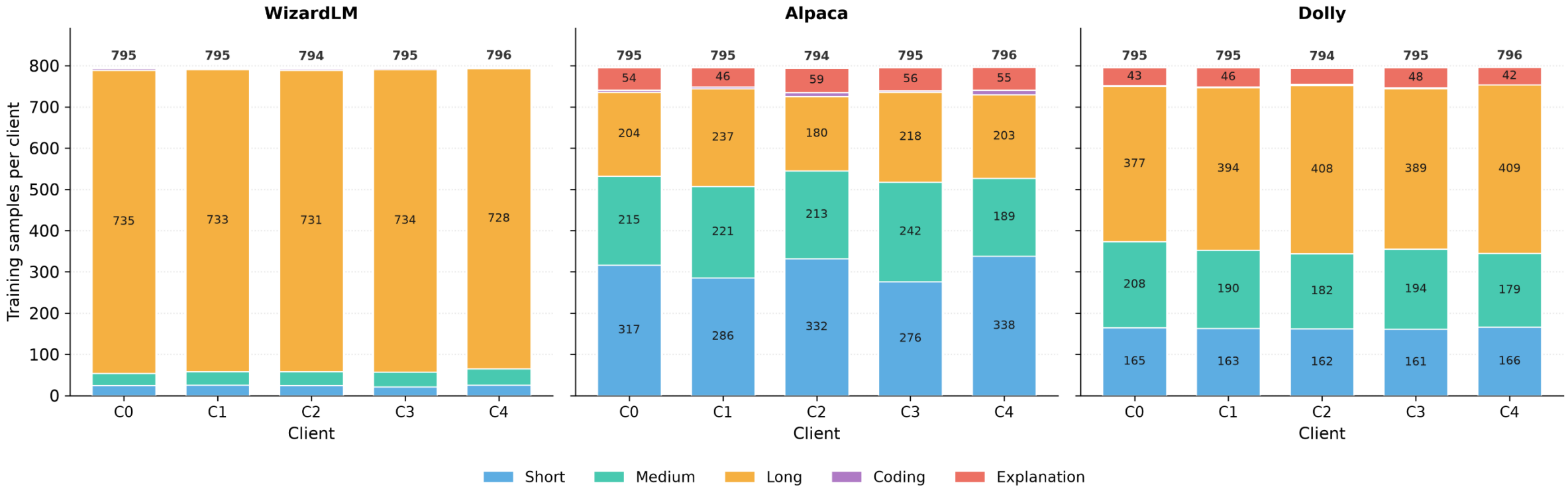
Proposed Idea



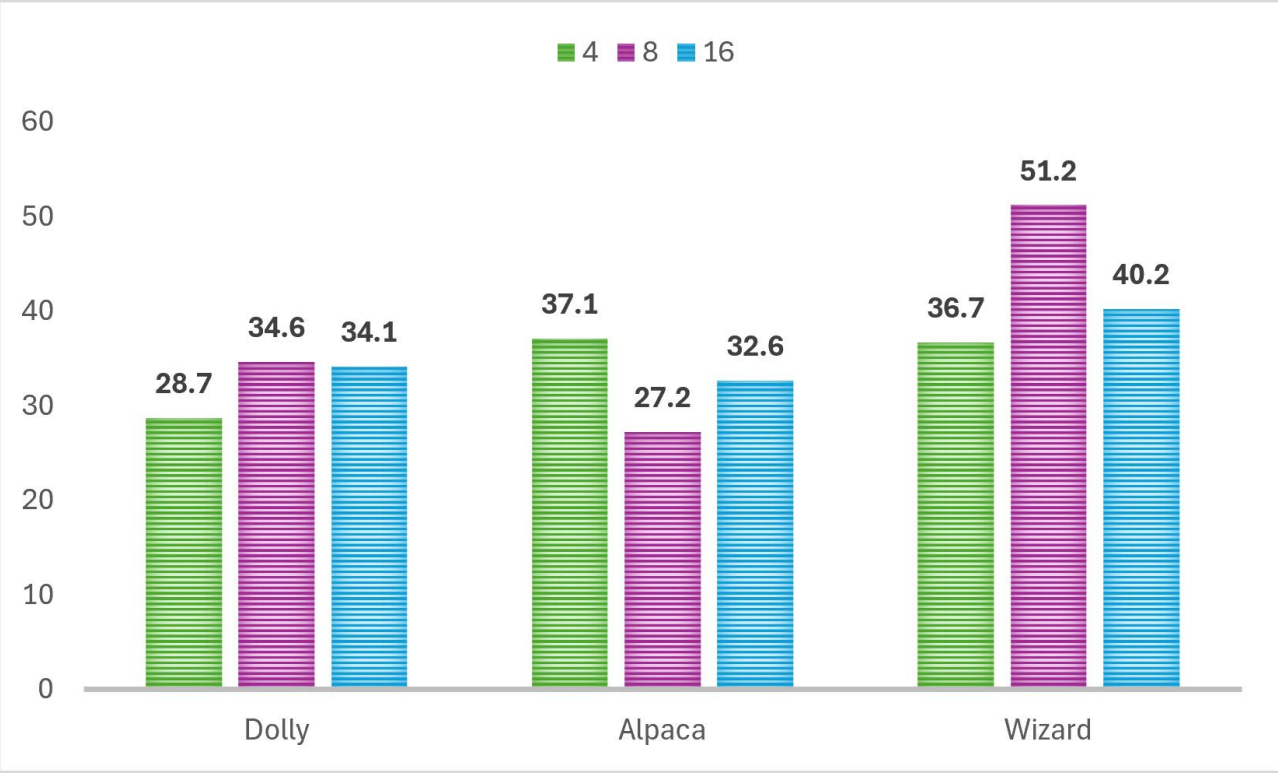
FedLoRA-XAI: Trustworthy Explainable Federated LLM Finetuning



Client data distribution

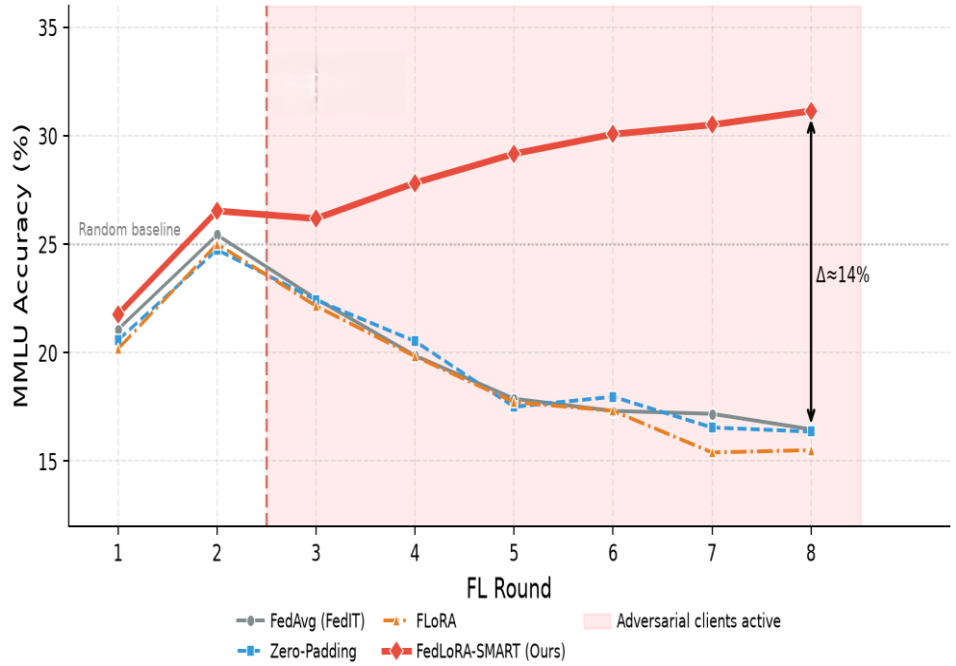


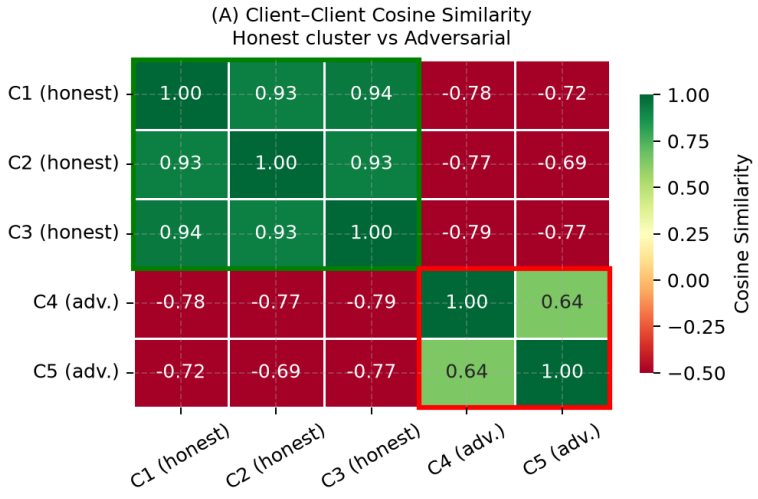
Results: MMLU Score

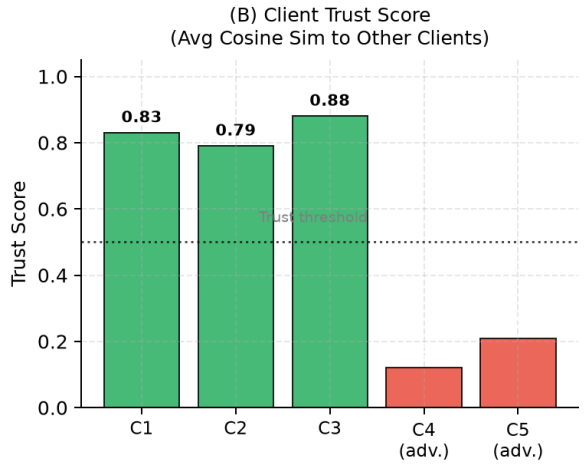


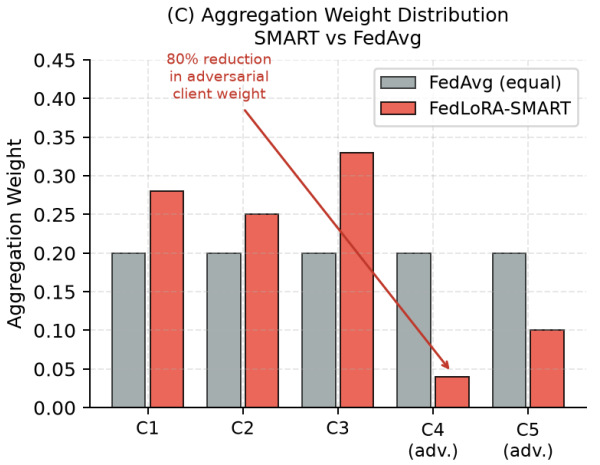
Setting	Method	Dolly	Alpaca	Wizard
Homo	FedIT	16.35	30.20	42.51
	FLoRA	30.80	31.92	43.87
	FedLoRAX	**34.60**	**37.10**	**51.20**
Hetero	Zero Padding	15.76	29.56	40.79
	FLoRA	18.45	29.69	41.48
	FedLoRAX	**28.97**	**29.76**	38.24

Byzntine attack









Illustrative schematic: per-client cosine / trust / weight values are stylized to show the mechanism. The measured robustness outcome is the 81% vs 70% MMLU retained under a 2/5-Byzantine attack (see the adversarial-robustness figure).

- Submitted:
  - **FedLoRAX: Trustworthy** Heterogenous **LoRA aggregation for Federated LLM finetuning**
- In progress:
  - Personalized Federated Learning for smart buildings.



Thank you for your attention

MODAL