A Framework for Double-Blind Federated Adaptation of Foundation Models

Nurbek Tastan ♠ Karthik Nandakumar ♠,♣
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) ♠ Michigan State University (MSU) ♣
♠ nurbek.tastan@mbzuai.ac.ae ♣ nandakum@msu.edu

Problem Definition

We consider a federated setting where:

- A **learning service provider (LSP)** holds a pre-trained FM.
- Multiple **data owners (clients)** hold task-specific labeled datasets.
- The goal is to **adapt the FM** for a downstream task collaboratively.

Double-Blind Constraints: (i) **Model Privacy:** Clients cannot access the FM. (ii) **Data Privacy:** The LSP cannot access the local data.

The goal is to jointly train a **side adapter** A_θ and **classification head** H_η to maximize performance while preserving privacy.

Contributions

- We introduce the **first framework for double-blind adaptation** of foundation models, using fully homomorphic encryption (FHE) and secure multi-party computation (MPC).
- We design a **modular adaptation pipeline**: (i) distilling the FM into FHE-friendly blocks (ii) interactive encrypted inference with a privacy-preserving permutation scheme (iii) local training using low-rank parallel adapters (iv) secure MPC-based aggregation.
- Strong performance on four datasets, robust under heterogeneity.

Methodology

Our pipeline consists of four stages:

- **FHE-friendly Distillation.** The FM distilled into FHE-compatible transformer blocks; nonlinear ops approximated via polynomials.
- **Encrypted Inference with Permutations.** Clients send encrypted inputs; server processes and permutes intermediate outputs.
- **Local Learning via Parallel Adapters.** Clients decrypt the permuted outputs and train adapters/classifier without FM backprop.
- **Secure Aggregation via MPC.** Server aggregates client updates securely without gradient leakage.

Privacy Guarantee: The protocol maintains double-blind privacy throughout the collaborative learning process.

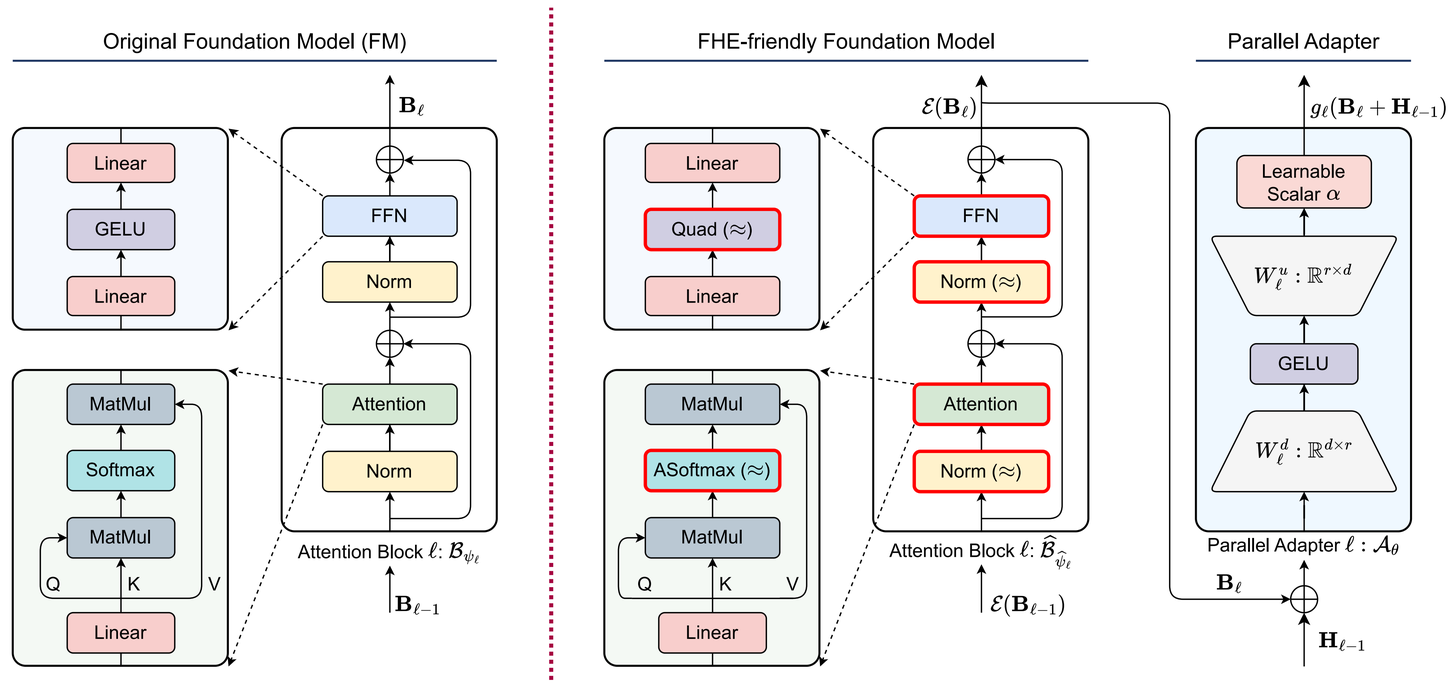
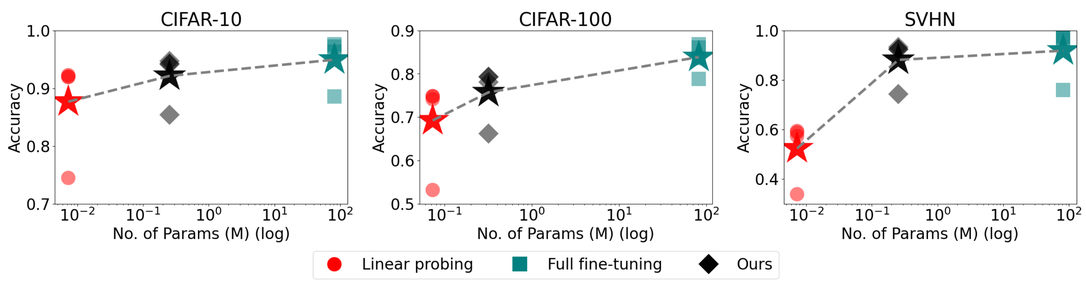


Illustration of the block decomposition and non-linear functions that need to be approximated (in red).

Experiments & Results

Accuracy Highlights. **Our method** outperforms linear probing, especially under strong heterogeneity. We achieve $\sim 94\%$ on CIFAR-10 with $K = 5$, close to full fine-tuning but with **300× fewer parameters.**



Accuracy vs. no. of trainable parameters trade-off (x-axis in log scale), illustrating the performance of the methods across three datasets.

Metric	Full fine-tuning	Ours	Linear probing
Trainable Parameters	82M	∼0.25M	<0.01M
Latency / Sample	47 ms	16 ms	15 ms
GPU Memory	18 GB	9 GB	9 GB

Efficiency metrics compared to Full fine-tuning.

Public dataset	Methods	Centralized	Federated
Fed-ISIC2019 (center=0) (InD)	Linear probing	0.6599	0.5856
	Full fine-tuning	0.7811	0.6752
	Ours	0.7090	0.6679
Tiny-Imagenet (OOD)	Linear probing	0.6372	0.5789
	Full fine-tuning	0.7817	0.6985
	Ours	0.7051	0.6481

Fed-ISIC2019 Results. Performance comparison of our method with baseline approaches on the Fed-ISIC2019 dataset with five clients, using two auxiliary datasets.

Methods	$K = 10$	$K = 20$	$K = 50$
Linear probing	0.9167	0.9142	0.9007
Full fine-tuning	0.9739	0.9513	N/A
Ours	0.9446	0.9422	0.9287

Scalability analysis of the proposed method to baseline approaches on CIFAR-10 under a Dirichlet parameter of 1.0 for data partitioning.

Takeaways

- **First double-blind FL framework** enabling adaptation of FMs while protecting both model and data.
- **Secure, scalable, and practical:** achieves strong results on 4 datasets with up to 50 clients.
- **Efficient:** requires $\sim 300\times$ fewer parameters than full fine-tuning.

Paper Link Webpage

Nurbek Tastan
MBZUAI – PhD Candidate in Machine Learning
Abu Dhabi, United Arab Emirates