Efficient Representation Learning for Heterogeneous Systems and Continual Task Adaptation

Fatema Siddika[1] (PhD Candidate), **Ali Jannesari**[1] (PhD Advisor)

Collaborators: Md Anwar Hossen[1], Wensheng Zhang[1], Anuj Sharma[1], Tanwi Mallick[2], Ravi Madduri[2], Zilinghan Li[2] J. Pablo Muñoz[3], Tanya Roosta[4]

[1]Iowa State University [2]Argonne National Laboratory [3]Intel Labs [4]UC Berkeley/Amazon/AMD

IOWA STATE UNIVERSITY

Department of Computer Science

Dissertation Statement

Large models can learn continually across heterogeneous, privacy-constrained clients when adaptation is confined to compact components that separate shared from task- and client-specific knowledge. Sparse experts in parameter space, and interventions and prototypes in representation space, provide this separation, letting models learn new tasks without forgetting, share knowledge without sharing data, and exchange only small updates that combine without interference. This dissertation aims to enable large models to keep learning after deployment, adapting to new tasks and new data across institutions that cannot share data, without retraining from scratch or forgetting prior knowledge.

This thesis studies and optimizes adaptation across three axes of heterogeneity:

- 4 studies: FedProtoKD, FedReFT, SETA, FedSEAM
- 4 mechanisms: adaptive prototype margins, All-But-Me ReFT aggregation, sparse expert splitting, subspace expert agreement merging

Challenges

- Heterogeneous clients: FedAvg averages misaligned updates; prototype margins shrink and classes collapse
- PEFT in FL: weight-space LoRA misses semantics; naïve aggregation breaks under task heterogeneity
- Sequential tasks: new learning overwrites old parameters, catastrophic forgetting in a shared subspace
- Federated continual learning: client experts learned on private task streams point in conflicting directions, so naïve merging erases client-specific and prior-task knowledge.

Our Solutions

- Trainable global prototypes with adaptive class-wise margins learned via contrastive generation (FedProtoKD)
- Federate representation fine-tuning: share only low-rank intervention parameters {W, R, b} with All-But-Me aggregation (FedReFT)
- Gradient-aware block-sparse experts that structurally split shared vs. task-specific knowledge (SETA)
- Subspace expert agreement merging: combine client experts only along shared SVD directions, keep disagreeing directions as private experts, and protect prior-task subspaces (FedSEAM).

Unified Thesis Framework

Frozen LLM backbone: all adaptation in representation space

Low-rank interventions: $\Phi(h) = h + R^T(Wh + b - Rh)$

FedProtoKD	FedReFT	SETA
Client heterogeneity	Federated tasks	Sequential tasks

FedSEAM: federated × continual, unified

Publications & CV

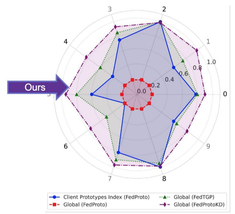
1. **FedReFT: Federated Representation Fine-Tuning with All-But-Me Aggregation.** EACL 2026
2. **Dual-Distilled Heterogeneous FL with Adaptive Margins for Trainable Global Prototypes.** IEEE CCGrid 2026.
3. **Split-on-Share: Mixture of Sparse Experts for Task-Agnostic Continual Learning.** Under Review
4. **FedSEAM , Federated Continual Learning of LLMs via Subspace Expert Agreement Merging.** Under Review
5. CV- Scan the QR code.

SwAPP Lab - Scan the QR code

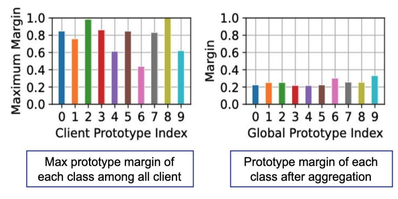
FedProtoKD: Heterogeneous FL

Trainable global prototypes with adaptive class-wise margins • IEEE CCGrid 2026

Problem: prototype margin shrink; weighted aggregation of client prototypes shrinks inter-class margins, so the global model becomes semantically confused.

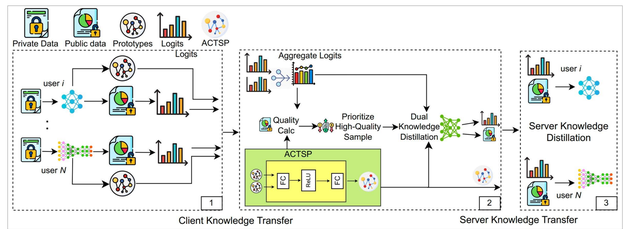


Prototype margin changes after generating global prototypes.

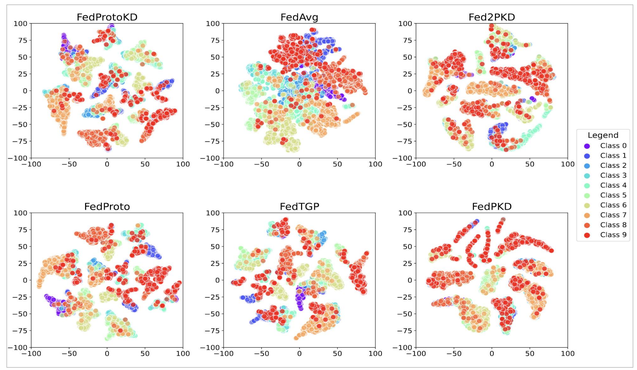


Max prototype margin of each class among all client

Prototype margin of each class after aggregation



FedProtoKD pipeline: dual distillation + contrastive prototype generation



Class prototypes stay separable; margins are enhanced rather than shrunk

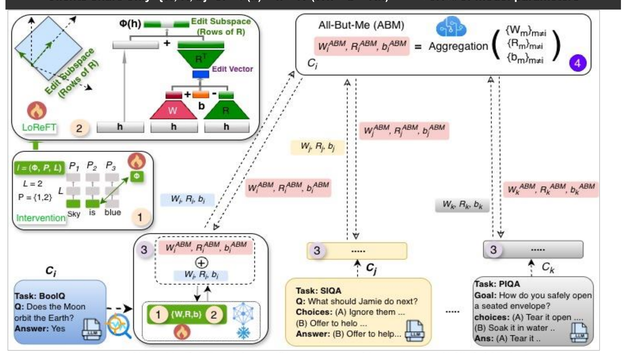
Key Insight: class geometry must be learned, not averaged, adaptive margins keep heterogeneous clients semantically aligned.

FedReFT: Federated ReFT

Representation fine-tuning with All-But-Me aggregation • EACL 2026

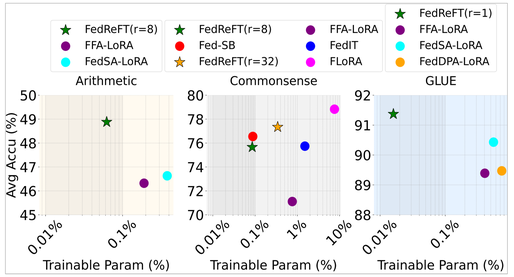
- **Problem:** full fine-tuning is infeasible on edge clients; PEFT: weight-space updates miss semantic richness, Representation Finetuning directly manipulates representations, outperforms PEFT

Clients share only {W, R, b} of $\Phi(h) = h + R^T(Wh + b - Rh)$ — < 0.1% of model parameters

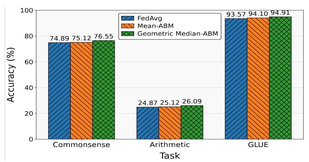


(1) local LoReFT updates → (3) server geometric median → (4) All-But-Me fusion

All-But-Me (ABM)	Geometric median	Test-Time Compute
each client receives an aggregate of the other clients, no self-averaging, personalization preserved	L_1 aggregation gives outlier clients near-zero weight; robust where FedAvg is pulled off-center	each client learns its own mixing weight α (mean ≈ 0.57) between local and global knowledge



Accuracy vs. trainable parameters (%): Arithmetic, Commonsense, GLUE benchmarks



Outlier clients receive near-zero weight automatically. FedAvg gets corrupted; Geometric Median stays anchored near the honest majority.

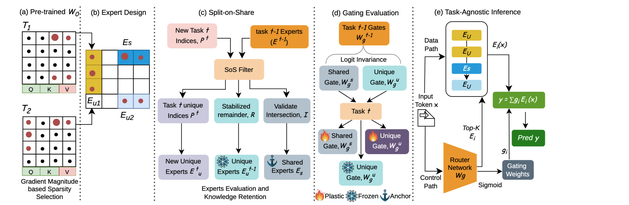
Key Insight : representations, not weights, are the right unit to federate, semantically aligned, heterogeneity-robust, and radically parameter-efficient.

SETA: Split-on-Share Sparse Experts

Task-agnostic continual learning for LLMs • Under Review

Problem: Sequential tasks compete for overlapping parameter subspaces: new-task gradients overwrite shared weights encoding prior knowledge, causing catastrophic forgetting.

Gradient magnitude reveals which parameter sub-blocks matter per task, so shared vs. task-specific knowledge can be separated structurally, not just penalized by a loss.



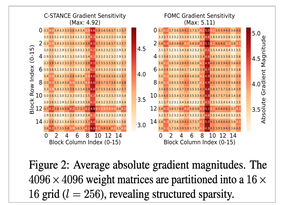


Figure 2: Average absolute gradient magnitudes. The 4096×4096 weight matrices are partitioned into a 16×16 grid ($l = 256$), revealing structured sparsity.

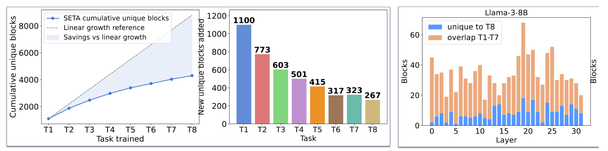
Gradient-aware attention block selection across layers

- **RQ1** How can sparse parameter patterns decouple new feature acquisition from historical retention? → *Subspace-based expert decomposition using gradient-magnitude block selection*
- **RQ2** Can a single framework automatically distinguish shared vs. task-specific knowledge? → *Split-on-Share (SoS) mechanism + Adaptive Elastic Anchoring*
- **RQ3** How can relevant experts be retrieved at inference without task IDs? → *Adaptive Gating Expansion with Logit Invariance*

Block-sparse experts	Split-on-Share	Task agnostic gating
top-k gradient blocks form each task's expert subspace	overlapping blocks split into shared vs. unique experts + elastic anchoring	adaptive gating expansion with logit invariance retrieves experts at inference

Table 2: Continual learning performance across a eight-task sequence using LLaMA-3 8B, LLaMA-2 7B, and Qwen3-4B. Performance is reported as the Learning Accuracy (LA), Overall Performance (OP), and Backward Transfer (BWT) after sequential training. Higher OP (↑) and BWT (↑) indicate better final task accuracy and less negative transfer. **Bold** values indicate the best performance in each column, underlined values denote the second-best.

LLaMA-3-8B: Continual Learning on TRACE Benchmarks									Metrics		
Method	C-STANCE	FOMC	MeetingBank	ScienceQA	NumGLUE-cm	20Minuten	Py150	NumGLUE-ds	LA ↑	OP ↑	BWT ↑
Seq-LoRA	1.5	1.8	2.0	33.1	42.0	39.5	44.8	58.5	53.31	27.90	−29.04
EWC	33.0	51.2	4.3	45.7	44.4	39.3	46.4	**59.7**	**54.08**	40.50	−15.51
A-GEM	33.3	47.0	6.0	64.3	40.7	**39.8**	48.1	54.8	53.60	41.75	−13.54
O-LoRA	6.4	11.1	14.2	72.7	30.9	37.5	13.0	45.8	51.00	28.95	**−25.20**
I-LoRA	13.1	21.6	10.9	14.1	33.3	39.2	47.6	56.3	53.28	29.51	−27.16
HiDe-LoRA	17.7	23.0	13.1	**72.6**	18.5	38.1	11.9	5.5	44.79	25.05	−22.56
SETA (ours)	**48.4**	**60.3**	**35.6**	63.4	**50.0**	38.6	**48.4**	59.4	53.78	**50.52**	−5.96
LLaMA-2-7B: Continual Learning on TRACE Benchmarks									**Metrics**		
Method	C-STANCE	FOMC	MeetingBank	ScienceQA	NumGLUE-cm	20Minuten	Py150	NumGLUE-ds	LA ↑	OP ↑	BWT ↑
Seq-LoRA	2.0	2.6	0.7	4.4	14.8	38.4	38.2	**48.6**	43.30	18.71	−28.10
EWC	33.1	48.8	4.0	40.8	14.8	39.1	42.1	46.2	44.96	33.61	−12.97
GEM	33.2	43.1	2.5	32.8	13.6	39.0	**43.4**	47.1	44.38	31.84	−14.33
O-LoRA	30.2	35.1	10.1	54.0	17.3	35.0	11.7	35.7	44.50	28.60	**−18.10**
I-LoRA	18.0	43.8	11.1	19.8	23.5	**39.3**	38.4	45.5	43.45	29.93	−15.46
HiDe-LoRA	23.8	26.2	10.2	45.0	11.1	38.1	15.2	0.0	31.36	21.20	−11.61
SETA (ours)	**43.6**	**57.4**	**30.3**	**60.2**	**28.6**	38.7	43.2	47.4	**49.25**	**43.67**	−6.38
Qwen3-4B: Continual Learning on TRACE Benchmarks									**Metrics**		
Method	C-STANCE	FOMC	MeetingBank	ScienceQA	NumGLUE-cm	20Minuten	Py150	NumGLUE-ds	LA ↑	OP ↑	BWT ↑
Seq-LoRA	0.0	39.0	4.8	60.0	51.9	38.3	**49.7**	**57.0**	54.30	37.59	−19.10
EWC	0.5	43.0	5.8	63.5	**54.3**	39.3	47.1	56.0	**54.76**	38.69	−18.37
GEM	0.0	54.0	6.8	70.0	48.1	**39.8**	47.0	56.0	53.36	40.21	−15.03
O-LoRA	18.0	22.0	11.7	**85.0**	13.6	38.0	10.2	0.0	31.18	24.81	−7.27
I-LoRA	8.0	20.0	10.4	66.0	30.9	39.3	49.1	**57.0**	53.44	35.09	−20.97
HiDe-LoRA	31.0	25.0	12.6	**85.0**	19.8	38.0	9.9	0.0	37.79	27.66	−11.57
SETA (ours)	**47.4**	**59.1**	**25.9**	73.8	38.5	37.2	48.9	56.3	53.69	**48.40**	**−6.05**

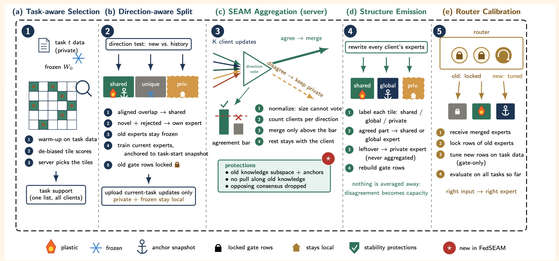


(A) cumulative count against a linear growth; (B) new unique blocks contributed by each task;(C) Per-layer block composition for the final task T8

Key Insight : sparsity is structure, splitting shared from unique subspaces decouples plasticity from stability.

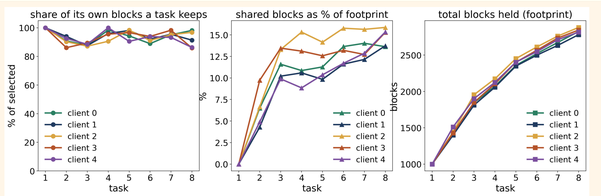
- **Problem:** conflicting expert merge: averaging client experts trained on private task streams mixes update directions that disagree, which cancels client-specific knowledge and overwrites prior-task subspaces, so the global model forgets.

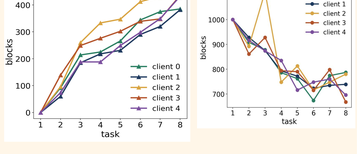
Subspace agreement	Shared vs. private	Prior-task protection
per-block SVD subspaces from each client; merge only directions whose agreement passes a chance-calibrated threshold	agreed directions form global experts; disagreeing directions stay as client-private experts	prior-task subspaces shielded from overwrite, with local router calibration on each client



FedSEAM: Federated SEAM

Federated Continual Learning for LLMs Subspace Expert Agreement Merging • Under Review





Key Insight : agreement is geometric; merging client experts only where their subspaces align lets clients gain global knowledge without erasing local or prior-task knowledge.

Acknowledgments: Joint work with Md Anwar Hossen (equal contribution on FedReFT), Ali Jannesari (advisor, Iowa State University),Wensheng Zhang, and Anuj Sharma (Iowa State University); Tanwi Mallick (Argonne National Laboratory); J. Pablo Muñoz (Intel Labs); and Tanya Roosta (UC Berkeley / Amazon/AMD). FedSEAM Ravi Madduri, **Zilinghan Li**, (Argonne National Laboratory)

References: 1. Wu et al. ReFT: Representation Fine-Tuning for Language Models. NeurIPS 2024.