FLAME: FEDERATED LAPLACE-ALIGNED MIXTURE OF EXPERTS FOR PERSONALIZED LLM FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized federated fine-tuning of large language models calls for adaptation to heterogeneous client tasks together with reliable uncertainty estimates under limited local data. Existing personalized federated low-rank adaptation (LoRA) methods expose only a single level of cross-client locality, average the LoRA factors A and B identically despite their structural asymmetry, and return *point* estimate adapters whose predictions remain poorly calibrated. To address these limitations, we propose **FLAME**, a **F**ederated **L**aplace-**A**ligned **M**ixture of **E**xperts whose three design choices are each grounded in the geometry of the learned parameter space. To capture hierarchical locality across client adaptations, FLAME jointly routes among self, intra-cluster, and cross-cluster experts, with the trainable individual expert active in every forward pass. To resolve the LoRA basis ambiguity that Euclidean aggregation introduces, FLAME aligns the input-side factor A on the Grassmann manifold and keeps the output-side factor B in Euclidean space, where the entry-level task-specific signals live. For *calibrated* per-client predictive distributions, FLAME fits a local post-hoc Kronecker-factored Laplace posterior over each client's individual expert, leaving training and communication untouched. We further establish a non-asymptotic convergence guarantee for FLAME and show that the Grassmannian aggregator tightens the aggregation-drift more than Euclidean averaging. Experimental results across 12-client MCQA classification benchmark and 8-client FLAN NLG benchmarks spanning IID, cross-dataset, and label-skew heterogeneity show that FLAME consistently improves both accuracy and calibration (ECE, NLL) over all baselines.

1 INTRODUCTION

Modern data sources such as hospitals, banks, and edge devices generate text at scale, yet privacy regulations and institutional policies prevent these data from being centralized for training (McMahan et al., 2017). Federated learning (FL) addresses this constraint by allowing collaborative training across distributed clients while keeping raw data local (McMahan et al., 2017; Tan et al., 2022). Deploying FL on top of large language models (LLMs) (Devlin et al., 2019; Brown et al., 2020; Touvron et al., 2023) is a natural next step, since adapting LLMs to specialized domains typically relies on fine-tuning with task-specific data (Minaee et al., 2024) that is itself federated by nature. Full fine-tuning of an LLM on every client is, however, computationally prohibitive and incurs per-round communication on the order of the entire backbone, which has motivated the federated adoption of parameter-efficient fine-tuning (PEFT) methods (Ding et al., 2023). Among PEFT methods, low-rank adaptation (LoRA) (Hu et al., 2022; Li et al., 2026) has become the de facto standard: LoRA inserts trainable low-rank matrices A and B alongside frozen pretrained weights and updates far fewer parameters per adapted layer, reducing both local compute and per-round communication by orders of magnitude relative to full fine-tuning, which makes it the natural adapter to use under FL constraints (Zhang et al., 2024; Kuang et al., 2024; Ye et al., 2024; Fan et al., 2023).

This federated LoRA setup introduces a complication absent from centralized fine-tuning: in realistic LLM deployments, clients often hold data from genuinely different downstream tasks rather than mere label shifts of a shared task, and a single globally averaged LoRA degrades per-client performance under such task heterogeneity. *Personalized* federated LoRA addresses this by allowing each client to retain some adapter state, with FedDPA (Yang et al., 2024) blending a global LoRA with a local LoRA at inference, FedALT (Bian et al., 2026) freezing the rest-of-world (RoW) LoRA during local training and routing it against a per-client individual LoRA, and FedLEASE (Wang et al., 2026)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

replacing the single RoW adapter with a clustered mixture of experts. Despite their different routing structures, the server-side aggregator of these methods remains FedAvg-style and averages A and B identically in Euclidean space.

A closer look at the trained federated LoRA parameter space shows that the server-side aggregator is poorly matched to the geometry it operates on. Three structural properties go unexploited by current designs, each raising a research question for personalized federated LoRA. We measure all three properties directly on the trained federated LoRA parameter space in Section 2.2, and preview them here as *locality structure*, *gauge ambiguity and asymmetry*, and *calibration gap*. *(1) Locality structure.* Existing personalized routers expose only a single level of cross-client locality, whether binary (Yang et al., 2024; Bian et al., 2026) or per-cluster (Wang et al., 2026), while the trained A-factors carry non-redundant locality at both intra-cluster and cross-cluster scales, with row-space structure shared among intra-cluster peers (Collins et al., 2021) but absent across cluster boundaries. How should an architecture expose both scales jointly with each client's self-adaptation, and how should the cluster count be set without manual tuning? *(2) Gauge ambiguity and asymmetry.* The shared aggregator is doubly mismatched to LoRA. The factorization (Hu et al., 2022) is non-unique up to an invertible change of basis on the rank-r space, so independently trained clients converge to different factor pairs and Euclidean averaging drifts in this null space (Sun et al., 2024; Meng et al., 2026). The two factors also carry the signal asymmetrically (Tian et al., 2024; Zhang et al., 2023), and the asymmetry *inverts* across scales. Shared row-space structure across intra-cluster peers lives on A, whereas task identity is encoded in the entries of B. What is the right aggregation operator on each factor, and which factor should drive the clustering? *(3) Calibration gap.* Architecture and aggregation aside, per-client adaptation on limited local data is the regime where calibrated predictive uncertainty matters most for selective prediction and downstream decision-making (MacKay, 1992; Wilson & Izmailov, 2020), yet uncertainty quantification has remained largely unexplored for personalized federated LLMs, with existing federated UQ work (Achituve et al., 2021; Kassab & Simeone, 2022) targeting sub-LLM-scale models. Centralized Bayesian LoRA methods do exist (Daxberger et al., 2021; Ritter et al., 2018), though they either modify the training loop (Wang et al., 2024a; Yang et al., 2023) or add per-step memory cost (Wang et al., 2023; Cong et al., 2024) that does not scale across federated clients. Can a Bayesian recipe deliver calibrated uncertainty for personalized federated LLMs while leaving training or communication intact?

We propose **FLAME** (**F**ederated **L**aplace-**A**ligned **M**ixture of **E**xperts), a personalized federated LoRA framework with three components, each matched to one of the properties above. For locality, FLAME assigns each client a trainable individual adapter together with a set of frozen experts at the intra-cluster and cross-cluster scales, and routes among them while the trainable individual adapter takes part in every forward pass. To address the gauge ambiguity on A and the A/B asymmetry jointly, FLAME applies a *selective Grassmannian aggregator* that aggregates A through a gauge-invariant Fréchet mean on the rank-r Grassmannian and B retain the Euclidean mean where the entry-level task signal lives. For calibration, FLAME fits a post-hoc Kronecker-factored (KFAC) Laplace posterior over each client's individual adapter, storing per-layer curvature in two compact Kronecker factors and leaving training and per-round communication untouched. Figure 3 contrasts FLAME with representative baselines along locality, aggregation geometry, and uncertainty treatment.

Our contributions are summarized as follows.

- We identify three structural properties of the trained federated LoRA parameter space that existing personalized federated LoRA designs leave unexploited: cross-client locality at three non-redundant scales, an A/B asymmetry that inverts between subspace and entry scales, and a calibration gap that persists across routing architectures.
- We propose FLAME with three matched components: (i) a three-granularity (self / intra-cluster / cross-cluster) expert hierarchy with guaranteed per-token participation of the individual adapter; (ii) a *selective Grassmannian aggregator* resolving LoRA's gauge ambiguity on A via a Fréchet mean on the rank-r Grassmannian while keeping Euclidean averaging on B; (iii) a per-client post-hoc KFAC Laplace posterior fitted without extra communication.
- We provide a non-asymptotic convergence analysis for FLAME's training in which each design component maps to an identifiable term of the bound, with the Grassmannian aggregator strictly tightening the aggregation-drift term over Euclidean averaging.
- We show on a 12-client MCQA classification benchmark and three 8-client FLAN NLG benchmarks under varying heterogeneity that FLAME outperforms competitive baselines in every setting, with the calibration improvement unmatched by scaling the LoRA rank of a point-estimate baseline.

2 PRELIMINARIES AND MOTIVATIONS

2.1 LORA AND FEDERATED LORA

Consider a federated setting with N clients. Each client n ($n = 1, \ldots, N$) possesses a local dataset $\mathcal{D}_n$ that may originate from a different downstream task. All clients share a frozen pretrained LLM whose weight at every adapted linear layer is denoted $W^0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, where d_{in} and d_{out} are the input and output dimensions. LoRA (Hu et al., 2022) fine-tunes a low-rank residual $\Delta W = BA$ with $A \in \mathbb{R}^{r \times d_{\text{in}}}$, $B \in \mathbb{R}^{d_{\text{out}} \times r}$, and $r \ll \min(d_{\text{in}}, d_{\text{out}})$, so that the adapted weight is $W = W^0 + BA$. The input-side factor A is initialized by Kaiming sampling, while the output-side factor B is initialized to zero, so that the residual at the start of training is exactly zero. The personalized federated fine-tuning objective is to learn a personalized adapted model $W_n = W^0 + B_n A_n$ that fits $\mathcal{D}_n$,

$$\min_{\{(A_n, B_n)\}_{n=1}^N} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_n\big(A_n, B_n; \mathcal{D}_n\big), \tag{1}$$

where $\mathcal{L}_n$ is the local task loss for client n. The standard FedAvg (McMahan et al., 2017) aggregator returns

$$\bar{A} = \frac{1}{N} \sum_{n=1}^{N} A_n, \qquad \bar{B} = \frac{1}{N} \sum_{n=1}^{N} B_n. \tag{2}$$

The aggregator in equation 2 has two structural issues. Averaging A and B independently does not commute with the matrix product, so $\bar{B}\bar{A} \neq \frac{1}{N}\sum_n B_n A_n$ in general, which introduces a bilinear aggregation error that is absent when full-weight FedAvg is used (Meng et al., 2026). The factorization $\Delta W_n = B_n A_n$ is also non-unique up to a change of basis on the rank-r latent space: for any invertible $R \in \mathrm{GL}(r)$, where $\mathrm{GL}(r)$ is the group of $r \times r$ invertible real matrices, the pair $(RA_n, B_n R^{-1})$ yields the same effective update $(B_n R^{-1})(RA_n) = B_n A_n$. Two clients can therefore represent the same ΔW_n with completely different (A, B) pairs related by some such R, and Euclidean averaging of either factor across clients drifts in this gauge null space, a phenomenon we refer to as *decomposition drift*. These two issues, together with the asymmetric initialization (Kaiming for A, zero for B) that drives qualitatively different update trajectories during local training, indicate that A and B are not interchangeable under aggregation, and that any aggregator applying the same operator to both inherits these errors. To decide whether and how the two factors should be aggregated differently, we examine the structure that emerges in each from the trained federated LoRA parameter space, leading to the three observations of Section 2.2 that motivate FLAME's design.

2.2 OBSERVATIONS

We report three observations on the trained federated LoRA parameter space that motivate FLAME's design. The studies use the MCQA-12c benchmark (defined in Section 5.1), with $N = 12$ clients spanning 4 downstream tasks at 3 clients per task and LoRA rank $r = 8$ at every adapted linear layer. Each client first runs one round of local-only fine-tuning of its own (A_n, B_n) on $\mathcal{D}_n$ without aggregation, and the trained factors are then analyzed.

Observation 1: A carries shared structure at the subspace scale only. We measure two properties of the trained A-factors against the random rank-r baseline. The largest principal angle between client A-subspaces (Figure 1(a)) shows tight alignment at the intra-cluster scale and substantially weaker alignment across distinct task clusters. The eigengap ratio λ_r/λ_{r+1} of the average A-subspace projector (Figure 1(b)), where a larger ratio indicates a better-separated rank-r subspace, is well-resolved at the intra-cluster scale and collapses by one to two orders of magnitude at both cross-cluster and global peer sets. The per-layer cosine similarity of A at the entry scale (Figure 1(c), A panel) shows no separation between same-task and different-task client pairs at any layer. The signal on A therefore lives in its row space, and only at the intra-cluster scale does that row space line up across peers. A Grassmannian aggregator on A at the cluster level matches this geometry directly.

Observation 2: B carries identifying signal at the entry scale only. We apply the same two measurements to B. At the subspace scale (Figure 1(a), B bars), the largest principal angle between client B-subspaces sits at or beyond the random rank-r baseline at both intra-cluster and cross-cluster peers, leaving the row space of B statistically indistinguishable from a random subspace and offering nothing for an alignment step to exploit. At the entry scale (Figure 1(c), B panel), the per-layer cosine

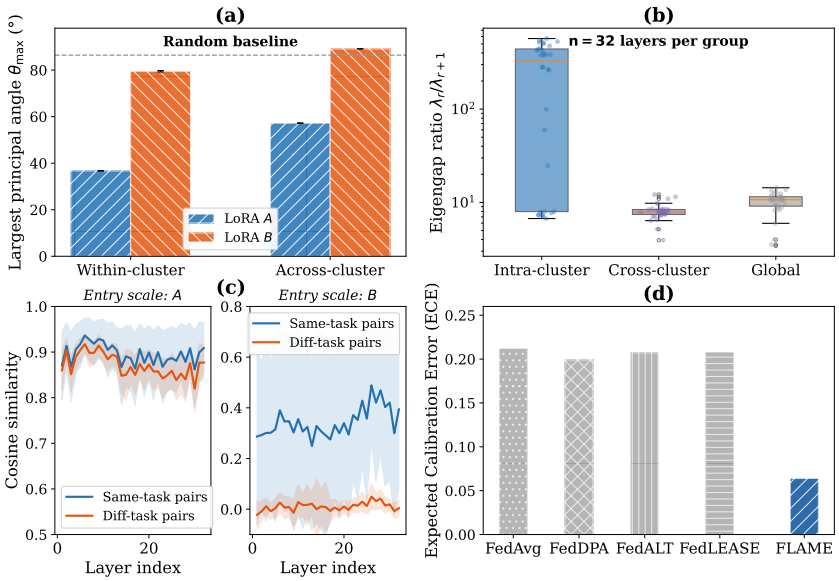

Figure 1: Three observations on the trained federated LoRA parameter space. *(a)* Largest principal angle between client A/B subspaces at intra- and cross-cluster granularities, against the random rank-r baseline (dashed). *(b)* Eigengap ratio λ_r/λ_{r+1} of the average A-subspace projector at intra-cluster, cross-cluster, and global peer sets (n=32 layers per group); larger ratio $\rightarrow$ better-separated rank-r subspace. *(c)* Per-layer entry-scale cosine similarity on A and B between same-task and different-task client pairs. *(d)* Aggregate ECE of four federated LoRA approaches against FLAME's full pipeline.

similarity cleanly separates same-task from different-task client pairs at every adapted layer, so the task signal on B's lives in its entries. Comparing these results with Observations 1, the federated signal sits on A and B at inverted scales. On A, it lives in the row-space and only among intra-cluster peers. On B, it lives directly in the entries. A single aggregator applied to both factors at the same scale therefore reads one signal as the other's noise, which motivates a selective aggregator that aligns A on the Grassmann manifold while averaging B in Euclidean space.

Observation 3: the calibration gap is invariant to the routing approach. Figure 1(d) reports the aggregate expected calibration error (ECE) of four federated LoRA approaches spanning the personalization spectrum, from a single global LoRA (FedAvg) to a clustered mixture of experts (FedLEASE). The four approaches reach similar ECE levels despite substantial differences in routing architecture, and the component ablation in Section 5.3 (Table 2) shows that FLAME's own architecture without any Bayesian treatment also lies in the same range, while FLAME's full pipeline brings ECE substantially lower. Each approach returns a maximum-a-posteriori (MAP) adapter whose output is a single softmax over data already seen during training. The calibration gap is, therefore, a property of point estimation rather than of any particular routing layer, and closing it calls for a Bayesian treatment that is orthogonal to the routing.

3 FLAME: Federated Laplace-Aligned Mixture of Experts

The three observations of Section 2.2 identify three properties of the trained federated LoRA parameter space that no existing personalized federated LoRA method jointly addresses. We next present FLAME, a framework that introduces three components matched to these three properties. The end-to-end FLAME pipeline is presented in Figure 2.

3.1 Personalized Clustered Mixture of Experts

The federated training in FLAME consists of three stages. The first stage initiates a brief local warmup to produce task-specific representations from which the server can infer cluster structure. The second stage is a one-shot clustering step at the server, and the third stage involves clustered mixture-of-experts (MoE) training for the remaining rounds. During the warmup, every client

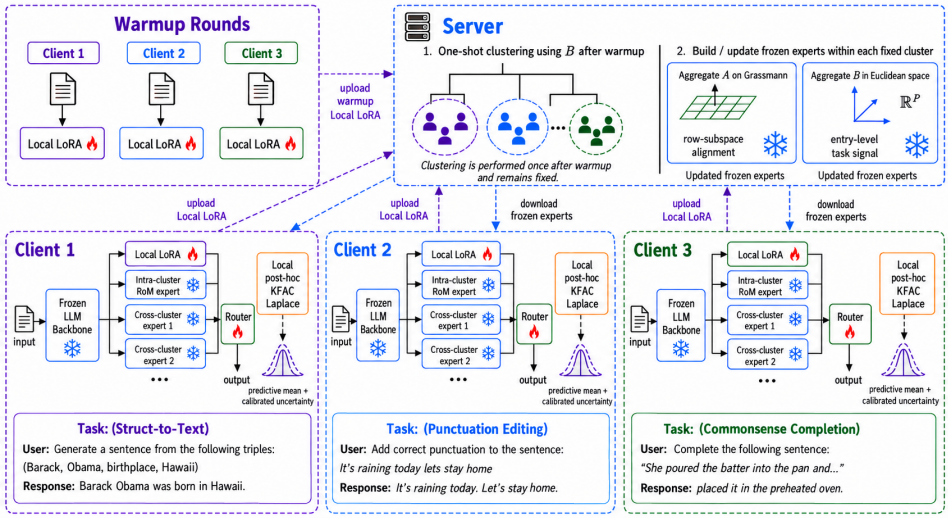

Figure 2: Overview of FLAME. Clients first perform a local warmup and upload their local LoRA factors to the server. After warmup, the server performs one-shot clustering using the output-side factor B, then builds frozen experts within each fixed cluster by aggregating A on the Grassmann manifold and B in Euclidean space. During personalized federated training, each client combines its trainable local LoRA with intra-cluster and cross-cluster frozen experts through a learned router. After training converges, each client fits a local post-hoc KFAC Laplace approximation on its individual expert to produce a predictive mean and calibrated uncertainty.

trains its own LoRA (A_n, B_n) on $\mathcal{D}_n$ for T_w warmup rounds, after which the warmed-up factors are uploaded once to the server. The server then forms M clusters by hierarchical agglomerative clustering (Wang et al., 2026) on the pairwise distance matrix

$$D_{n,m} \;=\; \frac{1}{L}\sum_{\ell=1}^{L}\Big(1-\cos\big(\mathrm{vec}(B_n^{(\ell)}),\mathrm{vec}(B_m^{(\ell)})\big)\Big), \tag{3}$$

where $\mathrm{vec}(\cdot)$ flattens a matrix into a vector, $\cos(\cdot,\cdot)$ is the cosine, L is the number of adapted linear layers, and $B_n^{(\ell)}$ is client n's output-side factor at layer ℓ. The number of clusters M is selected by silhouette score over $\{2,\dots,M_{\max}\}$ with average linkage, where $M_{\max}$ is a hyperparameter upper bound on the cluster count, and the result is the partition $\{\mathcal{C}_j\}_{j=1}^{M}$ in which $\mathcal{C}_j$ collects all clients assigned to the j-th cluster. Clustering is performed on B rather than A, indicated by Observation 2 that B carries entry-level task identity whereas A's entries are dominated by the shared Kaiming initialization at this stage and provide no task-discriminative signal.

After clustering, every client n in cluster $\mathcal{C}_j$ is assigned $M+1$ LoRA experts spanning three disjoint granularities, matched one to one with the locality levels of Observation 1. The *individual expert* at granularity 0 holds the client's own trainable parameters (A_n, B_n). The *intra-cluster RoW expert* at granularity 1 is the frozen leave-one-out aggregate $(\bar{A}_{\mathcal{C}_j\setminus\{n\}}, \bar{B}_{\mathcal{C}_j\setminus\{n\}})$ (Bian et al., 2026) over the client's cluster peers, with the aggregation operator described in Section 3.2. The $M-1$ *cross-cluster experts* at granularity 2 are frozen aggregates $(\bar{A}_{\mathcal{C}_{j'}}, \bar{B}_{\mathcal{C}_{j'}})$ for each peer cluster $j' \neq j$ over all of its members. The leave-one-out construction at granularity 1 ensures that the frozen signal received by client n is genuinely out-of-self and therefore complements the trainable individual expert without redundancy.

We adopt the enlarged top-k router of (Wang et al., 2026), where the individual expert (A_n, B_n) occupies multiple output positions while each frozen expert occupies a single position. With $k = M+1$, this construction guarantees that the individual expert participates in every forward pass. Specifically, ϕ_n is a per-layer trainable linear layer, $\mathcal{T}(x)$ denotes the index set of the top $M+1$ outputs of $\phi_n(x)$, and $\pi(i) \in \{0,1,\dots,M\}$ maps each output position i to its wired expert, with $\pi(i)=0$ for the individual. The LoRA update at every adapted layer is

$$\Delta W(x) \;=\; \sum_{i\in\mathcal{T}(x)} g_i(x)\, B^{(\pi(i))}A^{(\pi(i))}x, \qquad g_i(x) \;=\; \frac{\exp(\phi_n(x)_i)}{\sum_{j\in\mathcal{T}(x)}\exp(\phi_n(x)_j)}, \tag{4}$$

where $(A^{(0)}, B^{(0)}) = (A_n, B_n)$ is the trainable individual expert and $(A^{(e)}, B^{(e)})$ for $e \in \{1, \ldots, M\}$ are frozen aggregates. The total per-token gate weight assigned to the individual expert,

$$g_I(x) := \sum_{i \in \mathcal{T}(x):\, \pi(i)=0} g_i(x), \tag{5}$$

admits a structural lower bound $g_I(x) \geq g_{i^\star}(x) > 0$ at every x, where $i^\star$ is any selected position wired to the individual expert. Throughout this clustered training stage, gradients flow only into (A_n, B_n) and the router ϕ_n, while the frozen experts and W^0 remain unchanged; at the end of every round, client n uploads only (A_n, B_n) to the server, and the per-layer router ϕ_n stays on the client.

3.2 Grassmannian alignment of A at the server side

With the MoE-based clustering mechanism, we now specify how the frozen experts in equation 4 are constructed from the LoRA factors uploaded by clients. To avoid the decomposition drift observed in Observation 1 for the input-side factor, we aggregate A at every level via the chordal-distance Fréchet mean on the Grassmann manifold $\mathrm{Gr}(r, d_{\mathrm{in}})$, while keeping the standard Euclidean mean for B at every level, in accordance with the asymmetry of Observation 2.

Let $\mathcal{P} \subseteq \{1, \ldots, N\} \setminus \{n\}$ denote a peer set whose factors are to be aggregated: $\mathcal{P} = \mathcal{C}_j \setminus \{n\}$ for the intra-cluster RoW expert of client $n \in \mathcal{C}_j$, and $\mathcal{P} = \mathcal{C}_{j'}$ for some $j' \neq j$ for cross-cluster expert $e \in \{2, \ldots, M\}$. The Grassmannian aggregator is computed in three steps. First, every peer's A_m is orthonormalized via thin QR decomposition of $A_m^\top$ to give $Q_m = \mathrm{orth}(A_m) \in \mathbb{R}^{r \times d_{\mathrm{in}}}$ whose rows form an orthonormal basis of $\mathrm{rowspan}(A_m)$, and the average orthogonal projector onto these row spaces is computed as

$$\bar{P}_{\mathcal{P}} = \frac{1}{|\mathcal{P}|} \sum_{m \in \mathcal{P}} Q_m^\top Q_m, \tag{6}$$

where each $Q_m^\top Q_m \in \mathbb{R}^{d_{\mathrm{in}} \times d_{\mathrm{in}}}$ is invariant under the gauge transformation $A_m \mapsto R A_m$, so $\bar{P}_{\mathcal{P}}$ depends only on the row spaces themselves rather than on any specific basis. Second, the consensus basis is read off as the top-r eigenvectors of $\bar{P}_{\mathcal{P}}$,

$$\bar{Q}_{\mathcal{P}} = \big[\, \text{top-}r\text{-eigvecs}(\bar{P}_{\mathcal{P}}) \,\big]^\top \in \mathbb{R}^{r \times d_{\mathrm{in}}}, \tag{7}$$

which is the chordal-distance Fréchet mean on $\mathrm{Gr}(r, d_{\mathrm{in}})$ in closed form (Absil et al., 2004). Third, each peer's coefficients are projected onto the consensus basis and averaged,

$$\bar{A}_{\mathcal{P}} = \left(\frac{1}{|\mathcal{P}|} \sum_{m \in \mathcal{P}} A_m \bar{Q}_{\mathcal{P}}^\top \right) \bar{Q}_{\mathcal{P}}. \tag{8}$$

The output-side factor uses the standard Euclidean mean

$$\bar{B}_{\mathcal{P}} = \frac{1}{|\mathcal{P}|} \sum_{m \in \mathcal{P}} B_m. \tag{9}$$

The frozen experts in equation 4 are then $(A^{(1)}, B^{(1)}) = (\bar{A}_{\mathcal{C}_j \setminus \{n\}}, \bar{B}_{\mathcal{C}_j \setminus \{n\}})$ for the intra-cluster RoW and $(A^{(e)}, B^{(e)}) = (\bar{A}_{\mathcal{C}_{j_e}}, \bar{B}_{\mathcal{C}_{j_e}})$ for cross-cluster expert $e \geq 2$, where j_e enumerates the other clusters.

The eigengap measurement in Figure 1(b) constrains where this aggregator may be applied. The intra-cluster eigengap is well-resolved, while choosing the peer set $\mathcal{P}$ to span multiple clusters collapses the eigengap by one to two orders of magnitude and leaves $\bar{Q}_{\mathcal{P}}$ underdetermined. FLAME therefore restricts the Grassmannian aggregator to single-cluster peer sets at every level of the hierarchy: the intra-cluster RoW expert uses $\mathcal{P} = \mathcal{C}_j \setminus \{n\}$, which is a leave-one-out subset of one cluster, and each cross-cluster expert uses $\mathcal{P} = \mathcal{C}_{j'}$ for one peer cluster $j' \neq j$, rather than letting $\mathcal{P}$ span multiple clusters. Both peer sets sit in the high-eigengap regime of Figure 1(b), so $\bar{Q}_{\mathcal{P}}$ is well-resolved at every frozen expert FLAME constructs in Section 3.1, and the gauge invariance of the projector $Q_m^\top Q_m$ removes the LoRA basis ambiguity that the Euclidean mean fails to handle.

3.3 LOCAL POST-HOC KFAC LAPLACE AT LOCAL AGENTS

The MoE mechanism in Section 3.1 returns a maximum-a-posteriori (MAP) adapter. According to Observation 3 in Section 3, this point estimate produces a calibration gap that no further reorganization of the routing layer can close. To recover calibrated predictive distributions, we add a Bayesian treatment over the only set of parameters that has received a local gradient signal during clustered MoE training, namely, each client's individual expert (A_n, B_n). After federated training has converged, every client independently and locally fits a Kronecker-factored Laplace posterior (Daxberger et al., 2021; Martens & Grosse, 2015; Ritter et al., 2018) over (A_n, B_n). The Laplace stage modifies neither the training loop nor the per-round communication, and adds no per-step memory cost beyond a one-shot accumulation of Kronecker factors.

Consider an adapted linear layer with effective individual-expert weight $W_n = B_n A_n \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ on client n. Let $a_{n,t} \in \mathbb{R}^{d_{\text{in}}}$ and $\delta_{n,t} \in \mathbb{R}^{d_{\text{out}}}$ denote the layer's input activation and output gradient at sample $t \in \{1, \dots, T_n\}$ of client n's local dataset $\mathcal{D}_n$, where $T_n = |\mathcal{D}_n|$, collected by forward and backward hooks attached only to the individual expert. The hooks are evaluated under the per-client task loss equation 1 on $\mathcal{D}_n$, with the frozen experts and the trained router ϕ_n active in the forward pass equation 4 but only (A_n, B_n) contributing to the Fisher accumulation. The Kronecker factors at client n are

$$\Omega_n \;=\; \frac{1}{T_n} \sum_{t=1}^{T_n} a_{n,t}\, a_{n,t}^{\top}, \qquad \Lambda_n \;=\; \frac{1}{T_n} \sum_{t=1}^{T_n} \delta_{n,t}\, \delta_{n,t}^{\top}, \tag{10}$$

where $\Omega_n \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$ is the input activation Gram and $\Lambda_n \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$ is the output gradient Gram. Computing (Ω_n, Λ_n) amounts to a single forward-backward pass over $\mathcal{D}_n$ with no gradient updates, the cost of which is below one training epoch.

At test time, S posterior weight samples are drawn at every adapted linear layer from the matrix-normal posterior implied by the Kronecker factors at that layer,

$$W_{n,s} \;=\; W_{n,\text{MAP}} + \big(\Lambda_n + \sqrt{\alpha_n}\, I\big)^{-1/2} Z_{n,s} \big(\Omega_n + \sqrt{\alpha_n}\, I\big)^{-1/2}, \qquad Z_{n,s} \overset{\text{iid}}{\sim} \mathcal{N}(0, I), \tag{11}$$

for $s = 1, \dots, S$, where $W_{n,\text{MAP}} = B_n A_n$ is the trained individual-expert weight at the end of federated optimization. The prior precision $\alpha_n > 0$ is selected by maximizing the Kronecker-factored log marginal likelihood $\log p(\mathcal{D}_n; \alpha_n)$ over a one-dimensional grid, and the posterior precision is approximated by $\Lambda_n \otimes \Omega_n + \alpha_n I$ with the standard symmetric KFAC damping in which each Kronecker factor is shifted by $\sqrt{\alpha_n}\, I$ (Martens & Grosse, 2015). Equations equation 10-equation 11 are stated for a generic adapted linear layer and run independently at every adapted linear layer of the LLM. Each per-sample collection of layerwise weights replaces the individual expert in the MoE forward pass equation 4 while the frozen experts and trained router remain fixed at their post-training values, so the K-sample predictive distribution at input x on client n is

$$p_n(y \mid x) \;\approx\; \frac{1}{S} \sum_{s=1}^{S} \text{softmax}\big(f(x; W_{n,s})\big), \tag{12}$$

where, with mild abuse of notation, $f(x; W_{n,s})$ denotes the model logits at the predicted token position with the per-layer individual experts collectively replaced by their layerwise samples indexed by s, and y ranges over the answer-token vocabulary for multiple-choice classification cast as next-token prediction or over the full vocabulary for free-form generation. Every quantity in equation 6-equation 12 is computed from local data on client n, so the post-hoc Laplace stage transmits nothing to the server and adds no communication cost beyond the federated training of Sections 3.1-3.2.

Remark (Why Laplace fits FLAME?) The recipe of equation 10-equation 12 relies on three preconditions, all satisfied by FLAME's mechanism. *MAP validity*: the individual expert is the only parameter group trained against a forward pass in which the frozen experts remain fixed, providing the local minimum required for Laplace's quadratic expansion to be well-defined (Daxberger et al., 2021). *Fisher well-definedness*: only the individual expert receives a local gradient signal during clustered MoE training, so the factors (Ω_n, Λ_n) in equation 10 accumulate posterior precision on the parameters being made Bayesian and on no others (Kunstner et al., 2019). *BMA leverage*: the lower bound on $g_I(x)$ guaranteed by the router equation 4 keeps the individual expert's posterior contribution bounded away from zero at every token, ensuring non-vanishing Bayesian model averaging weight (Wilson & Izmailov, 2020); FedALT's unconstrained two-expert mixer admits no such bound, and $g_I(x)$

can collapse on tokens dominated by the RoW expert. We confirm this empirically in Section 5.3 (Table 10): equipping FedALT with the same post-hoc Laplace stage still leaves a calibration gap to FLAME.

4 Theoretical Analysis

We analyze the convergence of FLAME's clustered MoE training stage (Section 3.1), in which the trainable individual expert (A_n, B_n) is the only parameter group that receives a local gradient signal. Let $K = TE$ be the total number of local SGD steps across T rounds with E steps per round, $\mathcal{L}_n^t$ the round-t local loss on client n, and $\Delta_T = \sum_{t=1}^{T-1} \|\bar{\mathbf{E}}^{t+1} - \bar{\mathbf{E}}^t\|$ the cumulative inter-round drift of the frozen-expert configuration. Here, we only state the main result, with the full descriptions of assumptions, notation, and proofs delegated to Appendix C;

Theorem 1 (**Convergence of FLAME**). *Under the assumptions in Appendix C.2, the local SGD iterates produced by FLAME with constant step size $\eta \in (0, 1/L_\Phi]$, for every client n, satisfy*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\big\|\nabla \mathcal{L}_n^{t(k)}(\theta_n^{(k)})\big\|^2 \;\le\; \frac{2\big(\mathcal{L}_n^1(\theta_n^{(0)}) - \mathcal{L}_n^\star\big)}{\eta K} \;+\; L_\Phi\,\eta\,G_\Phi^2 \;+\; \frac{2C_S\Delta_T}{\eta K}. \tag{13}$$

With $\eta = \min\big(1/L_\Phi, \sqrt{2D_0/(L_\Phi G_\Phi^2 K)}\big)$ and $D_0 = \mathcal{L}_n^1(\theta_n^{(0)}) - \mathcal{L}_n^\star + \mathcal{C}_S\Delta_T$, the right-hand side decays at rate $\mathcal{O}\big(K^{-1/2}\big)$.

Each term on the right side of equation 13 can be traced to a component of FLAME. The smoothness and second-moment constants L_Φ and G_Φ^2 in the first two terms come from the gate ceiling $g_{n,I} \le 1$, since the per-layer gradient on the individual expert factors as $g_{n,I}\,\Phi_n^{(\ell)}$ in FLAME's architecture (Appendix Lemma 1). The drift constant Δ_T in the third term is smaller under FLAME's Grassmannian aggregator on A than under Euclidean averaging whenever the LoRA gauge varies across rounds (Appendix Lemma 6), so the bound is strictly tighter than its FedAvg-style counterpart in this regime. The gate floor $g_{\min} > 0$, enforced by FLAME's enlarged top-$(M{+}1)$ router but not by FedALT's two-expert mixer, does not appear in equation 13; it instead supplies the per-token Fisher mass on (A_n, B_n) that the post-hoc Laplace stage of Section 3.3 requires (Appendix Corollary 1).

5 Experiments

We evaluate FLAME against personalized federated LoRA baselines and Bayesian baselines, and ablate the contribution of each FLAME component.

5.1 Setup

We evaluate on four federated benchmarks: *MCQA-12c*, a 12-client FLAN classification setup over four tasks (3 IID clients each) on Llama-3.2-1B; and three 8-client NLG generation setups on Llama-2-7B under increasing heterogeneity, namely *FLAN* (IID replicas across four tasks), *FLAN-XD* (cross-dataset within each task), and *FLAN-LS* (disjoint label partitions within each task). All methods use LoRA rank $r = 8$ at `q_proj` and `v_proj`, $T = 10$ rounds with $T_w = 1$ warmup, and the post-hoc KFAC Laplace runs once per client with $S = 10$ posterior draws and adds no communication. We compare against FedAvg (McMahan et al., 2017), FedDPA (Yang et al., 2024), FedALT (Bian et al., 2026), FedLEASE (Wang et al., 2026), and FLoRG (Meng et al., 2026) under matched training budget, and report accuracy / ROUGE-1, ECE, and NLL averaged across clients with mean $\pm$ standard deviation across seeds. Full setup details appear in Appendix B.3.

5.2 Main results

Table 1 reports the full comparison across the four benchmarks. FLAME achieves the highest accuracy / ROUGE-1, the lowest ECE, and the lowest NLL on every benchmark. The accuracy lead is consistent with the hierarchical locality of Observation 1 that existing personalized routers expose at a single level only. The calibration lead is substantially larger than the accuracy lead, in line with Observation 3 that the calibration gap is a property of point estimation rather than of any specific routing architecture. The ROUGE-1 lead persists across IID, cross-dataset, and

Table 1: Main results across four federated benchmarks. MCQA-12c is a 12-client classification setup (Llama-3.2-1B); FLAN, FLAN-XD, and FLAN-LS are 8-client NLG setups (Llama-2-7B) under IID, cross-dataset, and label-skew heterogeneity, respectively. The primary metric is accuracy on MCQA-12c and ROUGE-1 on the three FLAN benchmarks; ECE and NLL are reported on all four. All methods use LoRA rank $r = 8$. **Bold** denotes the best value in each column.

		MCQA-12c	NLG-8c		
Metric	Method	FLAN	FLAN	FLAN-XD	FLAN-LS
Acc / R-1 (↑)	FedAvg	73.58 ± 0.55	67.40 ± 3.71	69.15 ± 6.80	63.36 ± 0.54
	FedDPA	78.72 ± 1.71	70.25 ± 3.87	75.33 ± 1.77	70.34 ± 3.25
	FedLEASE	79.00 ± 0.98	73.17 ± 3.37	77.56 ± 1.63	74.69 ± 1.04
	FedALT	79.06 ± 2.22	73.03 ± 4.66	77.01 ± 0.37	73.76 ± 2.13
	FLoRG	72.61 ± 0.51	69.81 ± 0.41	72.51 ± 0.28	65.44 ± 0.57
	FLAME	$\mathbf{80.86 \pm 0.35}$	$\mathbf{76.03 \pm 3.49}$	$\mathbf{79.68 \pm 0.55}$	$\mathbf{77.28 \pm 0.40}$
ECE (↓)	FedAvg	0.212 ± 0.008	0.147 ± 0.004	0.167 ± 0.001	0.189 ± 0.004
	FedDPA	0.199 ± 0.016	0.195 ± 0.011	0.174 ± 0.021	0.199 ± 0.003
	FedLEASE	0.208 ± 0.009	0.176 ± 0.001	0.159 ± 0.013	0.205 ± 0.001
	FedALT	0.207 ± 0.023	0.181 ± 0.011	0.130 ± 0.015	0.184 ± 0.002
	FLoRG	0.212 ± 0.004	0.163 ± 0.015	0.161 ± 0.012	0.170 ± 0.010
	FLAME	$\mathbf{0.064 \pm 0.004}$	$\mathbf{0.065 \pm 0.000}$	$\mathbf{0.050 \pm 0.002}$	$\mathbf{0.071 \pm 0.014}$
NLL (↓)	FedAvg	1.28 ± 0.08	1.864 ± 0.078	1.747 ± 0.056	1.880 ± 0.054
	FedDPA	1.65 ± 0.19	1.828 ± 0.160	1.747 ± 0.140	1.941 ± 0.300
	FedLEASE	2.96 ± 0.25	2.216 ± 0.130	1.998 ± 0.187	2.265 ± 0.100
	FedALT	2.76 ± 0.39	2.087 ± 0.280	2.125 ± 0.182	2.196 ± 0.360
	FLoRG	1.294 ± 0.016	1.836 ± 0.021	1.793 ± 0.015	1.940 ± 0.024
	FLAME	$\mathbf{0.791 \pm 0.016}$	$\mathbf{0.745 \pm 0.054}$	$\mathbf{0.682 \pm 0.016}$	$\mathbf{0.727 \pm 0.000}$

label-skew heterogeneity regimes, indicating that the design choices motivated by the MCQA-12c parameter-space study transfer to generation tasks at a larger backbone scale.

5.3 Ablations

We isolate the contribution of each FLAME component on MCQA-12c (Appendix B.2Table 2): removing the post-hoc Laplace (*w/o Lap.*) leaves accuracy unchanged but inflates ECE and NLL sharply, confirming Observation 3 that the calibration gap is closed only by the Bayesian model average; replacing it with MC Dropout (*MCD*) does not recover FLAME's calibration; and further removing the Grassmannian aggregator (*w/o Grass, w/o Lap.*) drops accuracy and inflates cross-seed variance, consistent with the gauge-invariance argument of Section 3.2. Additional ablations corroborating these findings are reported in the appendix: hierarchy and aggregator ablations confirming Observations 1-2 (Appendix B.8, Table 8), sanity controls ruling out random frozen experts and high-rank FedALT (Appendix B.9, Table 9), Laplace universality across baselines (Appendix B.10, Table 10), and cluster recovery on FLAN-XD (Appendix B.11). **Additional experiments.** Additional experiments in Appendix B include the Local-only and additional federated LoRA comparisons (Table 3), temperature-scaling controls (Table 4), system-overhead measurements (Tables 5 and 6), and a 16-client scalability study (Table 7).

6 Conclusion

We introduced FLAME, a personalized federated LoRA framework whose three components, a clustered mixture of experts with a guaranteed individual path, a selective Grassmannian aggregator on the input-side factor, and a post-hoc Kronecker-factored Laplace posterior, are each motivated by a measured property of the trained parameter space. A non-asymptotic convergence analysis attributes a distinct term of the stationary-point bound to each component. Across a 12-client MCQA benchmark and three FLAN NLG benchmarks under varying heterogeneity, FLAME improves accuracy and calibration over strong personalized federated LoRA baselines without adding per-round communication. The post-hoc Laplace stage is currently restricted to a Kronecker-factored Gaussian over the individual expert, and the Grassmannian aggregator assumes a shared rank across clients. Extending the posterior family and integrating the aggregator with heterogeneous-rank federated LoRA are natural directions for future work.

Reproducibility Statement

The complete algorithm is provided in Algorithm 1. Experimental settings and additional results are provided in Appendix B, and the complete convergence analysis is provided in Appendix C.

Ethics Statement

This work evaluates federated fine-tuning methods on existing language-model benchmarks.

AI Use Statement

Generative AI tools were used to assist with language editing and manuscript organization.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Riemannian geometry of grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematicae*, 2004.

Idan Achituve, Aviv Shamsian, Aviv Navon, Gal Chechik, and Ethan Fetaya. Personalized federated learning with gaussian processes. In *Proc. of Adv. Neural Inf. Process. Syst.*, 2021.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *Proc. of Adv. Neural Inf. Process. Syst.*, 2024.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Fedalt: Federated fine-tuning through adaptive local training with rest-of-world lora. In *Proc. AAAI Conf. Artif. Intell.*, 2026. arXiv:2503.11880.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proc. of Adv. Neural Inf. Process. Syst.*, volume 33, pp. 1877–1901, 2020.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *Proc. Intl. Conf. Mach. Learn.*, 2021.

Bai Cong, Nico Daheim, Yuesong Shen, Daniel Cremers, Rio Yokota, Mohammad Emtiyaz Khan, and Thomas Möllenhoff. Variational low-rank adaptation using ivon. *arXiv preprint arXiv:2411.04421*, 2024.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. In *Proc. of Adv. Neural Inf. Process. Syst.*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, 2019.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, et al. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *Proc. of Annu. Meet. Assoc. Comput. Linguist.*, pp. 1932–1945, 2024.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Proc. of Adv. Neural Inf. Process. Syst.*, 2020.

Tao Fan, Yan Kang, Guoqiang Ma, Weijing Chen, Wenbin Wei, Lixin Fan, and Qiang Yang. Fate-llm: A industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049*, 2023.

Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. On the effectiveness of parameter-efficient fine-tuning. In *Proc. AAAI Conf. Artif. Intell.*, 2023.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. Intl. Conf. Learn. Repr.*, 2022.

Rahif Kassab and Osvaldo Simeone. Federated generalized bayesian learning via distributed stein variational gradient descent. *IEEE Transactions on Signal Processing*, 2022.

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. In *Proc. of ACM SIGKDD Conf. Knowl. Discov. Data Min.*, 2024.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. *Proc. of Adv. Neural Inf. Process. Syst.*, 32, 2019.

Bingcong Li, Yilang Zhang, and Georgios B Giannakis. Low-rank adaptation redux for large models. *arXiv preprint arXiv:2604.21905*, 2026.

Q. Liu, X. Wu, X. Zhao, Y. Zhu, D. Xu, F. Tian, and Y. Zheng. When MOE meets LLMs: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1104–1114, 2024. doi: 10.1145/3626772.3657722.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proc. Intl. Conf. Mach. Learn.*, 2015.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proc. Intl. Conf. Artif. Intell. Stat.*, 2017.

Chuiyang Meng, Ming Tang, and Vincent WS Wong. Florg: Federated fine-tuning with low-rank gram matrices and procrustes alignment. *arXiv preprint arXiv:2602.17095*, 2026.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *Proc. Intl. Conf. Learn. Repr.*, 2018.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving lora in privacy-preserving federated learning. 2024.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. In *Proc. of Adv. Neural Inf. Process. Syst.*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Lei Wang, Jieming Bian, Letian Zhang, and Jie Xu. Adaptive lora experts allocation and selection for federated fine-tuning. In *Proc. of Adv. Neural Inf. Process. Syst.*, 2026. arXiv:2509.15087.

Xi Wang, Laurence Aitchison, and Maja Rudolph. Lora ensembles for large language model fine-tuning. *arXiv preprint arXiv:2310.00035*, 2023.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. Blob: Bayesian low-rank adaptation by backpropagation for large language models. In *Proc. of Adv. Neural Inf. Process. Syst.*, 2024a.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In *Proc. of Adv. Neural Inf. Process. Syst.*, 2024b.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Proc. of Adv. Neural Inf. Process. Syst.*, 33:4697–4708, 2020.

Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111*, 2023.

Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. Dual-personalizing adapter for federated foundation models. In *Proc. of Adv. Neural Inf. Process. Syst.*, 2024.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proc. of ACM SIGKDD Conf. Knowl. Discov. Data Min.*, 2024.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the federatedgpt: Federated instruction tuning. In *Proc. of Intl. Conf. Acoust. Speech Signal Process.*, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. In *Proc. Intl. Conf. Learn. Repr.*, 2023.

Supplementary File

A Related Work

Parameter-Efficient Fine-Tuning and LoRA. As LLMs grow in scale, full-parameter fine-tuning becomes prohibitive in compute and memory (Hadi et al., 2023). Parameter-efficient fine-tuning methods (Ding et al., 2023; Fu et al., 2023) reduce the trainable parameter footprint by introducing a small set of additional parameters while keeping the pretrained backbone frozen. LoRA (Hu et al., 2022; Li et al., 2026) approximates the fine-tuning update by a low-rank matrix BA, with A Kaiming-initialized and B zero-initialized so that the residual is exactly zero at the start of training. Subsequent work improves LoRA along multiple axes, including asymmetric architectures (Tian et al., 2024), mixture-of-experts variants (Liu et al., 2024; Dou et al., 2024), and rank-adaptive methods (Zhang et al., 2023). These advances are developed primarily for centralized training, and the question of how to deploy LoRA across clients with heterogeneous data distributions remains open.

Federated Learning and Federated Fine-Tuning. Federated Averaging (FedAvg) (McMahan et al., 2017) is the foundational FL algorithm and aggregates locally trained models on the server before redistributing to clients. Personalized federated learning (Tan et al., 2022; Collins et al., 2021; Fallah et al., 2020) addresses statistical heterogeneity by allowing clients to retain client-specific parameters, but most existing methods target small models and label-distribution skew rather than LLMs and task heterogeneity. The intersection of FL and PEFT has produced a growing federated fine-tuning literature (Zhang et al., 2024; Kuang et al., 2024; Ye et al., 2024; Fan et al., 2023), with FedIT (Zhang et al., 2024) pioneering federated LoRA, FlexLoRA (Bai et al., 2024), and FLoRA (Wang et al., 2024b) addressing heterogeneous client ranks, and a separate line analyzing the bilinear aggregation error of separately averaging A and B (Meng et al., 2026; Sun et al., 2024). These methods either share a single global LoRA across all clients or address aggregation under the assumption of label-shift heterogeneity, and do not target the task-heterogeneous regime that is the natural deployment scenario for LLMs.

Personalized Federated LoRA and Uncertainty Quantification. The most relevant prior works personalize federated LoRA under task heterogeneity. FedDPA (Yang et al., 2024) maintains a global LoRA and a local LoRA per client and blends them at inference with a fixed mixing weight. FedALT (Bian et al., 2026) replaces the fixed weight with a learned softmax router and treats the average over all other clients as a frozen RoW LoRA. FedLEASE (Wang et al., 2026) replaces the single RoW by a clustered mixture of experts whose number is selected by silhouette score on the cosine similarity of LoRA B-matrices. These designs each operate at a single level of cross-client locality, either binary or per-cluster, and apply Euclidean averaging to both LoRA factors regardless of how each factor carries signal across clients. None of them addresses the calibration of the resulting per-client adapters. Centralized Bayesian LoRA methods (Wang et al., 2024a; Yang et al., 2023; Wang et al., 2023; Cong et al., 2024) provide calibrated predictive distributions in the centralized regime, but each modifies the training loop or adds per-step memory cost that does not scale across federated clients. Federated UQ work (Achituve et al., 2021; Kassab & Simeone, 2022) targets sub-LLM scale models and does not adapt to the LoRA regime. Personalized federated LLMs therefore inherit no usable uncertainty quantification recipe, which is the gap our local post-hoc Laplace closes. FLAME extends this line of work along three axes that prior personalized federated LoRA designs leave unaddressed. Where FedALT and FedLEASE expose only a binary or per-cluster split, FLAME routes among the trainable individual adapters, an intra-cluster aggregate, and cross-cluster aggregates jointly, exposing the hierarchical locality the trained parameter space exhibits. Where prior designs apply Euclidean averaging to both LoRA factors, FLAME aligns A on the Grassmann manifold to remove the basis ambiguity while keeping B in Euclidean space. Where prior designs return point-estimate adapters, FLAME equips each client with a post-hoc Kronecker-factored Laplace posterior whose computation is local and adds no communication.

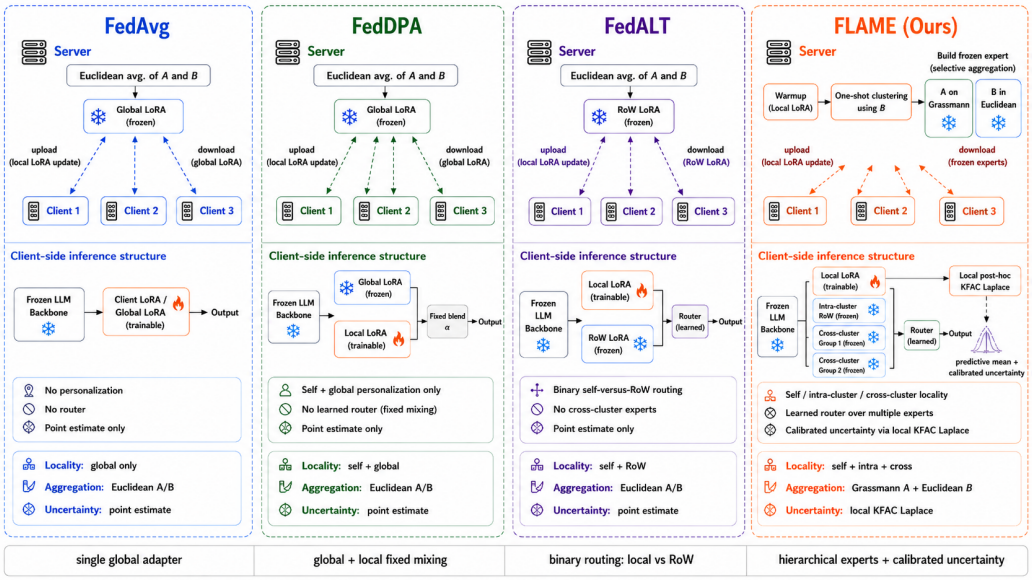

Figure 3: Comparison of personalized federated LoRA methods. FedAvg uses a single global adapter shared across clients. FedDPA combines a global adapter with a local adapter through fixed mixing. FedALT routes between a client-specific local adapter and a single Row adapter. FLAME jointly exploits self, intra-cluster, and cross-cluster locality, aggregates the input-side factor A on the Grassmann manifold while keeping Euclidean aggregation on the output-side factor B, and adds local post-hoc KFAC Laplace to obtain calibrated predictive uncertainty.

B Additional Experimental Details and Results

B.1 Full Setup Details

Datasets and clients. *MCQA-12c* is a 12-client FLAN classification setup with four tasks (3 IID clients each): SNLI (3-class NLI), HellaSwag (4-class commonsense), BoolQ (2-class reading comprehension), and Sentiment140 (2-class sentiment). The remaining three benchmarks are 8-client FLAN generation setups (NLG-8c) with four tasks (2 clients each) under increasing heterogeneity. *FLAN* assigns two IID replicas per task across four tasks: Sentiment140 (sentiment), WebNLG (struct-to-text), FixPunct (text editing), and HellaSwag (commonsense). *FLAN-XD* replaces each IID replica with a distinct source dataset from the same task category, yielding eight datasets: Sentiment140/SST2, WebNLG/DART, FixPunct/WordSegment, and HellaSwag/PIQA. *FLAN-LS* uses the same four datasets as FLAN but splits each task into two clients with disjoint label partitions (e.g., positive-only vs negative-only for Sentiment140). Each client holds 600 training and 100 test samples; data are disjoint across clients in all settings.

Backbone and adapter. MCQA-12c uses Llama-3.2-1B; the three NLG-8c benchmarks use Llama-2-7B. Both backbones are loaded in 8-bit quantization via `bitsandbytes` (LLM.int8), with LoRA weights computed in `float16`. LoRA rank $r = 8$ is applied to the `q_proj` and `v_proj` layers of every transformer block, with $\alpha = 16$ and dropout 0.05.

Federated protocol. The local warmup runs $T_w = 1$ round of local-only training without aggregation. The server then clusters clients by agglomerative clustering on the entry-level B-similarity distance equation 3, selecting M by silhouette score (recovering $M = 4$ on all benchmarks, matching the underlying number of tasks). Clustered MoE training runs $T - T_w = 9$ additional rounds (total $T = 10$ communication rounds), with $E = 10$ local epochs per round on MCQA-12c and $E = 1$ on NLG-8c. The post-hoc KFAC Laplace stage runs once per client on its training data with $S = 10$ posterior draws; no additional communication is required.

Table 2: Component ablation on MCQA-12c. We compare the full FLAME against variants that swap or disable components: *w/o Grass, w/o Lap.* uses Euclidean aggregation and MAP prediction; *w/o Lap.* adds Grassmannian A but keeps MAP; *MCD* replaces the KFAC Laplace with MC Dropout (K=10). All variants use LoRA rank r=8. **Bold** denotes the best in each column.

Variant	Acc (%) ↑	ECE ↓	NLL ↓
FLAME (w/o Grass, w/o Lap.)	79.83 ± 1.41	0.212 ± 0.007	2.500 ± 0.291
FLAME (w/o Lap.)	80.86 ± 0.35	0.204 ± 0.013	2.359 ± 0.217
FLAME (MCD)	77.62 ± 1.88	0.202 ± 0.015	2.166 ± 0.151
FLAME	$\mathbf{80.86 \pm 0.35}$	$\mathbf{0.064 \pm 0.004}$	$\mathbf{0.791 \pm 0.016}$

Optimization and random seeds. We use AdamW with learning rate 10^{-4}, weight decay 0.05, and batch size 4. All main experiments and ablations are reported over three random seeds.

Baselines. We compare against FedAvg (McMahan et al., 2017), FedDPA (Yang et al., 2024), FedALT (Bian et al., 2026), FedLEASE (Wang et al., 2026), and FLoRG (Meng et al., 2026), all under the same LoRA rank, training rounds, and per-round local epochs.

Metrics. On MCQA-12c, we report classification accuracy, expected calibration error (ECE), and negative log-likelihood (NLL). On the three NLG-8c benchmarks, we report ROUGE-1 and per-token ECE and NLL. Each metric is computed per client and averaged across clients. Performances are reported as mean $\pm$ standard deviation across multiple random seeds.

B.2 Component Ablation

B.3 Dataset Specifications

We evaluate on four federated benchmarks spanning classification and generation. *MCQA-12c* is a 12-client FLAN classification setup with four tasks (3 IID clients each): SNLI (3-class NLI), HellaSwag (4-class commonsense), BoolQ (2-class reading comprehension), and Sentiment140 (2-class sentiment). The remaining three benchmarks are 8-client FLAN generation setups (NLG-8c) with four tasks (2 clients each) under increasing heterogeneity. *FLAN* assigns two IID replicas per task across four tasks: Sentiment140 (sentiment), WebNLG (struct-to-text), FixPunct (text editing), and HellaSwag (commonsense). *FLAN-XD* replaces each IID replica with a distinct source dataset from the same task category, yielding eight datasets: Sentiment140/SST2, WebNLG/DART, FixPunct/WordSegment, and HellaSwag/PIQA. *FLAN-LS* uses the same four datasets as FLAN but splits each task into two clients with disjoint label partitions (e.g., positive-only vs negative-only for Sentiment140). Each client holds 600 training and 100 test samples; data are disjoint across clients in all settings.

B.4 Local-Only and Additional Federated LoRA Baselines

We extend the MCQA-12c comparison with a matched Local-only reference and two additional federated LoRA baselines, FFA-LoRA and LoRM. For Local-only, each client trains only its individual LoRA on its own 600 samples. Communication, clustering, routing, and all shared experts are removed, while the initialization, optimizer, round boundaries, local epochs, batch size, and total local optimization budget are matched to the federated methods.

The single-global-adapter methods, including FFA-LoRA and LoRM, underperform Local-only by 4.87–17.23 accuracy points. FedDPA, FedLEASE, and FedALT exceed Local-only accuracy by only 0.27, 0.55, and 0.61 points, respectively, which are small relative to their across-seed variation. None improves both calibration metrics over Local-only. In contrast, FLAME improves accuracy by 2.41 points while reducing ECE from 0.203 to 0.064 and NLL from 1.62 to 0.791. The comparison shows that federation is useful in this setting when the method preserves client-specific adaptation and selectively shares compatible cross-client information.

B.5 Temperature Scaling

We evaluate temperature scaling (TS) on MCQA-12c under two calibration protocols.

Table 3: Local-only and additional federated LoRA baselines on MCQA-12c.

Method	Acc. (%) ↑	ECE ↓	NLL ↓
FedAvg	73.58 ± 0.55	0.212 ± 0.008	1.28 ± 0.08
FLoRG	72.61 ± 0.51	0.212 ± 0.004	1.29 ± 0.02
FFA-LoRA	61.22 ± 0.56	0.216 ± 0.011	1.04 ± 0.01
LoRM	69.69 ± 0.87	0.242 ± 0.003	1.30 ± 0.08
Local-only	78.45 ± 1.59	0.203 ± 0.018	1.62 ± 0.24
FedDPA	78.72 ± 1.71	0.199 ± 0.016	1.65 ± 0.19
FedLEASE	79.00 ± 0.98	0.208 ± 0.009	2.96 ± 0.25
FedALT	79.06 ± 2.22	0.207 ± 0.023	2.76 ± 0.39
FLAME	80.86 ± 0.35	0.064 ± 0.004	0.791 ± 0.016

Protocol A: matched label access. The temperature T is fitted on the same local training labels available to FLAME's evidence maximization, and every method is evaluated on the full test split. Values are means over the three training seeds.

Table 4: Temperature scaling on MCQA-12c under matched label access.

Method	ECE per-cl. T	ECE global T	NLL per-cl. T	NLL global T
FedAvg	0.098	0.116	0.612	0.636
FLoRG	0.098	0.107	0.629	0.652
FFA-LoRA	0.087	0.123	0.748	0.779
LoRM	0.090	0.149	0.686	0.742
FedDPA	0.210	0.209	14.50	10.37
FedLEASE	0.203	0.203	3.88	4.57
FedALT	0.209	0.208	2.63	2.13
Local-only	0.214	0.215	5.52	5.92

Under this matched comparison, the best temperature-scaled baseline reaches ECE 0.087, compared with FLAME's 0.064.

Protocol B: held-out sensitivity check. The temperature T is fitted on half of each client's test split and the temperature-scaled baselines are evaluated on the disjoint half. FLAME uses neither half for calibration. The deterministic 50/50 client-level split uses a fixed RNG seed and is identical across methods and training seeds.

The best per-client temperature-scaled ECE/NLL results are $0.086/0.562$ for FedALT, $0.087/0.537$ for FedLEASE, $0.104/0.536$ for FedDPA, $0.105/0.534$ for Local-only, $0.099/0.598$ for FLoRG, and $0.108/0.584$ for FedAvg. The best global-T ECE is 0.123.

Because Protocol B changes both calibration-label access and the evaluated subset, these half-test results are reported as a sensitivity check rather than compared directly with FLAME's full-test result. FLAME's Laplace stage uses only the training data already available to each client and requires no separate calibration set. Under the matched training-label protocol, FLAME reaches ECE 0.064, compared with 0.087 for the best train-fitted temperature-scaled baseline.

B.6 System Overhead

We profile communication, storage, computation, memory, and inference latency on MCQA-12c using Llama-3.2-1B on an RTX 6000 Ada.

Communication and storage. These figures use rank-8 fp16 LoRA states. FLAME uploads one (A_n, B_n) pair per client, giving the same 1.62 MB uplink as FedAvg, FedALT, and FedDPA. Its 6.50 MB downlink and 9.25 MB persistent model state arise from the routed expert set and are close to FedLEASE's 6.50 MB and 9.00 MB. The post-hoc Laplace stage is entirely local and adds no communication. At fixed cluster count M, the per-client expert footprint is $\mathcal{O}(M)$ rather than $\mathcal{O}(N)$.

Table 5: Communication and persistent model state on MCQA-12c.

Method	Uplink	Downlink	Persistent model state
Local-only	0	0	1.62 MB
FFA-LoRA	0.62 MB	0.62 MB	1.62 MB
LoRM	0.93 MB	1.62 MB	1.62 MB
FedAvg / FLoRG	1.62 MB	1.62 MB	1.62 MB
FedALT	1.62 MB	1.62 MB	1.62 MB
FedDPA	1.62 MB	1.62 MB	3.25 MB
FedLEASE	1.62 MB	6.50 MB	9.00 MB
FLAME	1.62 MB	6.50 MB	9.25 MB

Table 6: Single-forward MAP inference latency on MCQA-12c.

Method	Structure	ms/sample
LoRM	Single adapter	9.85
FedAvg	Single adapter	9.98
FLoRG / Local-only	Single adapter	10.04
FedDPA	Single adapter	10.06
FedALT	RoW-routed	10.33
FFA-LoRA	Single adapter	10.38
FedLEASE	MoE	12.09
FLAME	MoE	12.36

Profiling and inference latency. FLAME's MAP path is 24% slower than FedAvg and close to FedLEASE. The one-time post-hoc KFAC fitting stage takes 233.7 s per client and consists of one forward-backward profiling sweep without parameter updates. Its measured temporary fp32 memory peak is 784 MB.

In the offline evaluation, the local posterior state is retained through posterior prediction and released afterward. Continuous calibrated serving would instead retain the local factors or eigendecompositions, or cache posterior samples; this state remains local but is additional to the 9.25 MB MAP model state reported above.

Posterior-predictive latency grows approximately linearly with the sample count S. Latencies are 13.1, 26.3, 65.7, and 131.9 ms per prediction for $S = 1, 2, 5$, and 10, respectively, averaged over the 12 clients. Latency is profiled using seed 36. The main experiments use $S = 10$, for which FLAME reports ECE 0.064 ± 0.004 and NLL 0.791 ± 0.016 over three seeds.

B.7 Client-Count Scalability

We extend MCQA from 12 to 16 clients while keeping Llama-3.2-1B, 600 local training samples per client, and the training protocol unchanged. Results are means and standard deviations over three seeds.

Table 7: MCQA results with 16 clients under the same backbone, per-client data budget, and training protocol as MCQA-12c.

Method	Acc. (%) ↑	ECE ↓	NLL ↓
FedAvg	73.13 ± 1.24	0.210 ± 0.013	1.236 ± 0.078
FedALT	78.75 ± 0.84	0.210 ± 0.008	3.179 ± 0.126
FedLEASE	79.67 ± 0.62	0.201 ± 0.006	3.206 ± 0.158
FLAME	80.58 ± 0.61	0.063 ± 0.003	0.761 ± 0.026

FLAME retains the highest accuracy and the lowest ECE and NLL when increasing the client count from 12 to 16. Calibration remains stable relative to MCQA-12c, where FLAME reports ECE 0.064 ± 0.004 and NLL 0.791 ± 0.016.

B.8 Architecture Ablations

Table 8 drills into each architectural choice on MCQA-12c. The hierarchy ablation in Table 8(a) removes one expert level at a time. Every removal degrades accuracy and inflates cross-seed variance, most sharply when the cross-cluster experts are removed; all three levels contribute, and none is redundant, in line with the three-granularity locality of Observation 1. The aggregator ablation in Table 8(b) crosses Grassmannian and Euclidean aggregation on A and B independently. Grassmannian on A paired with Euclidean on B outperforms every other combination, while the symmetric Grass-Grass and Eucl-Eucl variants both lag, confirming the asymmetry of Observation 2: A-subspaces carry shared structure that benefits from gauge-invariant alignment, whereas Grassmannian projection on B disrupts the entry-level task identity it encodes.

Table 8: Architecture ablations on MCQA-12c. **(a) C1: hierarchy ablation** removes one expert level at a time from FLAME, retaining the Grassmannian A aggregator. **(b) C2: aggregator ablation** crosses Grassmannian (Grass) and Euclidean (Eucl) aggregation on A and B independently. All variants use LoRA rank r=8. **Bold** denotes the best in each panel.

(a) C1: hierarchy ablation

Variant	Acc (%) ↑
FLAME (all three levels)	$\mathbf{80.86 \pm 0.35}$
w/o cross-cluster experts	79.47 ± 1.71
w/o intra-cluster RoW	79.00 ± 0.76
w/o individual expert	79.06 ± 0.43

(b) C2: aggregator

A agg.	B agg.	Acc (%) ↑
Grass	**Eucl**	$\mathbf{80.86 \pm 0.35}$
Grass	Grass	79.64 ± 0.51
Eucl	Grass	79.14 ± 1.21
Eucl	Eucl	79.83 ± 1.41

B.9 Sanity Controls

Two sanity controls in Table 9 address natural alternative explanations for FLAME's gain. The *Random experts* variant replaces the content of FLAME's frozen experts with norm-matched random weights while keeping the routing architecture identical, and underperforms FLAME by a wide margin, isolating the gain to the content of the cluster-aware aggregates rather than the routing architecture alone. The *FedALT (r=40)* variant raises FedALT's LoRA rank by a factor of five over its default; this lifts its accuracy to FLAME's level but does not close the calibration gap, so the gap cannot be closed by adding adapter parameters to a point-estimate routing scheme.

Table 9: Sanity controls on MCQA-12c. *Random experts* replaces the content of FLAME's frozen experts with norm-matched random weights while keeping the routing architecture identical to FLAME, isolating the value of the cluster-aware aggregates from the routing architecture. *FedALT (r=40)* is a high-capacity baseline that raises FedALT's LoRA rank by a factor of five. **FLAME** is reported as the reference point.

Variant	Acc (%) ↑	ECE ↓	NLL ↓
Random experts	78.46 ± 0.65	0.215 ± 0.008	3.14 ± 0.11
FedALT (r=40)	80.25 ± 0.82	0.197 ± 0.006	2.58 ± 0.11
FLAME	$\mathbf{80.86 \pm 0.35}$	$\mathbf{0.064 \pm 0.004}$	$\mathbf{0.791 \pm 0.016}$

B.10 Universality of Post-hoc Laplace

A natural follow-up to the component ablation in Section 5.3 is whether the Bayesian gain transfers to any personalized federated LoRA design, or whether it is bound to FLAME's specific architecture. Table 10 reports per-client best(MAP, Laplace) ECE and NLL after equipping every baseline with the same post-hoc Laplace stage on MCQA-12c. The Laplace stage improves the calibration of every baseline, but FLAME retains the lowest post-Laplace ECE. Among the baselines, FedALT is the closest architectural match to FLAME, sharing the trainable-individual-plus-frozen design but operating at a single locality scale; the persistent gap of FLAME over FedALT post-Laplace confirms that the calibration gain comes from the conjunction of FLAME's three-granularity routing, the

selective Grassmannian aggregator, and the post-hoc Laplace, and is not reproducible by appending Laplace to a single-locality-scale individual-plus-frozen routing architecture alone.

Table 10: Universality of post-hoc Laplace on MCQA-12c. Each row reports per-client best (MAP, Laplace) ECE and NLL. The Laplace stage improves calibration for every method, but **FLAME** achieves the lowest post-Laplace ECE.

Method	Best ECE ↓	Best NLL ↓
FedAvg	0.079 ± 0.006	0.821 ± 0.005
FedDPA	0.086 ± 0.007	0.823 ± 0.005
FedALT	0.081 ± 0.004	0.821 ± 0.002
FedLEASE	0.086 ± 0.013	0.817 ± 0.005
FLAME (w/o Grass)	0.074 ± 0.007	0.821 ± 0.010
FLAME	$\mathbf{0.064 \pm 0.004}$	$\mathbf{0.791 \pm 0.016}$

B.11 Cluster Recovery under Cross-Dataset Heterogeneity

We further test the server-side clustering of Section 3.1 on FLAN-XD, where each task is represented by two distinct source datasets rather than IID replicas. Figure 4 shows that the entry-level B-similarity matrix exhibits clean 2×2 task-pair blocks despite the surface-form shift between datasets within each pair, and the silhouette score peaks at $k = 4$, recovering the underlying task partition exactly rather than splitting into eight dataset-specific singletons.

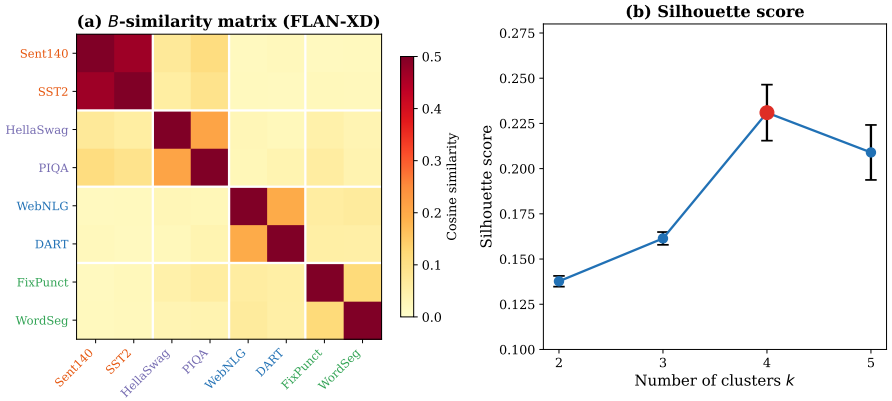

Figure 4: Cluster recovery on FLAN-XD. *(a)* B-similarity matrix equation 3; block structure tracks the four-task partition, not the eight source datasets. *(b)* Silhouette peaks at $k = 4$.

C Convergence Analysis of FLAME

We analyze the convergence of FLAME's clustered MoE training stage (Phase 3 of Section 3.1). The trainable individual expert is the only parameter group that receives a local gradient signal during this stage, and the per-step descent of its loss is what the bound below tracks. The Laplace stage of Section 3.3 is a test-time procedure applied to converged MAP weights, so it lies outside the optimization scope; we return to it in Section C.6, where the gate floor of the enlarged router is shown to supply the per-token Fisher mass it requires.

C.1 Proof-Specific Setting

We adopt the notation of Section 3 and introduce only quantities local to the proof.

Algorithm 1 FLAME: Federated Laplace-Aligned Mixture of Experts

Require: Clients $\{1, \dots, N\}$ with local data $\{\mathcal{D}_n\}_{n=1}^N$; frozen LLM weights W^0; LoRA rank r; number of pre-clustering rounds T_w; total rounds T; local epochs per round E; cluster-count upper bound $M_{\max}$; posterior samples K.

Local warmup ▷ rounds $1, \dots, T_w$

for $t = 1, \dots, T_w$, each client n **in parallel do**
  Train (A_n, B_n) on $\mathcal{D}_n$ for E epochs without aggregation.
end for

Clustering at the server

Form the pairwise distance matrix equation 3 and partition clients into $\{\mathcal{C}_1, \dots, \mathcal{C}_M\}$ by agglomerative clustering with $M \le M_{\max}$ selected by silhouette score.

Clustered MoE training ▷ rounds $T_w{+}1, \dots, T$

for $t = T_w + 1, \dots, T$ **do**
  Server: for each client $n \in \mathcal{C}_j$, build the intra-cluster RoW expert with $\mathcal{P} = \mathcal{C}_j \setminus \{n\}$ and the $M - 1$ cross-cluster experts with $\mathcal{P} = \mathcal{C}_{j'}$ for $j' \neq j$, in each case using equation 6-equation 8 for A and equation 9 for B.
  Client n **in parallel:** load the frozen RoW and cross-cluster expert states; train (A_n, B_n) and the router ϕ_n on $\mathcal{D}_n$ under equation 4 for E epochs; upload (A_n, B_n) to the server.
end for

Post-hoc KFAC Laplace ▷ per client, no communication

for each client n **in parallel do**
  Populate the KFAC factors (Ω_n, Λ_n) via equation 10 from forward-backward hooks under the per-client task loss equation 1, with hooks attached only to the individual expert and the frozen experts and trained router active.
  Select the prior precision α_n by maximizing $\log p(\mathcal{D}_n; \alpha_n)$ on a one-dimensional grid.
  At test time, draw K posterior weight samples via equation 11 at every adapted linear layer and predict via equation 12.
end for

Iteration index. Phase 3 runs for T rounds with E local SGD steps per round, for $K = TE$ total steps. The global step is $k \in \{0, \dots, K-1\}$ and its round is $t(k) \in \{1, \dots, T\}$, with steps $(t-1)E, \dots, tE - 1$ belonging to round t.

Individual expert and frozen experts as Euclidean vectors. The trainable individual factors are stacked across layers into $\theta_n^{(k)} = (A_n^{(k),(\ell)}, B_n^{(k),(\ell)})_{\ell=1}^L \in \mathbb{R}^{P_{\text{ind}}}$ with $P_{\text{ind}} = Lr(d_{\text{in}} + d_{\text{out}})$, equipped with the Frobenius norm. The round-t frozen-expert configuration $\bar{\mathbf{E}}^t = \{(\bar{A}^{(e),t}, \bar{B}^{(e),t})\}_{e=1}^M$ is treated as a vector with the natural Frobenius norm.

Per-token gate on the individual expert. The router output is $\phi_n(x) = G_n x$ with $G_n \in \mathbb{R}^{(2(M+1)-1)\times d_{\text{in}}}$, mapped to experts by π as in equation 4. The per-token gate weight on the individual expert at layer ℓ is

$$g_{n,I}^{(\ell)}(x; G_n) := \sum_{i \in \text{TopK}(\phi_n(x), M+1),\, \pi(i)=0} g_{n,i}^{(\ell)}(x; G_n). \tag{14}$$

We will use repeatedly that $g_{n,I}^{(\ell)}$ depends on x and G_n but not on θ_n, because $\phi_n(x) = G_n x$ does not involve θ_n.

Local loss and SGD update. Holding $\bar{\mathbf{E}}^t$ and G_n fixed throughout round t, the per-sample loss is $F_n(\theta_n; \xi; \bar{\mathbf{E}}^t, G_n)$ and the local population loss is $\mathcal{L}_n^t(\theta_n) := \mathbb{E}_{\xi \sim \mathcal{D}_n}[F_n(\theta_n; \xi; \bar{\mathbf{E}}^t, G_n)]$. With learning rate $\eta > 0$, the local SGD update is

$$\theta_n^{(k+1)} = \theta_n^{(k)} - \eta \nabla F_n(\theta_n^{(k)}; \xi_n^{(k)}; \bar{\mathbf{E}}^{t(k)}, G_n), \qquad \xi_n^{(k)} \sim \mathcal{D}_n, \tag{15}$$

with $\mathbb{E}[\nabla F_n(\theta_n^{(k)}; \xi_n^{(k)}) \mid \theta_n^{(k)}] = \nabla \mathcal{L}_n^{t(k)}(\theta_n^{(k)})$. The router G_n is fixed within a round and absorbed into the loss notation throughout the analysis.

C.2 ASSUMPTIONS

Assumption 1 (Per-layer core gradient)**.** There exist $L_\Phi, G_\Phi > 0$ such that the $g_{n,I}$-stripped Jacobian factors $\Phi_n^{(\ell)}$ defined in equation 20 below satisfy, for every $n, \xi, G_n, \bar{\mathbf{E}}$ visited during training and every θ, θ',

$$\text{(i)} \sum_{\ell=1}^{L} \|\Phi_n^{(\ell)}(\xi;\theta) - \Phi_n^{(\ell)}(\xi;\theta')\|^2 \le L_\Phi^2 \|\theta - \theta'\|^2, \quad \text{(ii)} \sum_{\ell=1}^{L} \|\Phi_n^{(\ell)}(\xi;\theta)\|^2 \le G_\Phi^2.$$

Assumption 2 (Bounded loss)**.** There exists $\mathcal{L}_n^\star \in \mathbb{R}$ with $\mathcal{L}_n^t(\theta) \ge \mathcal{L}_n^\star$ for every θ and every t.

Assumption 3 (Gate floor and ceiling)**.** There exists $g_{\min} \in (0, 1]$ such that $g_{\min} \le g_{n,I}^{(\ell)}(x; G_n) \le 1$ for every n, G_n, ℓ and every x in the support of $\mathcal{D}_n$.

The ceiling ≤ 1 is automatic since $g_{n,I}^{(\ell)}$ is a sum of softmax weights. The floor $g_{\min} > 0$ is structural for FLAME: the top-$(M+1)$ rule on $2(M+1)-1$ positions with $M+1$ wired to the individual expert forces, by pigeonhole, at least one individual position into every selection, and the post-softmax weight on the selected position is bounded away from 0 on any compact range of router logits (preserved by standard regularization throughout training).

Assumption 4 (Lipschitz in frozen experts)**.** There exists $\mathcal{C}_S > 0$ such that $|F_n(\theta;\xi;\bar{\mathbf{E}}, G_n) - F_n(\theta;\xi;\bar{\mathbf{E}}', G_n)| \le \mathcal{C}_S \|\bar{\mathbf{E}} - \bar{\mathbf{E}}'\|$ for every n, θ, ξ, G_n and every two configurations visited during training.

Assumption 5 (Slowly-varying frozen experts)**.** Define $\delta_t := \|\bar{\mathbf{E}}^{t+1} - \bar{\mathbf{E}}^t\|$, the cumulative drift $\Delta_T := \sum_{t=1}^{T} \delta_t$ is finite.

C.3 PER-STEP LEMMAS

Lemma 1 (Per-layer Jacobian factorization)**.** *For every $n, \xi = (x, y), \bar{\mathbf{E}}, G_n$ and every adapted layer ℓ, the per-layer gradient admits the exact factorization*

$$\nabla_{\theta_n^{(\ell)}} F_n(\theta_n;\xi;\bar{\mathbf{E}}, G_n) \;=\; g_{n,I}^{(\ell)}(x; G_n)\, \Phi_n^{(\ell)}(\xi;\theta_n;\bar{\mathbf{E}}, G_n), \tag{16}$$

where $\Phi_n^{(\ell)}$ is the per-layer Jacobian factor obtained by pulling the gate scalar $g_{n,I}^{(\ell)}$ out of the gradient.

Proof. By equation 4, the layer-ℓ output is

$$y_{\text{out}}^{(\ell)}(x;\theta_n,\bar{\mathbf{E}}, G_n) \;=\; W^{0,(\ell)}x \;+\; g_{n,I}^{(\ell)}(x; G_n)\, B_n^{(\ell)} A_n^{(\ell)}\, x \;+\; S^{(\ell)}(x;\bar{\mathbf{E}}, G_n), \tag{17}$$

where $S^{(\ell)}(x;\bar{\mathbf{E}}, G_n) = \sum_{e=1}^{M} g_{n,e}^{(\ell)}(x; G_n) \bar{B}^{(e),(\ell)} \bar{A}^{(e),(\ell)} x$ collects the frozen-expert contributions and depends on $x, \bar{\mathbf{E}}, G_n$.

Let $h_\ell := \nabla_{y_{\text{out}}^{(\ell)}} F_n$ denote the upstream gradient at layer ℓ. Applying the chain rule to equation 17 and using that $g_{n,I}^{(\ell)}(x; G_n)$ does not depend on θ_n, so it can be pulled out as a scalar,

$$\nabla_{A_n^{(\ell)}} F_n = g_{n,I}^{(\ell)}(x; G_n)\, (B_n^{(\ell)})^\top h_\ell\, x^\top, \tag{18}$$

$$\nabla_{B_n^{(\ell)}} F_n = g_{n,I}^{(\ell)}(x; G_n)\, h_\ell\, (A_n^{(\ell)} x)^\top. \tag{19}$$

Define

$$\Phi_n^{(\ell)}(\xi;\theta_n;\bar{\mathbf{E}}, G_n) \;:=\; \big((B_n^{(\ell)})^\top h_\ell x^\top,\; h_\ell (A_n^{(\ell)} x)^\top\big); \tag{20}$$

stacking equation 18–equation 19 gives equation 16. □

Remark 1 (No router-Jacobian remainder). Because $g_{n,I}^{(\ell)}(x; G_n)$ is independent of θ_n, no remainder term arises from differentiating the gate weights, so equation 16 is an exact equality. This holds because the router parameters G_n and the LoRA factors θ_n are disjoint in FLAME.

A direct consequence of Lemma 1 is the global identity

$$\|\nabla F_n(\theta_n;\xi;\bar{\mathbf{E}},G_n)\|^2 \;=\; \sum_{\ell=1}^{L}\big(g_{n,I}^{(\ell)}(x;G_n)\big)^2\big\|\Phi_n^{(\ell)}(\xi;\theta_n;\bar{\mathbf{E}},G_n)\big\|^2. \tag{21}$$

Lemma 2 (Effective L_Φ-smoothness)**.** *Under Assumptions 1 and 3, $\mathcal{L}_n^t$ is L_Φ-smooth for every fixed t.*

Proof. By Lemma 1 and the θ-independence of $g_{n,I}^{(\ell)}$,

$$\|\nabla F_n(\theta;\xi)-\nabla F_n(\theta';\xi)\|^2=\sum_{\ell}(g_{n,I}^{(\ell)})^2\|\Phi_n^{(\ell)}(\xi;\theta)-\Phi_n^{(\ell)}(\xi;\theta')\|^2\le L_\Phi^2\|\theta-\theta'\|^2,$$

using the ceiling $g_{n,I}^{(\ell)}\le 1$ and Assumption 1(i). Jensen's inequality $\|\mathbb{E}Z\|\le\mathbb{E}\|Z\|$ extends the bound to $\nabla\mathcal{L}_n^t$. □

Lemma 3 (Bounded second moment)**.** *Under Assumptions 1 and 3, $\|\nabla F_n(\theta;\xi)\|^2\le G_\Phi^2$ pointwise, hence $\mathbb{E}_\xi\|\nabla F_n(\theta;\xi)\|^2\le G_\Phi^2$.*

Proof. From equation 21 and $g_{n,I}^{(\ell)}\le 1$, $\|\nabla F_n(\theta;\xi)\|^2\le\sum_\ell\|\Phi_n^{(\ell)}\|^2\le G_\Phi^2$ by Assumption 1 (ii). □

Lemma 4 (Per-step descent inequality)**.** *Under Assumptions 1, 2, 3, and the SGD update equation 15 with $\eta\in(0,1/L_\Phi]$, every step k within round t satisfies*

$$\mathbb{E}\Big[\mathcal{L}_n^t(\theta_n^{(k+1)})\,\Big|\,\theta_n^{(k)}\Big]\;\le\;\mathcal{L}_n^t(\theta_n^{(k)})\;-\;\frac{\eta}{2}\big\|\nabla\mathcal{L}_n^t(\theta_n^{(k)})\big\|^2\;+\;\frac{L_\Phi\eta^2}{2}G_\Phi^2. \tag{22}$$

Proof. Lemma 2 and the standard descent lemma for L_Φ-smooth functions give, deterministically,

$$\mathcal{L}_n^t(\theta_n^{(k+1)})\le\mathcal{L}_n^t(\theta_n^{(k)})-\eta\big\langle\nabla\mathcal{L}_n^t(\theta_n^{(k)}),\nabla F_n(\theta_n^{(k)};\xi_n^{(k)})\big\rangle+\tfrac{L_\Phi\eta^2}{2}\|\nabla F_n(\theta_n^{(k)};\xi_n^{(k)})\|^2. \tag{23}$$

Take conditional expectation given $\theta_n^{(k)}$ on both sides. The cross term equals $\eta\|\nabla\mathcal{L}_n^t(\theta_n^{(k)})\|^2$ by unbiasedness. The quadratic term is bounded via $\mathbb{E}\|Z\|^2=\|\mathbb{E}Z\|^2+\mathrm{Var}(Z)$ with $\mathrm{Var}(Z)\le\mathbb{E}\|Z\|^2\le G_\Phi^2$ by Lemma 3, giving $\le\|\nabla\mathcal{L}_n^t\|^2+G_\Phi^2$. Substituting,

$$\mathbb{E}[\mathcal{L}_n^t(\theta_n^{(k+1)})\mid\theta_n^{(k)}]\le\mathcal{L}_n^t(\theta_n^{(k)})-\eta\big(1-\tfrac{L_\Phi\eta}{2}\big)\|\nabla\mathcal{L}_n^t(\theta_n^{(k)})\|^2+\tfrac{L_\Phi\eta^2}{2}G_\Phi^2.$$

Since $\eta\le 1/L_\Phi\Rightarrow 1-L_\Phi\eta/2\ge 1/2$, this yields equation 22. □

C.4 Frozen-Expert Drift Under Grassmannian Aggregation

We now bound the cumulative drift Δ_T of Theorem 8 under FLAME's choice of aggregator. The Grassmannian construction equation 6–equation 8 can be rewritten as a projection of the Euclidean mean onto the consensus row-space (Equation equation 25 below). This rewriting separates the per-round change of $\bar{A}_\mathcal{P}$ into a basis-frame component, which the Grassmannian projection removes, and a row-space component, which both aggregators retain.

Projection form of the Grassmannian aggregate. Let $\mathcal{P}\subseteq\{1,\dots,N\}\setminus\{n\}$ be a peer set (intra-cluster RoW or cross-cluster), with consensus basis $\bar{Q}_\mathcal{P}^t$ from equation 7 at round t. Define the consensus projector

$$\Pi_\mathcal{P}^t\;:=\;(\bar{Q}_\mathcal{P}^t)^\top\bar{Q}_\mathcal{P}^t\;\in\;\mathbb{R}^{d_{\text{in}}\times d_{\text{in}}}, \tag{24}$$

which is a rank-r orthogonal projector onto the consensus row-space. Combining equation 8 with the definition of $\bar{A}_\mathcal{P}^{\text{Eucl}}=\frac{1}{|\mathcal{P}|}\sum_{m\in\mathcal{P}}A_m$ yields the explicit factorization

$$\bar{A}_\mathcal{P}^{\text{Grass}}(t)\;=\;\bar{A}_\mathcal{P}^{\text{Eucl}}(t)\,\Pi_\mathcal{P}^t, \tag{25}$$

i.e., the Grassmannian aggregator is the Euclidean mean projected onto the consensus row-space.

Lemma 5 (Gauge invariance of the consensus projector)**.** *For any peer transformation $A_m \mapsto R_m A_m$ with $R_m \in \mathrm{GL}(r)$, the matrices $Q_m^\top Q_m$, $\bar{P}_\mathcal{P}$, and $\Pi_\mathcal{P}$ are invariant.*

Proof. $\mathrm{rowspan}(R_m A_m) = \mathrm{rowspan}(A_m)$, so the orthogonal projector $Q_m^\top Q_m$ depends only on the row-space, not on the basis. The remaining quantities are constructed from $\{Q_m^\top Q_m\}_{m\in\mathcal{P}}$ via equation 6 and equation 24 and inherit gauge invariance. □

Lemma 6 (Drift decomposition: Grass dominates Eucl)**.** *For any peer set $\mathcal{P}$ and consecutive rounds $t, t+1$,*

$$\big\|\bar{A}_\mathcal{P}^{\mathrm{Grass}}(t+1) - \bar{A}_\mathcal{P}^{\mathrm{Grass}}(t)\big\|_F \;\le\; \big\|\bar{A}_\mathcal{P}^{\mathrm{Eucl}}(t+1) - \bar{A}_\mathcal{P}^{\mathrm{Eucl}}(t)\big\|_F \;+\; \big\|\bar{A}_\mathcal{P}^{\mathrm{Eucl}}(t)\big\|_F \,\big\|\Pi_\mathcal{P}^{t+1} - \Pi_\mathcal{P}^{t}\big\|_{\mathrm{op}}. \tag{26}$$

In particular, when the consensus row-space is stable across rounds ($\Pi_\mathcal{P}^{t+1} = \Pi_\mathcal{P}^t$),

$$\big\|\bar{A}_\mathcal{P}^{\mathrm{Grass}}(t+1) - \bar{A}_\mathcal{P}^{\mathrm{Grass}}(t)\big\|_F \;\le\; \big\|\bar{A}_\mathcal{P}^{\mathrm{Eucl}}(t+1) - \bar{A}_\mathcal{P}^{\mathrm{Eucl}}(t)\big\|_F, \tag{27}$$

with strict inequality whenever the Euclidean residual has a nonzero component outside the consensus row-space.

Proof. Using equation 25, add and subtract $\bar{A}_\mathcal{P}^{\mathrm{Eucl}}(t)\,\Pi_\mathcal{P}^{t+1}$:

$$\bar{A}_\mathcal{P}^{\mathrm{Grass}}(t+1) - \bar{A}_\mathcal{P}^{\mathrm{Grass}}(t) = \big(\bar{A}_\mathcal{P}^{\mathrm{Eucl}}(t+1) - \bar{A}_\mathcal{P}^{\mathrm{Eucl}}(t)\big)\Pi_\mathcal{P}^{t+1} + \bar{A}_\mathcal{P}^{\mathrm{Eucl}}(t)\big(\Pi_\mathcal{P}^{t+1} - \Pi_\mathcal{P}^{t}\big).$$

Apply the triangle inequality $\|XY\|_F \le \|X\|_F \|Y\|_{\mathrm{op}}$ and enforce $\|\Pi_\mathcal{P}^{t+1}\|_{\mathrm{op}} = 1$ to the first term, we obtain equation 26. When $\Pi_\mathcal{P}^{t+1} = \Pi_\mathcal{P}^t = \Pi$, the second term vanishes, and the bound reduces to equation 27, with strict inequality iff the Euclidean residual has a nonzero component in $\Pi^\perp$. The Euclidean aggregator $\bar{A}_\mathcal{P}^{\mathrm{Eucl}}$ has no analogous projection step and incurs the residual in full. □

Consequence for Δ_T under FLAME's empirical regime. Section 2.2 of the main text (cf. Figure 1(a)) documents that within-cluster client A-row-spaces are tightly aligned across the converged training trajectory; equivalently, the consensus projectors $\Pi_\mathcal{P}^t$ change slowly across rounds, and the second term of equation 26 is dominated by the first component. By equation 27, the per-round contribution of the A-factor to δ_t is strictly smaller under Grass than under Eucl whenever the Euclidean residual carries a basis-frame component outside the consensus row-space, i.e., the LoRA gauge ambiguity of Section 2.2. The per-round contribution of the B-factor is identical under both aggregators (both use Euclidean averaging on B, by Observation 2). Hence

$$\Delta_T^{\mathrm{Grass}} \;\le\; \Delta_T^{\mathrm{Eucl}}, \tag{28}$$

with strict inequality whenever the LoRA gauge varies across rounds.

Lemma 7 (Loss continuity across rounds)**.** *Under Assumption 4, for every θ and every two consecutive rounds $t, t+1$,*

$$\big|\mathcal{L}_n^{t+1}(\theta) - \mathcal{L}_n^t(\theta)\big| \;\le\; \mathcal{C}_S\,\delta_t. \tag{29}$$

Proof. Take expectation over $\xi \sim \mathcal{D}_n$ of the pointwise bound from Assumption 4 with $\|\bar{\mathbf{E}}^{t+1} - \bar{\mathbf{E}}^t\| = \delta_t$, and apply $|\mathbb{E}[X - Y]| \le \mathbb{E}|X - Y|$. □

C.5 Main Convergence Theorem

Theorem 8 (Convergence of FLAME's individual expert)**.** *Under Assumptions 1–5, the SGD update equation 15 runs for $K = TE$ steps with learning rate $\eta \in (0, 1/L_\Phi]$, for every client n,*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\big\|\nabla\mathcal{L}_n^{t(k)}(\theta_n^{(k)})\big\|^2 \;\le\; \underbrace{\frac{2(\mathcal{L}_n^1(\theta_n^{(0)}) - \mathcal{L}_n^\star)}{\eta K}}_{\textit{initial gap}} + \underbrace{L_\Phi\,\eta\,G_\Phi^2}_{\textit{noise floor}} + \underbrace{\frac{2C_S\Delta_T}{\eta K}}_{\textit{aggregation drift}}. \tag{30}$$

With $D_0 := \mathcal{L}_n^1(\theta_n^{(0)}) - \mathcal{L}_n^\star + \mathcal{C}_S\Delta_T$ and the choice

$$\eta = \min\Big(\tfrac{1}{L_\Phi}, \sqrt{\tfrac{2D_0}{L_\Phi G_\Phi^2 K}}\Big), \tag{31}$$

equation 30 simplifies to

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big\|\nabla\mathcal{L}_n^{t(k)}(\theta_n^{(k)})\big\|^2 \;\le\; 2\sqrt{\tfrac{2L_\Phi G_\Phi^2 D_0}{K}}+\tfrac{2L_\Phi D_0}{K} \;=\; \mathcal{O}\big(K^{-1/2}\big). \tag{32}$$

The constant $\mathcal{C}_S\Delta_T$ in equation 30 is strictly smaller under FLAME's Grassmannian-on-A aggregator than under Euclidean-on-A aggregation, by equation 28.

Proof. **Step 1: Within-round telescope.** Sum the per-step descent equation 22 over $k = (t-1)E, \dots, tE-1$ and take total expectation:

$$\frac{\eta}{2}\sum_{k=(t-1)E}^{tE-1}\mathbb{E}\big\|\nabla\mathcal{L}_n^{t}(\theta_n^{(k)})\big\|^2 \;\le\; \mathbb{E}\big[\mathcal{L}_n^{t}(\theta_n^{((t-1)E)})\big]-\mathbb{E}\big[\mathcal{L}_n^{t}(\theta_n^{(tE)})\big]+\tfrac{EL_\Phi\eta^2}{2}G_\Phi^2. \tag{33}$$

Step 2: Cross-round bridge. The iterate $\theta_n^{(tE)}$ is shared between the end of round t and the start of round $t+1$, but the loss switches from $\mathcal{L}_n^t$ to $\mathcal{L}_n^{t+1}$. Lemma 7 gives

$$\mathbb{E}\big[\mathcal{L}_n^{t+1}(\theta_n^{(tE)})\big] \;\le\; \mathbb{E}\big[\mathcal{L}_n^{t}(\theta_n^{(tE)})\big]+\mathcal{C}_S\delta_t. \tag{34}$$

Step 3: Cross-round telescope. Sum equation 33 over $t = 1, \dots, T$ and chain via equation 34: at each round boundary, $-\mathbb{E}[\mathcal{L}_n^t(\theta_n^{(tE)})]$ is replaced by $-\mathbb{E}[\mathcal{L}_n^{t+1}(\theta_n^{(tE)})]+\mathcal{C}_S\delta_t$, then cancels the matching term that opens round $t+1$. Only the global endpoints survive:

$$\frac{\eta}{2}\sum_{k=0}^{K-1}\mathbb{E}\big\|\nabla\mathcal{L}_n^{t(k)}(\theta_n^{(k)})\big\|^2 \le \mathcal{L}_n^1(\theta_n^{(0)})-\mathbb{E}\big[\mathcal{L}_n^{T}(\theta_n^{(TE)})\big]+\sum_{t=1}^{T-1}\mathcal{C}_S\delta_t+\tfrac{KL_\Phi\eta^2}{2}G_\Phi^2. \tag{35}$$

Step 4: Lower bound and rearrange. Apply Assumption 2 ($\mathbb{E}[\mathcal{L}_n^T(\theta_n^{(TE)})] \ge \mathcal{L}_n^\star$), use $\sum_{t=1}^{T-1}\delta_t \le \Delta_T$, and divide by $\eta K/2$ to obtain equation 30.

Step 5: Closed-form rate. The unconstrained minimizer of $\frac{2D_0}{\eta K}+L_\Phi\eta G_\Phi^2$ is $\eta^\star = \sqrt{2D_0/(L_\Phi G_\Phi^2 K)}$, attaining $2\sqrt{2L_\Phi G_\Phi^2 D_0/K}$. We split into two cases for equation 31:

- If $\eta^\star \le 1/L_\Phi$ (large K), $\eta = \eta^\star$ gives bound $2\sqrt{2L_\Phi G_\Phi^2 D_0/K}$; the second term in equation 32 is non-negative slack.
- If $\eta^\star > 1/L_\Phi$ (small K, equivalent to $K < 2L_\Phi D_0/G_\Phi^2$), $\eta = 1/L_\Phi$ gives initial-gap term $2L_\Phi D_0/K$ and noise-floor term G_Φ^2. The condition $K < 2L_\Phi D_0/G_\Phi^2$ implies $G_\Phi^2 < 2\sqrt{2L_\Phi G_\Phi^2 D_0/K}$, so the bound in equation 32 dominates.

Step 6: Comparison to Euclidean aggregation. The cumulative drift Δ_T enters equation 30 (and D_0) linearly through the additive term $\frac{2C_S\Delta_T}{\eta K}$, so the bound is monotone in Δ_T; equation 28 gives $\Delta_T^{\mathrm{Grass}} \le \Delta_T^{\mathrm{Eucl}}$ with strict inequality whenever the LoRA gauge varies across rounds. The Grassmannian-aggregator bound is therefore strictly tighter in this regime.

C.6 Role of the Gate Floor: Fisher Mass

Theorem 8 uses the gate ceiling but not the floor. The floor $g_{\min} > 0$ enters elsewhere: it gives a uniform per-token lower bound on the gradient on the individual expert, and hence on the Fisher mass accumulated by the post-hoc Laplace stage of Section 3.3.

Lemma 9 (Per-step gradient lower bound)**.** *Under Assumption 3, for every $n, \theta, \xi = (x, y), G_n, \bar{\mathbf{E}}$ and every ℓ,*

$$\big\|\nabla_{\theta_n^{(\ell)}}F_n(\theta;\xi;\bar{\mathbf{E}},G_n)\big\|^2 \;\ge\; g_{\min}^2\,\big\|\Phi_n^{(\ell)}(\xi;\theta;\bar{\mathbf{E}},G_n)\big\|^2. \tag{36}$$

Proof. Lemma 1 gives $\|\nabla_{\theta_n^{(\ell)}}F_n\| = g_{n,I}^{(\ell)}\|\Phi_n^{(\ell)}\| \ge g_{\min}\|\Phi_n^{(\ell)}\|$ by the floor; squaring yields equation 36. □

Corollary 1 (Uniform Fisher contribution)**.** *The Kronecker-factored Fisher equation 10 accumulated by FLAME's post-hoc Laplace stage receives a contribution from every token ξ in the support of $\mathcal{D}_n$ with multiplicative weight at least $g_{\min}^2$, uniformly in ξ.*

Proof. The KFAC factors are gradient outer products; equation 36 provides a uniform pointwise lower bound on each token's contribution. □

Why FLAME and not FedALT. For FLAME's enlarged-gate router, $g_{\min} > 0$ is structurally guaranteed: pigeonhole on the $2(M+1)-1$-dimensional softmax forces at least one individual position into every top-$(M+1)$ selection, with post-softmax weight bounded away from 0 on the compact range of router logits. For FedALT's unconstrained two-expert mixer, no such bound holds: the gate weight on the individual expert can be driven arbitrarily small on tokens where the mixer prefers the Row expert, so $g_{\min} = 0$ in general. Corollary 1 therefore applies to FLAME but not to FedALT, formalizing the qualitative contrast of Section 3.3.

Comparison to prior personalized federated LoRA analyses. The convergence analyses in FedALT (Bian et al., 2026) and FedLEASE (Wang et al., 2026) establish a parameter-sequence stability bound of the form $\|Z_k^{t+1} - Z_k^t\| \to \mathcal{O}(\epsilon/(1-\beta))$ under per-client strong convexity and smoothness, where β is the Lipschitz constant of the optimal-parameter map with respect to the frozen aggregate. Theorem 8 works in the nonconvex SGD setting and provides an $\mathcal{O}(K^{-1/2})$ rate to a stationary point, so the two analyses target different quantities and are not directly comparable as numerical statements. What is specific to our bound is that each additive term in equation 30 traces back to a single FLAME component: the smoothness and variance constants L_Φ and G_Φ^2 are induced by the gate ceiling of the enlarged router, the cumulative drift is controlled by the Grassmannian aggregator through Lemma 6, and the Fisher-mass condition used by the Laplace stage is supplied by the gate floor through Corollary 1. The FedALT and FedLEASE bounds depend on β alone, which is consistent with neither method introducing a manifold-aware aggregator or a Bayesian post-processing step that would require this kind of decomposition.

□