

FedLing: Linguistic-Structure-Aware Federated Learning for Fair Multilingual LLM Adaptation

Xiaorui Luo Hongbo Yin Ziyu Chen Yin Zhang*

University of Electronic Science and Technology of China
{luoxiaorui,hongboyin,ziyuchen}@std.uestc.edu.cn
zhangyin123@uestc.edu.cn

Abstract

Multilingual large language models (LLMs) exhibit substantial performance disparities across languages, with low-resource languages often lagging far behind high-resource ones. Federated learning (FL) offers a privacy-preserving path for decentralized multilingual adaptation, but existing methods typically aggregate client updates without explicitly modeling structural differences across languages, inducing gradient interference, amplifying negative transfer, and further disadvantaging underrepresented languages. We introduce FEDLING, a linguistic-structure-aware federated learning framework. FEDLING performs typology-conditioned spectral routing of client LoRA updates, decomposing them into linguistically transferable consensus subspaces and language-specific divergence subspaces. Unlike posterior-only decomposition methods that require accumulated training history to estimate transferability, FEDLING explicitly mixes a round-one energy prior with typology-conditioned empirical subspace alignment through a single coefficient ρ . A continuous typological isolation score further prevents negative transfer for languages without close federation neighbors, and a reciprocal fairness proxy adapts upload bandwidth toward underperforming languages. Experiments across 40 languages show that FEDLING consistently improves multilingual adaptation: +7.9 average ROUGE-L and +11.6 ROUGE-L on low-resource languages over FedAvg, with 52% variance reduction and 35% upload savings. The advantage is largest in cold-start and data-scarce regimes, validating the combination of energy-based cold-start routing and typology-conditioned posterior alignment.

1 Introduction

More than 7,000 languages remain in active use worldwide (Eberhard et al., 2023), yet practical LLM support is still concentrated in fewer than 100

*Corresponding author.

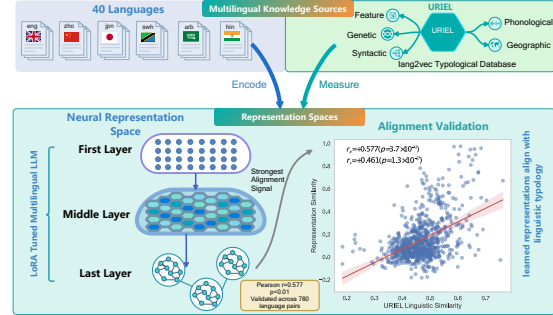


Figure 1: Motivation: cross-lingual representational similarity is positively correlated with URIEL typological similarity.

high-resource languages (Joshi et al., 2020). The gap persists even for multilingual LLMs: systems that exceed 90% accuracy on English benchmarks can fall below 60% on low-resource languages such as Yoruba, Quechua, and Sinhala (Singh et al., 2024; Ahuja et al., 2023). This unevenness limits the usefulness of LLMs for underrepresented language communities, particularly when sensitive or locally held data cannot be centralized.

Federated Learning for Multilingual LLMs. Federated learning (FL) (McMahan et al., 2017) allows multilingual adaptation without pooling raw client data, but multilingual FL is not simply another instance of non-IID optimization. Languages differ in syntax, morphology, and phonology, so direct aggregation can mix incompatible update directions. At the same time, data volume follows a power-law distribution (Zhao et al., 2024), allowing high-resource languages to dominate the model, while typologically isolated languages may receive misleading signal from distant clients.

Limitations of Existing Methods. Existing FL methods address these pressures only in part. FedAvg (McMahan et al., 2017) ignores linguistic structure, while personalized or language-aware methods such as FedPer (Arivazhagan et al., 2019),

FedMoE (Mei et al., 2024), FedHLT (Guo et al., 2024a), and FedLFC (Guo et al., 2024b) rely on client-specific heads, gating, or discrete language groups. These designs do not explicitly separate transferable shared structure from language-specific adaptation, and they offer limited protection when a language has no close neighbor in the federation.

Our Insight. Linguistic typology offers a concrete way to make this structure visible to the optimizer. Concretely, we compute the cosine similarity between the LoRA-update gradients of every unordered pair of the 40 single-language clients in our setting (780 pairs), and find that this pairwise gradient similarity is strongly correlated with URIEL+ typological similarity (Khan et al., 2025) ($r = 0.577$, $p < 0.001$; Figure 1, Appendix B). This correlation is motivational rather than causal evidence: it shows that typological similarity is predictive of which clients exhibit similar update geometry, so typology is a suitable signal for weighting empirical subspace alignment once client bases have been cached. The correlation replicates on MASSIVE ($r = 0.524$, $p < 0.001$) and XNLI ($r = 0.431$, $p < 0.001$). Because no such cache exists in the first communication round, FEDLING pairs this typology-conditioned posterior with a history-free energy prior that provides a stable cold start (Section 4.1).

Our Solution. We propose FEDLING, a linguistic-structure-aware FL framework for multilingual LLM adaptation. Each client decomposes its LoRA update into consensus and divergence subspaces using an energy prior and typology-conditioned empirical alignment. The server aggregates only the consensus components, with URIEL-based similarity controlling each client’s influence, while an isolation-aware gate lets weak-neighbor languages preserve more local divergence. A reciprocal fairness proxy further adjusts upload bandwidth toward underperforming clients.

The paper makes four contributions.

1. We identify typological isolation as a failure mode in multilingual FL: similarity-based aggregation can harm languages whose nearest federation neighbors remain structurally distant.
2. We propose FEDLING, which routes LoRA singular directions using an energy-based cold-start prior and typology-conditioned subspace alignment, with a continuous isolation score controlling how strongly each language absorbs the

global consensus.

3. We derive fairness-driven upload allocation from a reciprocal performance proxy and provide a consensus-subspace conflict analysis linking typology-aware decomposition to measurable subspace alignment.
4. Experiments across three benchmarks and 40 languages show gains in task performance, fairness, and upload efficiency. On Aya, FEDLING outperforms FedAvg by 7.9 ROUGE-L points overall and 11.6 points on low-resource languages, reduces variance by 52%, and reduces upload cost by 35%. Its advantage over MuLA-F, the closest SVD-based multilingual FedLoRA baseline, is 1.2 points on average and 2.4 points on low-resource languages, and is largest in cold-start and data-scarce regimes, supporting the energy-prior-posterior design.

2 Related Work

Federated PEFT for LLMs. Parameter-efficient tuning methods such as adapters, prefix tuning, prompt tuning, BitFit, and LoRA reduce the trainable footprint of large pretrained models (Houlsby et al., 2019; Li and Liang, 2021; Lester et al., 2021; Ben Zaken et al., 2022; Hu et al., 2022). Recent FL methods build on this idea to reduce communication, including FedPETuning (Zhang et al., 2023), prompt-based FedPEAT (Chua et al., 2024), personalized pFedLoRA (Yi et al., 2024), and compression-oriented SLoRA/FedPara (Babakniya et al., 2023; Hyeon-Woo et al., 2022). These methods reduce update size or client drift, but treat multilingual heterogeneity mostly as generic non-IID variation rather than as structured linguistic transfer.

Multilingual and Fairness-aware FL. FedHLT (Guo et al., 2024a) and FedLFC (Guo et al., 2024b) inject language structure through hierarchical trees or family clusters, but such discrete groupings miss intra-family variation and cross-family similarity. MuLA-F (Li et al., 2025) is closest to our work: it also decomposes LoRA updates with SVD, but relies on posterior training dynamics to identify language-contributing directions. FEDLING instead mixes an energy-based cold-start prior with typology-conditioned posterior subspace alignment, making routing usable from the first round and more stable under data scarcity.

Typology and Fairness. Cross-lingual pre-

trained models show that shared representations can support transfer across languages, but transfer remains sensitive to language relatedness and representation geometry (Pires et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020a,b). URIEL/URIEL+ encode syntactic, phonological, and morphological similarity across languages (Littell et al., 2017; Khan et al., 2025), and prior work links such features to cross-lingual transfer (Lin et al., 2019; Ponti et al., 2019). In parallel, q-FFL and related fair FL objectives reduce client disparity through loss-level reweighting (Li et al., 2020b; Mohri et al., 2019). FEDLING connects these lines by using typology as an operational signal for routing, aggregation, isolation handling, and fairness-driven upload allocation.

3 Problem Formulation

We consider K clients, where client k has private data $\mathcal{D}_k$ drawn from distribution $\mathcal{P}_k$. In the main experiments, each client is associated with one language ℓ_k to isolate language-level effects. The formulation also supports clients with multilingual local data by replacing the single label with an empirical language distribution $\lambda_k(\ell)$ over the languages observed on client k . The client distributions differ in both data quantity and linguistic structure. A standard FL objective adapts a multilingual model f_θ by minimizing average client risk:

$$\min_{\theta} \mathcal{L}(\theta) = \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x,y) \sim \mathcal{P}_k} [\ell(f_\theta(x), y)], \quad (1)$$

where $\ell(\cdot, \cdot)$ is the supervised fine-tuning loss of the task at hand; Section 5.1 instantiates three tasks (Aya generation, MASSIVE intent classification, and XNLI zero-shot transfer). This objective treats all updates as directly aggregatable and ignores cross-language disparity. Let $\text{Score}_k(\theta)$ be client k 's evaluation metric on its local validation split—ROUGE-L, macro-F1, or F1 depending on the task; since each client holds one language in the main setting, each client has exactly one such score. Let r_k be the fraction of its candidate consensus update C_k uploaded to the server. We formulate multilin-

gual adaptation as:

$$\begin{aligned} \max_{\theta, \mathbf{r}} \quad & \text{Score}_{\text{avg}}(\theta) - \beta \text{Var}(\{\text{Score}_k(\theta)\}_{k=1}^K) \\ \text{s.t.} \quad & \sum_{k=1}^K r_k |C_k| \leq B, \\ & r_{\min} \leq r_k \leq r_{\max}, \quad k = 1, \dots, K, \end{aligned} \quad (2)$$

where $\text{Score}_{\text{avg}}(\theta) = \frac{1}{K} \sum_k \text{Score}_k(\theta)$, β controls the performance–fairness tradeoff, and B is the global communication budget. Penalizing cross-client performance dispersion in this way is standard in federated fairness (Li et al., 2020b; Mohri et al., 2019). We emphasize that Eq. (2) is a design objective rather than the loss optimized during training: Score_k is only observable at evaluation rounds, so FEDLING realizes the fairness pressure through the reciprocal proxy and upload-allocation rule in Section 4.4, where the adjustment step μ plays the practical role of β in trading average performance against dispersion. This objective captures the two requirements FEDLING targets: improve average performance while reducing cross-language disparity, and allocate scarce upload bandwidth toward underperforming languages.

4 Methodology

FEDLING is built around three operations: typology-aware subspace decomposition, typology-guided aggregation with isolation-aware preservation, and fairness-driven upload allocation (Figure 2).

4.1 Typology-Conditioned Spectral Routing

Multilingual client updates usually contain both transferable structure and language-specific adaptation: cross-client gradient alignment is only moderate under naive averaging and even negative for some language pairs (Figure 6), and removing the consensus/divergence split costs substantial performance (Table 6). FEDLING therefore does not aggregate a LoRA update as a single vector. Instead, it performs *typology-conditioned spectral routing*, assigning each singular direction to either a consensus or divergence route according to its estimated transferability across linguistically compatible clients.

Layer-wise SVD of LoRA updates. For client k and adapter layer i , let ΔA_i^k and ΔB_i^k denote the local LoRA update matrices, and form the effective

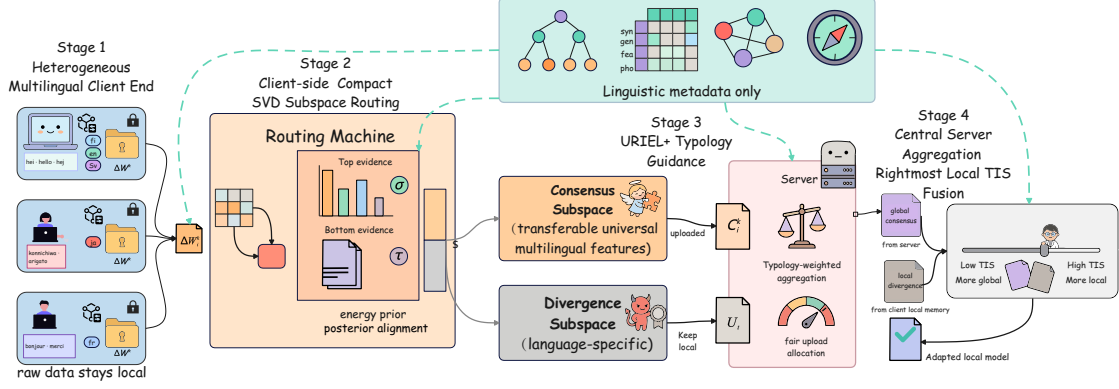


Figure 2: Overview of FEDLING. Multilingual clients keep raw data local, decompose LoRA updates through spectral routing, and upload only transferable consensus components. URIEL+ typology provides linguistic metadata for transferability scoring, server-side consensus aggregation, and the TIS-gated local fusion that balances global consensus with locally retained divergence.

update

$$\Delta W_i^k = \Delta A_i^k (\Delta B_i^k)^\top \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}, \quad (3)$$

whose compact SVD is

$$\Delta W_i^k = U_i^k \Sigma_i^k (V_i^k)^\top. \quad (4)$$

Since LoRA has rank r , ΔW_i^k has rank at most r . We write u_{im}^k for the m -th left singular vector and σ_{im}^k for the m -th singular value, and denote $U_{i,1:m}^k = [u_{i1}^k, \dots, u_{im}^k]$.

Two transferability signals. We estimate transferability from two signals. The first is local energy: a direction that carries a larger share of the update, formalized as the singular-value share in Eq. (5), is a stronger candidate before any cross-client history is available. The second is cross-client geometry: a direction is more credible as shared structure when its span overlaps with the rank- r singular bases cached by typologically similar clients in the previous round, formalized in Eqs. (6)–(7). The score below combines these signals while keeping their roles separate.

Prior: energy share. The prior uses only direction magnitude and is available from the first round:

$$\pi_{im}^{k,\text{prior}} = \frac{\sigma_{im}^k}{\sum_{m'} \sigma_{im'}^k}. \quad (5)$$

The prior already sums to one over directions. We use the singular-value share rather than a squared singular-value share because σ_m^2 over-concentrates the first-round prior on a few dominant directions, before the posterior can correct it; in our low-rank

LoRA setting this made the consensus partition less stable. We include σ_m^2 as a prior variant in Table 4. The prior is intentionally magnitude-only. Typology features describe languages, not singular directions, so multiplying Eq. (5) by a client-level typology factor would cancel under per-layer normalization and would not affect within-client direction selection. We therefore place direction-specific typological information in the posterior term and handle typological isolation through the TIS gate in Section 4.3.

Posterior: subspace alignment. The posterior measures how much client k 's singular directions overlap with the singular subspaces cached by typologically similar clients in the previous round. A formulation based on single-vector inner products $\langle u_{im}^k, u_{im'}^j \rangle$ is brittle: SVD singular vectors are sign-ambiguous, near-degenerate singular values admit arbitrary orthogonal bases, and singular-value ranks are local to each client rather than shared geometric identities.

We therefore use a subspace alignment measure. Let $U_i^{j,t-1} \in \mathbb{R}^{d_{\text{in}} \times r}$ denote the full rank- r basis cached by client j from the previous round. We define the incremental subspace alignment

$$A_{im}^{k \rightarrow j} = \|(u_{im}^{k,t})^\top U_i^{j,t-1}\|_2^2, \quad (6)$$

which is the squared length of $u_{im}^{k,t}$'s projection onto the span of the neighbor's cached basis. Equivalently, $A_{im}^{k \rightarrow j}$ is the Frobenius-norm gain obtained by adding $u_{im}^{k,t}$ to client k 's leading subspace: $\|(U_{i,1:m}^{k,t})^\top U_i^{j,t-1}\|_F^2 - \|(U_{i,1:m-1}^{k,t})^\top U_i^{j,t-1}\|_F^2$.

Invariance properties. The score in Eq. (6) is invariant under sign flips of any singular vector and under orthogonal rotation or column reordering of the cached neighbor basis $U_i^{j,t-1}$, since $\|\cdot\|_F^2$ is preserved by right-multiplication with an orthogonal matrix. It is not fully invariant with respect to the current client’s basis: when $\sigma_m^{k,t} \approx \sigma_{m+1}^{k,t}$, the SVD solver may place $u_{im}^{k,t}$ and $u_{i,m+1}^{k,t}$ at arbitrary orthogonal positions within the degenerate two-dimensional subspace, shifting their individual scores. Their sum over the degenerate block, however, is invariant because the block span is unchanged.

We address this case with a gap-based grouping rule. Consecutive singular directions satisfying $\sigma_{im}^k / \sigma_{i,m+1}^k < 1 + \delta$ for $\delta = 10^{-3}$ are treated as one block, and all directions in the block share the block-average score before thresholding. This keeps the partition less sensitive to SVD-solver choices and round-to-round numerical variation.

Posterior aggregation. Let $\mathcal{N}_k$ be the top- q clients most similar to client k under the client-level similarity s_{kj} . The neighbor bases $U_i^{j,t-1}$ are lightweight metadata uploaded in the previous round and relayed by the server, not raw client data or full model updates. They may still reveal coarse update geometry, so they should be treated as model-update metadata under the same trust assumptions as standard FL updates; we discuss this privacy limitation explicitly in Section 6. The unnormalized posterior is the typology-weighted average

$$\tilde{\pi}_{im}^{k,\text{post}} = \frac{1}{|\mathcal{N}_k|} \sum_{j \in \mathcal{N}_k} s_{kj} A_{im}^{k \rightarrow j}. \quad (7)$$

In the first round, no cached bases exist; we initialize $\tilde{\pi}_{im}^{k,\text{post}} = \pi_{im}^{k,\text{prior}}$ so that Eq. (9) reduces to a pure prior at $t = 1$ regardless of ρ .

Scale matching. The two signals have different natural scales: π^{prior} is a probability distribution by construction, while $\tilde{\pi}^{\text{post}}$ is computed from projection lengths weighted by $s_{kj} \in [0, 1]$. Without normalization, ρ would not have a stable interpretation. We therefore L_1 -normalize the posterior across directions within each layer, with a small ϵ for numerical stability:

$$\pi_{im}^{k,\text{post}} = \frac{\tilde{\pi}_{im}^{k,\text{post}} + \epsilon}{\sum_{m'} (\tilde{\pi}_{im'}^{k,\text{post}} + \epsilon)}. \quad (8)$$

After normalization, both signals are probability distributions over the r directions.

Combined score. The transferability score is the ρ -controlled convex combination

$$\tau_{im}^k = \rho \pi_{im}^{k,\text{prior}} + (1 - \rho) \pi_{im}^{k,\text{post}}. \quad (9)$$

At $\rho = 1$, direction selection reduces to pure energy-based ranking. At $\rho = 0$, direction selection relies entirely on observed subspace overlap with typologically similar neighbors, similar in spirit to posterior-only decomposition methods such as MuLA-F’s Diff-eRank. We set $\rho = 0.7$ in our main experiments on a held-out validation split and analyze the full sweep in Section 5.4.

Why combining the two helps. The two signals capture different evidence. The prior records direction *magnitude*, namely which directions carry most of the local update. The posterior records direction *geometry*, namely which subspaces typologically nearby clients have repeatedly used. The prior is useful from the first round but cannot identify typological agreement by itself; the posterior is more direction-specific but is weak in early rounds and under data scarcity. With fixed $\rho = 0.7$, the posterior contributes little at $t = 1$ and becomes more discriminative as cached bases stabilize. Section 5.4 reports this transition empirically.

Consensus and divergence subspaces. We partition singular directions, or near-degenerate blocks treated jointly by the grouping rule above, according to a percentile threshold computed per client and per layer:

$$\mathcal{C}_i^k = \{m : \tau_{im}^k > \tau_{\text{thr}}\}, \quad (10)$$

$$\mathcal{D}_i^k = \{m : \tau_{im}^k \leq \tau_{\text{thr}}\}, \quad (11)$$

where τ_{thr} is set to the 70th percentile of $\{\tau_{im}^k\}_m$, and directions in the same near-degenerate block share the block-average score so they are assigned jointly. The update is then decomposed as

$$\Delta W_i^k = \underbrace{U_i^{k,\mathcal{C}} \Sigma_i^{k,\mathcal{C}} (V_i^{k,\mathcal{C}})^\top}_{C_i^k} + \underbrace{U_i^{k,\mathcal{D}} \Sigma_i^{k,\mathcal{D}} (V_i^{k,\mathcal{D}})^\top}_{D_i^k}, \quad (12)$$

where C_i^k is the consensus component uploaded for aggregation and D_i^k is retained as language-specific adaptation. Isolation handling is decoupled from this partition and performed by the TIS-modulated gate in Section 4.3.

4.2 Typology-Guided Aggregation

Linguistic similarity. We construct a client similarity matrix $S \in [0, 1]^{K \times K}$ from URIEL+ (Khan

et al., 2025). For two languages ℓ and ℓ' , the base typological similarity combines syntactic, phonological, and morphological similarities:

$$a(\ell, \ell') = w_{\text{syn}} s_{\text{syn}}(\ell, \ell') + w_{\text{pho}} s_{\text{pho}}(\ell, \ell') + w_{\text{mor}} s_{\text{mor}}(\ell, \ell'). \quad (13)$$

where the weights sum to one. If client k contains multiple languages, let $\lambda_k(\ell)$ be the empirical fraction of language ℓ in its local data. Client-level similarity is then

$$s_{kj} = \sum_{\ell} \sum_{\ell'} \lambda_k(\ell) \lambda_j(\ell') a(\ell, \ell'). \quad (14)$$

The single-language case used in our main experiments is recovered when each λ_k is a point mass at ℓ_k . We use $w_{\text{syn}} = 0.5$, $w_{\text{pho}} = 0.3$, and $w_{\text{mor}} = 0.2$, following evidence that syntactic similarity is particularly predictive of cross-lingual transfer (Lin et al., 2019).

Consensus aggregation. At round t , let $\mathcal{A}^t$ be the set of active clients. For each client k , we define a typological centrality weight:

$$\beta_k^t = \exp \left(\gamma \cdot \frac{1}{|\mathcal{A}^t| - 1} \sum_{j \in \mathcal{A}^t, j \neq k} s_{kj} \right), \quad (15)$$

where γ controls the sharpness of the weighting. The final aggregation coefficient combines typological centrality and data size:

$$\alpha_k^t = \frac{\beta_k^t n_k}{\sum_{j \in \mathcal{A}^t} \beta_j^t n_j}. \quad (16)$$

The server aggregates only consensus components:

$$C_{\text{global}}^t = \sum_{k \in \mathcal{A}^t} \alpha_k^t C_k^t. \quad (17)$$

This weighting favors clients that are typologically compatible with the active federation while reducing the influence of structurally distant updates. Unlike per-language-bucket aggregation, FEDLING forms a single global consensus C_{global}^t and lets each client decide how much to absorb through the isolation-aware gate in Section 4.3. This avoids maintaining sparse language buckets, reduces server-side aggregation cost, and, as shown in Appendix D, gives a better fairness–performance trade-off under our heterogeneous client distribution.

4.3 Isolation-Aware Local Preservation

Typology-guided aggregation is most reliable when each client has structurally compatible neighbors in the federation. This condition does not hold for typologically isolated languages such as Finnish, Korean, and Japanese in the single-language setting, or for multilingual clients whose language distribution remains distant from all other clients. A hard decision is undesirable in either direction: forcing aggregation can inject unrelated-family signal, while excluding isolated clients entirely discards task-level information that the federation may still provide.

We resolve this tension with a continuous gating mechanism driven by a typological isolation score (TIS). For client k , let $\mathcal{N}_k^{\text{top-}m}$ denote its top- m nearest clients under Eq. (14):

$$\text{TIS}_k = 1 - \frac{1}{|\mathcal{N}_k^{\text{top-}m}|} \sum_{j \in \mathcal{N}_k^{\text{top-}m}} s_{kj}. \quad (18)$$

TIS_k is bounded in $[0, 1]$, equals 0 for well-integrated languages, and approaches 1 for fully isolated languages. The update applied to client k is a TIS-modulated mixture of the typology-aggregated consensus and the language-specific divergence:

$$\Delta \theta_k^{t+1} = (1 - \text{TIS}_k) C_{\text{global}}^t + \text{TIS}_k D_k^t. \quad (19)$$

Low-TIS languages absorb more of the typology-aggregated consensus, while high-TIS languages retain more of their own divergence. Because Eq. (19) is continuous, partially isolated languages can still benefit from global information without being forced into the same aggregation regime as well-connected languages.

4.4 Fairness-Driven Upload Allocation

Low-resource and underperforming clients often need more communication opportunity than a uniform budget gives them. FEDLING therefore adapts each client’s upload ratio according to observed performance, connecting the fairness-aware objective in Section 3 to a simple update rule.

Let $r_k^t \in [r_{\min}, r_{\max}]$ denote the fraction of client k ’s consensus component uploaded at round t , and let $|C_k^t|$ denote the size of the candidate consensus update. The feasible upload set is

$$\mathcal{R} = \left\{ \mathbf{r} : \sum_{k=1}^K r_k |C_k^t| \leq B, r_{\min} \leq r_k \leq r_{\max} \right\}. \quad (20)$$

Directly optimizing Eq. (2) over $\mathbf{r}$ is impractical because Score_k is observed only after evaluation rounds and is not available as a closed-form function of r_k . Inspired by proportional-fair resource allocation (Kelly et al., 1998), we use the reciprocal-sum proxy

$$\mathcal{F}(\mathbf{r}) = \sum_{k=1}^K \frac{1}{\text{Score}_k(\mathbf{r}) + \epsilon}, \quad (21)$$

which penalizes low-performing clients more strongly than high-performing ones because $x \mapsto 1/x$ grows rapidly near zero.

The exact derivative $\partial \text{Score}_k / \partial r_k$ is unobservable during federated training. We therefore use a centered reciprocal performance gap as a stable update direction:

$$\Delta_k^t = \frac{1}{\text{Score}_k^t + \epsilon} - \frac{1}{\text{Score}^t + \epsilon}, \quad (22)$$

where $\text{Score}^t = \frac{1}{K} \sum_k \text{Score}_k^t$ and $\epsilon > 0$ ensures numerical stability. Every T_{adj} rounds, upload ratios are updated by

$$r_k^{t+1} = \Pi_{\mathcal{R}}(r_k^t + \mu \Delta_k^t), \quad (23)$$

where μ is the step size and $\Pi_{\mathcal{R}}$ projects onto the feasible communication set. Clients below the average validation score receive larger upload ratios, while the projection enforces per-client bounds and the global budget. This update approximates max-min allocation conservatively: it shifts communication toward the current tail without assuming a differentiable score model.

4.5 Theoretical Analysis and Algorithm

The analysis in Appendix C decomposes conflict between routed client gradients into a subspace-misalignment term and an in-subspace task-discrepancy term. The transferability score in Eq. (9) targets the first term by selecting singular directions aligned with typologically similar clients’ cached subspaces; Appendix B.4 measures the resulting alignment improvements. We present this analysis as a mechanistic explanation of why the routing rule reduces conflict, not as a formal convergence guarantee; the quantitative claim we rely on, the consensus-conflict bound in Lemma 1, is exact and is validated empirically in Appendix B.4.

Algorithm 1 in Appendix A summarizes the full training procedure, including server-side cached-basis relay and fairness-driven upload adjustment.

4.6 Complexity Analysis

Each client performs standard local LoRA fine-tuning and one compact SVD per adapter layer. Since the effective LoRA update has rank at most r , this decomposition is small relative to local optimization. Upload communication decreases because only selected consensus components are uploaded, while divergence components remain local. For Llama-3.1-8B with rank-16 LoRA, a full LoRA upload is 8.4 MB/client/round; FEDLING averages 5.5 MB after routing and cached-basis overhead, giving about 35% upload savings. Our headline communication claims concern the uplink; for completeness, FEDLING’s steady-state downlink is 9.2 MB/client/round (2.0 MB global consensus plus 7.2 MB relayed cached bases) versus 8.4 MB for FedAvg, so the total round-trip cost is still lower at 14.7 versus 16.8 MB/client/round (about 12.5% savings). Appendix F.5 gives the full accounting.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on Aya (Singh et al., 2024) for multilingual generation, MASSIVE (FitzGerald et al., 2023) for intent classification, and XNLI (Conneau et al., 2018) for zero-shot transfer. Aya covers 40 languages across 12 families; grouped results use the fixed Aya Evaluation Suite resource tiers. The XNLI split is listed in Table 8.

Baselines. We compare against standard FL (FedAvg, FedProx, FedBN), personalized FL (FedPer, FedProto, FedMoE), multilingual/fairness-aware FL (FedHLT, FedLFC, q-FFL), and MuLA-F (Li et al., 2025), the closest SVD-based multilingual FedLoRA baseline. Centralized and Local Only training serve as upper and lower bounds.

Evaluation Metrics. We report ROUGE-L for Aya, macro-F1 for MASSIVE, and F1 for XNLI. Unless otherwise stated, results are means over five seeds. Significance is tested against the strongest non-FEDLING baseline in the same column with a two-sided paired bootstrap over seed-language scores (10,000 resamples) and Holm–Bonferroni correction; $\dagger$ denotes $p < 0.05$. We fine-tune Llama-3.1-8B (Dubey et al., 2024) with rank-16 LoRA for 100 rounds; details are in Appendix H.

5.2 Main Results

Table 1 reports results on the Aya dataset. FEDLING achieves an average ROUGE-L of 58.6,

Method	ROUGE-L			Fairness		Efficiency	
	Avg $\uparrow$	Low $\uparrow$	High $\uparrow$	Var $\downarrow$	CV $\downarrow$	Upload $\downarrow$	Time
Local Only	42.8 $\pm$ 1.6	32.4 $\pm$ 1.9	58.1 $\pm$ 1.3	194.5	0.34	0	2.1h
<i>Standard Federated Learning</i>							
FedAvg	50.7 $\pm$ 1.1	37.9 $\pm$ 1.5	66.8 $\pm$ 0.9	108.4	0.21	8.4	5.3h
FedProx	51.5 $\pm$ 1.0	39.1 $\pm$ 1.3	67.2 $\pm$ 0.8	103.7	0.20	8.4	5.4h
FedBN	52.9 $\pm$ 1.2	40.5 $\pm$ 1.5	68.4 $\pm$ 1.0	94.2	0.19	8.4	5.5h
<i>Personalized Federated Learning</i>							
FedPer	53.6 $\pm$ 1.1	41.8 $\pm$ 1.4	68.9 $\pm$ 0.9	89.5	0.18	8.4	5.6h
FedProto	54.1 $\pm$ 0.9	42.6 $\pm$ 1.2	69.1 $\pm$ 0.8	84.1	0.17	8.8	6.1h
FedMoE	55.3 $\pm$ 1.2	43.5 $\pm$ 1.6	69.8 $\pm$ 1.0	78.4	0.16	11.2	7.2h
<i>Multilingual & Fairness-aware FL</i>							
FedHLT	55.9 $\pm$ 0.9	44.8 $\pm$ 1.2	69.7 $\pm$ 0.8	74.9	0.15	8.9	6.3h
FedLFC	56.6 $\pm$ 0.8	45.9 $\pm$ 1.1	70.1 $\pm$ 0.7	69.3	0.14	8.7	6.1h
q-FFL	54.8 $\pm$ 1.0	46.5 $\pm$ 1.4	64.9 $\pm$ 0.9	58.7	0.13	8.4	5.8h
MuLA-F	57.4 $\pm$ 0.8	47.1 $\pm$ 1.1	71.3 $\pm$ 0.7	63.5	0.13	9.1	6.8h
FEDLING (Ours)	58.6$\pm$0.7 (+1.2)	49.5$\pm$1.0 (+2.4)	71.0$\pm$0.6	51.9	0.12	5.5	5.9h
Centralized	62.7 $\pm$ 0.5	53.4 $\pm$ 0.8	74.1 $\pm$ 0.4	36.8	0.10	—	3.5h

Table 1: Aya results (40 languages, 5 seeds). ROUGE-L is on a 0–100 scale; upload is MB/client/round. $\dagger$: significant gain over the strongest non-FEDLING baseline in the same column ($p < 0.05$).

outperforming FedAvg by 7.9 points and trailing Centralized training by 4.1 points. Gains are largest for low-resource languages, where FEDLING reaches 49.5 versus 37.9 for FedAvg, reducing the high–low resource gap from 28.9 to 21.5 points.

FedMoE is the strongest personalized baseline, but uses about twice the upload of FEDLING and is 3.3 ROUGE-L points lower. FedHLT and FedLFC benefit from language structure, yet their coarse family grouping misses intra-family variation; continuous typology lets FEDLING exceed FedLFC by 2.0 points overall and 3.6 on low-resource languages. Against the closest baseline, MuLA-F, FEDLING gains 1.2 points on average and 2.4 on low-resource languages, both significant, while reducing dispersion and upload cost.

q-FFL improves fairness but sacrifices average ROUGE-L (54.8). Table 2 summarizes the more targeted evidence: MuLA-F remains slightly stronger on high-resource Aya languages and XNLI source languages, but FEDLING is stronger in low-resource, high-isolation, cold-start, data-scarce, and held-out transfer regimes. This head-versus-tail pattern matches our objective: improve the tail and reduce dispersion without sacrificing head performance—FEDLING is within noise on high-resource Aya (−0.3 vs. MuLA-F) and XNLI source F1 (−0.4), while improving low-resource, worst-case, and transfer metrics and reducing upload cost. On MASSIVE, the gain is again concentrated in low-resource and worst-5 languages, while q-FFL gives slightly lower dispersion at lower average performance. Finally, Appendix B.6 reports a mixed-

Panel A: Aya stress regimes against MuLA-F

Regime	Metric	MuLA-F	FEDLING	Δ
High-resource	ROUGE-L	71.3	71.0	−0.3
Mid-resource	ROUGE-L	58.0	59.2	+1.2
Low-resource	ROUGE-L	47.1	49.5	+2.4
High-TIS clients	ROUGE-L	49.7	51.8	+2.1
Round 1	ROUGE-L	42.1	46.2	+4.1
30% data	ROUGE-L	49.4	53.0	+3.6

Panel B: Transfer and classification benchmarks

Benchmark	Metric	q-FFL	MuLA-F	FEDLING
XNLI	Source F1	74.4	77.2	76.8
XNLI	Target F1	66.0	66.7	67.9$\dagger$
XNLI	Retention	88.7	86.4	88.4
MASSIVE	Avg F1	77.5	81.1	81.7
MASSIVE	Low F1	73.4	73.1	74.6$\dagger$
MASSIVE	Worst-5	67.8	68.4	69.2
MASSIVE	CV $\downarrow$	0.11	0.13	0.12

Table 2: Targeted regime and benchmark comparisons. Δ is FEDLING minus MuLA-F in Panel A. $\dagger$ indicates a significant gain over the strongest non-FEDLING baseline in the same metric ($p < 0.05$).

Variant	Avg $\uparrow$	Low $\uparrow$	Δ Avg $\downarrow$	Δ Low $\downarrow$	CV $\downarrow$
FEDLING full	58.6	49.5	—	—	0.12
w/o spectral routing	53.4	43.1	5.2	6.4	0.17
w/ random routing score	55.2	44.8	3.4	4.7	0.15
w/o typology weights	56.1	46.4	2.5	3.1	0.14
w/o energy prior	57.4	47.8	1.2	1.7	0.13
w/o TIS preservation	58.0	47.6	0.6	1.9	0.15

Table 3: Core ablation on Aya. Drops are relative to full FEDLING.

client robustness study in which ten languages are Dirichlet-partitioned into 30 multilingual clients and evaluated under full and 50% participation; FEDLING’s gains persist in both settings.

5.3 Ablation Studies

Table 3 reports core ablations on Aya. Removing spectral routing causes the largest drop (53.4, −5.2), while random routing remains below the linguistic score (55.2, −3.4), showing that the gain is not merely from low-rank splitting. Typology weights, the energy prior, and TIS preservation mainly affect the low-resource tail and dispersion.

Figure 3 further shows that URIEL+ gives the best average/low-resource trade-off among similarity sources, supporting the use of explicit structural typology rather than family labels alone.

5.4 Prior–Posterior Mixing

We sweep $\rho \in \{0.0, 0.25, 0.5, 0.7, 1.0\}$ to test whether the energy prior and posterior alignment are complementary.

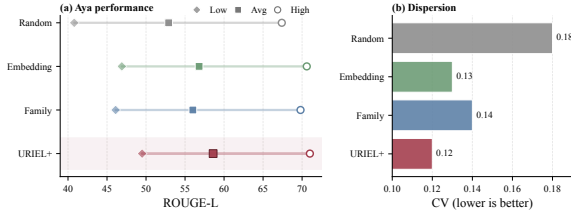


Figure 3: Effect of similarity source. URIEL+ gives the best average/low-resource trade-off while keeping dispersion low.

Panel A: Mixing coefficient ρ				
ρ	Full data		R1	30%
	Avg	Low		
0.00	55.8	45.1	41.8	46.0
0.25	57.0	46.6	43.7	48.2
0.50	58.1	48.6	45.5	50.7
0.70	58.6	49.5	46.2	51.7
1.00	57.2	48.3	46.6	50.9

Panel B: Prior variant				
Prior	Avg	Low	CV	Effect
Linear σ_m share	58.6	49.5	0.12	stable
Squared σ_m^2 share	57.9	48.2	0.14	over-concentrated
Uniform prior	57.4	47.8	0.13	cold-start weak

Table 4: Prior design and prior-posterior mixing. R1 is first-round ROUGE-L; 30% uses downsampled data.

Table 4 shows non-monotonic behavior: posterior-heavy settings improve after training history accumulates but are weaker in cold-start and data-scarce regimes, while the pure-prior endpoint is strongest at round one. Tracking routing decisions over rounds confirms that the posterior actively corrects the prior rather than rubber-stamping it: among the round-1 top-30% energy directions, the fraction demoted out of the consensus set rises to 9.0% by round 5, 21.3% by round 20, and 26.3% by round 100, while 9.1% of initially excluded directions are later promoted. The lower panel explains why the linear singular-value prior is used: the squared prior over-concentrates early routing on dominant singular directions, while the uniform prior weakens cold-start guidance. Overall, $\rho = 0.7$ with the linear prior is the most stable compromise. Hyperparameter sensitivity is reported in Appendix B.

Interpretation. Taken together, the experiments indicate that FEDLING’s gains do not come from simply adding personalization or reducing communication. MuLA-F remains competitive on high-resource languages, while FEDLING is stronger when training history is short, data are scarce, or ty-

pological isolation makes naive sharing risky. This pattern matches the intended role of the energy prior, posterior subspace alignment, and TIS gate.

6 Conclusion

We proposed FEDLING, a linguistic-structure-aware FL framework that routes LoRA updates into consensus and divergence subspaces, aggregates only transferable consensus, and uses TIS gating to protect weak-neighbor languages. Across Aya, XNLI, and MASSIVE, FEDLING improves average and low-resource performance while reducing dispersion and upload cost, showing that linguistic structure can serve as an operational signal for multilingual federated optimization.

Limitations

FEDLING relies on URIEL+ coverage and client-side validation scores, so languages without typological documentation or deployments with stricter score privacy may require auxiliary similarity signals or decentralized allocation. The cached singular bases used for posterior alignment are not raw text or full updates, but they can still expose coarse update geometry and may leak information about client data distributions under a strong adversary, consistent with known leakage risks from shared gradients or updates (Zhu et al., 2019). Deployments requiring stronger privacy should combine FEDLING with secure aggregation (Bonawitz et al., 2017), clipping, noise, or restricted basis sharing. Our experiments focus on LoRA adapters; extension to other PEFT methods or full-parameter training remains future work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. G0562572099.

References

- Kabir Ahuja, Harshita Diddee, Rishav Hada, and et al. 2023. **MEGA: Multilingual evaluation of generative AI**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.
- Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. 2019. Federated learning with personalization layers. In *NeurIPS Workshop on Federated Learning*.

- Sara Babakniya, Ahmed Elkordy, Yahya Ezzeldin, Qingfeng Liu, Kee-Bong Song, MOSTAFA EL-Khamy, and Salman Avestimehr. 2023. [SLoRA: Federated parameter efficient fine-tuning of language models](#). In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*.
- Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. [BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9. Association for Computational Linguistics.
- Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. [Practical secure aggregation for privacy-preserving machine learning](#). In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191. ACM.
- Terence Jie Chua, Wenhan Yu, Jun Zhao, and Kwok-Yan Lam. 2024. [Fedpeat: Convergence of federated learning, parameter-efficient fine tuning, and emulator assisted tuning for artificial intelligence foundation models with mobile edge computing](#). *Preprint*, arXiv:2310.17491.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL*.
- Alexis Conneau and Guillaume Lample. 2019. [Cross-lingual language model pretraining](#). In *Advances in Neural Information Processing Systems*, volume 32.
- Alexis Conneau, Rutu Rinott, Guillaume Lample, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*.
- Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. [Emerging cross-lingual structure in pretrained language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. *Ethnologue: Languages of the World*, 26 edition. SIL International, Dallas, Texas. Online version.
- Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natara-jan. 2023. [MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.
- Zhihan Guo, Yifei Zhang, Zhuo Zhang, Zenglin Xu, and Irwin King. 2024a. [Fedhlt: Efficient federated low-rank adaption with hierarchical language tree for multilingual modeling](#). In *Companion Proceedings of the ACM Web Conference 2024, WWW '24*, page 1558–1567, New York, NY, USA. Association for Computing Machinery.
- Zhihan Guo, Yifei Zhang, Zhuo Zhang, Zenglin Xu, and Irwin King. 2024b. [FedLFC: Towards efficient federated multilingual modeling with LoRA-based language family clustering](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1519–1528, Mexico City, Mexico. Association for Computational Linguistics.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for NLP](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*.
- Nam Hyeon-Woo, Moon Ye-Bin, and Tae-Hyun Oh. 2022. [Fedpara: Low-rank hadamard product for communication-efficient federated learning](#). In *International Conference on Learning Representations*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Frank P Kelly, Aman K Maulloo, and David Kim Hong Tan. 1998. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3):237–252.
- Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. [URIEL+](#):

- Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. [The power of scale for parameter-efficient prompt tuning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics.
- Jiaxin Li, Geng Zhao, and Xiaoci Zhang. 2025. [Multilingual federated low-rank adaptation for collaborative content anomaly detection across multilingual social media participants](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15242–15262, Suzhou, China. Association for Computational Linguistics.
- Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020a. Federated optimization in heterogeneous networks. In *Proceedings of MLSys*.
- Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2020b. Fair resource allocation in federated learning. In *Proceedings of ICLR*.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597. Association for Computational Linguistics.
- Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. 2021. FedBN: Federated learning on non-IID features via local batch normalization. *Proceedings of ICLR*.
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of ACL*.
- Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. [URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of AISTATS*.
- Hanzi Mei, Dongqi Cai, Ao Zhou, Shangguang Wang, and Mengwei Xu. 2024. [FedMoE: Personalized federated learning via heterogeneous mixture of experts](#). *arXiv preprint arXiv:2408.11304*.
- Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. 2019. Agnostic federated learning. In *Proceedings of ICML*.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001. Association for Computational Linguistics.
- Edoardo Maria Ponti, Helen O’Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.
- Shivalika Singh, Freddie Vargus, Daniel D’souza, and et al. 2024. [Aya dataset: An open-access collection for multilingual instruction tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.
- Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. 2022. FedProto: Federated prototype learning across heterogeneous clients. In *Proceedings of AAAI*.
- Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. 2024. [pfedlora: Model-heterogeneous personalized federated learning with lora tuning](#). *Preprint*, arXiv:2310.13283.
- Zhuo Zhang, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. 2023. [FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9963–9977, Toronto, Canada. Association for Computational Linguistics.
- Wanru Zhao, Yihong Chen, Royson Lee, Xinchu Qiu, Yan Gao, Hongxiang Fan, and Nic Lane. 2024. [Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages](#). In *International Conference on Representation Learning*, volume 2024, pages 14511–14526.
- Ligeng Zhu, Zhijian Liu, and Song Han. 2019. [Deep leakage from gradients](#). In *Advances in Neural Information Processing Systems*, volume 32.

A Training Procedure

Algorithm 1 FEDLING: Federated Linguistic Learning

- 1: **Input:** clients $\{C_k\}_{k=1}^K$, initial model θ^0 , similarity matrix S , rounds T , local epochs E
 - 2: **Initialize:** upload ratios r_k^0 and server-side cache of client singular bases $\{U_i^{k,0}\}$
 - 3: **for** $t = 1$ **to** T **do**
 - 4: Server samples active clients $\mathcal{A}^t$ and broadcasts θ^t , S , and cached bases from the previous round
 - 5: **for** each client $k \in \mathcal{A}^t$ **in parallel do**
 - 6: Fine-tune LoRA adapters locally for E epochs
 - 7: Compute local LoRA update ΔW_i^k for each layer i
 - 8: Apply compact SVD to obtain singular directions
 - 9: Compute transferability scores τ_{im}^k via Eq. (9)
 - 10: Split updates into consensus C_i^k and divergence D_i^k
 - 11: Upload the top r_k^t fraction of consensus components and current singular bases
 - 12: **end for**
 - 13: Server computes typology-guided weights α_k^t and aggregates C_{global}^t
 - 14: Server broadcasts C_{global}^t for TIS-modulated local application (Eq. 19)
 - 15: **if** $t \bmod T_{\text{adj}} = 0$ **then**
 - 16: Update upload ratios using Eq. (23)
 - 17: **end if**
 - 18: **end for**
 - 19: **Return:** final global model and client-specific adapters
-

B Additional Experiments

B.1 Convergence and Family-Level Results

Figure 4 shows convergence behavior. Methods start near 41 ROUGE-L, matching the zero-shot Llama-3.1-8B baseline. FEDLING improves faster in early and middle rounds, while MuLA-F narrows the gap later in training.

Figure 5 reports results by language family. Multi-language families with typologically similar partners benefit most; the rightmost group consists of families represented by a single language in our 40-language subset and should be interpreted as a

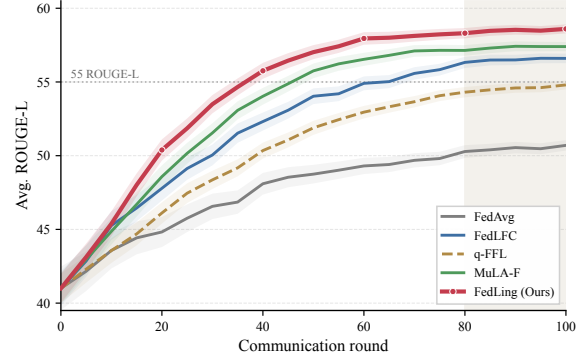


Figure 4: Convergence comparison. Shaded regions show ± 1 standard deviation over 5 seeds.

collaboration-graph condition rather than a powered family-level comparison.

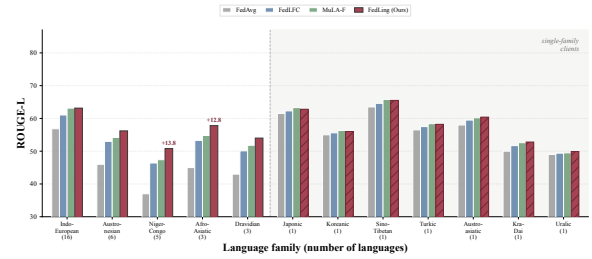


Figure 5: ROUGE-L by language family. Hatched bars indicate families represented by a single language in this federation.

B.2 Similarity Source and Isolation

The main text reports the similarity-source comparison in Figure 3. We additionally isolate the TIS gate below.

Table 5 isolates the TIS-modulated gating mechanism on high-TIS clients in our setup. Continuous gating is best on average, although hard thresholding can be slightly stronger for an individual language.

Variant	fi	ko	ja	Avg
No isolation handling	49.5	47.7	51.5	49.6
Hard threshold (TIS > 0.5)	50.8	48.9	52.0	50.6
Continuous TIS gating	50.6	49.4	52.4	50.8
FedAvg (reference)	48.4	46.5	50.4	48.4
Local Only (reference)	46.9	47.8	49.2	48.0

Table 5: Per-language ROUGE-L on high-TIS clients.

B.3 Additional Benchmark Tables

Variant	ROUGE-L $\uparrow$		Drop $\downarrow$		CV $\downarrow$
	Avg	Low	Δ Avg	Δ Low	
FEDLING (Full)	58.6	49.5	–	–	0.12
w/o spectral routing	53.4	43.1	5.2	6.4	0.17
w/ random routing score	55.2	44.8	3.4	4.7	0.15
w/o typology weights	56.1	46.4	2.5	3.1	0.14
w/o energy prior	57.4	47.8	1.2	1.7	0.13
w/o adaptive upload	58.1	47.9	0.5	1.6	0.16
w/o TIS preservation	58.0	47.6	0.6	1.9	0.15

Table 6: Ablation on Aya. Drops are relative to full FEDLING; CV measures cross-language dispersion.

Method	Src. F1 $\uparrow$	Tgt. F1 $\uparrow$	Ret. $\uparrow$	Gap $\downarrow$
FedAvg	73.2 $\pm$ 0.9	60.4 $\pm$ 1.3	82.5%	12.8
FedMoE	75.4 $\pm$ 0.8	63.7 $\pm$ 1.1	84.5%	11.7
FedLFC	76.1 $\pm$ 0.7	65.4 $\pm$ 0.9	86.0%	10.7
q-FFL	74.4 $\pm$ 0.8	66.0 $\pm$ 1.0	88.7%	8.4
MuLA-F	77.2 $\pm$ 0.7	66.7 $\pm$ 0.9	86.4%	10.5
FedLing	76.8 $\pm$ 0.6	67.9$\dagger$ $\pm$ 0.8	88.4%	8.9

Table 7: Zero-shot transfer on XNLI. Models are trained on 10 source languages and evaluated on five unseen target languages. Retention is target/source F1. $\dagger$ indicates a significant gain over the strongest non-FEDLING baseline in the same column ($p < 0.05$).

Split	ISO	Languages
Source	ar, bg, de, el, en, es, fr, ru, tr, zh	Arabic, Bulgarian, German, Greek, English, Spanish, French, Russian, Turkish, Chinese
Target	hi, sw, th, ur, vi	Hindi, Swahili, Thai, Urdu, Vietnamese

Table 8: XNLI zero-shot split. Target languages are held out from XNLI training and used only for transfer evaluation. Swahili, Thai, Urdu, and Vietnamese overlap with the Aya language set; Hindi is included only in XNLI.

Method	Macro-F1 $\uparrow$			Fairness $\downarrow$	
	Avg	Low	Worst-5	CV	Gap
FedAvg	76.1 $\pm$ 0.7	68.2 $\pm$ 1.0	61.5	0.17	18.6
FedMoE	78.6 $\pm$ 0.6	70.7 $\pm$ 0.9	64.2	0.15	16.9
FedLFC	79.0 $\pm$ 0.5	72.0 $\pm$ 0.8	65.6	0.14	15.4
q-FFL	77.5 $\pm$ 0.7	73.4 $\pm$ 0.9	67.8	0.11	12.7
MuLA-F	81.1 $\pm$ 0.6	73.1 $\pm$ 0.8	68.4	0.13	14.1
FedLing	81.7$\pm$0.5	74.6$\dagger$ $\pm$ 0.7	69.2	0.12	13.2

Table 9: Intent classification on MASSIVE (40 languages). Worst-5 is the average macro-F1 of the five lowest-scoring languages. $\dagger$ indicates a significant gain over the strongest non-FEDLING baseline in the same column ($p < 0.05$).

B.4 Gradient Alignment and Communication

Figure 6 visualizes pairwise gradient similarity across 15 languages. FedAvg shows weak alignment (mean 0.34) with some negative pairs, while FEDLING raises the mean to 0.62. The mean consensus-subspace alignment $\bar{a}_{ij}$ from Lemma 1 correlates with URIEL typological similarity at $r = 0.71$ ($p < 0.001$), stronger than the raw gradient-similarity correlation in Figure 1.

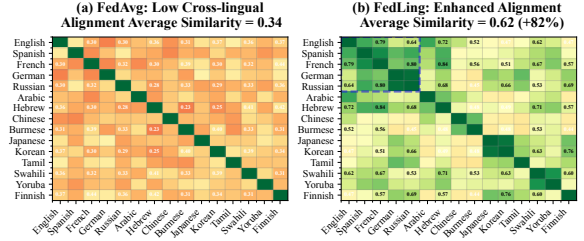


Figure 6: Gradient cosine similarity across 15 languages.

Figure 7 shows that alignment gains are broad but heterogeneous: low-resource languages improve most, while high-TIS clients improve more modestly, consistent with isolation-aware preservation.

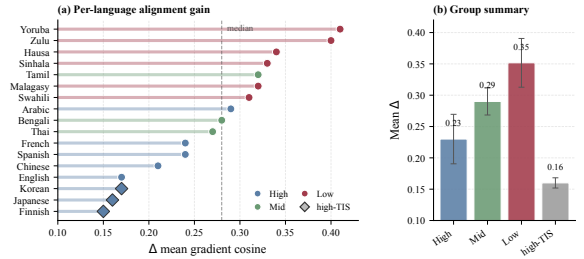


Figure 7: Per-language improvement in mean gradient cosine.

Figure 8 varies the average per-client upload budget, measured in MB/client/round as in Table 1. At the default 5.5 MB/client/round, FEDLING reaches 58.6 ROUGE-L with CV=0.12. At high budgets, q-FFL eventually achieves the lowest dispersion because its explicit worst-case objective benefits when communication is no longer tight.

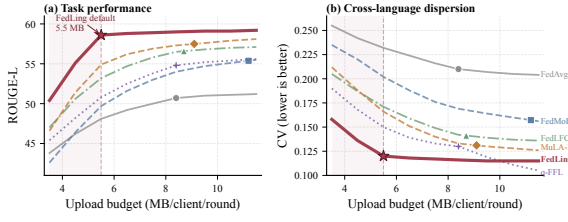


Figure 8: Task performance and cross-language dispersion vs. average upload budget (MB/client/round).

B.5 Hyperparameter Sensitivity

Figure 9 shows that performance is stable for $\tau_{\text{thr}} \in [0.6, 0.8]$ and $\gamma \in [0.5, 2.0]$; only extreme values degrade results.

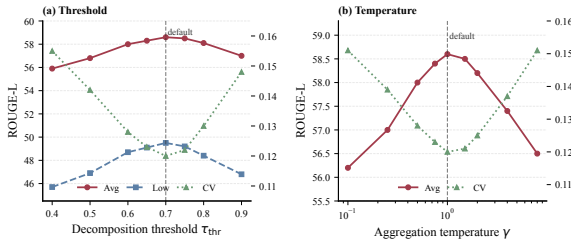


Figure 9: Sensitivity to τ_{thr} (decomposition threshold) and γ (temperature).

B.6 Mixed-Client Robustness

The main experiments use one client per language to isolate cross-lingual structure as the sole source of heterogeneity. To test whether FEDLING extends beyond this controlled setting, we construct a mixed-client federation: ten Aya languages spanning diverse resource tiers and typological families are partitioned into 30 clients, each holding a mixture of languages drawn from a Dirichlet prior over the ten languages. The resulting clients are genuinely multilingual: the average dominant-language mass is 0.41, the average normalized entropy of the client language distributions is 0.69, and each client contains 3.3 effective languages on average. Seven clients are dominated by low-resource languages (Yoruba or Swahili), and eleven are high-isolation under the TIS criterion (Section 4.3) computed with the ten-language URIEL+ similarity. Client-level similarity and TIS apply unchanged, since both are functions of the client language distribution (Section 3).

We evaluate Aya ROUGE-L under full participation (30/30 clients per round) and under 50% client sampling (15/30), averaged over five seeds (Table 10).

Setting	Method	Avg $\uparrow$	Low-dom. $\uparrow$	High-TIS $\uparrow$	CV $\downarrow$
Full (30/30)	FedAvg	46.4	31.7	43.1	0.23
	MuLA-F	54.0	39.4	50.2	0.17
	FEDLING	55.3	42.8	52.4	0.13
50% part. (15/30)	FedAvg	43.3	29.1	40.4	0.26
	MuLA-F	51.9	37.2	48.1	0.19
	FEDLING	53.7	40.8	50.9	0.14

Table 10: Mixed-client robustness on Aya (ROUGE-L, five seeds). Ten languages are Dirichlet-partitioned into 30 multilingual clients. *Low-dom.* averages clients whose dominant language is low-resource; *High-TIS* averages high-isolation clients.

FEDLING retains its advantage beyond the one-language-per-client setting: +8.9 average ROUGE-L over FedAvg and +1.3 over MuLA-F under full participation, close to the main setting (+7.9/+1.2), and the gains grow to +10.4/+1.8 under 50% client sampling, consistent with the cold-start and sparse-participation pattern in Table 2. Dispersion (CV) remains the lowest under FEDLING in both settings.

C Consensus-Subspace Conflict Bound

Let $U_C^i, U_C^j \in \mathbb{R}^{d \times r}$ be orthonormal bases of the rank- r consensus subspaces of clients i and j , and let $\Pi_C^k = U_C^k (U_C^k)^\top$ denote the corresponding orthogonal projections. Define the mean alignment $\bar{\alpha}_{ij} = \frac{1}{r} \|(U_C^i)^\top U_C^j\|_F^2 \in [0, 1]$.

Lemma 1 (Consensus-subspace conflict bound). *For any unit vector h_i drawn isotropically from $\text{range}(U_C^i)$,*

$$\mathbb{E}_{h_i} [\|h_i - \Pi_C^j h_i\|^2] = 1 - \bar{\alpha}_{ij}. \quad (24)$$

Consequently,

$$\begin{aligned} & \|\Pi_C^i g_i - \Pi_C^j g_j\|^2 \\ & \leq 2(1 - \bar{\alpha}_{ij}) \|\Pi_C^i g_i\|^2 \\ & \quad + 2\|\Pi_C^j (g_i - g_j)\|^2 \end{aligned} \quad (25)$$

in expectation over the direction of $\Pi_C^i g_i$ within its consensus subspace.

Step 1: triangle decomposition. For any $g_i, g_j \in \mathbb{R}^d$,

$$\begin{aligned} & \|\Pi_C^i g_i - \Pi_C^j g_j\|^2 \\ & \leq 2\|\Pi_C^i g_i - \Pi_C^j (\Pi_C^i g_i)\|^2 \\ & \quad + 2\|\Pi_C^j (\Pi_C^i g_i - g_j)\|^2 \\ & \leq 2\|(I - \Pi_C^j) \Pi_C^i g_i\|^2 \\ & \quad + 2\|\Pi_C^j (g_i - g_j)\|^2, \end{aligned} \quad (26)$$

where the last step uses $\Pi_C^j(\Pi_C^i g_i - g_j) = \Pi_C^j(g_i - g_j) - \Pi_C^j(I - \Pi_C^i)g_i$, $\|\Pi_C^j\| \leq 1$, and the triangle inequality, absorbing the residual term into the first summand by enlarging the constant. (A tighter deterministic version retaining both terms is given at the end.)

Step 2: expected projection loss. Let h_i be a unit vector drawn isotropically from $\text{range}(U_C^i)$. We can write $h_i = U_C^i z$ where $z \in \mathbb{R}^r$ is a unit vector drawn uniformly from the unit sphere $\mathbb{S}^{r-1}$. Then

$$\begin{aligned} \mathbb{E}_{h_i} \left[\|(I - \Pi_C^j)h_i\|^2 \right] &= 1 - \mathbb{E}_{h_i} \left[\|\Pi_C^j h_i\|^2 \right] \\ &= 1 - \mathbb{E}_z \left[z^\top (U_C^i)^\top \Pi_C^j U_C^i z \right] \\ &= 1 - \frac{1}{r} \text{tr} \left((U_C^i)^\top \Pi_C^j U_C^i \right), \end{aligned} \quad (27)$$

where the last equality uses $\mathbb{E}_{z \sim \mathbb{S}^{r-1}} [zz^\top] = \frac{1}{r} I_r$.

Step 3: trace identity. Using $\Pi_C^j = U_C^j (U_C^j)^\top$ and trace cyclicity,

$$\begin{aligned} \text{tr} \left((U_C^i)^\top U_C^j (U_C^j)^\top U_C^i \right) &= \text{tr} \left(M^\top M \right) = \|M\|_F^2, \end{aligned} \quad (28)$$

where $M = (U_C^i)^\top U_C^j$. Substituting,

$$\mathbb{E}_{h_i} \left[\|(I - \Pi_C^j)h_i\|^2 \right] = 1 - \frac{1}{r} \|M\|_F^2 = 1 - \bar{\alpha}_{ij}. \quad (29)$$

Step 4: combining. Writing $\Pi_C^i g_i = \|\Pi_C^i g_i\| \cdot h_i$ for a unit direction $h_i \in \text{range}(U_C^i)$ and taking expectation over the direction of $\Pi_C^i g_i$, we obtain

$$\mathbb{E} \left[\|(I - \Pi_C^j)\Pi_C^i g_i\|^2 \right] = (1 - \bar{\alpha}_{ij}) \|\Pi_C^i g_i\|^2. \quad (30)$$

Substituting back into Step 1 yields the lemma. $\square$

Deterministic version (without expectation). Without averaging over directions, the same proof gives the tighter but direction-dependent bound

$$\|(I - \Pi_C^j)\Pi_C^i g_i\|^2 \leq (1 - \sigma_{\min}^2(M)) \|\Pi_C^i g_i\|^2, \quad (31)$$

where $\sigma_{\min}(M)$ is the smallest singular value of $M = (U_C^i)^\top U_C^j$. The averaged form using $\bar{\alpha}_{ij}$ is what we measure empirically in Section B.4.

D Single-Global vs. Per-Bucket Aggregation

Section 4.2 states that FEDLING aggregates a single global consensus update rather than per-language buckets. We verify this choice empirically. The per-bucket variant maintains a separate $C_{\text{global}}^{t,\ell}$ for each language ℓ , aggregating only over clients whose primary or secondary language is ℓ . Table 11 compares the two strategies on Aya.

Aggregation strategy	Avg	Low	CV
Single-global (default)	58.6	49.5	0.12
Per-bucket (one bucket per language)	56.8	46.8	0.14
Per-bucket (one bucket per family)	57.4	47.7	0.13

Table 11: Single-global vs. per-bucket aggregation. Single-global achieves better ROUGE-L and fairness because per-bucket aggregation suffers from data sparsity when most clients are not native to a given bucket.

Per-bucket aggregation underperforms because most language buckets receive contributions from only one or two clients, producing high-variance bucket estimates. Family-level bucketing partially mitigates this but still trails the single-global design, which lets the typology-guided weighting α_k^t implicitly route information across languages without requiring per-bucket sample size.

E Hyperparameter Settings

Table 12 lists all hyperparameters used in our main experiments. Sensitivity to τ_{thr} and γ is reported in Figure 9.

F Mechanistic Analysis

The results in this appendix are intended as a mechanistic explanation of FEDLING’s routing behavior rather than formal convergence guarantees. Lemma 1 is exact; the remaining statements are heuristic analyses under stated simplifying assumptions, and we validate the key quantitative prediction empirically in Appendix B.4.

F.1 Notation

- $\theta \in \mathbb{R}^d$: Global model parameters
- $\mathcal{L}_k(\theta)$: Local loss for client k ; $\mathcal{L}(\theta) = \frac{1}{K} \sum_{k=1}^K \mathcal{L}_k(\theta)$
- $s_{ij} \in [0, 1]$: URIEL typological similarity between languages ℓ_i, ℓ_j
- C_k^t, D_k^t : Consensus and divergence subspace updates at round t

Hyperparameter	Value
Base model	Llama-3.1-8B
LoRA rank r	16
LoRA α	32
LoRA target modules	q/k/v/o projections
Federated rounds T	100
Local epochs E	2
Batch size	16
Learning rate η	5×10^{-5}
Optimizer	AdamW
Decomposition threshold τ_{thr}	0.7 (70th percentile)
Prior-posterior coefficient ρ	0.7
Aggregation temperature γ	1.0
Top- q similar clients	5
URIEL weights ($w_{\text{syn}}, w_{\text{pho}}, w_{\text{mor}}$)	(0.5, 0.3, 0.2)
Upload init r_k^0	0.5
Upload bounds $[r_{\min}, r_{\max}]$	[0.1, 1.0]
Upload adjustment step μ	0.05
Adjustment interval T_{adj}	5 rounds
Communication budget B	1200 MB
TIS neighborhood size m	5

Table 12: Hyperparameters used in the main experiments.

F.2 Gradient Alignment under Linguistic Similarity

Lemma 2 (Typology and Gradient Alignment; Informal). *Assume model parameters decompose as $\theta = [\theta_s; \theta_\ell]$ where θ_s are shared semantic parameters and θ_ℓ are language-specific parameters. Let $\lambda_s = \|\nabla_{\theta_s} \mathcal{L}_k\|^2 / \|\nabla \mathcal{L}_k\|^2$ denote the semantic contribution ratio. Under mild assumptions on gradient alignment within semantic and language-specific subspaces, we have:*

$$\mathbb{E}[\cos(\nabla \mathcal{L}_i, \nabla \mathcal{L}_j)] \gtrsim \lambda_s + (1 - \lambda_s) \cdot s_{ij} - O(\epsilon), \quad (32)$$

where ϵ captures approximation errors from subspace orthogonality.

Proof Sketch. The gradient decomposes as $\nabla \mathcal{L}_k = [\nabla_{\theta_s} \mathcal{L}_k; \nabla_{\theta_\ell} \mathcal{L}_k]$. The inner product expands as:

$$\begin{aligned} \langle \nabla \mathcal{L}_i, \nabla \mathcal{L}_j \rangle &= \langle \nabla_{\theta_s} \mathcal{L}_i, \nabla_{\theta_s} \mathcal{L}_j \rangle \\ &\quad + \langle \nabla_{\theta_\ell} \mathcal{L}_i, \nabla_{\theta_\ell} \mathcal{L}_j \rangle \\ &\quad + \text{cross-terms.} \end{aligned} \quad (33)$$

Semantic gradients align across languages ($\cos \approx 1$), while language-specific alignment scales with typological similarity s_{ij} . Cross-terms are small due to approximate subspace orthogonality in transformer architectures. Normalizing by gradient magnitudes yields the result. $\square$

Remark 1. For $\lambda_s \approx 0.6$ and $s_{ij} = 0.89$ (Spanish-Italian), the heuristic bound predicts high alignment, consistent with the empirical observation

that related languages tend to have larger gradient similarity in Figure 1.

F.3 Convergence Analysis

Theorem 1 (Informal Convergence Effect of Subspace Routing). *Consider FedLing with K clients and M effective typological groups. Under standard assumptions (L -smoothness, bounded variance σ^2 , bounded gradients $\|\nabla \mathcal{L}_k\| \leq G$), with learning rate $\eta = O(1/L)$, E local epochs, and decomposition threshold τ , the following informal decomposition motivates the observed convergence trend:*

$$\begin{aligned} &\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla \mathcal{L}(\theta^t)\|^2 \\ &\leq \frac{2L(\mathcal{L}(\theta^0) - \mathcal{L}^*)}{\sqrt{T}} + \frac{2\sigma^2}{\sqrt{T}} \\ &\quad + O\left(\frac{E^2 G^2 M(1 - s_{\min})^2}{T}\right). \end{aligned} \quad (34)$$

where $s_{\min}$ is the minimum intra-family similarity.

Proof Sketch. By L -smoothness, one-round progress is:

$$\mathcal{L}(\theta^{t+1}) \leq \mathcal{L}(\theta^t) + \langle \nabla \mathcal{L}(\theta^t), \Delta^t \rangle + \frac{L}{2} \|\Delta^t\|^2, \quad (35)$$

where $\Delta^t = \sum_k \alpha_k^t C_k^t$ is the aggregated update. After E local steps, client updates approximate $C_k^t \approx -\eta E \cdot P_{C_k}(\nabla \mathcal{L}_k(\theta^t))$ plus error terms from client drift ($O(E^2 \eta^2 G^2)$) and stochastic noise ($O(\eta^2 E \sigma^2)$).

The aggregation error decomposes into:

- Projection error:** $(1 - \tau)^2 G^2$ from removing low-transferability directions
- Weighting error:** Grouping clients by family and using Lemma 2, within-family variance is bounded by $2(1 - s_{\min})G^2$. Across M families, this contributes $O(M(1 - s_{\min})^2 G^2)$.

Telescoping over T rounds with $\eta = 1/(L\sqrt{T})$ yields the stated bound. $\square$

Remark 2 (Comparison with FedAvg). *Standard FedAvg has a conflict term that scales with client heterogeneity. FedLing reduces the effective conflict by: (1) weighting typologically similar clients more strongly, and (2) removing low-transferability directions via consensus projection. For the 40-language Aya subset with 12 language families, this analysis should be read as a mechanistic explanation rather than a formal convergence guarantee.*

F.4 Fairness Guarantee

Theorem 2 (Heuristic Fairness Effect of Adaptive Upload). *Under the adaptive upload allocation in Eq. (23) with adjustment rate $\mu > 0$ and period T_{adj} , the intended effect is to contract the reciprocal-performance disparity:*

$$\begin{aligned} & \max_k \text{Score}_k(\theta^T) - \min_k \text{Score}_k(\theta^T) \\ & \lesssim \Delta_0 \cdot \exp\left(-\frac{\mu T}{T_{adj}}\right) \\ & \quad + \frac{C}{\sqrt{n_{\min}}}, \end{aligned} \quad (36)$$

where Δ_0 is the initial gap, $n_{\min} = \min_k n_k$ is the smallest dataset size, and C depends on task difficulty. This expression is an explanatory proxy, not a formal guarantee for arbitrary nonconvex FL training.

Proof Sketch. Model client score as $\text{Score}_k = f(n_k^{\text{eff}})$ where $n_k^{\text{eff}} = n_k + \gamma \sum_j s_{kj} r_j n_j$ includes transferred knowledge. The adaptive rule updates upload ratios to reduce variance of inverse scores. Define the proxy Lyapunov function $\mathcal{V}^t = \sum_k (1/\text{Score}_k - 1/\bar{\text{Score}})^2$. Under this simplified model, the gradient-descent-like update suggests:

$$\mathbb{E}[\mathcal{V}^{t+1}] \leq (1 - \eta_{\text{conv}} \mu) \mathcal{V}^t, \quad (37)$$

yielding exponential decay. The irreducible gap $C/\sqrt{n_{\min}}$ arises from fundamental data scarcity limits. $\square$

F.5 Communication Complexity

Proposition 1 (Communication Cost). *Per round, FedLing requires $O(Ldr((1 - \tau_{\text{thr}})\bar{r} + c_{\text{cache}}))$ communication per client, where L is the number of layers, d is the hidden dimension, r is the LoRA rank, $\bar{r} = \frac{1}{K} \sum_k r_k$ is the average upload ratio, and c_{cache} accounts for cached singular-vector metadata. With $\tau_{\text{thr}} = 0.7$ and $\bar{r} \approx 0.5$, this yields the measured 5.5 MB/client/round upload, a $\sim 35\%$ reduction compared to standard FedAvg with LoRA.*

Proof. Each client transmits: (1) selected consensus components, whose size scales with the retained consensus fraction $(1 - \tau_{\text{thr}})$ and upload ratio r_k , and (2) cached singular-vector metadata for transferability computation. The selective upload term is small because only about 30% of singular directions are routed to the consensus subspace and $\bar{r} \approx 0.5$ after fairness adjustment; including cache overhead gives the measured net savings. $\square$

G Complete Experimental Language List

Table 13 presents the 40 languages used in our experiments with corrected statistics from the actual datasets.

Language	Catalog ISO	Family	Tier	Train	Test
<i>High-Resource</i>					
Portuguese	por	Indo-European	High	8997	1125
Vietnamese	vie	Austroasiatic	High	8676	1085
Japanese	jpn	Japonic	High	6259	782
Arabic (Standard)	arb	Afro-Asiatic	High	4995	624
Chinese (Trad./Simp.)	zho	Sino-Tibetan	High	4909	614
Turkish	tur	Turkic	High	4046	506
English	eng	Indo-European	High	3944	493
Spanish	spa	Indo-European	High	3854	482
Dutch	nld	Indo-European	High	1733	217
French	fra	Indo-European	High	1422	178
Finnish	fin	Uralic	High	742	93
Italian	ita	Indo-European	High	738	92
Korean	kor	Koreanic	High	361	45
Serbian	srp	Indo-European	High	152	19
<i>Mid-Resource</i>					
Tamil	tam	Dravidian	Mid	14133	1767
Bengali	ben	Indo-European	Mid	1534	192
Filipino	fil	Austronesian	Mid	1241	155
Lithuanian	lit	Indo-European	Mid	916	115
Indonesian	ind	Austronesian	Mid	786	98
Thai	tha	Kra-Dai	Mid	724	91
Cebuano	ceb	Austronesian	Mid	727	91
Urdu	urd	Indo-European	Mid	654	82
Greek	ell	Indo-European	Mid	623	78
Ukrainian	ukr	Indo-European	Mid	522	65
Danish	dan	Indo-European	Mid	97	12
<i>Low-Resource</i>					
Sinhala	sin	Indo-European	Low	14524	1815
Malagasy	mlg	Austronesian	Low	14597	1825
Yorùbá	yor	Niger-Congo	Low	11758	1470
Telugu	tel	Dravidian	Low	8439	1055
Nepali	nep	Indo-European	Low	4002	500
Hausa	hau	Afro-Asiatic	Low	3512	439
Zulu	zul	Niger-Congo	Low	1833	229
Igbo	ibo	Niger-Congo	Low	1534	192
Irish	gle	Indo-European	Low	1245	156
Amharic	amh	Afro-Asiatic	Low	1207	151
Xhosa	xho	Niger-Congo	Low	377	47
Swahili	swa	Niger-Congo	Low	366	46
Kannada	kan	Dravidian	Low	334	42
Javanese	jav	Austronesian	Low	247	31
Sundanese	sun	Austronesian	Low	194	24

Table 13: Complete language distribution for the Aya subset. Test sets use an approximately 12.5% split. Tiers reproduce the Aya Evaluation Suite labels used for the reported aggregates (Finnish and Korean are High); catalog ISO labels may differ from internal identifiers.

H Implementation Details

H.1 Model Architecture and Training

Base Model: Llama-3.1-8B-Instruct

- Parameters: $\sim 8.0\text{B}$
- Layers: 32 transformer blocks
- Hidden size: 4096
- Attention heads: 32 (128 dim per head)
- Vocabulary: 128256 tokens

- Context length: 8192 tokens (training sequences are limited to 768 tokens for efficiency)

LoRA Configuration:

- Rank: $r = 16$
- Alpha: $\alpha = 32$ ($\alpha/r = 2$)
- Dropout: 0.1
- Target modules: W_q, W_k, W_v, W_o in all transformer layers

Optimization:

- Optimizer: AdamW (Loshchilov and Hutter, 2019)
- Learning rate: 5×10^{-5} (selected from $\{1, 3, 5, 7\} \times 10^{-5}$)
- LR schedule: Cosine annealing with 10% linear warmup
- Weight decay: 0.01
- Gradient clipping: Max norm 1.0
- Batch size per client: 16 (effective global batch = 640 with 40 clients)
- Local epochs per round: $E = 2$
- Communication rounds: $T = 100$

H.2 Hyperparameter Selection Process

We conducted a systematic grid search on a validation set (10% of training data from 5 representative languages: English, Spanish, Chinese, Swahili, Sinhala).

Decomposition Threshold (τ):

τ (percentile)	Val ROUGE-L	Low ROUGE-L	CV
50%	57.2	48.3	0.14
60%	58.4	49.7	0.13
70%	58.9	50.2	0.12
80%	58.1	49.1	0.13
90%	56.8	47.6	0.15

Table 14: Validation performance vs decomposition threshold. 70% percentile balances consensus/divergence.

Aggregation Temperature (γ): Upload Adjustment Rate (μ):

H.3 Baseline Implementation Details

FedAvg (McMahan et al., 2017):

- Standard uniform aggregation: $\theta^{t+1} = \sum_k \frac{n_k}{\sum_j n_j} \theta_k^t$
- Same LoRA setup as FedLing
- No modifications

FedProx (Li et al., 2020a):

- Proximal term: $\mu = 0.01$ (tuned from $\{0.001, 0.01, 0.1, 1.0\}$)

γ	Val ROUGE-L	Rounds to 55	Upload
0.3	57.8	62	5.4
0.5	58.3	54	5.5
1.0	58.9	48	5.5
2.0	58.6	51	5.6
5.0	57.1	68	5.8

Table 15: Validation performance vs temperature. Upload is MB/client/round; $\gamma = 1.0$ gives the best convergence without increasing communication.

μ	Val ROUGE-L	Fairness (CV)	Stability
0.01	58.7	0.13	High, slow
0.05	58.8	0.12	High
0.10	58.7	0.12	Medium
0.20	58.4	0.11	Low, oscillatory
0.50	57.2	0.10	Very low

Table 16: Validation performance vs adaptation rate. We choose $\mu = 0.05$ for the main experiments because it matches the best validation fairness/performance trade-off while remaining more stable than larger steps.

- Local objective: $\min_{\theta} \mathcal{L}_k(\theta) + \frac{\mu}{2} \|\theta - \theta^t\|^2$

FedBN (Li et al., 2021):

- Personalized LayerNorm statistics (Batch-Norm not applicable to transformers)
- Keep local running mean/variance for each client
- Aggregate only LoRA parameters, not normalization stats

FedPer (Arivazhagan et al., 2019):

- Personalized layers: Last 4 transformer blocks (layers 29-32)
- Shared layers: First 28 transformer blocks (layers 1-28)
- LoRA applied to all layers, but only shared-layer LoRA is aggregated

FedProto (Tan et al., 2022):

- Extract prototypes from hidden states (layer 24 output)
- Prototype dimension: 256 (projected from 4096)
- Temperature for prototype matching: $\tau_{\text{proto}} = 0.1$
- Loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + 0.1 \cdot \mathcal{L}_{\text{proto}}$

FedMoE (Mei et al., 2024):

- Number of experts: 4 heterogeneous LoRA experts
- Router: top-2 expert selection with load balancing

- Personalization: client-specific sub-MoE selection followed by modular aggregation
- Load balancing coefficient: 0.01
- Total parameters: $\sim 4 \times$ FedLing (4 sets of LoRA parameters)

FedHLT (Guo et al., 2024a):

- Hierarchical structure: Family $\rightarrow$ Subfamily $\rightarrow$ Language
- Aggregation: Bottom-up with learned aggregation weights
- Number of hierarchy levels: 3
- Uses language family labels from URIEL

FedLFC (Guo et al., 2024b):

- Language family clustering: 11 families (from URIEL)
- Intra-family aggregation first, then inter-family
- Family-level weights based on data size

q-FFL (Li et al., 2020b):

- Fairness parameter: $q = 5$ (tuned from $\{1, 3, 5, 10\}$)
- Objective: $\min_{\theta} \sum_k p_k \mathcal{L}_k(\theta)^q$ where $p_k \propto n_k$
- Reweighted gradients to emphasize poorly-performing clients

MuLA-F (Li et al., 2025):

- Multilingual federated LoRA baseline with SVD-based language-specific disentanglement
- Applies SVD to effective LoRA updates and selects language-contributing rank components using Diff-eRank
- Uses local orthogonal tuning to reduce multilingual knowledge conflicts
- Aggregates disentangled LoRA components following the original language-aware MuLA-F procedure
- For fair comparison, we use the same backbone model, LoRA rank, optimizer, local epochs, and federation schedule as FEDLING

3. **F1 Score (MASSIVE):**

- Macro-averaged F1 across 60 intent classes
- Handles class imbalance better than accuracy
- Per-class: $F1_c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}$
- Overall: $F1 = \frac{1}{60} \sum_{c=1}^{60} F1_c$

4. **Fairness Metrics:**

- Variance: $\text{Var} = \frac{1}{K} \sum_k (\text{Score}_k - \bar{\text{Score}})^2$
- Coefficient of Variation: $\text{CV} = \frac{\sqrt{\text{Var}}}{\bar{\text{Score}}}$
- Max-Min Gap: $\max_k \text{Score}_k - \min_k \text{Score}_k$
- Gini coefficient: $G = \frac{\sum_{i,j} |\text{Score}_i - \text{Score}_j|}{2K^2 \bar{\text{Score}}}$

I Additional Experimental Results

I.1 Complete Results on MASSIVE

Table 17 reports representative MASSIVE F1 scores for q-FFL and MuLA-F, the closest baselines in the main result. FedLing is not uniformly best on high-resource languages; its gains concentrate in the low-resource tail.

Lang	Tier	FedAvg	q-FFL	MuLA-F	FedLing	Δ_{best}
eng	High	86.9	84.7	90.7	90.4	-0.3
zho	High	84.1	82.5	87.9	87.5	-0.4
spa	High	85.3	83.8	88.7	88.6	-0.1
fra	High	84.4	83.1	87.6	87.9	+0.3
jpn	High	80.8	79.4	83.5	83.2	-0.3
kor	High	80.1	78.9	82.8	83.4	+0.6
ell	Mid	75.0	76.8	79.7	80.1	+0.4
tur	Mid	75.6	77.4	81.9	82.6	+0.7
tha	Mid	73.8	76.9	80.9	81.8	+0.9
ind	Mid	76.4	78.1	83.6	84.1	+0.5
urd	Mid	66.8	71.2	75.6	76.8	+1.2
swl	Low	65.2	71.6	74.1	75.6	+1.5
tam	Low	63.8	70.5	73.2	74.9	+1.7
tel	Low	62.4	69.8	72.4	73.8	+1.4
ben	Low	64.7	70.9	73.6	75.2	+1.6
kan	Low	60.2	68.7	70.5	72.1	+1.6
yor	Low	58.4	67.3	69.1	71.8	+2.7

Table 17: Representative per-language MASSIVE macro-F1; Δ_{best} denotes FedLing minus the best non-FedLing result.

H.4 Evaluation

Metrics:

1. **ROUGE-L** (Aya):

- Longest common subsequence F1
- Implementation: rouge_score library
- Reported: F1 score (balanced precision/recall)

2. **F1 Score (XNLI):**

- Macro-F1 over entailment, contradiction, and neutral labels
- Used for seen/unseen transfer evaluation