RelFed: Relative-Representation Federation of Heterogeneous Small Language Models for Multi-Site AIoT Sensing

Anonymous authors
Paper under double-blind review

Abstract

Sites of one organization observe the same events through different sensor subsets while running small language models of different sizes, and such fleets cannot be federated by parameter averaging: the tensors have no common shape, and resize-and-average workarounds degrade accuracy. RelFed federates them exactly. Each client re-expresses its pooled features as cosine similarities to a shared public anchor set, which makes the representation dimension identical for every backbone width, and a single shared classifier head over this space is aggregated by plain FedAvg at roughly sixteen kilobytes per client per round. Across text and sensor benchmarks under literal feature partitions, RelFed beats the paired local-only floor in 37 of 41 seed comparisons, reaching +3.69 points on 20 Newsgroups with a bootstrap CI of [+2.98, +4.11]. Gains peak near twenty federation rounds and stay positive at longer horizons, concentrate where local training is weakest, and vanish under label-disjoint partitions, delimiting the method to what it actually transfers: feature-view knowledge.

1 Introduction

Factories instrument themselves faster than they standardize. Sites of the same firm monitor the same machines and the same failure modes, yet each site is wired with its own subset of sensors, accumulated over years of retrofits, and no two sites observe the process through the same channels. Small language models have meanwhile become practical consumers of exactly this kind of stream: they ground raw telemetry in text on consumer AIoT devices (Aboya Messou et al., 2026a), forecast electrical load over long horizons (Aboya Messou et al., 2026e), reprogram spectral measurements into language for equipment monitoring (Aboya Messou et al., 2026c;d), and fill the gaps that radio dropouts leave in multivariate sensor windows (Aboya Messou et al., 2026b). Each site therefore ends up with what its hardware can afford: its own small language model, sized to its own budget, adapted to its own sensors. The experience those models accumulate is complementary by construction, and pooling it is the obvious next step.

Three constraints stand in the way of that pooling, and together they disqualify standard federated learning (McMahan et al., 2017). The site models are independently pretrained at different scales, so their parameter tensors have incompatible shapes and cannot be averaged; we call this *architecture heterogeneity*. Each site sees only its own feature subset, at training time and at deployment, because a production system only ever has its own sensors; we call this *feature-view heterogeneity*, and it is distinct from the well-studied label-skew setting, since every client covers all classes but none observes all evidence. Finally, raw observations never leave a site, so only auxiliary parameters and outputs on public data may be shared.

A tempting fix is weight-space surgery: resize each client's tensors to a common shape, average, and resize back. Our pre-registered campaigns show that this family of operators does not beat the local-only floor, which matches the underlying intuition that averaging deltas of independently pretrained bases injects noise rather than knowledge. Distillation over a public probe set fails for a different reason: when every client is competent on all classes but each sees different evidence, the logit consensus carries little that any single client lacks. Appendix A.4 quantifies both failures as adapted baselines (Table 9).

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.

What heterogeneous clients *can* share exactly is geometry relative to common reference points. That observation is the opportunity this paper pursues. Building on relative representations (Moschella et al., 2023), each client re-expresses its pooled hidden state as cosine similarities to a shared public anchor set embedded under its own current adapter. The dimension of this representation is the number of anchors, identical for every client regardless of backbone width. A single linear head over this space is then federated by plain sample-weighted FedAvg, with no resizing, no interpolation, and no distillation target. To our knowledge this is the first exact parameter federation across architecture-heterogeneous language models.

Contributions.

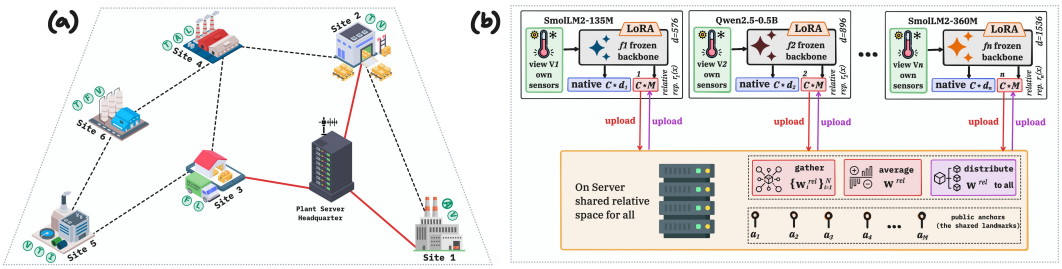
- **Exact federation without weight surgery.** Anchor-cosine relative representations give every client an identically shaped feature space, so one shared linear head is federated by plain FedAvg. The entire payload is at most 4,020 parameters per client per round for a 20-class task, independent of backbone size, and the mechanism improves the paired local-only floor by +3.69 pp on 20 Newsgroups with a bootstrap CI of $[+2.98, +4.11]$ and 5/5 seeds positive.
- **The featureskew protocol.** Clients observe overlapping feature subsets of the same underlying events, which for sensor data are literal channels. This fills a gap between label-skew and modality-split benchmarks and makes the industrial scenario testable on public data.
- **Two modalities, one pipeline.** The same system handles text and numerical time series, the two ingredients of the mixed logs modern machines emit; 37 of 41 paired comparisons improve.
- **A measured operating envelope.** Gains peak near twenty rounds and stay positive at larger budgets, concentrate on the partitions with the weakest local floors, and vanish under label-disjoint partitions. Static rel-space alignment does not predict the gain; we tested that natural hypothesis and reject it.

2 Related work

Federated learning under heterogeneity. FedAvg (McMahan et al., 2017) and its optimization-robust variants (Li et al., 2020) assume a shared architecture. HeteroFL (Diao et al., 2021) supports width-scaled submodels of one master network, which does not cover independently pretrained backbones of different widths. Federated fine-tuning of language models through LoRA has been demonstrated for intelligent automotive systems (Chen et al., 2025) and now spans privacy-hardened, rank-heterogeneous, and resource-adaptive adapter aggregation at LLM scale (Sun et al., 2024; Cho et al., 2024; Bai et al., 2024; Ye et al., 2024), while uncertainty-guided evidential distillation improves generalization under data heterogeneity (Chen et al., 2026); all, however, merge adapters of one architecture or into a common-rank object. Knowledge-distillation approaches sidestep parameter sharing entirely: FedMD (Li & Wang, 2019) and FedDF (Lin et al., 2020) exchange logits on public data, and prototype methods refine the shared object into class centroids (Tan et al., 2022; Zhang et al., 2024). These transfer label-space knowledge and are effective under label skew, but under feature-view heterogeneity, where all clients are competent on all classes and each sees different evidence, the ensemble consensus adds little, which our experiments confirm. The setting also differs from vertical federation, where parties hold complementary features of shared samples under one coordinated protocol (Liu et al., 2024); here each site trains and deploys alone on its own view.

Relative representations and feature-space federation. Moschella et al. (2023) showed that expressing latents relative to anchor samples makes independently trained spaces comparable zero-shot, a line since extended to translation between spaces (Maiorca et al., 2023) and to richer invariance products (Cannistraci et al., 2024). FedFA (Zhou & Konukoglu, 2023) federates feature statistics rather than weights. We combine the two ideas into a trainable, FedAvg-able object: a shared classifier head over the anchor-relative space of heterogeneous LM backbones, refreshed as LoRA adaptation (Hu et al., 2022) shifts each client's geometry.

Language models for numerical data. Serialization lets frozen language models consume numerical sequences as text (Gruver et al., 2023), and a growing line of work runs sensor workloads



C : *class* **d** : *dimension* **a** : *anchor* **V** : *view* ***f*** : *pretrained model* ***r*** : *relative representation*

Figure 1: **RelFed overview.** (a) The deployment world. Sites of one industry observe different sensor subsets, each runs its own small language model on an edge device, and only the tiny shared head travels to the plant server; raw data and model weights never leave a site. Right: the mechanism behind it. Clients run frozen backbones of different widths d_i, mixed SmolLM2 and Qwen2.5 scales in our experiments, with LoRA adapters, and observe only their own feature views. Each client keeps a native head ($C\times d_i$, never shared) and re-expresses pooled features as cosine similarities to the M public anchors, feeding a shared rel-head ($C\times M$ plus bias, one column per anchor), identical in shape for every client; the server gathers the client copies, averages them exactly, and distributes the result.

through language backbones directly: sensor-text grounding on consumer devices (Aboya Messou et al., 2026a), long-horizon load forecasting (Aboya Messou et al., 2026e), spectral reprogramming for equipment monitoring (Aboya Messou et al., 2026c;d), and frozen-backbone time-series imputation (Aboya Messou et al., 2026b). Here, serialized UCI-HAR windows (Anguita et al., 2013) let the identical federation pipeline run on sensor data, where the feature-view split is physically literal because a client's view *is* its sensor channels.

3 Method

Figure 1 shows the mechanism end to end, and Algorithm 1 states one round.

3.1 Setting

N clients hold private datasets D_i of examples (x, y), an input x with its label y from a task-wide set of C classes, drawn from the same task distribution but observed through client-specific feature views V_i, which are subsets of F feature groups and, for sensor data, physical channels. Client i runs a frozen, independently pretrained backbone f_i of width d_i, whose pooled output $f_i(x) \in \mathbb{R}^{d_i}$ is the representation of x; a LoRA adapter with trainable parameters θ_i; and a native classifier head $W_i \in \mathbb{R}^{C\times d_i}$. Widths d_i differ across clients, so no native parameter is shape-compatible for averaging. What the setting therefore demands is a meeting point that no client owns: a reference frame every backbone can reach from its own coordinates, cheap enough to ship to the humblest site, and public enough to threaten no one's data. The next subsection builds that frame.

3.2 Relative representations over a public anchor set

A public anchor set $A = \{a_1, \ldots, a_M\}$ of M task-domain texts, containing no private data, is shared by all clients. Once per round, client i embeds the anchors under its current adapter into

$$E_i = \left[f_i(a_1);\ \ldots;\ f_i(a_M) \right] \in \mathbb{R}^{M\times d_i}, \tag{1}$$

whose j-th row $E_i^{(j)}$ is the embedding of anchor a_j. For an input x the relative representation over equation 1 is

$$r_i(x) = \left(\cos(f_i(x), E_i^{(1)}), \ldots, \cos(f_i(x), E_i^{(M)}) \right) \in \mathbb{R}^M, \tag{2}$$

where, for vectors u and v with $\langle \cdot, \cdot \rangle$ the inner product and $\|\cdot\|_2$ the Euclidean norm, the similarity is

$$\cos(u, v) = \frac{\langle u, v \rangle}{\|u\|_2 \, \|v\|_2}, \tag{3}$$

applied row-wise against equation 1, so the representation in equation 2 is identical in dimension for every client regardless of d_i.

Two properties of this construction carry the method. First, equation 2 is invariant to any transformation of a client's feature space that preserves angles to the anchor embeddings, so two backbones that encode the same evidence in different bases produce comparable coordinates without the bases ever being compared. Second, every quantity in equation 1 and equation 2 is computed under the client's own adapter and tokenizer, so a site reads the public anchors exactly the way it reads its private stream and the shared frame adds no deployment plumbing. Because the adapter moves during training, E_i is refreshed once per round at the cost of M extra forwards, and the anchors are a seed-keyed draw of M=200 from a fixed public file of 300, so the per-seed consistency reported later also spans anchor-set variation.

The ingredients are known separately; their federation role is not. Relative representations were introduced to stitch and compare independently trained encoders after the fact, with both models fixed and fully observed (Moschella et al., 2023); here they become a training-time substrate, recomputed as adapters drift, observed only through each site's partial view, and carrying a live federated parameter rather than a post-hoc comparison. Classifier-sharing methods conversely assume what this setting denies, a common feature width for the shared head (Yi et al., 2023) or exchangeable prototypes in the native space (Tan et al., 2022), so under mixed d_i both need exactly the resize surgery whose failure Appendix A.4 documents. The construction above supplies the missing piece: a shared space that exists by construction rather than by assumption. Over it, a single shared head maps the representation to logits,

$$z_i(x) = W^{rel} r_i(x) + b^{rel}, \qquad W^{rel} \in \mathbb{R}^{C\times M},\ b^{rel} \in \mathbb{R}^{C}, \tag{4}$$

whose width is the anchor count by construction, one head column per anchor, which is the one M this paper uses everywhere (local epochs are written M_{loc}). Linearity here is a design decision rather than a concession: each site's representational capacity stays in its backbone and adapter, and the shared object is deliberately the smallest model that can carry class evidence over the common frame, one weight per class-anchor pair and one bias per class. A linear map over an identical input space is also what makes averaging meaningful in equation 6 below, since the mean of such maps is a map of the same kind over the same space, whereas means of full networks in incompatible coordinates are precisely the surgery our negative controls reject.

3.3 Training and federation

The frame of equation 2 is only useful if local training feeds it. Each round, clients therefore refresh anchors and train locally for M_{loc} epochs with the joint loss

$$\mathcal{L} = \mathrm{CE}(W_i f_i(x), y) + \mathrm{CE}(W^{rel} r_i(x), y), \tag{5}$$

where CE is the cross-entropy to the label y and a single backbone forward feeds both terms of equation 5; the server then aggregates the returned client copies W_i^{rel} as

$$W^{rel} \leftarrow \sum_i \tfrac{n_i}{n} W_i^{rel}, \tag{6}$$

and broadcasts, with n_i the number of training examples at client i and $n = \sum_i n_i$. The aggregation in equation 6 is exact in the sense that matters here: every client copy shares one shape and one input space, so the server performs plain sample-weighted parameter averaging with no resizing, no projection, and no distillation target, and is stateless across rounds, holding nothing but the current head. Inference ensembles the native and relative softmax into the site's predictive distribution

$$p_i(x) = \tfrac{1}{2}\Big[\mathrm{softmax}\big(W_i f_i(x)\big) + \mathrm{softmax}\big(z_i(x)\big)\Big], \tag{7}$$

the predictor every site deploys. The two terms of equation 7 play complementary roles: the native softmax scores full-resolution local features that never leave the site, while the relative softmax carries whatever the fleet has accumulated in the shared frame, and §4.5 separates their contributions empirically. The server never touches private data, and the only shared parameters are the weights and bias of equation 4, whose shape depends only on the class and anchor counts and on no backbone width. In the factory reading, a site uploads sixteen kilobytes once per round and receives the fleet's

pooled evidence back in the same currency; nothing about its backbone, its adapter, or its sensors ever leaves the premises.

Algorithm 1 RelFed: one federation round

given public anchors $A = \{a_1, \ldots, a_M\}$, shared head $W^{rel} \in \mathbb{R}^{C \times M}$, clients $i = 1..N$ with frozen backbones f_i (width d_i), LoRA θ_i, native heads W_i
for each client i **in parallel do**
  $E_i \leftarrow \text{normalize}([f_i(a_1); \ldots; f_i(a_M)])$ ▷ anchor refresh equation 1 under current θ_i
  for M_{loc} local epochs over private data (x, y) **do**
    $h \leftarrow f_i(x); \quad r \leftarrow \cos(h, E_i) \in \mathbb{R}^M$ ▷ relative representation equation 2 via equation 3
    $z \leftarrow W_i^{rel} r + b_i^{rel}$ ▷ shared-head logits equation 4
    $\mathcal{L} \leftarrow \text{CE}(W_i h, y) + \text{CE}(z, y)$ ▷ joint loss equation 5; backprops into θ_i
    update θ_i, W_i, W_i^{rel}
  end for
  upload W_i^{rel} ▷ payload equation 8; nothing else moves
end for
server: $W^{rel} \leftarrow \sum_i \frac{n_i}{n} W_i^{rel}$; broadcast to all clients ▷ exact FedAvg equation 6
each site serves predictions with the ensemble equation 7

Algorithm 1 assembles the round end to end. What remains is the empirical question the mechanism cannot answer for itself: whether so small a shared object actually carries feature-view knowledge between sites. The next section fixes a protocol in which every claim is a paired same-seed comparison, then measures where the exchange pays, how fast, at what cost, and where it has nothing to offer.

4 Experiments

4.1 Setup

Clients are five: two SmolLM2-135M instances plus SmolLM2-360M (Ben Allal et al., 2025), and Qwen2.5-0.5B with Qwen2.5-1.5B (Qwen Team, 2024), so $d_i \in \{576, 896, 960, 1536\}$, all with frozen backbones and LoRA r=8 on the attention projections. The tasks are 20-way 20 Newsgroups (20NG), 10-way Yahoo Answers, and 6-way UCI-HAR, whose nine inertial channels are serialized as text and serve as the feature groups. The protocol is K=20 rounds with one local epoch per round (M_{loc}=1) and M=200 unmasked anchors; every comparison is a paired same-seed delta against the local-only floor over $n \geq 5$ seeds (the K=100 row uses n=3), with bootstrap 95% CIs from 10,000 resamples and per-seed extremes reported; the appendix lists the full roster and dataset details in Tables 6 and 8.

A note on the choice of data is in order. The mechanism has additionally been piloted on a proprietary multi-site industrial dataset, where both the local floors and the federated accuracies fall in the 60 to 80% range; confidentiality prevents disclosing that data or those runs. Every experiment reported in this paper therefore uses public benchmarks under the documented protocol, and all claims rest exclusively on these reproducible results.

4.2 Main results

Table 1 reports the paired comparison on every task; Figure 2 shows every pair raw, and Figure 3 shows when and for whom the gains arise (Figure 8 in the appendix pools all deltas). On 20 Newsgroups the gain is +3.69 pp (CI [+2.98, +4.11], 5/5 seeds) and survives replacing the mixed pool with Qwen-only clients (+2.79, CI [+2.04, +3.31], 5/5), so the effect is not an artifact of one architecture mix. Yahoo (+1.27, CI [−0.10, +2.70], 3/5) and HAR (+0.94, CI [−0.78, +2.42], 6/8) trend positive but their intervals graze zero at current seed counts; we report them as consistent trends, not confirmed effects. Pooling all 41 featureskew pairs, 37 improve (sign test p=5.1×10^{-8}): the effect is not carried by a few fortunate cells. Per backbone, every backbone gains on 20NG and the largest backbone gains most, as Table 2 shows; federation here is not a small-model subsidy. For a multi-site deployment the sign consistency matters as much as the mean: a site cannot choose its

Table 1: **Main results** (paired same-seed comparison, $n \geq 5$ seeds): mean accuracy of the local-only floor vs RelFed, paired delta with bootstrap 95% CI, per-seed min/max, and sign consistency. Protocol: K=20, M_{loc}=1, M=200 unmasked anchors, ensemble eval.

Task	n	Floor	RelFed	Δ (pp)	95% CI	range	sign
UCI-HAR (numerical)	8	55.94 ±2.56	56.88 ±2.00	+0.94	[-0.78, +2.42]	[-4.10, +3.60]	6/8
20 Newsgroups	5	34.29 ±1.46	**37.98 ±1.71**	**+3.69**	[+2.98, +4.11]	[+2.30, +4.20]	5/5
20 Newsgroups (Qwen)	5	34.62 ±2.22	**37.41 ±1.78**	**+2.79**	[+2.04, +3.31]	[+1.35, +3.40]	5/5
Yahoo Answers	5	47.96 ±2.84	49.23 ±2.54	+1.27	[-0.10, +2.70]	[-0.75, +3.60]	3/5

Table 2: **Per-backbone gains** (K=20, paired, pp with bootstrap 95% CI; bold = CI excludes zero). Federation is not a small-model subsidy: on 20NG the largest backbone gains most.

Task	SmolLM-135M	Qwen-0.5B	SmolLM-360M	Qwen-1.5B
20NG	**+2.53 [+0.85,+4.15]**	**+3.25 [+0.60,+5.85]**	**+3.15 [+2.80,+3.50]**	**+7.00 [+4.30,+10.55]**
Yahoo	-0.97 [-2.62,+0.82]	+2.50 [-0.10,+5.00]	+0.85 [-0.95,+2.95]	**+4.95 [+1.85,+9.10]**
HAR	+0.31 [-1.45,+2.05]	+1.28 [-1.28,+3.62]	**+2.91 [+1.34,+4.31]**	-0.09 [-2.94,+3.03]

Table 3: Round-scaling rows (20NG featureskew, protocol as Table 1 except the varied K and M_{loc}).

Rounds	n	Floor	RelFed	Δ (pp)	95% CI	range	sign
$K = 4, M_{loc} = 5$	5	34.65 ±1.74	**36.40 ±1.40**	**+1.75**	[+1.03, +2.71]	[+0.95, +3.50]	5/5
$K = 8, M_{loc} = 5$	5	35.57 ±1.63	**37.28 ±2.18**	**+1.71**	[+0.89, +2.79]	[+0.50, +3.75]	5/5
$K = 20, M_{loc} = 1$	5	34.29 ±1.46	**37.98 ±1.71**	**+3.69**	[+2.98, +4.11]	[+2.30, +4.20]	5/5
$K = 50, M_{loc} = 1$	5	35.39 ±1.34	**37.97 ±1.65**	**+2.58**	[+0.69, +4.58]	[+0.10, +6.25]	5/5
$K = 100, M_{loc} = 1$	3	35.17 ±0.25	**37.60 ±1.84**	**+2.43**	[+0.90, +4.75]	[+0.90, +4.75]	3/3

seed, so 37 improvements in 41 paired runs is the number that tells an operator what to expect from enrolling.

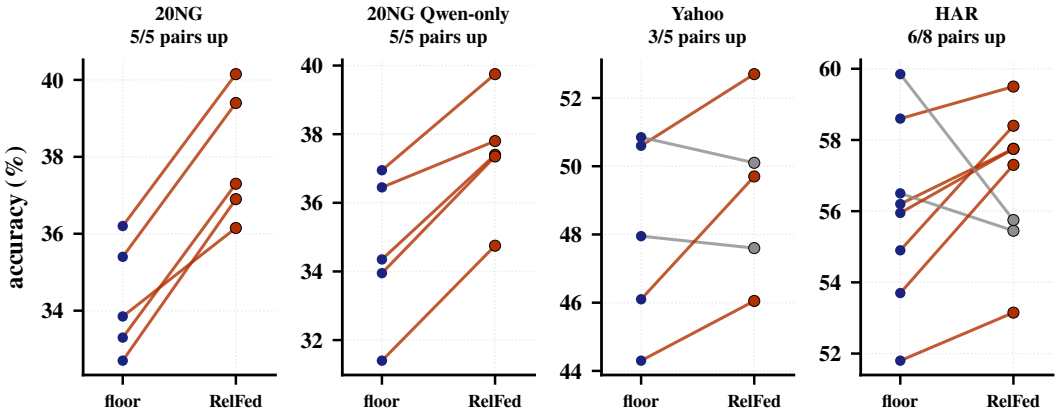


Figure 2: **Every pair, individually**: floor → RelFed per seed for the four K=20 configurations (20NG, 20NG-Qwen, Yahoo, HAR; vermillion: improved; grey: degraded). The paired design behind every statistic in the paper, shown raw.

4.3 Round scaling

Having established that the effect is real at one budget, the next question is how it depends on the one resource federation actually spends: rounds of aggregation. Table 3 lists the rows; Figure 10 in

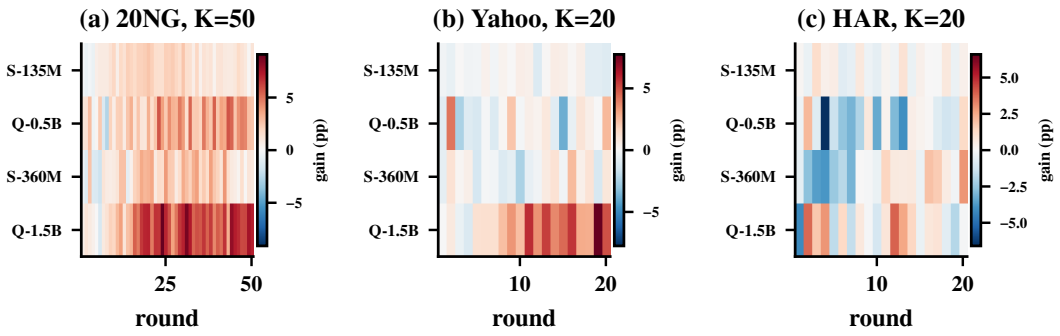


Figure 3: **When and for whom**: per-client paired gain over rounds (mean over seeds; red = RelFed above floor). On 20NG and Yahoo gains build steadily and concentrate in the largest backbone late in training; on HAR early rounds lag (blue) before the shared head catches up.

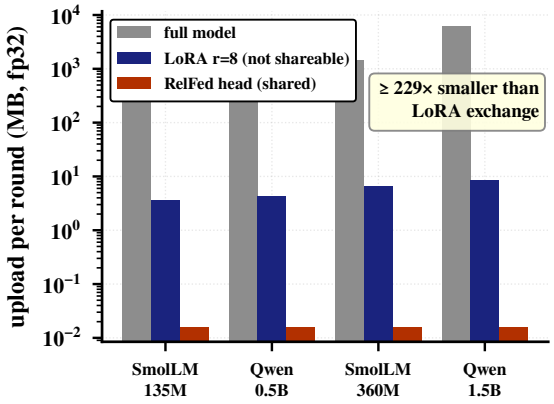


Figure 4: **Per-round upload per client** (log scale, fp32, exact model configurations). RelFed's only federated payload is the shared rel-head, $C{\times}200{+}C$ = 4,020 parameters for 20NG: 229–542× smaller than exchanging the shape-incompatible LoRA adapters and 4–6 orders of magnitude below full-model exchange.

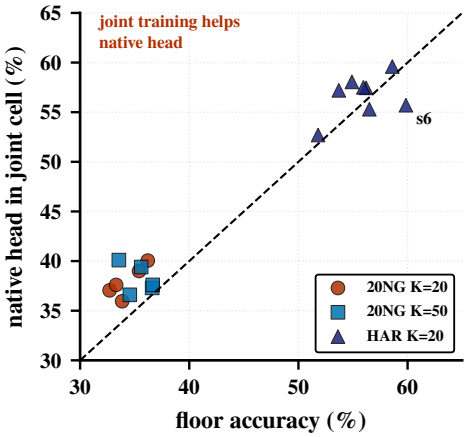


Figure 5: **The joint-training effect on the native head.** Native-head accuracy under the joint loss equation 5 vs the paired floor, per configuration and seed. On 20NG the rel gradients *improve* the native path through the shared LoRA (points above the diagonal): representation shaping, not mere ensembling. On HAR the coupling can become a tax for individual seeds (seed 6, labeled).

the appendix shows the per-backbone trajectories at the horizons $K \in \{20, 40, 50, 100\}$, alongside the per-backbone view across tasks and the summary curve (Figures 9 and 7). Two observations. First, at *equal local compute* (20 local epochs; Figure 6b) federating often beats federating rarely: K=20/M_{loc}=1 doubles the gain of K=4/M_{loc}=5 (+3.69 vs +1.75). Second, gains peak at $K{\approx}20$ and settle lower yet clearly positive at longer horizons (+2.58 at K=50, +2.43 at K=100): modest round budgets suffice, and the operative lever is the *frequency* of head aggregation, not the total amount of local training.

4.4 Communication efficiency

Frequent aggregation only matters if a round is cheap, which raises the question of what a round costs. Because clients share only the head of equation 4, the per-site upload B in single precision is

$$B \;=\; 4\,(CM + C) \text{ bytes per round,} \tag{8}$$

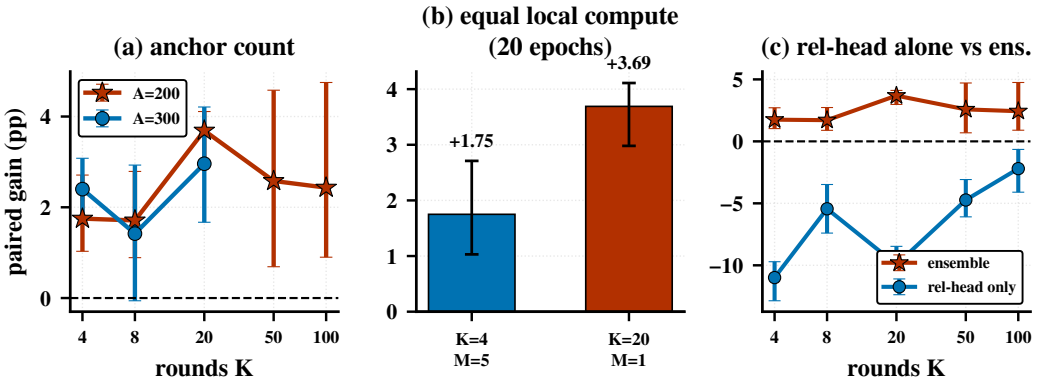


Figure 6: **Ablations**: (a) anchor count (A300 helps at small K, A200 wins at K=20); (b) equal local compute, where federating often doubles the gain of federating rarely; (c) the rel-head alone converges toward floor with K; the ensemble makes it a gain.

independent of backbone size: 16,080 bytes at C=20, M=200, smaller than a single Transformer layer's LoRA factors (Figure 4 plots equation 8 against the alternatives).

4.5 Ablations

Each ablation below removes or varies one design choice and names what its result changes about the paper's claims. **Anchor count** (Figure 6a). More anchors help when federation is scarce (A=300 beats A=200 at K=4: +2.40 vs +1.75) but not at the paper protocol (A=200 wins at K=20: +3.69 vs +2.96; Table 5): a richer static coordinate system partially substitutes for federation steps, and becomes unnecessary once aggregation is frequent. **Anchor publicity.** Masking the anchors through each client's own feature view, the fully-literal reading of the sensor scenario, destroys the method: cross-client rel-representation correlation collapses to 0.08–0.22 and rel-only accuracy falls toward chance, because anchor j then means something different on every client. Anchors must be a *public shared reference*; this failure was caught in a pre-registered smoke test, and the unmasked protocol was adopted before any battery ran. **Eval mode.** The shared head alone starts far below the floor, −11 pp at K=4, and closes most of that gap by K=100, −2 pp, though not monotonically (Figure 6c): the deployed gain requires the ensemble of equation 7, while the rel-only trajectory isolates how much transferable knowledge the shared head itself accumulates. **Model family.** The Qwen-only pool of §4.2 preserves the effect, so the gain does not depend on any one architecture mix. **Federation attribution.** A control trains the identical rel-head but never aggregates it (the head stays local; mean cross-client head distance grows to 2.2 by round twenty instead of returning to zero). The unfederated rel-head adds +1.33 pp over the paired floor (CI [+0.27, +2.22], 4/5), the joint-objective effect of Figure 5; aggregation by equation 6 adds a further +2.36 pp (CI [+1.80, +2.92], 5/5). The components sum to the headline +3.69 pp exactly; the larger exists only under federation.

4.6 Scope: where the method does not help

A method that claims to transfer feature-view knowledge should fail where there is none to transfer; this section checks exactly that (Table 4), and Figure 5 shows the mechanism behind the gains it does deliver.

Label-disjoint partitions are null at both K=4 and K=20 (CIs straddle zero): the shared rel-head transfers feature-view evidence, not absent-class knowledge. Per-partition gain variation is predicted by floor strength (Pearson $r = -0.45$ over 41 pairs), not by static cross-client rel-space correlation ($r = +0.04$; hypothesis tested and rejected).

A campaign of roughly one hundred and twenty further paired cells sharpens this anti-correlation into an operating envelope (Appendix A.5). Longer inputs, more local data, or wider feature views on the mixed fleet buy absolute accuracy at the cost of the paired gain; backbone quality preserves it: a Qwen-only fleet reaches 60.2% on the sensor task at +1.39 pp, interval [+0.64, +2.14], five seeds

of five, and 58.7% on movie-review sentiment at +2.83 pp, interval [+0.95, +4.71]. Stacking data on that fleet lifts floors to 63.6% at 800 windows per site and 68.0% at 1,400 while the gain thins to +1.22 pp, interval [−1.44, +4.59], and then into noise, a third lever confirming the envelope inside a single task. For calibration, federated fine-tuning of RoBERTa-base with full text views reports 62 to 65% on 20 Newsgroups (Wu et al., 2026), so the floors above sit under an observation model withholding six of ten feature groups.

5 Conclusion

We showed that architecture-heterogeneous SLMs observing partial feature views of a common task can be federated *exactly*, with no weight surgery and no distillation, by training a shared classifier head over anchor-relative representations, at a communication cost of 4,020 parameters per client per round. Across text and serialized sensor data the method beats the local-only floor in 37 of 41 paired comparisons, with gains that peak near twenty rounds of federation and stay positive at longer horizons, concentrate where local training is weakest, and survive a single-family client pool.

Limitations. Yahoo and HAR are positive trends whose intervals cross zero at current seed counts; the method transfers no absent-class knowledge (label-disjoint null); and the main protocol uses N=5 clients with 400 examples each, with per-site data varied to 1,400 windows by the envelope campaign but client-count scaling untested. Open next steps are the residual per-partition gain variation (predicted by floor strength, not rel-space alignment), regression heads for imputation, and *dynamic* federations where clients and features arrive over time, the natural extension of the industrial setting here.

Reproducibility statement

Every number, table, and figure regenerates deterministically from raw run artifacts via the analysis scripts in the supplementary material, which also contains the complete training pipeline, the public anchor files, an exact environment lockfile, and a README that reproduces the headline cell and its paired floor with one command each. All datasets are public and load from the identifiers baked into the loaders; the protocol, seeds, and feature-view construction are specified in §4.1 and the appendix, and the hardware and software stack in Table 7. Training kernels are GPU-nondeterministic, so replication reproduces the reported distributions and conclusions rather than bit-identical trajectories, which is why every claim is a paired same-seed comparison with a bootstrap confidence interval.

AI usage statement

Generative AI tools were used to aid and polish the writing of this manuscript and for retrieval and discovery of related work. All method design decisions, experimental protocols, and claims are the authors' own, and every reported number is produced by deterministic analysis scripts from raw experiment artifacts.

References

Franck Junior Aboya Messou, Jinhua Chen, Tong Liu, P. Suresh, Chinmay Chakraborty, and Keping Yu. Lightweight and secure real-time sensor-text grounding for consumer AIoT devices via small language models. *IEEE Consumer Electronics Magazine*, pp. 1–12, 2026a. doi: 10.1109/MCE.2026.3707950. Early Access.

Franck Junior Aboya Messou, Jinhua Chen, Tong Liu, Shilong Zhang, Weiyu Wang, Tao Yu, and Keping Yu. LLM4Imp: Leveraging frozen large language models with spectral prompts for time-series imputation. In *IEEE International Conference on Communications (ICC)*, pp. 1–6, 2026b. doi: 10.1109/ICC59461.2026.11586827.

Franck Junior Aboya Messou, Jinhua Chen, Tong Liu, Shilong Zhang, Weiyu Wang, Tao Yu, and Keping Yu. Spec2LLM: A spectral-to-language reprogramming framework for power transformer temperature forecasting. In *2026 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–6, 2026c. doi: 10.1109/ICCE67443.2026.11449688.

Franck Junior Aboya Messou, Jinhua Chen, Tao Yu, Keping Yu, Osama Alfarraj, and Amr Tolba. Spec2Llama: A spectral-aware forecasting approach for versatile time series analysis. *Expert Systems with Applications*, 319:131986, 2026d. doi: 10.1016/j.eswa.2026.131986.

Franck Junior Aboya Messou, Jinhua Chen, Shilong Zhang, Tong Liu, Amr Tolba, Osama Alfarraj, and Keping Yu. TAPformer: A transformer with adaptive attention and temporal projection encoding for long-term electricity load forecasting in IoT systems. *Internet of Things*, 37:101895, 2026e. ISSN 2542-6605. doi: 10.1016/j.iot.2026.101895.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *European Symposium on Artificial Neural Networks (ESANN)*, pp. 437–442, Bruges, Belgium, 2013.

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 *Advances in Neural Information Processing Systems*, 2024.

Loubna Ben Allal et al. SmolLM2: When Smol goes big – data-centric training of a small language model, 2025.

Irene Cannistraci, Luca Moschella, Marco Fumero, Valentino Maiorca, and Emanuele Rodolà. From bricks to bridges: Product of invariances to enhance latent space communication. In *International Conference on Learning Representations*, 2024.

Jinhua Chen, Franck Junior Aboya Messou, Shilong Zhang, Tong Liu, Keping Yu, and Dusit Niyato. Federated fine-tuning of large language models for intelligent automotive systems with low-rank adaptation. In *2025 IEEE 101st Vehicular Technology Conference (VTC2025-Spring)*, pp. 1–6, 2025. doi: 10.1109/VTC2025-Spring65109.2025.11174441.

Jinhua Chen, Franck Junior Aboya Messou, Shilong Zhang, Tao Yu, Tong Liu, Weiyu Wang, and Keping Yu. FedUED: Enhancing generalization of federated learning via uncertainty-guided evidential distillation in consumer electronics. In *2026 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–6, 2026. doi: 10.1109/ICCE67443.2026.11449922.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous LoRA for federated fine-tuning of on-device foundation models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12903–12913, 2024.

Enmao Diao, Jie Ding, and Vahid Tarokh. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. In *International Conference on Learning Representations*, 2021.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, 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 *International Conference on Learning Representations*, 2022.

Daliang Li and Junpu Wang. FedMD: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, pp. 429–450, 2020.

Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *Advances in Neural Information Processing Systems*, 2020.

Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-Qin Zhang, and Qiang Yang. Vertical federated learning: Concepts, advances, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3615–3634, 2024.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. In *Advances in Neural Information Processing Systems*, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023.

Qwen Team. Qwen2.5 technical report, 2024.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *International Conference on Learning Representations*, 2024.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. FedProto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8432–8440, 2022.

Fei Wu, Jia Hu, Geyong Min, and Shiqiang Wang. Preventing rank collapse in federated low-rank adaptation with client heterogeneity. *arXiv preprint arXiv:2602.13486*, 2026.

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 *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6137–6147, 2024.

Liping Yi, Gang Wang, Xiaoguang Liu, Zhuan Shi, and Han Yu. FedGH: Heterogeneous federated learning with generalized global header. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8686–8696, 2023.

Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. FedTGP: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16768–16776, 2024.

Tianfei Zhou and Ender Konukoglu. FedFA: Federated feature augmentation. In *International Conference on Learning Representations*, 2023.

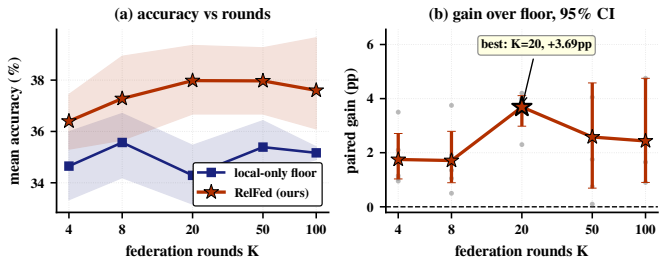


Figure 7: **Round scaling** (20NG featureskew): (a) mean accuracy of floor vs RelFed with bootstrap CI bands; (b) paired gain with 95% CI whiskers and per-seed points. Gains grow to $K \approx 20$ and settle lower yet positive beyond it: federation frequency, not local epochs, is the operative lever.

A Additional experimental detail

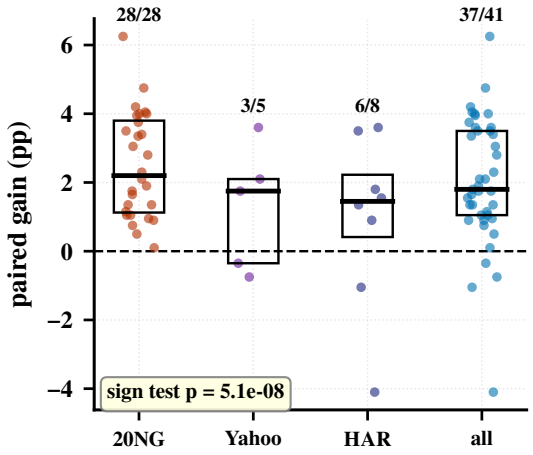


Figure 8: **Robustness**: every paired featureskew delta in the paper, by task and pooled: 37/41 positive, one-sided sign test p=5.1×10^{-8}. The effect is not carried by a few fortunate cells.

A.1 Extended convergence and analysis figures

Figures 7, 8, and 9 extend the round-scaling, robustness, and per-backbone views of Section 4.

Table 4: Scope: featureskew vs label-disjoint (20NG, paired, protocol as Table 1). Label-disjoint deltas are null at both K=4 and K=20, so the head transfers feature-view evidence, not absent-class knowledge.

Split	n	Floor	RelFed	Δ (pp)	95% CI	range	sign
featureskew K=20	5	34.29 ±1.46	**37.98 ±1.71**	**+3.69**	[+2.98, +4.11]	[+2.30, +4.20]	5/5
featureskew K=4	5	34.65 ±1.74	**36.40 ±1.40**	**+1.75**	[+1.03, +2.71]	[+0.95, +3.50]	5/5
label-disjoint K=20	5	15.31 ±0.30	15.12 ±0.88	-0.19	[-0.90, +0.37]	[-1.55, +0.50]	2/5
label-disjoint K=4	5	15.49 ±0.66	15.33 ±0.46	-0.16	[-0.61, +0.21]	[-0.95, +0.30]	3/5

A.2 Feature-view (featureskew) construction

For text tasks, each document's words are assigned to F=10 interleaved groups (word $j \rightarrow$ group j mod F), so every group spans the whole document sparsely rather than one group owning the informative opening. Client i keeps a contiguous cyclic window of 4 groups from a seed-permuted

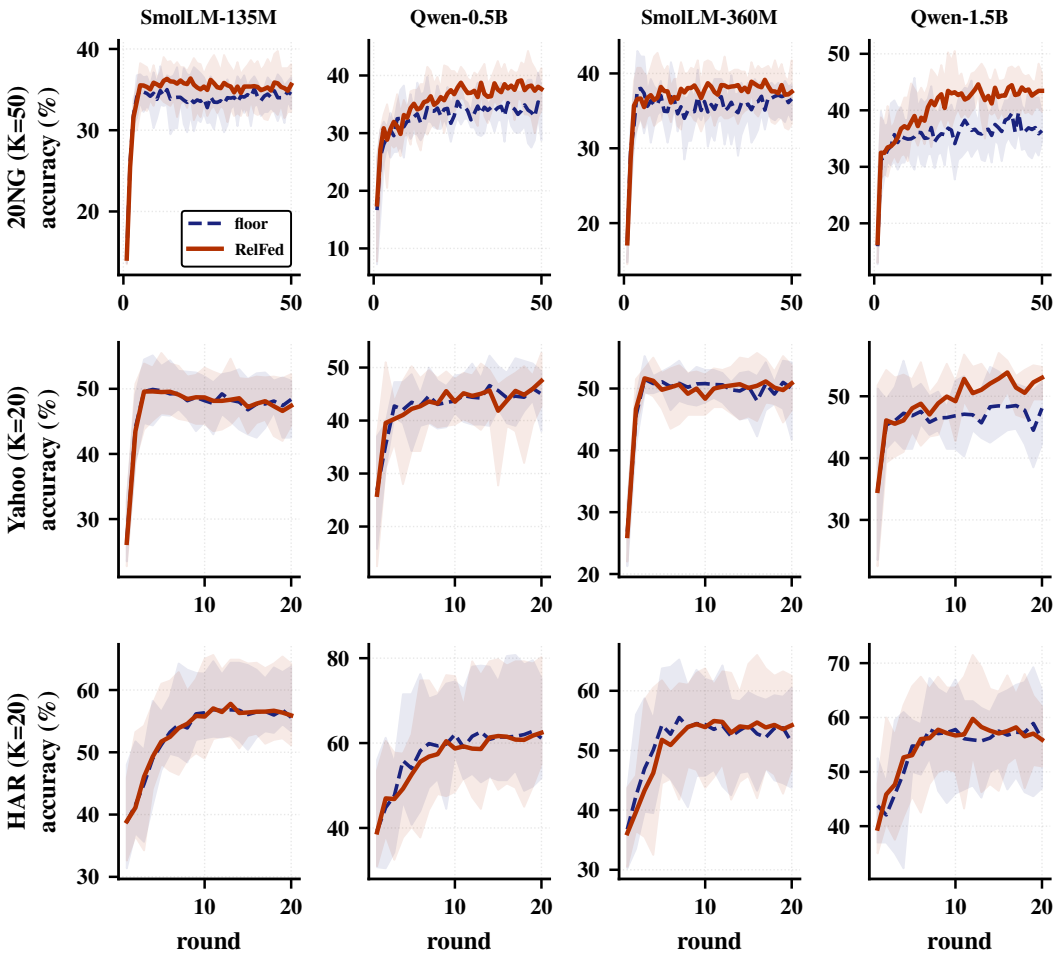


Figure 9: **Per-client accuracy over rounds**, one panel per backbone (mean over seeds; bands: seed min/max; dashed: floor, solid: RelFed). Gains emerge in the second half of training as the FedAvg'd head accumulates cross-client evidence, and no client is sacrificed for the mean.

Table 5: Anchor-count ablation (20NG featureskew, pp with 95% CI).

Rounds	A=200	A=300
K=4	**+1.75 [+1.03,+2.71]**	**+2.40 [+1.74,+3.08]**
K=8	**+1.71 [+0.89,+2.79]**	+1.42 [-0.06,+2.93]
K=20	**+3.69 [+2.98,+4.11]**	**+2.96 [+1.73,+4.16]**

group order, applied to *both* train and test text (a branch's deployment sees only its own sensors). For UCI-HAR the groups are the 9 physical inertial channels (`body_acc`/`body_gyro`/`total_acc` × x/y/z); each client observes 4 channels. Windows are downsampled ×4 (128→32 steps) and serialized per channel as `chi:` v_1 v_2 ... with two-decimal precision, tokenized by each client's own tokenizer (sequence length 320).

A.3 Anchors and refresh schedule

Anchors are M=200 lines drawn under the run seed from a fixed 300-line public probe file per task, task-domain text and, for HAR, serialized windows from the public train split; every seed therefore uses a different anchor draw, so the reported per-seed consistency also spans anchor-set variation. Anchor embeddings E_i are recomputed once per round under the client's current LoRA (drift control)

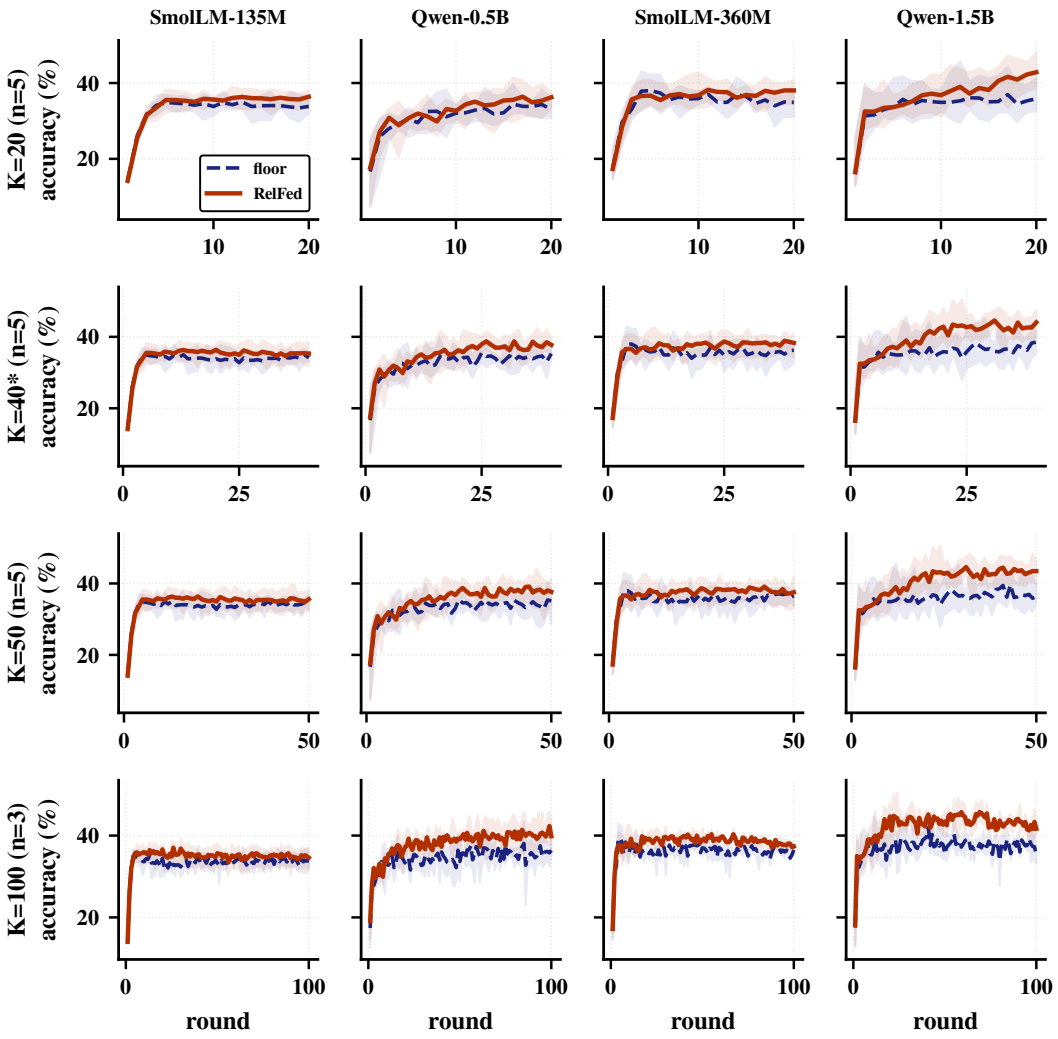


Figure 10: **20NG per-backbone trajectories across round horizons** $K \in \{20, 40, 50, 100\}$ (bands: seed min/max; K=40* truncates the K=50 runs, exact since no quantity depends on the total rounds budget). The RelFed–floor separation widens with the horizon and is largest for the largest backbone.

Table 6: Client roster and training configuration (heads/KV: attention heads / key-value heads). Widths d span 576 to 1536, so no two backbones are shape-compatible for averaging.

Backbone	params	d	layers	heads/KV	LoRA params
SmolLM2-135M (×2)	135M	576	30	9/3	0.92M
Qwen2.5-0.5B	494M	896	24	14/2	1.08M
SmolLM2-360M	362M	960	32	15/5	1.64M
Qwen2.5-1.5B	1.54B	1536	28	12/2	2.18M

LoRA r=8, α=16 on q/k/v/o_proj; AdamW, lr 2×10^{-4}, weight decay 0.01, batch 8, grad-clip 1.0; heads init $\mathcal{N}(0, 0.02^2)$; rel-head $C\times200$, FedAvg'd each round.

and L2-normalized; within a round they are constant, so the per-batch cost of equation 3 against all M anchors is one matrix product. Anchors are *not* masked by the client's feature view: masked anchors break cross-client rel-space comparability (rel-only accuracy collapses toward chance), an ablation quantified in §4.5.

Table 7: Device and environment specification for every experiment; the complete package lockfile ships in the supplementary material.

Component	Specification
GPU	NVIDIA RTX PRO 6000 Blackwell Server Edition (96 GB)
GPU driver / CUDA	575.57.08 / 12.9
Operating system	Ubuntu 20.04.6 LTS
Python	3.11.15
PyTorch	2.11.0 (cu128 build)
Transformers	5.14.1
PEFT (LoRA)	0.20.0
Datasets	5.0.1
NumPy	2.4.6

Table 8: Dataset and feature-view details (per client, N=5).

Dataset	modality	classes	train/client	test	feature groups	seq len
20 Newsgroups	text	20	400	400	10 word-groups	64
Yahoo Answers	text	10	400	400	10 word-groups	64
UCI-HAR	sensor (9 ch.)	6	400	400	9 channels	320

Feature views: interleaved word-groups (word $j \rightarrow$ group j mod F) for text; physical inertial channels for HAR (×4 downsampled, 2-decimal serialization). Each client keeps a cyclic window of 4 groups, applied to train AND test. Anchors: 200 lines drawn under the run seed from the fixed 300-line public per-task probe file, unmasked.

Table 9: Adapted baselines. Existing heterogeneous-federation families do not apply to unequal widths without modification; each row states the adaptation our pre-registered campaigns tested, on the same tasks, clients, and paired-seed protocol as the main results, and the paired outcome.

Family	Canonical methods	Adaptation to unequal widths	Δ vs floor (pp)
Weight resize- and-average	FedAvg (McMahan et al., 2017)	resize LoRA tensors to a common shape (SVD, interpolation, DCT, wavelet, random projection), average, resize back	+0.16 (best of 11 operators; noise)
Shared-classifier federation	FedGH (Yi et al., 2023), FedProto (Tan et al., 2022)	native heads have width d_i, so head averaging requires the same resize family applied to $C \times d_i$ heads	−0.55 to −2.53 (3 settings, 11 runs)
Logit-consensus distillation	FedMD (Li & Wang, 2019)	domain-matched public probe set; per-client KL to the consensus	−2.11, CI [−3.09, −1.22]
Server ensemble distillation	FedDF (Lin et al., 2020)	trainable server pivot distilled from the client-logit ensemble on the public probes	−1.82
Class-aware consensus	weighted FedMD variant	consensus weighted by client label histograms	null (CIs straddle zero)
RelFed (ours)	–	none needed: the shared head is $C \times M$ by construction	**+3.69**, CI [+2.98, +4.11]

A.4 Negative controls: the approaches that do not work

The claims in §1 that weight-space surgery and logit distillation fail in this regime rest on our own pre-registered campaigns, run on the same tasks, clients, and paired-seed protocol as the main results, summarized as adapted baselines in Table 9: **(i) weight-space resize-and-average** (SVD truncation, Procrustes-anchored SVD, bicubic/bilinear/area interpolation, DCT, wavelets, random projections applied to LoRA tensors): across a 756-cell campaign the best operator beat the floor by +0.16 pp on average, which is noise, and a single-seed +11.8 pp outlier collapsed to +3.96 pp under 5-seed

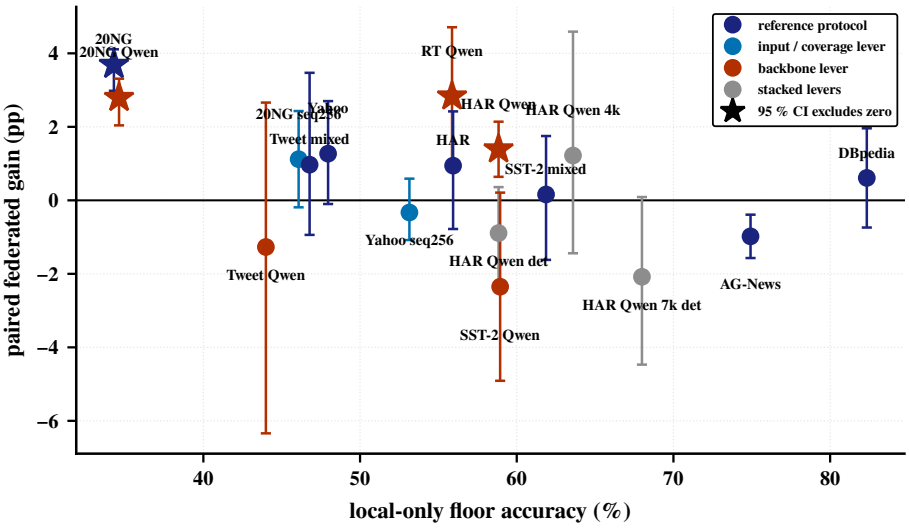


Figure 11: The measured operating envelope. Each point is one configuration with at least five paired seeds: mean local-only floor against mean paired federated gain, with bootstrap 95% intervals; stars mark configurations whose interval excludes zero. Gains concentrate where partial views leave local training short of the task ceiling, and the backbone lever moves configurations up and to the right where the other levers trade one axis for the other.

validation of the same configuration. **(ii) classifier-head federation** by the same resize family: configuration means of −0.55 to −2.53 pp across three settings spanning 11 paired runs, no setting positive. **(iii) logit-consensus distillation** (FedMD/DS-FL-style): with domain-matched public probes, −2.11 pp (CI [−3.09, −1.22]) under featureskew; null under label-disjoint. **(iv) server-side distillation** (FedDF-style trained pivot): mean −1.82 pp, worse than naive averaging. **(v) class-aware consensus weighting**: a pre-registered gate found weighting effects statistically indistinguishable from zero on both splits, closing the distillation family. Under feature-view heterogeneity every client is competent on every class, so logit consensus carries little any client lacks; and resized weight deltas of independently pretrained backbones do not share a coordinate frame. What heterogeneous clients *can* share exactly is geometry relative to public reference points, the observation that led to RelFed.

A.5 The operating envelope across configuration levers

After the main campaign we ran a pre-registered configuration search of roughly one hundred and twenty further paired cells to ask which levers move absolute accuracy and the paired gain together, and which trade one for the other (Figure 11, Table 10). Three levers were tested on top of the reference protocol. Quadrupling the input budget from 64 to 256 tokens with one thousand training examples per site raises the 20 Newsgroups floor from 34.3 to 46.1% while the gain compresses from +3.69 to +1.12 pp, and the same configuration turns Yahoo slightly negative. Widening views from four to six of ten feature groups behaves the same way in single-seed probes, which we report as direction only. Upgrading the fleet to Qwen-class backbones is the exception: the gain survives at +2.79 pp on 20 Newsgroups, at +1.39 pp with interval [+0.64, +2.14] at 60.2% accuracy on the sensor task, and at +2.83 pp with interval [+0.95, +4.71] at 58.7% on movie-review sentiment, while the same fleet is null or negative on two short-text tasks whose views retain little complementary evidence. The reference protocol tells the same high-floor story on its easy tasks: AG-News is significantly negative at −0.98 with interval [−1.57, −0.39] and DBpedia is null at +0.61 against an 82.3% floor, so past the local ceiling the exchange has nothing left to carry. The envelope also delimits what long resisted every lever: across the input and coverage dials no setting combined seventy percent absolute accuracy with a preserved gain, because a view that withholds most of the evidence caps local training exactly where partner sites still hold something worth teaching. Stacking data on the backbone lever confirms the envelope rather than escaping it: at 800 training windows per site the sensor floors rise to 63.6% with the gain thinned to +1.22 pp and interval [−1.44, +4.59], and at 1,400 windows floors reach 68.0% with the gain gone and one seed in five failing its run twice under

Table 10: All post-freeze frontier configurations with five or more paired seeds, sorted by gain. Every row regenerates from raw run artifacts; single-seed probes are excluded.

Configuration	Floor	Ours	Δ (pp)	95% CI	Sign
20NG	34.3	38.0	+3.69	[+2.98, +4.11]	5/5
RT Qwen	55.9	58.7	+2.83	[+0.95, +4.71]	5/5
20NG Qwen	34.6	37.4	+2.79	[+2.04, +3.31]	5/5
HAR Qwen	58.8	60.2	+1.39	[+0.64, +2.14]	5/5
Yahoo	48.0	49.2	+1.27	[−0.10, +2.70]	3/5
HAR Qwen 4k	63.6	64.8	+1.22	[−1.44, +4.59]	3/5
20NG seq256	46.1	47.2	+1.12	[−0.19, +2.43]	4/5
Tweet mixed	46.8	47.8	+0.97	[−0.94, +3.47]	3/5
HAR	55.9	56.9	+0.94	[−0.78, +2.42]	6/8
DBpedia	82.3	83.0	+0.61	[−0.74, +1.96]	3/5
SST-2 mixed	61.9	62.0	+0.16	[−1.62, +1.75]	3/5
Yahoo seq256	53.1	52.8	-0.33	[−1.08, +0.59]	2/5
HAR Qwen det	58.8	58.0	-0.89	[−2.27, +0.36]	2/5
AG-News	74.9	73.9	-0.98	[−1.57, −0.39]	0/5
Tweet Qwen	44.0	42.7	-1.27	[−6.34, +2.66]	2/5
HAR Qwen 7k det	68.0	65.9	-2.08	[−4.47, +0.09]	1/5
SST-2 Qwen	58.9	56.6	-2.35	[−4.91, +0.21]	1/5

a mechanical collapse-exclusion rule recorded with the artifacts. Detaching the relative loss from the adapters at the same operating point inverts the gain to −0.89 with interval $[-2.27, +0.36]$, a control consistent with the joint objective of equation 5 being the channel that carries the transfer. For calibration against full-view federation, RoBERTa-base fine-tuned federatively on 20 Newsgroups with complete documents reports 62 to 65% (Wu et al., 2026), and prototype federation across heterogeneous clients spans 16 to 60% on CIFAR-100 depending on skew (Tan et al., 2022).