EQUITABLE UNIFIED LEARNING FOR LLM APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning large language models (LLMs) is resource-intensive: it requires a large centralized corpus and substantial compute and memory. This raises three problems. First, such fine-tuning is out of reach for many institutions with limited resources. Second, architecture heterogeneity: institutions choose different architectures suited to their tasks, so models cannot be averaged. Third, data heterogeneity: each institution holds private data that is not independently and identically distributed (non-IID) and must remain local. To address these problems, we propose Unified Learning with Language Models (UniLearnLM), an equitable approach to LLM approximation that fine-tunes small language models (SLMs) through GNN-based unified federated learning. UniLearnLM represents each client's architecture as a typed graph and federates one shared graph neural network (GNN) whose output modulates each frozen backbone through a private adapter, so neither private data nor backbone weights leave a client, client architectures need not match, and under 0.07% of any client's parameters is communicated. Experiments on five language tasks (named entity recognition, question answering, topic classification and two summarization tasks) with four SLM families, and on educational question and answer generation, including a live federation in Tunisia, Germany, Colombia and Peru, show that UniLearnLM is on par with the strongest published heterogeneous federation on single-task federations and yields deeper questions than text-consensus distillation on two testbeds, with depth comparable to local training. Code will be released upon acceptance.

1 INTRODUCTION

Large language models (LLMs) set the quality bar for text generation and understanding, but fine-tuning them is resource-intensive: it requires a large centralized corpus and substantial compute and memory. Many institutions that hold text therefore cannot use them. Data protection rules or per-token cost rule out sending the text to a hosted model, and hardware rules out running a model of that size locally. These institutions instead fine-tune a small language model (SLM) on their own data (Lewis et al., 2020; Raffel et al., 2020; Beltagy et al., 2020; Chung et al., 2024; Abdin et al., 2024; Gemma Team et al., 2024; Nguyen et al., 2025). This raises three problems.

First, limited resources. An SLM trained in isolation is bounded by the data of one institution and by the model that institution can afford, so institutions with less data and smaller models obtain weaker approximations of an LLM, and joint learning across institutions that keep their data and their own SLM is required. Federated learning (FL) (McMahan et al., 2017; Li et al., 2020a; Kairouz et al., 2021) trains a shared model without moving data, but fairness in FL is stated as a property of the aggregate and pursued by reweighting the objective (Li et al., 2020c), and participants can gain nothing from a federation while its mean improves (Yu et al., 2020). *Second, architecture heterogeneity.* Institutions choose different architectures suited to their tasks, so their models cannot be averaged. Width- and depth-elastic federations train sub-models of one backbone (Diao et al., 2021; Alam et al., 2022; Kim et al., 2023a), and LoRA federations aggregate adapters of client-specific rank on one backbone (Cho et al., 2024; Bai et al., 2024; Wang et al., 2024b), so both still require one backbone, which surveys identify as the limiting assumption of this line (Ye et al., 2024). Hypernetwork-based federations remove the shape constraint by sharing a generator instead of a tensor: personalized federated hypernetworks (pFedHN) map a learned client embedding to that

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.

client's weights (Shamsian et al., 2021), and graph hypernetworks condition generated parameters on a model's computation graph (Knyazev et al., 2021; Litany et al., 2022). However, these generate parameters that replace a pretrained backbone, so the pretrained weights that make an SLM useful are discarded. *Third, data heterogeneity.* Each institution holds private data that is not independently and identically distributed (non-IID) and must remain local. Distillation federations exchange outputs on a public dataset instead of parameters (Li & Wang, 2019; Lin et al., 2020; Wu et al., 2022), federated mutual knowledge transfer (FedMKT) aligns output distributions across tokenizers (Fan et al., 2025), and federated heterogeneous neural networks (FedHeNN) align client representations on a shared anchor set (Makhija et al., 2022). These require a public corpus close to the private data or an alignment stage between vocabularies, so a shared corpus is reintroduced. Therefore, a federation of pretrained SLMs must modulate each backbone rather than replace it, without a shared corpus (Fig. 1), and it must be evaluated client by client against local training.

To address these problems, we propose Unified Learning with Language Models (UniLearnLM), an equitable approach to LLM approximation that fine-tunes SLMs through GNN-based unified federated learning and shares a model of architecture rather than weights or data. First, each client converts its SLM into a typed architecture graph whose nodes are the modules of the network. Second, a single graph neural network (GNN), the unification encoder, reads every client's graph and emits one embedding per block; it is the only component averaged across clients, so heterogeneous parameter shapes never meet. Third, a private adapter maps each block embedding to a feature-wise scale and shift (Perez et al., 2018) on the output of the client's frozen, LoRA-adapted block (Hu et al., 2022), so the pretrained weights are modulated rather than replaced. Moreover, to keep the first rounds identical to local training, the adapter's gate is initialized at zero (Zhang et al., 2023). More importantly, by comparing each client with its own locally trained model, we measure whether the federation leaves any participant worse off. UniLearnLM extends unified learning over model graphs (Pala & Rekik, 2025), whose neuron-level graphs do not fit in memory for language models, to graphs with one node per sub-module of a pretrained SLM.

Four hypotheses anchor our evaluation:

Hypothesis 1 (equity). A resource-constrained institution generates more pedagogically deep and contextually grounded educational QAs by participating in UniLearnLM than by training in isolation, with gains that track curriculum difficulty rather than merely model capacity and without ever accessing other institutions' private curricula.

Hypothesis 2 (LLM approximation). A unified GNN-orchestrated federation of heterogeneous SLMs produces educational QA pairs whose quality is comparable to that of a single LLM.

Hypothesis 3 (sufficiency). Sharing only a model of client architecture matches federations that exchange representations, outputs or text, preserves the question depth that text-consensus distillation loses, and costs no accuracy relative to local training.

Hypothesis 4 (mechanism). The shared encoder carries knowledge through the architecture graph, so it transfers between clients with different tasks, and a client index or an untyped graph does worse.

Our main contributions are as follows:

1. *On a methodological level.* We propose a federation of heterogeneous pretrained SLMs in which a shared GNN of 87,744 parameters reads a typed architecture graph of each client and conditions a private FiLM adapter on a frozen, LoRA-adapted backbone, with no shared vocabulary, label space, public corpus or private data.
2. *On an educational level.* We apply the framework to question and answer (QA) generation from teaching material, measure question depth on the revised Bloom taxonomy (Anderson et al., 2001; Krathwohl, 2002) for every institution separately, and run the federation live across institutions that hold private curricula.
3. *On a generic level.* We evaluate the framework on five standard natural language processing tasks and a four-task federation with four SLM families and non-IID partitions, against local training, text-consensus and published heterogeneous federations.

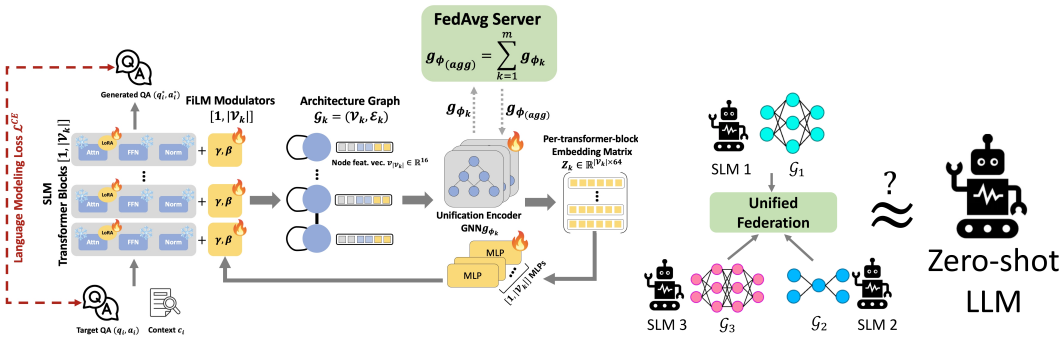


Figure 1: Overview of UniLearnLM. **Left:** the federated framework. First, each client converts its SLM into a typed architecture graph $\mathcal{G}_k$ whose nodes carry structural features and running statistics of the block and LoRA weights. Second, the shared unification encoder $g(\cdot; \phi)$ emits per-block embeddings $\mathbf{Z}_k$, and only ϕ is sent to the server, averaged and broadcast. Third, a private FiLM adapter maps each embedding to a scale γ and shift β on the frozen, LoRA-adapted block. Snowflake = frozen; flame = trainable. **Right:** the comparison between the federation of SLMs and one LLM prompted zero-shot.

2 Related Work

Heterogeneous federated learning. Standard FL assumes one model for every client (McMahan et al., 2017; Li et al., 2020b; 2021). Width- and depth-elastic federations (Diao et al., 2021; Liu et al., 2022b; Alam et al., 2022; Kim et al., 2023a) and LoRA federations with client-specific ranks (Cho et al., 2024; Bai et al., 2024; Wang et al., 2024b), learned parameter-efficient configurations (Lee et al., 2026) or LoRA experts (Duan et al., 2026) relax this assumption within one backbone. Generator-based federations share a network that produces client parameters or representations: pFedHN (Shamsian et al., 2021), graph hypernetworks (Knyazev et al., 2021; Litany et al., 2022; Lim et al., 2024), FedHeNN (Makhija et al., 2022), FedProto (Tan et al., 2022), and C2A, which generates adapters for one shared pretrained language model (Kim et al., 2023b). Unified learning trains one GNN over neuron-level graphs of several models (Pala & Rekik, 2025); such graphs do not fit in memory for language models, so UniLearnLM represents each model at the sub-module level (one node per sub-module, Section 3), and its shared encoder emits modulation for frozen pretrained SLMs of different families rather than parameters.

Distillation on public data. Distillation federations exchange logits, output distributions or generated samples on a public dataset (Li & Wang, 2019; Lin et al., 2020; He et al., 2020; Cho et al., 2022; Zhu et al., 2021; Wu et al., 2022). They assume a public corpus close to the clients' data or a shared vocabulary, and FedMKT removes the second assumption with a token alignment table (Fan et al., 2025). We run two FedMD-style text-consensus federations and FedMKT as baselines; both see text from other clients that UniLearnLM never uses.

Equity and per-client evaluation. Fairness in FL is usually a property of the aggregate, such as uniform accuracy or a worst-case guarantee (Li et al., 2020c). We instead report whether each client is better off inside the federation than under local training. Bloom-aware scoring of generated questions evaluates single models (Scaria et al., 2024; Moore et al., 2024); we report its change under federation per client. Appendix B extends this section.

3 The UniLearnLM Framework

In this section, we detail the UniLearnLM framework (Fig. 1). In the first stage, each client builds an architecture graph of its SLM. In the second stage, a shared GNN encodes every graph into per-block embeddings. In the third stage, a private adapter converts each embedding into a modulation of the client's frozen backbone, and only the GNN is federated. Moreover, to isolate what a shared public corpus adds, we define an ablation that trains every client on the other clients' generations.

Problem statement. Consider m clients $k \in [m]$. Client k holds a private dataset $\mathcal{X}_k$ of input and target pairs (x, y) drawn from a client-specific, non-IID distribution $\mathbb{P}_k$, and a pre-trained model $f_k(\cdot; \mathbf{W}^{(k)}_{\text{base}})$ with its own tokenizer and hidden size $d^{(k)}$. Each client trains a private LoRA adapter $\mathbf{L}_k = (\mathbf{A}_k, \mathbf{B}_k)$ of rank r on its attention projections, so that $\mathbf{W}^{(k)}_{\text{eff}} = \mathbf{W}^{(k)}_{\text{base}} + (\alpha/r)\,\mathbf{B}_k\mathbf{A}_k$ with $\mathbf{W}^{(k)}_{\text{base}}$ frozen (Hu et al., 2022). Neither $\mathbf{W}^{(k)}_{\text{base}}$ nor $\mathbf{L}_k$ can be averaged across clients, because their shapes differ. The goal is to learn shared encoder parameters $\boldsymbol{\phi}^\star = \arg\min_{\boldsymbol{\phi}} \frac{1}{m}\sum_{k=1}^{m} F_k(\boldsymbol{\phi}; \mathbf{L}_k, \boldsymbol{\psi}_k)$, where F_k is client k's loss and $\boldsymbol{\psi}_k$ its private adapter parameters; $\mathbf{L}_k$ and $\boldsymbol{\psi}_k$ never leave the client (Appendix K).

Architecture graph. Each client builds, once and locally, an architecture graph $\mathcal{G}_k = (\mathcal{V}_k, \mathcal{E}_k)$ with one node per sub-module, which gives 228 to 499 nodes per client. Edges connect consecutive modules in traversal order in both directions, with self-loops, so the graph is a typed module sequence rather than a computation graph. Each node carries a 16-dimensional feature vector of layer-type indicators, size, normalized depth, role flags, and running statistics of the node's own weights, refreshed at every mini-batch (Table 12). The LoRA factors are nodes of their own, so these statistics follow the adapters as they train. The schema is shared by every model family (Appendix O).

Unification encoder. The federated component is a GNN $g(\cdot; \boldsymbol{\phi})$, a three-layer graph attention network (Veličković et al., 2018) with dimensions $16 \to 256 \to 256 \to 64$ (4, 4 and 1 heads, ELU, dropout 0.1). It outputs $\mathbf{Z}_k = g(\mathcal{G}_k; \boldsymbol{\phi}) \in \mathbb{R}^{|\mathcal{V}_k| \times 64}$, whose row $\mathbf{z}_\ell$ conditions block ℓ. The encoder has $|\boldsymbol{\phi}| = 87{,}744$ parameters, which is 0.063% of the smallest client model (BART-base, 140M), so the communication cost is independent of backbone size.

Client-local FiLM modulation. To inject the shared representation without replacing pretrained weights, each client holds a private feature-wise linear modulation (FiLM) adapter (Perez et al., 2018) $h(\cdot; \boldsymbol{\psi}_k)$, a one-hidden-layer ReLU network of width 128 that maps $\mathbf{z}_\ell$ to $(\boldsymbol{\gamma}_\ell, \boldsymbol{\beta}_\ell) \in \mathbb{R}^{d^{(k)}} \times \mathbb{R}^{d^{(k)}}$. The adapter acts as a residual on the output $\mathbf{H}_\ell$ of block ℓ:

$$\mathbf{H}_\ell \leftarrow \mathbf{H}_\ell + \alpha_k\big(\boldsymbol{\gamma}_\ell \odot \mathbf{H}_\ell + \boldsymbol{\beta}_\ell\big), \tag{1}$$

where α_k is a learnable per-client gate. The adapter is the only component whose width depends on $d^{(k)}$, which is why it stays private. Furthermore, to make the first update start from the LoRA-adapted SLM, we initialize α_k=0; the gradient with respect to α_k is non-zero from the first step, as for the zero convolutions of ControlNet (Zhang et al., 2023). For the T5 family and the Qwen client, the block class is not a node type in our implementation, so every block of those clients receives the modulation of the mean-pooled graph embedding; Appendix AF.1 reports a run with per-block nodes.

Unified federation. Each client minimizes the conditional language-modeling loss $\mathcal{L}_k = -\sum_{(x,y)\in\mathcal{X}_k} \log p_k(y \mid x)$ with $\mathbf{W}^{(k)}_{\text{base}}$ frozen, and gradients reach $\boldsymbol{\phi}$, $\mathbf{L}_k$ and $\boldsymbol{\psi}_k$ through Eq. (1). In round t, every client runs E local epochs and uploads its copy $\boldsymbol{\phi}^{[t]}_{(k)}$, and the server returns the sample-weighted average $\boldsymbol{\phi}^{[t]} = \sum_k (n_k / \sum_j n_j)\,\boldsymbol{\phi}^{[t]}_{(k)}$ with $n_k = |\mathcal{X}_k|$ (McMahan et al., 2017). LoRA and FiLM parameters are never aggregated, so differences in hidden size never enter the aggregation (Appendix K).

Public-pool variant (UniLearnLM+). To measure what a shared text channel adds, UniLearnLM+, the variant that additionally trains on the public pool, trains every client also on the other clients' generations for a public pool of about 180 passages per fold, drawn from the clients' training chunks and disjoint from every test set. After round $t-1$, every client generates one output per pool passage, and in round t each client adds these generations to its loss with weight $\beta = 0.5/K$. The pool is therefore not an external corpus, and UniLearnLM+ and the distillation baselines see text from other clients that UniLearnLM never sees (Appendix J).

4 Experimental Setup

We evaluate UniLearnLM in two settings. The generic benchmark tests the framework on standard tasks against baselines that other work can reproduce. The application, QA generation for education,

tests it on private curricula, where the data cannot be moved and the per-client comparison with local training is the question of interest.

4.1 Generic Benchmark

Tasks, clients and partitions. The benchmark covers named entity recognition on CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), extractive QA on SQuAD v1.1 (Rajpurkar et al., 2016), topic classification on AG News (Zhang et al., 2015), and summarization on XSum (Narayan et al., 2018) and SAMSum (Gliwa et al., 2019), plus a four-task federation in which each client holds a different one of the first four tasks. Every task is cast as generation with a fixed prompt, with the loss on target tokens and greedy decoding. Each setting has four clients running Flan-T5-base (Chung et al., 2024), BART-base (Lewis et al., 2020), Qwen2.5-0.5B (Qwen Team, 2024) and SmolLM2-360M (Ben Allal et al., 2025), so no two clients share a tokenizer or a hidden size, and they hold 40%, 30%, 20% and 10% of the training rows. Partitions draw a Dirichlet(0.3) distribution over labels or topics per client.

Evaluation and parameter setting. Metrics are entity F1 (CoNLL-2003), answer F1 (SQuAD), accuracy (AG News), ROUGE-L (Lin, 2004) (XSum, SAMSum), and the mean of the clients' task metrics (four-task federation). Every method trains for 20 rounds of 2 local epochs on the same partitions and prompts, and every method selects its model on validation data only (Appendix D). A cell is the mean over the four clients on the global test split. We report mean $\pm$ standard deviation across seeds or folds where a cell has several, and mark with † a comparator that UniLearnLM beats in a paired two-sided t-test across those seeds or folds ($p < 0.05$); single-seed cells are reported without variability.

Comparison methods. Indiv trains each client on its own data and is the reference each client must beat. **Centralized** trains one model on the pooled data, which the setting forbids, and serves as a privileged reference. TC-Best and TC-Ens are FedMD-style text-consensus federations (Li & Wang, 2019), adapted to heterogeneous SLMs, that distil each client towards the best or the ensemble of the other clients' generations on the public pool (Appendix S). **pFedHN** (Shamsian et al., 2021), **FedHeNN** (Makhija et al., 2022) and **FedMKT** (Fan et al., 2025) are the published heterogeneous federations closest to our setting, run under the same schedule and data (Appendix U). UniLearnLM+ is the public-pool ablation of Section 3.

4.2 Application: QA Generation for Education

Testbeds. An institution is a client, an input is a passage of teaching material of 150 to 400 words, and a target is a QA pair. The public testbed has three institutions built from open sources: $\mathcal{I}_0^{\text{pub}}$ = MIT 18.657 (Massachusetts Institute of Technology, 2015), $\mathcal{I}_1^{\text{pub}}$ = Stanford CS229 (Stanford University, 2024) and $\mathcal{I}_2^{\text{pub}}$ = arXiv machine learning papers, with 110 to 981 passages each and five GPT-4o-annotated QA pairs per passage (Appendix C.1). We hold out 15% of each institution as test data, run three-fold cross-validation on the rest, and pool the holdouts into a global test set of 916 rows over 179 contexts. The private testbed involves nine institutions in eight countries on three continents (Table 4; locations in Appendix E). Five institutions prepared private machine learning curricula and committed to running a client, and four dataset-only institutions contributed a private global test set of 200 contexts and 996 QA pairs on which no institution trains (Appendix C.2). In four live runs through a federation server, four of the five committed institutions connected. Because participation in the live runs was uneven, we re-ran the five curricula as a controlled federation with full participation and three-fold cross-validation (Appendix AE). No raw data, QA pairs or model weights leave an institution.

Evaluation and parameter setting. The client models are five encoder-decoder SLMs of 140M to 392M parameters (Table 5). Local training runs for up to 100 epochs, and UniLearnLM runs 33 rounds of 3 local epochs (Appendix I); every method is compared using its validation-selected model. We score eleven quality metrics on each generated triple: ROUGE-L, BLEU-4, BERTScore-F1, a Bloom-level BERT classifier B_{cls}, a Bloom LLM rater B_{LLM}, RAGAS answer relevancy (AnsR) and faithfulness (Faith), QAFactEval (QAFE), RQUGE, round-trip consistency (RTC), and a four-axis LLM judge (LLM-J; Appendix F). Both Bloom estimators are mapped to $\text{EVS}(q) = (\hat{b}(q) - 1)/5 \in$

Table 1: Generic benchmark: task metric on the global test split, mean over the four clients, Dirichlet(0.3) partitions, seed 42, validation-selected model for every method, 20 rounds of 2 local epochs for every federated method. pFedHN uses the inner learning rate 10^{-3} selected on CoNLL validation data from $\{5{\cdot}10^{-4}, 10^{-3}, 2{\cdot}10^{-3}, 5{\cdot}10^{-3}\}$ and carried to the other tasks. Bold = best non-privileged method per column. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019), pFedHN (Shamsian et al., 2021), FedHeNN (Makhija et al., 2022), FedMKT (Fan et al., 2025), CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), SQuAD (Rajpurkar et al., 2016), XSum (Narayan et al., 2018), SAMSum (Gliwa et al., 2019).

Method	CoNLL (Tjong Kim Sang & De Meulder, 2003)	SQuAD (Rajpurkar et al., 2016)	AG News (Zhang et al., 2015)	XSum (Narayan et al., 2018)	SAMSum (Gliwa et al., 2019)	Four task
Centralized	*0.8260*	*0.8031*	*0.9242*	*0.2705*	*0.4109*	*0.6372*
indiv	0.7237	0.7758	0.8374	0.2529	0.3909	0.3396
pFedHN (Shamsian et al., 2021)	0.5772	0.7785	0.7307	0.2321	0.3842	0.3133
FedHeNN (Makhija et al., 2022)	0.7680	0.7710	**0.8692**	0.2530	**0.3921**	0.3054
FedMKT (Fan et al., 2025)	0.7472	0.7652	0.7805	0.2296	0.3796	0.3036
TC-Best (Li & Wang, 2019)	0.7631	0.7738	0.8348	0.2457	0.3886	0.4065
TC-Ens (Li & Wang, 2019)	**0.7757**	0.7650	0.8275	0.2545	0.3905	0.3686
UniLearnLM	0.7652	0.7631	**0.8692**	**0.2546**	0.3904	0.2985
UniLearnLM+	0.7715	**0.7826**	0.7254	0.2412	0.3905	**0.4239**

$[0, 1]$, so one taxonomy level corresponds to 0.2 (Appendix M). Uncertainty on depth is a paired, context-clustered bootstrap over the 179 test contexts (5,000 resamples, 95% intervals).

Comparison methods. On the public testbed, the application uses the methods of Section 4.1, together with a client-index control that replaces the graph encoder by a learned client embedding in the same FiLM adapter (Appendix U). On the private testbed, the live and the controlled federations compare Indiv, TC-Ens, UniLearnLM and UniLearnLM+ (Appendix C.2). For the comparison with LLMs, the federation of UniLearnLM-trained SLMs is compared with six LLMs prompted zero-shot (Appendix AC).

5 Results

5.1 Generic Benchmark

Our experiments on the generic benchmark (Table 1) test *Hypothesis 3* and *Hypothesis 4*. The evidence validates *Hypothesis 3* on single-task federations and does not support *Hypothesis 4*.

Comparison with published heterogeneous federations. *Hypothesis 3* predicts that sharing the encoder alone reaches the strongest published heterogeneous federation, FedHeNN, which is above the other two on four of five tasks. It is validated for parity: in the tabled run UniLearnLM is above FedHeNN on one task, tied on one and below on three, and over three seeds a paired test finds no significant difference on either CoNLL-2003 or AG News (Table 20); UniLearnLM is above pFedHN and above FedMKT on four of five tasks each. The discovery is that parity is reached without the public pool on which FedHeNN and FedMKT exchange representations or outputs.

Comparison with local training. *Hypothesis 3* also predicts that federating the encoder costs no accuracy relative to training alone. Over three seeds (Table 17, Appendix T) this is validated on four of five tasks, where the mean of UniLearnLM is higher, with a significant gain on CoNLL (paired t-test over seeds, $p = 0.016$); SAMSum is the one task where its mean is lower.

Four-task federation. The prediction of *Hypothesis 4* that the shared encoder transfers between clients holding different tasks is not supported. UniLearnLM ranks last of the eight methods and falls behind local training at three of the four clients (Tables 1 and 17), while the three methods that exchange text on the public pool (UniLearnLM+, TC-Best, TC-Ens) are the only ones above local training. A shared encoder of architecture carries no task content, so it offers no channel through which a client can learn another client's task; the public pool does.

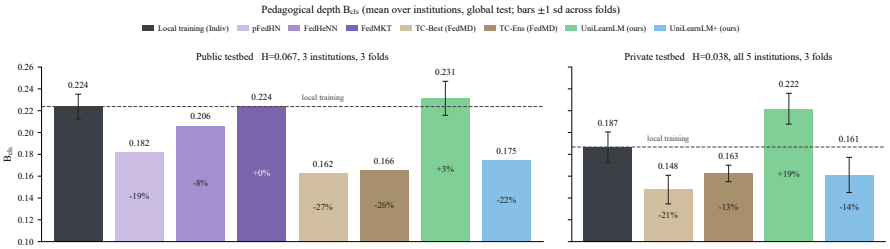


Figure 2: Question depth B_{cls} on the global test set of each testbed (mean over institutions) for local training, the text-consensus federations (Li & Wang, 2019), the published heterogeneous federations (Shamsian et al., 2021; Makhija et al., 2022; Fan et al., 2025), UniLearnLM and UniLearnLM+. Validation-selected model for every method; public testbed: three folds, one seed (Table 2); private testbed: full-participation federation, all five institutions, 3 folds (Table 45). Error bars = ±1 standard deviation across folds (none for UniLearnLM+ on the public testbed, whose per-fold values are not tabled). Dashed line = local training; percentages = change relative to it; H = heterogeneity index of the testbed; the published heterogeneous federations were evaluated on the public testbed only. UniLearnLM+ additionally trains on other clients' generations for the public pool and, like the text-consensus federations, gains surface overlap at the cost of depth (Table 2).

Three-seed ablations. *Hypothesis 4* also predicts that the benefit comes from the architecture graph and its node types, but neither a learned client index nor an untyped graph differs significantly from UniLearnLM in any of the four paired tests over three seeds (Table 20 and Figure 6, Appendix T). *Hypothesis 4* is therefore not supported on either count, and the parity with FedHeNN cannot be attributed to the graph structure.

Partition controls. As a check on *Hypothesis 3*, under Dirichlet(0.5) and IID partitions over three seeds, UniLearnLM is above local training in three of the six task and partition settings and below in three, and none of the six differences is significant (Table 19, Figure 5).

Table 2: Public testbed: UniLearnLM against local training (**Indiv**), the text-consensus federations TC-Best and TC-Ens (Li & Wang, 2019), the public-pool ablation UniLearnLM+, the published heterogeneous federations and the client-index control, on the global test set (916 rows over 179 contexts). Mean over the three institutions and three folds, one seed, validation-selected model for every method; per-institution values in Tables 35 and 22. Bold = best per metric; shading by rank among Indiv, TC-Best, TC-Ens, UniLearnLM and UniLearnLM+; *Centralized* is a privileged reference and is not ranked. Faith§ is the sentence-level variant; − = not computed. Metric abbreviations in Section 4.2. † marks a comparator that UniLearnLM beats in a paired two-sided t-test across the three folds ($p < 0.05$), tested for R-L, B-4, BERT and B_{cls} (per-fold values in Appendix V) and for LLM-J.

Inst.	Method	Lightweight				LLM-based / reference-free						
		R-L	B-4	BERT	B_{cls}	B_{LLM}	AnsR	Faith	QAFE	RQU	RTC	LLM-J
Mean across $\mathcal{I}^{pub}_{0,1,2}$	indiv	0.4316	0.1612†	0.8602	0.2238	0.2754	**0.5190**	0.2091	**0.1832**	4.0741	0.2049	0.6915†
	pFedHN (Shamsian et al., 2021)	0.4313	0.1624†	0.8614	0.1816†	–	–	0.1798§	0.1884	4.0576	0.2011	–
	FedHeNN (Makhija et al., 2022)	0.4372	0.1700	0.8636	0.2062†	–	–	0.1933§	0.1815	4.0666	0.2049	–
	FedMKT (Fan et al., 2025)	0.3997†	0.1494†	0.8524†	0.2243	–	–	0.1791§	0.1762	4.0496	0.1949	–
	TC-Best (Li & Wang, 2019)	0.4387	0.1714	0.8632	0.1623†	–	–	0.1966§	0.1724	4.0647	0.2045	–
	TC-Ens (Li & Wang, 2019)	0.4364	0.1717	0.8628	0.1655†	–	–	0.1893§	0.1729	4.0652	0.2023	–
	UniLearnLM	0.4385	0.1715	0.8640	**0.2313**	**0.2784**	0.5161	**0.2092**	0.1808	**4.0826**	**0.2130**	**0.6991**
	UniLearnLM+	**0.4450**	**0.1760**	**0.8653**	0.1746	0.2760	0.4665	0.2091	0.1784	4.0744	0.2069	0.6948
	Client index ctrl.	0.4402	0.1742	0.8645	0.2021	–	–	0.1905§	0.1724	4.0647	0.2057	–
	Centralized	0.4630	0.1923	0.8702	0.2122	0.2881	0.4775	0.2041	0.1773	4.0739	0.2111	0.7109

5.2 Application: Question Depth and Equity

Our experiments on QA generation test *Hypothesis 3* on question depth and *Hypothesis 1*. The evidence validates *Hypothesis 3* against text-consensus distillation on both testbeds and supports *Hypothesis 1* only in part.

Comparison with text-consensus distillation. *Hypothesis 3* predicts that training on consensus text from other clients lowers depth and that UniLearnLM, which moves no text, avoids the loss. It is validated. On the public testbed (Table 2, Fig. 2), UniLearnLM exceeds TC-Ens on B_{cls} in all three runs of Table 36, with every bootstrap interval excluding zero, and in the tabled run a paired t-test across folds gives $p < 0.05$ against four of the five federated comparators (TC-Ens, TC-Best, pFedHN, FedHeNN). The gaps correspond to 0.19 to 0.33 of a Bloom level. On the private testbed with full participation, UniLearnLM exceeds both text-consensus federations on every depth measure available, and on both testbeds both fall below local training on B_{cls} (Table 45). UniLearnLM+, which additionally trains on the public pool, ranks first on ROUGE-L, BLEU-4 and BERTScore on the public testbed but seventh of nine on depth, above only the two text-consensus federations (Table 2). The discovery is that training on other clients' text transfers surface form rather than depth (Gudibande et al., 2024), and a federation that moves no text avoids that loss.

Comparison with local training. *Hypothesis 1*, that every participant gains over training alone, is only partially validated on depth, and its clause on curriculum difficulty versus model capacity cannot be tested here because curriculum and backbone are confounded (Appendix AF). In the tabled run it is above local training on B_{cls} with a bootstrap interval that excludes zero, above it on 9 of 11 metrics, and significantly better on BLEU-4 and on the LLM judge (paired t-tests across folds, $p < 0.05$; $p = 0.022$ for the judge; Table 2), but the rerun and the second seed place it below local training on B_{cls}, with intervals that also exclude zero (Table 36). We therefore report depth relative to local training as comparable, not as a gain. Per institution, the change in B_{cls} is largest at the institution with the largest curriculum (Table 33), and removing that institution leaves the other two below local training (Table 41).

Comparison with published heterogeneous federations. On depth, *Hypothesis 3* is partially validated against these federations. In the tabled run it is significantly deeper than pFedHN and FedHeNN and level with FedMKT; in the rerun and the second seed FedMKT is significantly deeper and FedHeNN is level (Table 36). FedMKT, the one federation whose depth matches or exceeds that of UniLearnLM across runs, has the lowest ROUGE-L, BLEU-4 and BERTScore of all methods in Table 2, so its depth comes at the cost of surface quality.

Live deployment. The live runs test *Hypothesis 3* in a real deployment, with uneven participation, dropped connections and hardware from a workstation to a laptop, and validate it: the federation completed its rounds, and at the institutions where both methods ran, UniLearnLM exceeds the text-consensus federations on B_{LLM} in all 7 comparisons and on B_{cls} in 5 of 7, and exceeds local training on both depth measures at 3 of the 4 institutions it served (Table 30). On the Mean row it leads on the same three metrics as under full participation, B_{cls}, B_{LLM} and answer relevancy. The runs cover one fold with participant sets that differ between methods, so we compare institution by institution and read the counts as the direction of the effect, not its size (Appendix Y).

5.3 Comparison with Large Language Models

Hypothesis 2 is partially validated: it holds for grounding and not for lexical overlap or the holistic judge. Against six LLMs (Table 34), the federation of UniLearnLM-trained SLMs has the highest RAGAS faithfulness and QAFactEval of the seven systems. Both metrics rise as answers get shorter and more extractive (Ladhak et al., 2022), and the SLM answers are the shortest of the seven systems. When every system is held to the same word budget (Table 23), the SLMs stay first on both metrics at all four budgets, but at the tightest budget most of the faithfulness lead and part of the QAFactEval lead disappear. The discovery is that answer length explains much of the lead (Appendix Y).

6 Discussion

Across the generic benchmark and the two educational testbeds, the results validate *Hypothesis 3*, support *Hypothesis 1* and *Hypothesis 2* in part, and do not support *Hypothesis 4*. On single-task federations of four SLM families, UniLearnLM is level with the strongest published heterogeneous federation, with no significant difference over three seeds on either task tested (Tables 1 and 20), while exchanging only the graph encoder, and the graph ablations do not locate this parity in the graph structure.

The method pays off where institutions cannot move text and the objective rewards depth (*Hypothesis 3*). On both educational testbeds the text-consensus federations fall below local training on question depth, while UniLearnLM stays comparable to it: it has the highest B_{cls} of the five methods under full participation on the private testbed (Table 45) and exceeds TC-Ens in all three public-testbed runs (Table 36).

The federation also transfers knowledge between institutions without moving text, the part of *Hypothesis 1* that the private testbed supports. Under full-participation every receiver and source pair is defined, and on the Knowledge Transfer Index (Appendix R), the two unified variants attain the two highest means and UniLearnLM has the most pairs with positive transfer, 14 of 20 against 12 of 20 for TC-Ens and for UniLearnLM+ (Table 46). Nigeria, the most resource-constrained site, is the source in two of the four largest positive cells for UniLearnLM, so a low-resource site is a source of transfer, not only a receiver.

What the method does not provide is a channel for task content (*Hypothesis 4*) or an advantage over local training that holds in every run (*Hypothesis 1*). The rerun and a second seed place UniLearnLM significantly below local training on B_{cls} (Table 36).

Limitations

Most benchmark and application cells use one seed, the paired tests use only three seeds or three folds, and the depth result relative to local training does not survive a second seed or a rerun, so the supported claim is the gap to text-consensus distillation. The results are sensitive to the model-selection criterion, which differs between local training and the federated methods (Table 18). The T5-family and Qwen clients receive graph-level rather than per-block modulation (Section 3), and B_{cls} is a public classifier not validated on this domain. Appendix AF gives each limitation in full.

7 Conclusion

We proposed UniLearnLM, which federates small language models of different families through one shared graph encoder of their architectures and keeps every backbone, adapter and dataset private. We found, validating *Hypothesis 3*, that sharing a model of architecture, rather than text, outputs or representations, reaches parity with the strongest published heterogeneous federation on single-task federations, and that it preserves question depth where text-consensus distillation loses it: on a public and a private testbed UniLearnLM generates deeper questions than distillation, while its depth stays comparable to that of local training. This matters for institutions that cannot move text and run different architectures on modest hardware: the framework needs no public corpus, common architecture or exchange of generated text, and it ran live at four institutions on three continents. It does not transfer between clients that hold different tasks (*Hypothesis 4* is not supported), its depth relative to local training varies across runs (*Hypothesis 1* holds in part), and its lead over zero-shot LLMs is confined to grounding (*Hypothesis 2*). The next steps are a shared channel for task content that does not move raw text, and multi-seed evaluation of depth at more institutions.

Ethics Statement

This work involves no human subjects and no collection of personal data. The private testbed is built from teaching materials contributed by nine collaborating institutions; in every case the material was supplied by a faculty member who explicitly authorized its research use, and the consent and authorization procedure is documented in Appendix C.2. No student data, no assessment records, and no personally identifying information were used at any point. During federated training, contributed curricula never leave the contributing institution: the protocol exchanges only the shared graph encoder (0.088M parameters), so raw teaching material, model weights and LoRA adapters remain local by construction (Section 3). Before training, however, every institution sends its curriculum chunks to a third-party endpoint (GPT-4o) to create its reference QA pairs, the same class of egress that motivates the paper, and we state this here rather than only in Appendix C.2. Institutional identities in the private testbed are pseudonymized as $\mathcal{I}_0^{\text{priv}}$–$\mathcal{I}_4^{\text{priv}}$ and $\mathcal{D}_0$–$\mathcal{D}_3$ throughout, and reported only at country granularity.

The intended use of UniLearnLM is to give resource-constrained institutions a self-hosted source of formative self assessment items. Two risks follow. First, automatically generated QA pairs can be factually wrong or pedagogically shallow. The paper measures both (faithfulness, factual consistency, Bloom depth) and the picture is not reassuring in absolute terms: Table 3 places answer correctness at 3.75 to 3.91 out of 5 and educational value at 2.9 to 3.1 out of 5, the latter being the rubric band for basic recall and surface level facts. We give qualitative failure classes in Appendix Q but we do not report a per class error *rate*, which is a gap: the per sample judge outputs would support one and we did not compute it. The system is therefore positioned as an instructor in the loop generator of practice items, not as an autonomous assessment authority, and we would not recommend deploying it to students without review.

Second, because federation moves quality between unequal participants, an equity framing can mask harm to individual clients. Our own results show this happening: the depth gain is negative at the smallest public institution (Table 33) and six of twenty directed transfer pairs are negative under full participation (Table 46). We therefore report per institution results rather than only pooled ones, and we report negative transfer where it occurs. A deployment of this kind needs a per client monitor rather than a corpus average, and the suite here provides the per-institution reporting such a monitor would need.

Third, a fairness gap sits inside our own cohort. Two contributing institutions list bilingual teaching materials in Table 4, while the entire pipeline, including the reference QA pairs and every judge prompt, is English only. Both bilingual sites are dataset-only contributors: their material evaluates the models while they receive none. We do not measure what that costs those sites, and we should.

The public testbed is built from openly licensed course material and arXiv preprints, used in a manner consistent with their licenses.

Reproducibility Statement

The public testbed is constructed entirely from openly available sources, and every step needed to rebuild it, that is source URLs, chunking parameters, the GPT-4o annotation prompt, fold assignment, and dataset checksums, is given in Appendix N (Table 11) and Appendix G. The method is specified in Section 3, with the architecture-graph node features enumerated in Appendix O, the local objective and the federated optimization written out in Appendix K, and the full client/server procedure given as pseudocode in Appendix L. All hyperparameters, including the per-family LoRA targets, GAT and FiLM dimensions, and the federated round schedule, are in Appendix I (Tables 9–10); the baselines are specified in Appendix S. Every metric is defined operationally, with the exact scoring models, decoding settings, and any deviation from the published formulation, in Appendix M, and the verbatim LLM-judge prompt is reproduced in Appendix H. All public-testbed numbers are means over the same three cross-validation folds with fixed seeds, and the per fold values behind the headline table are reported in Appendix V so that the cross fold spread can be checked directly. The private testbed cannot be released, since the curricula are the contributing institutions' private property; the private-testbed protocol and participation record are documented in Appendix C.2 so that the setup can be reproduced on other private corpora. Source code for the framework, the evaluation suite, and the federation server is provided as anonymized supplementary material, together with the MD5 manifest for the public testbed splits described in Appendix N; the same code and the public testbed will be released publicly on acceptance.

We flag two gaps in what this statement can offer. Every headline run reported here uses a single seed (one second seed run of unified learning is reported in Appendix U), so a reader can reproduce our numbers exactly but cannot recover the seed level variance, and Appendix AF explains why that matters for the smaller effects. And the per sample metric scores behind Appendix Z.4 were not retained, only the generations and the aggregates, so the mixed model we recommend there requires re-scoring the stored generations rather than re-analyzing a saved table. The stored generations for every system and fold are part of the supplementary material so that this is possible. All result tables are traced to the stored evaluation files, and the figures are regenerated from the tables by script.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the paper, and to generate the synthetic reference question-answer pairs of the public testbed (Appendix N). LLMs also appear in this work as evaluated baselines and as the judge of the automatic evaluation, as described in the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Amr Abourayya, Jens Kleesiek, and Michael Kamp. Beyond parameter aggregation: Semantic consensus for federated fine-tuning of LLMs. *arXiv preprint arXiv:2605.11857*, 2026.

Samiul Alam, Luyang Liu, Ming Yan, and Mi Zhang. FedRolex: Model-heterogeneous federated learning with rolling sub-model extraction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6168–6173, 2019.

Lorin W. Anderson, David R. Krathwohl, Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James Raths, and Merlin C. Wittrock. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, 2001.

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*, volume 37, 2024.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, et al. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6521–6532, 2020.

Boxing Chen and Colin Cherry. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT)*, pp. 362–367, 2014.

Yae Jee Cho, Andre Manoel, Gauri Joshi, Robert Sim, and Dimitrios Dimitriadis. Heterogeneous ensemble knowledge transfer for training large models in federated learning. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2881–2887, 2022.

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 (EMNLP)*, pp. 12903–12913, 2024.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 181–189. SIAM, 2021. doi: 10.1137/1.9781611976700.21.

Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1342–1352, 2017.

Donghang Duan, Xu Zheng, Lizong Zhang, Chong Mu, and Meng Han. FedWeave: Rethinking the unit of specialization in heterogeneous federated MoE-LoRA. *arXiv preprint arXiv:2607.26618*, 2026.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, 2024.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of NAACL-HLT*, pp. 2587–2601, 2022.

Tao Fan, Guoqiang Ma, Yan Kang, Hanlin Gu, Yuanfeng Song, Lixin Fan, Kai Chen, and Qiang Yang. FedMKT: Federated mutual knowledge transfer for large and small language models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pp. 243–255, 2025.

Gemma Team, Thomas Mesnard, Cassidy Hardin, et al. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Wei Guo, Siyuan Lu, Yiqi Tong, Zhaojun Hu, Fuzhen Zhuang, Xiao Zhang, Tao Fan, and Jin Dong. H^2Tune: Federated foundation model fine-tuning with hybrid heterogeneity. In *ECAI 2025*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 3178–3185. IOS Press, 2025. doi: 10.3233/FAIA251183. arXiv:2507.22633.

Chaoyang He, Murali Annavaram, and Salman Avestimehr. Group knowledge transfer: Federated learning of large CNNs at the edge. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*, 2021.

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 (ICLR)*, 2022.

Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Ruofan Jia, Weiying Xie, Jie Lei, Jitao Ma, Haonan Qin, and Leyuan Fang. HeteroTune: Efficient federated learning for large heterogeneous models. *arXiv preprint arXiv:2411.16796*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Peter Kairouz, H. Brendan McMahan, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

Shivam Kalra, Junfeng Wen, Jesse C. Cresswell, Maksims Volkovs, and Hamid R. Tizhoosh. Decentralized federated learning through proxy model sharing. *Nature Communications*, 14(1):2899, 2023. doi: 10.1038/s41467-023-38569-4. arXiv:2111.11343.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. 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)*, pp. 565–576, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.47. arXiv:2106.04489.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. UnifiedQA-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*, 2022.

Minjae Kim, Sangyoon Yu, Suhyun Kim, and Soo-Mook Moon. DepthFL: Depthwise federated learning for heterogeneous clients. In *International Conference on Learning Representations (ICLR)*, 2023a.

Yeachan Kim, Junho Kim, Wing-Lam Mok, Jun-Hyung Park, and SangKeun Lee. Client-customized adaptation for parameter-efficient federated learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1159–1172, 2023b.

Boris Knyazev, Michal Drozdzal, Graham W. Taylor, and Adriana Romero-Soriano. Parameter prediction for unseen deep architectures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Boris Knyazev, Doha Hwang, and Simon Lacoste-Julien. Can we scale transformers to predict parameters of diverse ImageNet models? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17243–17259. PMLR, 2023. arXiv:2303.04143.

Miltiadis Kofinas, Boris Knyazev, Yan Zhang, Yunlu Chen, Gertjan J. Burghouts, Efstratios Gavves, Cees G. M. Snoek, and David W. Zhang. Graph neural networks for learning equivariant representations of neural networks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oO6FsMyDBt. arXiv:2403.12143.

David R. Krathwohl. A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4): 212–218, 2002.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1410–1421, 2022.

Royson Lee, Minyoung Kim, Da Li, Xinchi Qiu, Timothy Hospedales, Ferenc Huszár, and Nicholas D. Lane. FedL2P: Federated learning to personalize. In *Advances in Neural Information Processing Systems*, volume 36, pp. 14818–14836, 2023. doi: 10.52202/075280-0653. arXiv:2310.02420.

Royson Lee, Minyoung Kim, Fady Rezk, Rui Li, Stylianos I. Venieris, and Timothy Hospedales. FedP^{2}EFT: Federated learning to personalize PEFT for multilingual LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22805–22813, 2026.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7871–7880, 2020.

Daliang Li and Junpu Wang. FedMD: Heterogeneous federated learning via model distillation. In *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

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)*, 2020b.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations (ICLR)*, 2020c.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 109–123, 2022. doi: 10.52202/068431-0009. arXiv:2210.08823.

Derek Lim, Haggai Maron, Marc T. Law, Jonathan Lorraine, and James Lucas. Graph metanetworks for processing diverse neural architectures. In *International Conference on Learning Representations (ICLR)*, 2024.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74–81, 2004.

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 (NeurIPS)*, 2020.

Or Litany, Haggai Maron, David Acuna, Jan Kautz, Gal Chechik, and Sanja Fidler. Federated learning with heterogeneous architectures using graph hypernetworks. *arXiv preprint arXiv:2201.08459*, 2022.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1950–1965, 2022a. doi: 10.52202/068431-0142. arXiv:2205.05638.

Ruixuan Liu, Fangzhao Wu, Chuhan Wu, Yanlin Wang, Lingjuan Lyu, Hong Chen, and Xing Xie. No one left behind: Inclusive federated learning over heterogeneous devices. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3398–3406, 2022b.

Disha Makhija, Xing Han, Nhat Ho, and Joydeep Ghosh. Architecture agnostic federated learning for neural networks. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14860–14870, 2022.

Massachusetts Institute of Technology. MIT 18.657 — Mathematics of Machine Learning (opencourseware). https://ocw.mit.edu/courses/18-657-mathematics-of-machine-learning-fall-2015/, 2015.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. RQUGE: Reference-free metric for evaluating question generation by answering the question. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6845–6867, 2023.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Steven Moore, Eamon Costello, Huy A. Nguyen, and John Stamper. An automatic question usability evaluation toolkit. In *Artificial Intelligence in Education (AIED)*, 2024.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, 2018.

Chien Van Nguyen et al. A survey on small language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*, pp. 807–821, 2025.

Furkan Pala and Islem Rekik. GNN-based unified deep learning. In *Predictive Intelligence in Medicine: 8th International Workshop, PRIME 2025, Held in Conjunction with MICCAI 2025*, volume 16164 of *Lecture Notes in Computer Science*, pp. 34–46. Springer, 2025. doi: 10.1007/978-3-032-07904-6_4.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Jason Phang, Yao Zhao, and Peter J. Liu. Investigating efficiently extending transformers for long input summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. ProphetNet: Predicting future n-gram for sequence-to-sequence pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 2019.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. Automated educational question generation at different Bloom's skill levels using large language models: Strategies and evaluation. In *Proceedings of the 25th International Conference on Artificial Intelligence in Education (AIED)*, 2024. arXiv:2408.04394.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 9489–9502, 2021.

Tao Shen, Jie Zhang, Xinkang Jia, Fengda Zhang, Gang Huang, Pan Zhou, Kun Kuang, Fei Wu, and Chao Wu. Federated mutual learning. *arXiv preprint arXiv:2006.16765*, 2020.

Aliaksandra Shysheya, John F. Bronskill, Massimiliano Patacchiola, Sebastian Nowozin, and Richard E. Turner. FiT: Parameter efficient few-shot transfer learning for personalized and federated image classification. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9aokcgBVIj1. arXiv:2206.08671.

Stanford University. CS229 — Machine Learning course notes. https://cs229.stanford.edu/, 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.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. Generative language models for paragraph-level question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 670–688, 2022.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Jiaqi Wang, Chenxu Zhao, Lingjuan Lyu, Quanzeng You, Mengdi Huai, and Fenglong Ma. Bridging model heterogeneity in federated learning via uncertainty-based asymmetrical reciprocity learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52290–52308. PMLR, 2024a. arXiv:2407.03247.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. FLoRA: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature Communications*, 13(1):2032, 2022.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 447–460, 2022.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu, and Ying Xu. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 731–744, 2022.

Mang Ye, Xiuwen Fang, Bo Du, Pong C. Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2024.

Liping Yi, Han Yu, Chao Ren, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. Federated model heterogeneous matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66431–66454, 2024. doi: 10.52202/079017-2125. arXiv:2406.00488.

Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

Chen Zhang, Husheng Li, Xiang Liu, Linshan Jiang, and Danxin Wang. Hypernetworks for model-heterogeneous personalized federated learning. *arXiv preprint arXiv:2507.22330*, 2025.

Chris Zhang, Mengye Ren, and Raquel Urtasun. Graph HyperNetworks for neural architecture search. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rkgW0oA9FX. arXiv:1810.05749.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pp. 649–657, 2015.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xinyu Zhou, Boris Knyazev, Alexia Jolicoeur-Martineau, and Jie Fu. LoGAH: Predicting 774-million-parameter transformers using graph HyperNetworks with 1/100 parameters. *arXiv preprint arXiv:2405.16287*, 2024.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 12878–12889, 2021.

A Motivating Scenario

Two Students. Two Realities. One LLM Out of Reach. Consider two graduate students, **Alice** and **Bob**, both enrolled in introductory Machine Learning courses at different universities (Figure 3). **Alice** studies at **University A**, a well-funded research institution with an extensive digital library of ML pedagogical content. The university's computing resources easily support fine-tuning a relatively rich SLM, e.g., a 250M-parameter Flan-T5-base. Alice's locally fine-tuned automatic question and answer generation (AQAG) tool produces pedagogically rich, contextually grounded QAs that test deep conceptual understanding across the full ML curriculum. **Bob** studies at **University B**, a resource-constrained institution. Teaching materials are limited to a single set of introductory lecture slides and a handful of textbook chapters. Computational constraints restrict deployment to even smaller SLMs (e.g., BART-base at 140M parameters). Bob's locally fine-tuned AQAG tool produces repetitive, superficial QAs that test only surface-level recall. **The LLM neither can reach.** Both Alice and Bob would, in an ideal world, route their contexts through a frontier LLM such as GPT-4o. In reality, neither institution can afford the per-token cost at curriculum scale, and Bob's institution additionally has data-protection rules that forbid sending teaching materials to a third-party endpoint. The LLM is a quality *ceiling* they want to approach, not a deployable solution. **The UniLearnLM solution.** Both universities participate in a federated network alongside three others, each running its own private SLM (Flan-T5-base, BART-base, LED-base-16384, Pegasus-X-base, ProphetNet-large). Each institution fine-tunes a private LoRA adapter on its own curriculum and exposes its model to the federation only through a coarse architecture graph. A shared graph attention network, whose parameters are the *only* component federated, encodes each institution's graph into per-block embeddings that condition a local FiLM modulator. Pedagogical inductive bias learned through Alice's larger, richer configuration is implicitly transferred to Bob's lightweight model, while *no* curricular data ever leaves either institution. After federation, Bob's AQAG tool is the thing we

set out to build. Whether it also delivers what Bob wants is the question the paper answers, and the answer is mixed.

This scenario sets up four properties. Two of them UniLearnLM has, and two are hypotheses that our results qualify. *(i) Privacy of the federated protocol*: once the data exists, no raw curricular material and no LoRA weights cross institutional boundaries, and only the shared GAT parameters do. The qualification in Appendix C.2 matters here, since building the training targets did route curricula through a third-party endpoint. *(ii) Architectural heterogeneity*: different SLM families take part in one federation with no compatibility constraint, and Appendix AB adds a decoder-only client with no method change. *(iii) Uplift for the resource-constrained participant* (*Hypothesis 1*): delivered in part. On the public testbed the smallest institution improves on isolation on 8 of 11 metrics and keeps the depth that every distillation route removes, but its depth proxy itself does not rise, and in the leave one out study the two small clients together fall below local training on depth. *(iv) Approach to the LLM ceiling* (*Hypothesis 2*): delivered for grounding, where the federation leads every LLM also at a matched answer length, and not for fluency or the holistic judge. Bob gets privacy, heterogeneity, and questions that are deeper than a distillation baseline would give him. He does not, on this evidence, get Alice's quality or an LLM's.

B Extended Related Work

Educational QA generation and equity. Question generation moved from seq2seq models (Du et al., 2017) to QA pair generation with pedagogically grounded datasets (Xu et al., 2022; Yao et al., 2022). Bloom aware scoring (Scaria et al., 2024), item writing flaw rubrics (Moore et al., 2024), and QG-Bench (Ushio et al., 2022) evaluate single models; none reports the per-institution change in depth when a model joins a federation. Fairness in FL usually means accuracy parity across clients (Li et al., 2020c; Mohri et al., 2019; Du et al., 2021); we instead measure per-institution improvement over local training, since many participants gain nothing from federation (Yu et al., 2020), as two of our three public institutions confirm (Appendix B has the full discussion).

B.1 Where this sits in the literature

There are four routes around architectural heterogeneity in federated learning. The first three keep a shared object in a space the clients already agree on: a shared backbone whose sub-models differ only in width or depth (Diao et al., 2021; Alam et al., 2022; Kim et al., 2023a), a shared base model under heterogeneous LoRA ranks (Cho et al., 2024), or a shared label or vocabulary space over which logits or generations can be averaged (Li & Wang, 2019; Lin et al., 2020; Zhu et al., 2021). A fourth route drops that requirement. Hypernetwork federations share a generator rather than a parameter tensor: pFedHN conditions a hypernetwork on a client embedding (Shamsian et al., 2021), graph hypernetworks condition it on the client computation graph (Knyazev et al., 2021; Litany et al., 2022; Lim et al., 2024), and FedHeNN aligns intermediate representations on a shared anchor set instead of sharing parameters at all (Makhija et al., 2022). UniLearnLM belongs to this fourth route, and we make no claim to have introduced it. Two properties that this paper relies on are established there: the shared object need not have the shape of any client model, and the volume communicated is independent of client model size (Shamsian et al., 2021). Hypernetwork federations for heterogeneous models remain an active line (Zhang et al., 2025), and graph hypernetworks have grown from architecture search (Zhang et al., 2019) to predicting the parameters of large transformers (Knyazev et al., 2023; Zhou et al., 2024). Reading the weights of a network with a graph network is a separate line of its own (Kofinas et al., 2024); our node features summarize each module with three statistics, a much coarser signal.

Other model heterogeneous federations share a small homogeneous module next to private models (Shen et al., 2020; Kalra et al., 2023; Wang et al., 2024a; Yi et al., 2024), align factorized LoRA adapters across language model families and sizes (Guo et al., 2025; Jia et al., 2024), or federate a meta network that maps client statistics to personalization choices (Lee et al., 2023; 2026). Consensus over text generated on a shared prompt set (Abourayya et al., 2026) is the published relative of our text consensus baselines.

Our instantiation differs from that family in one respect that matters for language models. Graph hypernetworks generate the client's weights; even where this scales to transformers of several hundred

million parameters (Zhou et al., 2024), generated weights replace the pretrained ones, whereas the value of the institutions' SLMs lies in their pretrained weights. UniLearnLM instead keeps the backbone frozen and has the encoder emit a residual scale and shift per block, so the pretrained model is modulated rather than replaced. This is what makes the approach usable across BART, Flan-T5, LED, and a decoder-only Qwen client at once, with no aligned tokenizer and no shared teaching material. Representation alignment on anchors, the FedHeNN route, needs a shared anchor input set, which our setting does not have; we therefore run FedHeNN with the public pool as its anchor set (Appendix U). The ablations in Appendix Z test whether the graph is doing work: a parameter matched encoder that cannot read the graph, and a shared vector that is not conditioned on architecture, both fail to reproduce the effect, though we note in Section 5 that those ablations run on fewer folds than the headline table.

B.2 Heterogeneous Federated Learning

Standard federated learning (McMahan et al., 2017) and its proximal-term variants (Li et al., 2020b; 2021) assume that every client trains the same model with the same parameter shapes. This assumption has been relaxed along the capacity axis: width- and depth-elastic federations (Diao et al., 2021; Liu et al., 2022b; Alam et al., 2022; Kim et al., 2023a) let clients instantiate sub-models of different sizes, but always within a single model family. Surveys of model-heterogeneous FL (Ye et al., 2024) confirm that the shared backbone assumption remains the main limit of this line. On the language model side, HetLoRA (Cho et al., 2024) comes closest: it federates LoRA modules of heterogeneous *rank* on top of a shared base model.

A separate line removes the shared backbone entirely by federating a *generator* instead of a parameter tensor. pFedHN (Shamsian et al., 2021) trains a hypernetwork on the server that maps a learned client embedding to that client's weights; because the hypernetwork itself is never transmitted, the communicated volume is independent of client model size, and a single hypernetwork can serve clients whose target models differ in size. Graph hypernetworks push this further by conditioning on the client's computation graph rather than on a client index. Knyazev et al. (Knyazev et al., 2021) predict parameters for unseen architectures from an architecture graph, Litany et al. (Litany et al., 2022) bring that construction into federated learning and report a framework that does not restrict clients to a shared architecture type, uses no external data, and does not require clients to disclose their architecture, and graph metanetworks (Lim et al., 2024) generalize the encoder. FedHeNN (Makhija et al., 2022) takes a different route again and aligns intermediate representations on a shared anchor set, so clients need not share an architecture either, and a fifth route shares class prototypes or a generalized output header rather than a backbone (Tan et al., 2022). Closest to us in NLP, C2A (Kim et al., 2023b) uses a hypernetwork to generate client specific adapters for a pretrained language model under federated learning, conditioning on client information. C2A conditions on a client descriptor and assumes one shared pretrained backbone; the setting here differs in that clients run different backbone families with different tokenizers, and the conditioning is on a typed architecture graph rather than a descriptor. That is a narrow difference, and we state it plainly.

UniLearnLM belongs to this generator family and we claim no priority over it. Unified learning (Pala & Rekik, 2025) trains one GNN over neuron-level graphs of several models, which do not fit in memory for language models and which UniLearnLM replaces by sub-module-level graphs, and the idea of federating a generator that reads a client's computation graph is already established in the work cited above. What is added here is the application to pretrained language models with incompatible tokenizers, plus the deployment and the evaluation suite.

The mechanism we use for that application is also not new in isolation. Having a hypernetwork emit per block adapters or modulation for a frozen transformer is an established idea in multi task and parameter efficient transfer learning (Karimi Mahabadi et al., 2021), a learned per block scale and shift on a frozen backbone is itself a parameter efficient method (Lian et al., 2022; Liu et al., 2022a), and FiLM parameters have been federated for image models (Shysheya et al., 2023), so readers should treat our use of it as an application rather than an invention. What we have not found in that literature is the combination used here, namely conditioning on a typed architecture graph, across clients that run different pretrained families with disjoint tokenizers, in a federated setting. That combination is the residual contribution, and it is a modest one. FedHeNN needs a shared anchor input set, which our setting lacks, so we run it with the public pool as that set. The literature on heterogeneous federated learning specifically for educational QA generation across SLM families is

very limited, and this scarcity constrains the baselines available to our experiments, which we discuss in Section 4.

B.3 Knowledge Distillation on Public Data

A second line of work admits architectural heterogeneity by exchanging *knowledge*, in the form of logits, output distributions, or generated samples, rather than weights (Li & Wang, 2019; Lin et al., 2020; He et al., 2020; Cho et al., 2022; Zhu et al., 2021). FedKD (Wu et al., 2022) replaces public data using sample-size-weighted Federated Averaging (FedAvg) (McMahan et al., 2017) over gradients of a small shared mentee model compressed via singular value decomposition (SVD). FedKD is designed for heterogeneity: each client keeps a private, possibly different, mentor model and only the mentee is exchanged. What it does require is that clients agree on that shared mentee and on a common output space for the mutual distillation loss, which our tokenizer-disjoint setting does not provide without an alignment step. We do not implement it here. A common subroutine in this family is logit averaging, which presupposes a shared label or vocabulary space. Heterogeneous tokenizers do not satisfy that assumption directly, but they do not make the channel impossible: FedMKT (Fan et al., 2025) builds a token alignment table by minimum edit distance and then transfers logits between a server LLM and client SLMs with different vocabularies. We chose not to add that machinery, and instead adapt the public pool channel to text space, which is simpler at the cost of a coarser signal. This is a design choice and not a forced move, so we also run FedMKT itself as a baseline (Appendix U). We report two consensus rules as baselines (Appendix S): TC-Best, in which each client distills toward the single generation from another institution with the highest mean pairwise ROUGE-L, and TC-Ens, in which each client distills toward all other institutions' generations equally.

B.4 Equity-Aware FL and Educational NLP

Educational question generation began with attention seq2seq models for reading comprehension (Du et al., 2017) and matured into QA-pair generation with pedagogically grounded datasets (Xu et al., 2022; Yao et al., 2022). Pedagogically aware evaluation of generated questions is an active area. Scaria et al. (2024) generate and score questions by Bloom level using expert and LLM rubrics, SAQUET (Moore et al., 2024) checks generated items against an item writing flaw rubric, and QG-Bench (Ushio et al., 2022) provides a standard benchmark with manual evaluation. What none of these reports, because they study single model settings, is the *per-institution change* in depth when a model joins a federation, and that is the gap our suite fills. On the language model side, surveys of small language models (Nguyen et al., 2025) treat collaboration between independently owned SLMs as an area that is still open. Fairness in FL is typically formalized as accuracy parity across clients (Li et al., 2020c; Mohri et al., 2019; Du et al., 2021); we instead measure per institution *improvement* of federation over the local-only baseline, complemented by a pedagogical-depth metric derived from *Bloom's Revised Taxonomy* (Anderson et al., 2001), a six-level scale of cognitive depth (Remember, Understand, Apply, Analyze, Evaluate, Create) widely used in educational assessment. Our evaluation suite combines an LLM-as-judge rubric (Zheng et al., 2023) with reference-free QA-quality metrics (Mohammadshahi et al., 2023; Fabbri et al., 2022; Es et al., 2024).

Reporting improvement per-institution matters because federation does not help every participant by default. Yu et al. (2020) show that on standard tasks many participants gain nothing from federation, because the federated model is less accurate on their own data than a model they could train locally, and that privacy mechanisms widen that gap. Our results in Section 5 reproduce this pattern at two of three public institutions, which is why we report per institution deltas in the main text rather than only a mean over clients.

C Experiment Details

C.1 Public Testbed: Open ML Curricula

The public testbed comprises three institutions $(\mathcal{I}_0^{\text{pub}}, \mathcal{I}_1^{\text{pub}}, \mathcal{I}_2^{\text{pub}})$ whose curricula are built from openly available sources, so the entire pipeline can be reproduced by anyone with the corresponding files. Institution $\mathcal{I}_0^{\text{pub}}$ draws on *MIT 18.657: Mathematics of Machine Learning* (Massachusetts Institute of Technology, 2015), lecture notes on statistical learning theory and optimization (110

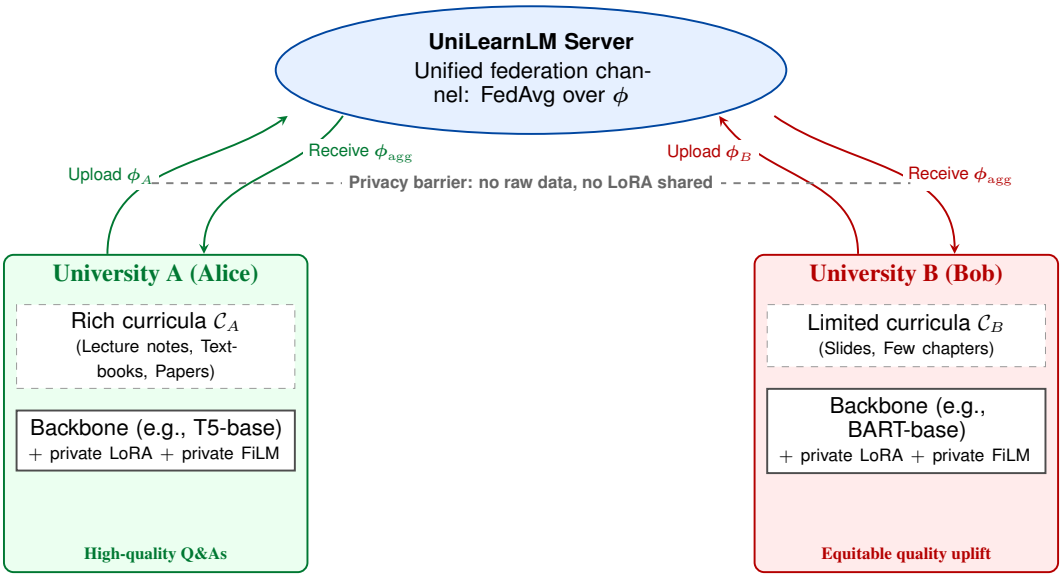


Figure 3: Overview of the UniLearnLM unified federated framework. Each institution exposes a coarse *architecture graph* of transformer blocks; a shared GNN based unification encoder (instantiated as a GAT, see §3) reads it, and only its parameters ϕ cross the privacy barrier; this is the unified federation channel. Per-block embeddings drive a client-local FiLM adapter that residually modulates hidden states. LoRA and FiLM weights, and all curricular data, remain client-local.

Table 3: LLM-judge per-dimension breakdown for UniLearnLM (public testbed; global test, validation-selected model, mean over 3 folds). 1–5 scale.

	Ctx.	Edu.	Corr.	Rel.
$\mathcal{I}_0^{\text{pub}}$	3.936	3.027	3.754	4.165
$\mathcal{I}_1^{\text{pub}}$	4.148	2.921	3.908	4.417
$\mathcal{I}_2^{\text{pub}}$	4.008	3.102	3.793	4.379

chunks, 547 GPT-4o-annotated QA pairs). Institution $\mathcal{I}_1^{\text{pub}}$ uses the *Stanford CS229* (Stanford University, 2024) undergraduate course notes, which span supervised and unsupervised learning, neural networks, and reinforcement learning (110 chunks, 559 QA pairs). Institution $\mathcal{I}_2^{\text{pub}}$ is the most demanding: a curated set of recent machine-learning arXiv preprints across vision, NLP, and reinforcement learning, an order of magnitude larger at 981 chunks and 5,037 QA pairs.

Data preparation. Source PDFs are converted to plain-text chunks of 150–400 words each (one coherent topic per chunk, boilerplate and figure captions stripped). For every chunk we request five (question, answer) pairs from GPT-4o; the stored splits contain between three and six pairs per chunk (global test: 154 chunks with five, 23 with six, 2 with four), because the model sometimes returned six and schema validation dropped malformed pairs. Each pair is labeled with its topic (a 2–4-word ML concept phrase), Bloom's Revised Taxonomy level (integer 1–6), difficulty (easy/medium/hard), and an answerable-from-context flag. Annotations are validated against a schema before splitting.

Splits. Splits are at the chunk level (not at the QA-pair level) to prevent context leakage. Each institution holds out 15% of chunks as a fixed test set and divides the remaining 85% into three equal CV folds. The split seed is 42 across all collaborators so that splits are bit-for-bit reproducible. We provide the checksums (Appendix N) to allow third-party verification.

Heterogeneity. Sample sizes differ by an order of magnitude across the public testbed ($\mathcal{I}_0^{\text{pub}}$, $\mathcal{I}_1^{\text{pub}}$: $\sim$110 chunks; $\mathcal{I}_2^{\text{pub}}$: 981 chunks), giving a controlled non-IID setting. The full per-source composition (chunk counts, QA-pair counts, backbones, and per fold split sizes) is reported in Table 11 (Appendix N).

C.2 PRIVATE TESTBED: REAL-WORLD FEDERATION

Distillation pool on the private testbed. The live TC-Best, TC-Ens and UniLearnLM+ federations and the three-fold controlled rerun all used the public testbed pool (the 180 row per fold pool built from the MIT OpenCourseWare, Stanford CS229 and arXiv chunks of Appendix C.1), served by the coordination server. No private chunk ever entered a distillation pool, and no raw private text crossed institutional boundaries; the baselines exchanged generated QA pairs on public contexts only. The statement in Section 3 that the distillation baselines condition on raw text from other institutions' curricula therefore holds on the public testbed, where the pool is peeled from the clients' own public data, and not on the private testbed, where the pool is an external public corpus, which is closer to FedMD's setting.

The private testbed instantiates UniLearnLM across **five real collaborating institutions** spanning five countries on three continents (Tunisia, Germany, Nigeria, Colombia, and Peru), each contributing its own private machine-learning curriculum. All five collaborators are university-level institutions teaching machine-learning curricula in English at the undergraduate or advanced-undergraduate level. The per-institution corpus statistics, privacy posture, material types, language, and education level are reported site by site in Table 4. These five institutions are the full-participation members $\mathcal{I}_0^{\text{priv}}$–$\mathcal{I}_4^{\text{priv}}$ defined in Section 4; the $\mathcal{I}_k^{\text{priv}}$ superscript distinguishes them from the public-testbed institutions $\mathcal{I}_k^{\text{pub}}$, with which they share no index. Concretely, each collaborator (i) runs the Phase 1 data-preparation pipeline locally on their own materials (PDF lectures, PPTX decks, textbook chapters, course assignments, research papers), producing a JSON file conforming to the same schema as the public testbed; (ii) validates their file with our schema validator; (iii) generates deterministic 3-fold CV splits with the shared seed $= 42$; and (iv) ships only the resulting checksum file to the coordinator for cross-collaborator reproducibility verification. *No raw curricular data, no QA pairs, and no model weights ever cross institutional boundaries.*

An important qualification to that sentence. Phase 1 builds each institution's reference QA pairs by calling GPT-4o on its chunks, so while nothing crosses between *institutions*, every participating institution does send its curriculum to a third party endpoint in order to create its training targets. That is the same class of egress our introduction gives as a reason institutions cannot simply use a hosted LLM, and we should not claim a privacy property we do not have. Three things are worth separating. The federated protocol itself is clean: once the data exists, training and inference are local and only the 0.088M parameter encoder is exchanged. The annotation step is not clean, and it is a property of how we built this study rather than of the method, since an institution could label its own corpus with a local model, with instructor written items, or with existing question banks. And the evaluation is not clean either, for the same reason (Appendix AF). A deployment that had to honor the constraint in our introduction end to end would need to replace the annotation step, and we have not demonstrated that this is possible without loss.

Two participation modes. An institution joins the private testbed in one of two ways. Under **full participation**, the institution runs the federation client on its own hardware and trains its own SLM with a private LoRA adapter, so it contributes both a curriculum and a model to the federation. The five $\mathcal{I}_k^{\text{priv}}$ institutions of Table 4 are full participants. Under **dataset-only participation**, the institution contributes its own curriculum but does not have the local compute to train a model; it provides its teaching and research materials and takes no part in federated training. The four dataset-only contributors, identified as $\mathcal{D}_0$–$\mathcal{D}_3$ (Table 4, Appendix C.2), supply materials that are pooled into a held-out cross institution test set.

Private global test set. The lecture and research materials contributed by the four dataset-only institutions $\mathcal{D}_0$–$\mathcal{D}_3$ are pooled and run through the same Phase-1 data-preparation pipeline as the rest of the testbed: PDFs and slide decks are converted to plain-text chunks of 150–400 words, and GPT-4o annotates each chunk with five (question, answer) pairs, each carrying a topic label, a Bloom level, a difficulty tag, and an answerability flag. This yields 200 contexts and 996 QA pairs. No institution trains on this set, so it is the private-testbed analog of the public global test set (Table 2) and supports the cross institution generalization comparison in Table 30.

The institutions are heterogeneous on two axes that are central to the equity hypothesis (Table 4, Appendix C.2). *Dataset size* varies from 50–200 pages ($\mathcal{I}_2^{\text{priv}}$) up to 1500–2000 pages ($\mathcal{I}_1^{\text{priv}}$), with

the remaining three institutions in the 200–700-page range. *Privacy posture* also varies: $\mathcal{I}_0^{\text{priv}}$ and $\mathcal{I}_4^{\text{priv}}$ contribute private institutional materials with research-use authorization, while $\mathcal{I}_1^{\text{priv}}$, $\mathcal{I}_2^{\text{priv}}$, and $\mathcal{I}_3^{\text{priv}}$ contribute a mix of public and private materials. Each institution is assigned a backbone family based on its declared compute resources and dataset size (Table 5, Appendix C). The federation involves no central training corpus, no shared evaluation pool of raw text, and no exchange of participant identities; the per-institution profiles in Table 4 are the only collaborator-level information that leaves any single institution.

Table 4: Private-testbed collaborators. Identifiers $\mathcal{I}_0^{\text{priv}}$–$\mathcal{I}_4^{\text{priv}}$ are full-participation institutions that run a federation client on their own private curriculum; identifiers $\mathcal{D}_0$–$\mathcal{D}_3$ are dataset-only contributors that supply curricular materials toward the cross institution private global test set (Table 30) but do not train. The "Dataset only" column marks the latter with ✓. All identifiers are pseudonymized; every contributor explicitly authorized our research use. "private (auth.)" marks private institutional materials contributed with research-use authorization. "n/a" = page count not available.

Inst.	Dataset only	Country	Pages	Privacy	Materials / level
$\mathcal{I}_0^{\text{priv}}$		Tunisia	200–400	private (auth.)	Lecture notes/slides; research papers (English; undergraduate advanced; pdf)
$\mathcal{I}_1^{\text{priv}}$		Germany	1500–2000	mixed	Lecture notes/slides; research papers; course assignments; textbook chapters (English; undergraduate; pdf/pptx)
$\mathcal{I}_2^{\text{priv}}$		Nigeria	50–200	mixed	Lecture notes
$\mathcal{I}_3^{\text{priv}}$		Colombia	500–700	mixed	Lecture notes/slides; textbook chapters (English; undergraduate; pdf)
$\mathcal{I}_4^{\text{priv}}$		Peru	300–500	private (auth.)	Lecture notes/slides; research papers (English; undergraduate advanced; pdf)
$\mathcal{D}_0$	✓	Ireland	n/a	private (auth.)	Lecture notes/slides; research papers (English; mixed academic)
$\mathcal{D}_1$	✓	United Kingdom	50–200	public	Lecture notes/slides; textbook chapters; research papers; course assignments; video lecture transcripts; online course materials (English; postgraduate Master's)
$\mathcal{D}_2$	✓	Senegal	50–200	private (auth.)	Lecture notes/slides; course assignments; online course materials (bilingual; undergraduate advanced)
$\mathcal{D}_3$	✓	Tunisia	50–200	private (auth.)	Lecture notes/slides; research papers (bilingual; undergraduate advanced)

C.3 Training Procedure

Table 5: SLMs used in the public and private testbeds. LoRA targets follow each model family's attention projection naming, and d_{model} is the hidden size used by the FiLM modulator.

Backbone	Family	LoRA targets	d_{m}	Params
BART-base (Lewis et al., 2020)	BART	query, value attention projections	768	140M
Flan-T5-base (Chung et al., 2024)	Flan-T5	query, value attention projections	768	250M
LED-base-16384 (Beltagy et al., 2020)	LED	query, value attention projections	768	162M
Pegasus-X-base (Phang et al., 2023)	Pegasus-X	query, value attention projections	768	272M
ProphetNet-large (Qi et al., 2020)	ProphetNet	query, value attention projections	1024	392M

Per-fold local training. Each institution trains its SLM with its private LoRA adapter for up to 100 epochs. The early-stopping criterion is ROUGE-L between the generated and the reference QA pair, computed by greedy decoding on a fixed 50-question subset of that institution's validation split (the same 50 questions are reused across all epochs, sampled once with the global random seed so that the score is comparable across epochs). Training stops when this validation ROUGE-L fails to improve by at least $\Delta_{\min} = 10^{-4}$ for 10 consecutive epochs. At the end of training, we keep two versions of the LoRA adapter for separate evaluation: the *validation-selected* model (the epoch at which the validation ROUGE-L was highest) and the *final-epoch* model (the last epoch).

Local optimization. AdamW with peak learning rate 3×10^{-4}, cosine schedule with 10% linear warm-up, gradient clipping 1.0, batch size 4, LoRA rank r=16 and α=32, LoRA dropout 0.1, mixed-precision BF16 autocast on the LM forward.

Federated optimization. UniLearnLM runs $T = 33$ federation rounds with 3 local epochs per round and sample-size-weighted FedAvg of the GAT parameters ϕ (Appendix K), so a client that

participates in every round sees up to 99 local epochs of data. The Indiv baseline trains for up to 100 epochs but early-stops well before that: its validation-selected model comes from epochs $11/13/16$ (MIT), $12/12/42$ (Stanford), and $34/28/13$ (arXiv) across the three folds, because additional epochs overfit the small institutional corpora. The validation-selected Indiv model is therefore *not* compute-starved relative to UniLearnLM; both are compared using the model selected on validation data (by validation ROUGE-L), and the gap is one of data and conditioning rather than of optimization budget.

C.4 Conditioning Regimes

When many QA pairs share a context, a model conditioned only on that context tends to emit near-duplicate questions. We mitigate this by conditioning the training and inference prompt on an extra field, in one of three regimes. The *baseline* regime asks for a QA pair from the context alone ("*Generate a question and answer pair from the following machine learning text:* [context]"); since the input does not vary across the five reference pairs of a chunk, evaluation uses diverse beam search to produce five generations per context and Hungarian-matches them to the references via ROUGE-L. The *topic* regime adds the chunk's topic label ("*. . . about* "[topic]" *from the following machine learning text:* [context]"), which makes each input unique and lets standard beam search suffice at evaluation. The *Bloom* regime instead anchors the target cognitive level ("*Generate a Bloom Level-[b] ([verb]) question . . .* "), where `[verb]` is the verb for level b (Remember/Understand/Apply/Analyze/Evaluate/Create). All reported results use the topic regime.

C.5 Baselines

The hypotheses require distinct comparators. *Hypothesis 1* is tested against federation-level methods that, like UniLearnLM, assume a federation of SLMs and evaluate per institution QA quality; *Hypothesis 2* is tested against single-model LLMs.

For *Hypothesis 1*. The lower-resource reference is **Indiv**: each institution trains its SLM with a private LoRA adapter on its own data and no federation, under the three conditioning regimes of Section C.4. The data-pooled reference is **Centralized**, its counterpart. Where Indiv trains each institution only on its own private data and the federated methods exchange only generated text, Centralized lifts the privacy constraint entirely: for each fold it concatenates the three institutions' fold-matched training and validation splits into one pooled corpus and trains a single model *per backbone architecture* (BART-base, Flan-T5-base, LED-base) on that union. Everything else (hyperparameters, topic conditioning, best-epoch early stopping on validation ROUGE-L, 3-fold cross-validation, and the eleven-metric evaluation with the GPT-4o-mini judge) is held identical to Indiv, so the only difference is the training data. The Centralized rows are therefore directly comparable to the Indiv rows in Table 2 and contextualize the federated results from above. The expected ordering Indiv (no data sharing) $\leq$ federated (text-only exchange) $\leq$ Centralized (full data pooling) holds on the aggregate token-overlap and LLM-judge metrics; the one exception is the three reference-free grounding metrics (Faithfulness, Answer Relevancy, QAFactEval), on which Indiv and UniLearnLM edge Centralized (Section 5.2).

Between these references sit two heterogeneous-federation distillation baselines, both adapted to a mix of SLM families. Because the SLM backbones (BART, Flan-T5, LED) have different tokenizers and vocabularies, the only well-defined cross-client signal is generated text, so both baselines exchange QA generations over a shared public pool. **TC-Ens** (adapted from FedMD (Li & Wang, 2019)) distils each client toward the *ensemble* of peers' public pool generations: each round, every client generates a QA pair per public pool context, and the ensemble of peer generations supplies an extra cross-entropy term for the next round. **TC-Best** is the contrasting *selective* variant of the same public pool channel: rather than distilling toward the whole ensemble, it picks the single per-context peer generation with the highest mean pairwise ROUGE-L to peers and distils toward that consensus alone. The selective vs ensemble axis mirrors the "single best peer" vs "averaged ensemble" distinction in the broader codistillation literature; we do not import FedKD's mentor/mentee mechanism here, which operates on each client's private data without a public pool and assumes homogeneous mentees that our heterogeneous backbones cannot supply. TC-Ens and TC-Best thus contrast ensemble against selective consensus over the same public pool. Both reuse the public-testbed chunks (Section C.1) as the distillation surface; algorithm and implementation details are in Appendix S. TC-Ens results are reported in Table 14.

We compare the federated mixture of all SLMs (UniLearnLM aggregated across institutions, not each institution in isolation) against six single-model LLMs prompted zero-shot to produce a (q, a) pair from each test context and run through the same evaluation suite: GPT-4o as the headline ceiling, GPT-4o-mini, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen-2.5-7B-Instruct (Qwen Team, 2024), Mistral-7B-Instruct (Jiang et al., 2023), and Gemini-2.5-Flash. The *Hypothesis 2* panel is reported separately (Section 5.3); these LLMs do not appear as rows in the per institution *Hypothesis 1* tables.

D Additional Setup Notes

Notation. We evaluate on two testbeds and use an explicit superscript to keep their institutions distinct. Public-testbed institutions are written $\mathcal{I}_k^{\text{pub}}$ and private-testbed full-participation institutions $\mathcal{I}_k^{\text{priv}}$; the two index sets are independent, so $\mathcal{I}_0^{\text{pub}}$ (MIT) and $\mathcal{I}_0^{\text{priv}}$ (Tunisia) are different institutions. A third symbol $\mathcal{D}_0, \ldots, \mathcal{D}_3$ denotes the dataset-only contributors of the private testbed, which provide curricular materials but do not run a federation client (Appendix C.2).

Why not host an open 7B model on site. A natural middle path between frontier APIs and federated SLMs is hosting an open-weight 7B model on site and prompting it zero-shot. Hosting, however, is not the binding constraint; adaptation is. Prompted zero-shot, such models do not learn the institution's curriculum, and in Table 34 they trail the fine-tuned federated mixture on both grounding metrics. Closing that gap would require on-site fine-tuning of a 7–8B model, which the compute available at the participating sites cannot support; LoRA fine-tuning of sub-500M backbones is the operating point these sites sustain. The framework is orthogonal to this choice: a locally hosted 7B can itself join the federation as a client, since only the graph encoder is exchanged.

Beyond encoder-decoder backbones. The five SLMs of the model zoo are encoder-decoder models, which is the family these sites can actually fine-tune on their hardware. To check that the architecture-graph representation is not specific to that family, we also swap the BART client at $\mathcal{I}_0^{\text{pub}}$ for a modern decoder-only SLM (Qwen2.5-1.5B-Instruct) and re-run the federation with no change to the graph construction, the encoder, or the schedule. The federated decoder-only client is competitive with the encoder-decoder configuration and leads it on both LLM-judged axes (Appendix AB).

What each evaluation axis asks. Table 6 lists the evaluation dimensions. Each axis answers a distinct pedagogical question: reference-based overlap asks whether generations match instructor-written references; Bloom-based depth asks whether questions demand more than recall ; reference-free relevancy, faithfulness, and factual consistency ask whether answers are grounded in the material students actually study; round-trip consistency asks whether a question is answerable from its context; and the holistic judge approximates an educator's overall reading. The equity metrics then ask how these qualities are distributed across institutions.

E Geographic Distribution of Collaborators

Figure 4 shows the geographic distribution of the nine collaborator institutions across three continents. Five are fully-participating members $\mathcal{I}_0^{\text{priv}}$–$\mathcal{I}_4^{\text{priv}}$ (Tunisia, Germany, Nigeria, Colombia, Peru) that trained their own SLMs under the live federations (TC-Best, UniLearnLM, UniLearnLM+, and TC-Ens). The remaining four are dataset-only contributors $\mathcal{D}_0$–$\mathcal{D}_3$ (Ireland, United Kingdom, Senegal, and a second Tunisia institution) whose lecture and research material was pooled into the global private test set $\mathcal{T}_{\text{priv}}^{\text{global}}$ used as the cross institution generalization benchmark (Section C.2, Table 30).

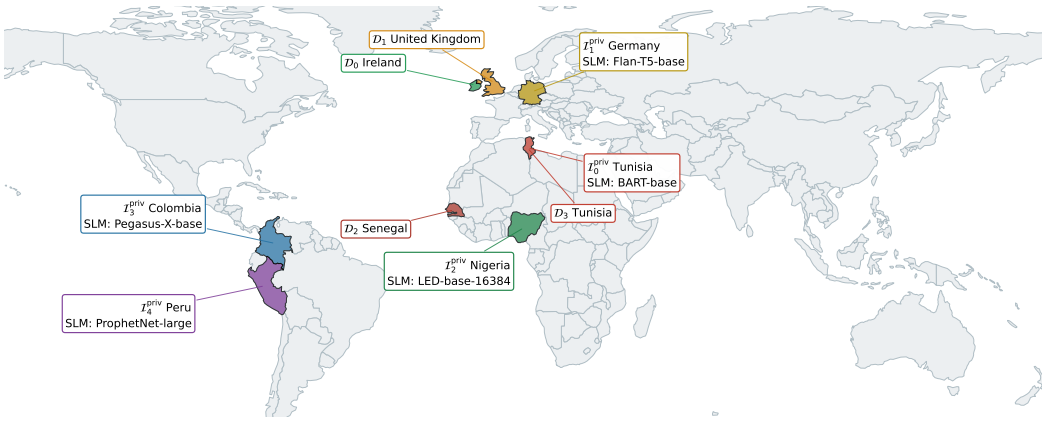


Figure 4: Geographic distribution of the 9 participating institutions in our *unified* federated framework.

Table 6: Evaluation dimensions for UniLearnLM. The eleven QA-quality metrics (R-L, B-4, BERT, B_{cls}, B_{LLM}, AnsR, Faith, QAFE, RQU, RTC, LLM-J) form the per-cell columns in every *Hypothesis 1* results table. The two equity scalars (PES, KTI) are computed from the LLM-judge overall score averaged across folds. Sources: GPT-4o (Hurst et al., 2024).

Dimension	**Metric(s)**
Lightweight reference-based	ROUGE-L; BLEU-4; BERTScore-F1
Question quality	RQUGE (Mohammadshahi et al., 2023)
Answer quality	RAGAS Faithfulness (Es et al., 2024); QAFactEval (Fabbri et al., 2022)
Joint QA quality	RTC (Alberti et al., 2019) via UnifiedQA (Khashabi et al., 2022); RAGAS Answer Relevancy (Es et al., 2024)
Pedagogical	Bloom's BERT classifier; Bloom's LLM judge (both normalized to $\mathrm{EVS}(q) = (\hat{b}(q)-1)/5$)
QA-quality LLM judge	GPT-4o-mini (Hurst et al., 2024) 4-axis rubric (context grounding, educational value, answer correctness, answer relevance, each $1-5$); per sample overall $=$ (mean axes$-1)/4 \in [0,1]$
Equity	Pairwise Equity Score (PES); Knowledge Transfer Index (KTI)

F Evaluation Suite: Per-Metric Definitions

F.1 Question, Answer, and Joint Quality

Question quality is measured by **RQUGE** (Mohammadshahi et al., 2023), a reference-free answerability score that runs a QA model on (q, c) and compares the predicted span to the generated answer; we use a BERTScore-based span scorer scaled to $[1, 5]$. **Answer quality** is measured by RAGAS **Faithfulness** (Es et al., 2024) (natural-language-inference (NLI) entailment rate of atomic answer claims against the context, via DeBERTa (He et al., 2021)) and an approximation of **QAFactE-**

val (Fabbri et al., 2022) based on UnifiedQA (Khashabi et al., 2022). **Joint QA quality** is measured by Round-Trip Consistency, inspired by Alberti et al. (2019) (token-F1 between the generated answer and a re-derived answer obtained by running a QA model on (q, c)), and by RAGAS **Answer Relevancy** (Es et al., 2024) (cosine similarity between the original question and embeddings of n reverse-generated questions). Full formulations are deferred to Appendix M.

F.2 Pedagogical Quality

We adopt **Bloom's Revised Taxonomy** (Anderson et al., 2001) (six levels: Remember, Understand, Apply, Analyze, Evaluate, Create) and assess cognitive depth via two independent classifiers: a publicly available **Bloom's BERT classifier** (`cip29/bert-blooms-taxonomy-classifier`) fine-tuned on Bloom's-labeled question data, and a **Bloom's LLM judge** (GPT-4o-mini queried with a structured prompt that supplies verb-anchored definitions of the six levels and returns a JSON $\{\texttt{level}, \texttt{reason}\}$ object). The Bloom's level is normalized to a continuous **Educational Value Score** in $[0, 1]$ via $\mathrm{EVS}(q) = (\hat{b}(q) - 1)/5$.

F.3 LLM QA-Quality Judge

A four-dimension LLM judge (GPT-4o-mini, JSON-object output) rates each (c, q, a) on a strict 1–5 rubric for *context grounding*, *educational value*, *answer correctness*, and *answer relevance*, with verb-anchored anchors at levels 1, 3, 5 (prompt in Appendix H). The overall score is the four-dimension mean normalized to $[0, 1]$ via $(\bar{s} - 1)/4$.

F.4 Equity Evaluation

Table 7: Equity metrics on the **public testbed** (Section F.4) for local training, UniLearnLM and UniLearnLM+. **Top:** Pairwise Equity Score $E_{kl}^{(m)}$ for the three institution pairs and the unweighted all-pairs mean $\bar{E}$; bold = best per row. **Bottom:** Knowledge Transfer Index $\mathrm{KTI}(l \leftarrow k;\, m) = Q_l^{(m)}(\mathcal{T}_{k\setminus l}) - Q_l^{(\mathrm{indiv})}(\mathcal{T}_{k\setminus l})$ for the six directed pairs (each federated method against the Indiv baseline); a positive value indicates that l's federated SLM gained quality on topics it never saw locally. Bold = best (most positive) per row. All Q values use the LLM-judge overall mean on the global test set, averaged over the 3 folds. The text-consensus reruns and the published heterogeneous federations were not scored by the LLM judge, so PES and KTI are not available for them.

PES pair $E_{kl}^{(m)}$	**Indiv**	**UniLearnLM**	**UniLearnLM+**
(MIT, Stanford)	3.171	**3.187**	3.146
(MIT, arXiv)	1.390	1.414	**1.433**
(Stanford, arXiv)	1.136	**1.148**	1.126
All-pairs mean $\bar{E}$	1.899	**1.916**	1.902
KTI $(l \leftarrow k;\, m)$		**UniLearnLM**	**UniLearnLM+**
arXiv ← MIT		$+0.037$	$\mathbf{+0.052}$
MIT ← Stanford		$+0.007$	$\mathbf{+0.058}$
arXiv ← Stanford		$+0.010$	$\mathbf{+0.027}$
MIT ← arXiv		$+0.006$	$\mathbf{+0.017}$
Stanford ← arXiv		$\mathbf{+0.001}$	-0.026
Stanford ← MIT		$\mathbf{-0.001}$	-0.074

Let $m \in \{\text{Indiv}, TC - Best, TC - Ens, UniLearnLM, UniLearnLM+\}$ denote the learning paradigm. $Q_k^{(m)}$ is institution k's aggregate QA quality under paradigm m (default: LLM-judge overall); $\mathcal{C}_k$ its curricular richness; $\mathcal{R}_k$ its compute resource (total parameter count of the deployed backbone, normalized by the largest in the federation). We assess equity through two quantities. The **Pairwise Equity Score (PES)** $E_{kl}^{(m)}$ divides the summed quality of a (k, l) pair by their gap in $(\mathcal{C}, \mathcal{R})$, normalizing pair quality by the resource asymmetry the federation has to bridge; the per-pair ordering of interest is $E_{kl}^{(m)} > E_{kl}^{(\mathrm{indiv})}$ for any federated m. Because the denominator is method-invariant for a fixed pair, PES re-ranks the methods exactly as pair-summed LLM-judge does and is a presentational lens rather than an independent metric (Appendix R). The **Knowledge Transfer Index** $\mathrm{KTI}(l \leftarrow k;\, m) = Q_l^{(m)}(\mathcal{T}_{k\setminus l}) - Q_l^{(\mathrm{indiv})}(\mathcal{T}_{k\setminus l})$ measures cross-client transfer on

topics exclusive to k's curriculum: a positive value indicates that l's federated SLM has gained quality on topics it never saw locally. Full definitions and derivations are in Appendix R. PES and KTI are reported on both the public testbed (Table 7) and the private testbed (Tables 31–32, Appendix X).

G QA-GENERATION PROMPT

Table 8: Per-client compute budget for the five SLM backbones under the two architecture configurations our methods occupy: **LoRA-only** (used by Indiv, TC-Best, TC-Ens, and the Centralized upper bound) versus **LoRA + FiLM + GAT** (used by UniLearnLM and UniLearnLM+). Numbers collected on a single accelerator with BF16 autocast, batch size 4, input/target sequence length 512→128, beam size 4. Training time is wall-clock for one epoch over a fixed 313-sample slice of the public fold-1 train split (same data for every row). Inference time is per sample over a fixed 100-sample val slice (beam-4, with two-batch warm-up before timing). Trainable parameters are reported as LoRA + FiLM + GAT; the GAT is the only component the federation exchanges each round. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019), BART (Lewis et al., 2020), Flan-T5 (Chung et al., 2024), LED (Beltagy et al., 2020), Pegasus-X (Phang et al., 2023), ProphetNet (Qi et al., 2020).

Backbone	Config	Trainable (M params)	Comm/round (M params)	Train/epoch (s)	Train peak mem (GB)	Inference (ms / sample)	Infer peak mem (GB)
BART-base (Lewis et al., 2020)	LoRA-only	0.885	0	6.4	1.77	613.8	0.97
BART-base	LoRA+FiLM+GAT	0.885+0.206+0.088=1.179	0.088	7.8	1.86	484.3	0.97
Flan-T5-base (Chung et al., 2024)	LoRA-only	1.769	0	15.8	5.04	745.9	1.89
Flan-T5-base	LoRA+FiLM+GAT	1.769+0.206+0.088=2.064	0.088	20.1	5.23	901.8	1.89
LED-base-16384 (Beltagy et al., 2020)	LoRA-only	0.590	0	8.3	1.52	671.7	1.24
LED-base-16384	LoRA+FiLM+GAT	0.590+0.206+0.088=0.884	0.088	19.1	4.04	520.9	1.24
Pegasus-X-base (Phang et al., 2023)	LoRA-only	1.769	0	23.2	6.17	882.1	1.78
Pegasus-X-base	LoRA+FiLM+GAT	1.769+0.206+0.088=2.064	0.088	23.7	6.36	805.8	1.79
ProphetNet-large (Qi et al., 2020)	LoRA-only	2.359	0	18.4	8.20	1401.2	2.36
ProphetNet-large	LoRA+FiLM+GAT	2.359+0.273+0.088=2.720	0.088	23.9	8.52	1426.5	2.36

Every model evaluated in this paper, the fine-tuned federated SLMs and all zero-shot LLM baselines, generates QA pairs from the single inference-time prompt below. The three conditioning regimes (Section C.4) differ only in the leading clause; all results in this paper use topic conditioning. The fine-tuned SLMs are trained to emit the plain target template; each zero-shot LLM baseline receives the same prompt followed by a one-line format directive so its free-form output parses into the identical shape.

QA-Generation Prompt

```
# Inference prompt

[baseline conditioning]
Generate a question and answer pair from the following machine learning text:
{context}

[topic conditioning, used for every reported result]
Generate a question and answer pair about "{question_topic}" from the
following machine learning text:
{context}

[bloom conditioning]
Generate a Bloom Level-{bloom_level} ({bloom_verb}) question and answer pair
from the following machine learning text:
{context}

# Target template the fine-tuned SLM learns to emit
Question: {question}
Answer:   {answer}

# Format directive appended for the zero-shot LLM baselines
```

```
# (GPT-4o, GPT-4o-mini, Llama-3.1-8B, Qwen-2.5-7B, Mistral-7B)
Respond with exactly:
Question: <one question>
Answer:   <one answer>
```

Relation to the prompt used to build the reference corpus. The inference prompt described here is deliberately not the prompt used to build the reference QA corpus. The reference QA pairs, which serve both as training targets for the SLMs and as the references for ROUGE-L, BLEU-4, and BERTScore, were produced once per institution by GPT-4o under a much richer structured prompt: it requests five QA pairs per context in a single response, together with a topic label, a Bloom level, a difficulty tag, and an answerability flag for each pair. The inference prompt, by contrast, asks any model under evaluation to return a single QA pair in plain question-then-answer form. Using this simpler inference prompt uniformly at test time keeps the SLM-versus-LLM comparison in Table 34 fair: no model under evaluation, including the GPT-4o row, is ever given the elaborate corpus-creation prompt, so every model is scored on the same task that the SLMs were trained for.

H LLM JUDGE PROMPT

This appendix reproduces verbatim the prompt sent to the GPT-4o-mini LLM QA-quality judge (Section F.3). The context, question, and answer placeholders are filled with the source context and the generated QA pair. The judge is run at temperature 0 and returns a structured response with one integer per rubric axis.

LLM Judge Prompt

```
You are an expert evaluator of educational question-answer pairs.

Given a CONTEXT (educational text) and a generated QUESTION-ANSWER pair, score
    the QA pair on FOUR dimensions, each on an integer 1-5 scale (1 = very poor,
    5 = excellent).

CONTEXT:
"""{context}"""

GENERATED QA:
Question: {question}
Answer: {answer}

Score each dimension. Use the rubrics below - be strict.

1. context_grounding: Is the question answerable using ONLY information in the
    context, and are the facts in the answer supported by the context?
   1 = Question or answer requires information not in the context (off-topic or
    hallucinated).
   3 = Partially grounded - some facts come from the context, but key parts are
    external.
   5 = Fully grounded - both question and answer are entirely supported by the
    context.

2. educational_value: Does this QA pair test meaningful understanding of the
    material?
   1 = Trivial, vague, or pointless (e.g. "What does the text say?", a yes/no
    with no insight).
   3 = Tests basic recall or surface-level facts, but not deeper understanding.
   5 = Tests genuine comprehension, application, analysis, or evaluation of the
    concept.
```

```
3. answer_correctness: Given the context, is the answer factually correct and
    complete?
   1 = Answer is wrong, contradicts the context, or contains hallucinations.
   3 = Answer is partially correct but missing key information or contains minor
    errors.
   5 = Answer is fully correct and complete given the context.

4. answer_relevance: Does the answer actually address the question that was asked
    ?
   1 = Answer ignores or evades the question.
   3 = Answer is tangentially related - touches on the topic but does not
    directly answer.
   5 = Answer directly and clearly addresses what was asked.

Return ONLY a single JSON object in this exact schema, with no extra text or
    markdown:
{
  "context_grounding": <int 1-5>,
  "educational_value": <int 1-5>,
  "answer_correctness": <int 1-5>,
  "answer_relevance": <int 1-5>,
  "justification": "<one or two sentences explaining the lowest-scoring dimension
    >"
}
```

I Hyperparameter Tables

Tables 9 and 10 report every hyperparameter used across the public-testbed runs, applied uniformly to all of them. Per-client backbone specifications are in Table 5; everything else is identical across the three institutions.

For the dataset-statistics side, the per fold sample counts on the public testbed are: $\mathcal{I}_0^{\text{pub}}$ (MIT, BART), fold 1 train/val $313/154$, fold 2 $310/157$, fold 3 $311/156$; $\mathcal{I}_1^{\text{pub}}$ (Stanford, Flan-T5), fold 1 $320/158$, fold 2 $321/157$, fold 3 $315/163$; $\mathcal{I}_2^{\text{pub}}$ (ML papers, LED), fold 1 $2854/1428$, fold 2 $2857/1425$, fold 3 $2853/1429$. Per-client local test sets are $|\mathcal{T}_0|$=80, $|\mathcal{T}_1|$=81, $|\mathcal{T}_2|$=755; the global test pool $\mathcal{T}^{\text{global}}$ is their concatenation (916 evaluation rows over 179 held-out contexts).

J Method Details

J.1 Conceptual overview of the unified federation

Figure 1 summarizes UniLearnLM: the left panel shows the per client architecture and federated training loop, and the right panel shows the comparison behind *Hypothesis 2*, in which a unified federation of small models is placed against a single LLM. Each institution holds an SLM together with its weights and curriculum, all kept strictly private. From that local SLM, the institution extracts once a coarse *architecture graph* where nodes represent transformer blocks, edges represent forward connections, and node features represent block typology and running statistics of the learnable weights. A *GNN based unification encoder* (instantiated as a graph attention network, see §3) reads each client's graph and outputs per-block embeddings that a local FiLM modulator translates into a block level scale and shift on hidden states, so the unified embedding becomes a small residual perturbation of the otherwise unchanged forward pass. Only the unification encoder GNN is federated (sample-size-weighted averaging across rounds); SLM weights, LoRA adapter, FiLM modulator, and curricular data stay client-local, decoupling the federation channel from the language model so any SLM family can plug in under one protocol. The ablation variant UniLearnLM+ adds an ensemble distillation pass over a shared public pool of unlabeled contexts, exposing an accuracy–equity trade-off characterized in Section 5.

Table 9: Public-testbed hyperparameters, part 1: shared, LoRA, individual-training, and Indiv-baseline evaluation settings. Identical across the three clients unless noted; the UniLearnLM-specific settings are in Table 10.

Block	Hyperparameter	Value
Shared	Random seed	42
	Batch size	4
	Max input length	512 tokens
	Max target length	128 tokens
	Optimizer	AdamW (weight decay 0.01)
	LR schedule	cosine with 10% linear warm-up
	Warm-up ratio	0.1
	Gradient clipping (L2 norm)	1.0
LoRA (all clients)	Rank r	16
	Scaling α	32
	Dropout	0.1
	Bias	none
Individual training	Max epochs	100
	Learning rate (LoRA)	3×10^{-4}
	Early-stop patience	10 epochs
	Early-stop $\Delta_{\min}$	10^{-4}
	Early-stop metric	greedy ROUGE-L on 50 fixed val samples
Eval (Indiv, baseline)	Decoding	diverse beam search
	Beams	10
	Beam groups	5
	Diversity penalty	1.0
	Returned sequences / context	5 (Hungarian-matched to refs)
	Eval suite	lightweight subset (R-L, BLEU-4, BERTScore-F1, Bloom-BERT EVS)

J.2 Warm start of the FiLM residual

Initializing $\alpha_k{=}0$ makes the FiLM residual identically zero at step zero, so the federation has no effect on the pretrained backbone at initialization and the model's first-step output is exactly that of the LoRA-adapted SLM alone. The initialization is not a freeze, however: the gradient $\partial\mathcal{L}/\partial\alpha_k$ is non-zero from step one whenever $(\boldsymbol{\gamma}_\ell, \boldsymbol{\beta}_\ell)$ and the language-modeling loss are non-zero, so α_k grows as the GNN based conditioning becomes useful. This is the warm-start mechanism used by ControlNet's "zero convolutions" (Zhang et al., 2023).

J.3 The public pool used by UniLearnLM+ and the distillation baselines

Public pool ablation (UniLearnLM+). UniLearnLM alone uses no public data: every client trains only on its own private curriculum. The UniLearnLM+ variant adds an auxiliary training signal that lets each client learn from the other institutions' generations on a shared public teaching material dataset (the *public pool*, N_{pool} samples, deterministic across folds; Appendix S). The pool in our experiments is not an external corpus. It is peeled from the union of the clients' own training and validation chunks, excluding anything in any client's test set, at roughly 180 rows per fold. Two consequences follow, and we state them here rather than in an appendix. First, this is a deviation from FedMD, which assumes a public dataset independent of the clients' private data. Second, and more importantly for reading Table 2, it means the distillation baselines and UniLearnLM+ condition on raw text drawn from other institutions' curricula while UniLearnLM never does. The comparison therefore gives those baselines strictly more information than UniLearnLM has, which makes their loss on pedagogical depth a stronger result than it would otherwise be, and at the same time means they could not lawfully be run by an institution operating under the privacy constraint this paper starts from. We report them as an upper bound on what a public pool channel could deliver, not as a deployable alternative. At the end of federation round $t{-}1$, every client generates one (q, a) pair per pool sample using its own SLM, and the server concatenates all K clients' generations into one set $\mathcal{P}^{[t-1]}$. In round t, each client then adds a cross-entropy term over $\mathcal{P}^{[t-1]}$ to its local loss:

Table 10: Public-testbed hyperparameters, part 2: the UniLearnLM-specific federated, GAT, FiLM, mixed-precision, and reproducibility settings, continuing Table 9.

Block	Hyperparameter	Value
Federated (UniLearnLM)	Rounds T	33
	Local epochs / round	3
	Learning rate (LoRA)	3×10^{-4}
	Learning rate (GAT)	1×10^{-3}
	Learning rate (FiLM)	1×10^{-3}
	Model-selection metric	avg greedy ROUGE-L on 50 val samples / client, post-aggregation
	FedAvg scope	sample-size-weighted; GAT only (LoRA + FiLM client-local)
GAT (UniLearnLM)	Layers	3
	Hidden	$16 \to 256 \to 256 \to 64$ (input $\to$ hidden $\to$ hidden $\to$ output)
	Heads (layers 1–2, concat)	4
	Heads (layer 3, no concat)	1
	Input feature dimension	16
	Activation / dropout	ELU / 0.1
FiLM (UniLearnLM)	Hidden	128
	Output dimension	$2\, d_{\text{model}}^{(k)}$ per client
	α_{init}	0 (identity at init)
Mixed precision	Base-model dtype	FP32 (frozen)
	LoRA adapter dtype	FP32
	Autocast	BF16 when supported, else FP32
	GradScaler	disabled (BF16 has full FP32 range)
Reproducibility	Global random seed (PyTorch, NumPy, Python)	42
	cuDNN deterministic mode	enabled
	cuDNN benchmark mode	disabled
	Deterministic algorithms enforced	yes (warn-only)
	Workspace-buffer environment override	set to a deterministic value
	Python hash-seed environment override	42

$\mathcal{L}_k^+ = \mathcal{L}_k^{\text{CE}} - \beta \sum_{(c,\tilde{q},\tilde{a}) \in \mathcal{P}^{[t-1]}} \log p_k(\tilde{q}, \tilde{a} \mid c)$, with $\beta = 0.5/K$ so the per-institution weight stays independent of the federation size K. The cost is one extra generation pass per client per round; in exchange, every client receives a cross-client training signal that travels through generated text. The full algorithm, loss derivation, and per round pseudocode (Algorithm 1) are in Appendix K.

K Local Objective and Federated Optimization

Local training objective. Every institution fine-tunes for AQAG with a standard conditional language-modeling loss:

$$\mathcal{L}_k(\boldsymbol{\phi}, \mathbf{L}_k, \boldsymbol{\psi}_k) = -\frac{1}{|\mathcal{X}_k|} \sum_{(c,q,a) \in \mathcal{X}_k} \log p_{(\boldsymbol{\phi}, \mathbf{L}_k, \boldsymbol{\psi}_k)}(q, a \mid c). \tag{2}$$

$\mathbf{W}_{\text{base}}$ is frozen; gradients flow to $\boldsymbol{\phi}$, $\mathbf{L}_k$, and $\boldsymbol{\psi}_k$ jointly through the FiLM hooks. Optimization uses AdamW with a cosine schedule and linear warm-up; mini-batches use BF16 autocast on the LM forward.

Algorithm 1 UniLearnLM: unified federated learning for AQAG.

Require: m institutions; private datasets $\{\mathcal{X}_k\}$; private LoRA $\{\mathbf{L}_k\}$; private FiLM $\{\boldsymbol{\psi}_k\}$; rounds T; initial $\boldsymbol{\phi}^{[0]}$; learning rate η.
Ensure: $\boldsymbol{\phi}^{[T]}$ and per client $(\mathbf{L}_k, \boldsymbol{\psi}_k)$.

for each institution k **do**
  Build $\mathcal{G}_k$ from local backbone (once).
end for
for $t = 1, \ldots, T$ **do**
  Server broadcasts $\boldsymbol{\phi}^{[t-1]}$.
  for each institution k in parallel **do**
    for each local mini-batch $(c, q, a) \in \mathcal{X}_k$ **do**
      Refresh the dynamic node features.
      $\mathbf{Z}_k \leftarrow g(\mathcal{G}_k; \boldsymbol{\phi})$.
      Update FiLM hooks; LM forward + backward.
      AdamW step on $(\boldsymbol{\phi}, \mathbf{L}_k, \boldsymbol{\psi}_k)$.
    end for
    Upload $\boldsymbol{\phi}_{(k)}^{[t]}$.
  end for
  Server: $\boldsymbol{\phi}^{[t]} \leftarrow \sum_k \frac{n_k}{\sum_j n_j} \boldsymbol{\phi}_{(k)}^{[t]}$.
end for
return $\boldsymbol{\phi}^{[T]}, \{(\mathbf{L}_k, \boldsymbol{\psi}_k)\}$.

Round structure. At the start of round t the server broadcasts $\boldsymbol{\phi}^{[t-1]}$. Each institution k runs its local update loop, producing $(\boldsymbol{\phi}_{(k)}^{[t]}, \mathbf{L}_k^{[t]}, \boldsymbol{\psi}_k^{[t]})$, and uploads *only* $\boldsymbol{\phi}_{(k)}^{[t]}$. The LoRA and FiLM updates are stored locally for the next round and are never transmitted.

Server aggregation. The server collects $\{\boldsymbol{\phi}_{(k)}^{[t]}\}_{k=1}^{m}$ and performs sample-size-weighted FedAvg (McMahan et al., 2017):

$$\boldsymbol{\phi}^{[t]} = \sum_{k=1}^{m} \frac{n_k}{\sum_j n_j} \boldsymbol{\phi}_{(k)}^{[t]}, \tag{3}$$

where $n_k = |\mathcal{X}_k^{\text{train}}|$. The broadcast cost per round is $\mathcal{O}(|\boldsymbol{\phi}|)$ floats: the GAT has $\sim 10^5$ parameters, independent of the backbone size. LoRA and FiLM are never aggregated, so heterogeneity in their dimension is irrelevant to the protocol.

Every institution participates in every round; no client subsampling is used.

L Federated Algorithm Pseudocode

Algorithm 1 formalizes one round of the unified federated optimization described in Appendix K: each institution refreshes its architecture-graph node features at every mini-batch, runs the shared GAT to obtain per-block embeddings, updates its private LoRA and FiLM parameters jointly with the local copy of the GAT parameters, and uploads only $\boldsymbol{\phi}$ to the server for sample-size-weighted FedAvg aggregation.

M Evaluation Metric Formulations

This appendix gives the formal definition of every QA-quality metric used in the paper, together with the intuition behind it, the exact computation, and any deviation from the original published formulation. Equity-metric derivations (PES, KTI) are in Appendix R.

Throughout, c denotes a context (a teaching material chunk on which the SLM is conditioned), q the question the model generates from c, a the answer the model generates jointly with q, and a_r the gold reference answer paired with the context in the dataset's reference QA file.

ROUGE-L (Lightweight, reference-based). ROUGE-L (Lin, 2004) measures the longest common subsequence (LCS) between the generated text and the reference text. Let g be the generated answer string, r the reference string, and $\mathrm{LCS}(g, r)$ the length of their longest common subsequence at the token level. The metric returns the F1 of LCS-derived precision and recall:

$$P_{\mathrm{LCS}} = \frac{\mathrm{LCS}(g,r)}{|g|}, \quad R_{\mathrm{LCS}} = \frac{\mathrm{LCS}(g,r)}{|r|}, \\ \text{ROUGE-L} = \frac{(1+\beta^2)\, P_{\mathrm{LCS}} R_{\mathrm{LCS}}}{R_{\mathrm{LCS}} + \beta^2 P_{\mathrm{LCS}}}, \tag{4}$$

with $\beta = 1$ in our setup (F1). ROUGE-L is high when the candidate and the reference share long sequences of tokens in the same order, regardless of intervening words.

BLEU-4 (Lightweight, reference-based). BLEU-4 (Papineni et al., 2002) averages modified n-gram precisions for $n \in \{1, 2, 3, 4\}$, multiplied by a brevity penalty BP that prevents the metric from rewarding very short outputs:

$$\text{BLEU-4} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{4} \tfrac{1}{4} \log p_n \right), \tag{5}$$

where p_n is the clipped count of n-grams in g that also appear in r divided by the total n-grams in g, and $\mathrm{BP} = \min(1, \exp(1 - |r|/|g|))$. We use the NLTK sentence-level implementation with the standard smoothing of Chen & Cherry (2014) (method 1). High BLEU-4 indicates token-level overlap with the reference at multiple n-gram orders.

BERTScore-F1 (Lightweight, reference-based). BERTScore (Zhang et al., 2020) replaces hard token matching with greedy matching in a contextual-embedding space. We use a DistilBERT encoder $\mathrm{enc}(\cdot)$ to embed each token of g and r, then for every token $g_i \in g$ pick its best-matching token $r_{j^\star} \in r$ by cosine similarity:

$$P_{\mathrm{B}} = \tfrac{1}{|g|} \sum_{g_i \in g} \max_{r_j \in r} \cos(\mathrm{enc}(g_i), \mathrm{enc}(r_j)), \\ R_{\mathrm{B}} = \tfrac{1}{|r|} \sum_{r_j \in r} \max_{g_i \in g} \cos(\mathrm{enc}(g_i), \mathrm{enc}(r_j)), \tag{6}$$

and report the F1, $\text{BERTScore-F1} = 2P_{\mathrm{B}}R_{\mathrm{B}}/(P_{\mathrm{B}} + R_{\mathrm{B}})$. BERTScore is high when the candidate and the reference are *semantically* similar even when they share few exact tokens.

RQUGE (Question quality, reference-free). RQUGE (Mohammadshahi et al., 2023) scores how well a generated question can be answered from the supplied context, without ever comparing the generated question to a gold question. Given the candidate question q, the context c, and the reference answer a_r already paired with this context, RQUGE first asks a QA model to read the context and answer the candidate question, then measures how close that QA-predicted answer is to the reference answer:

$$a_c = \mathrm{QA}(q, c), \qquad \kappa = S(q, a_c, a_r, c), \tag{7}$$

where $\mathrm{QA}(\cdot)$ is UnifiedQA-v2 (Khashabi et al., 2022), a_c is the answer extracted by the QA model from the context, a_r is the gold reference answer from the dataset, and $S(\cdot)$ is a span-quality scorer that returns a value in $[1, 5]$. The intuition: if the QA model can recover the reference answer from c using only q, the question is well-formed and answerable from the context; if not, the question is either unanswerable, off-topic, or ambiguous. The original paper trains $S(\cdot)$ on the MOCHA dataset (Chen et al., 2020); we approximate it with BERTScore-F1 to keep the metric cheap and dependency-free:

$$\kappa = 1 + 4 \cdot \text{BERTScore-F1}(a_c, a_r), \tag{8}$$

clamped to $[1, 5]$. This approximation preserves the ordinal interpretation (higher means more answerable) but is not numerically identical to the original MOCHA-trained scorer; the values reported in our tables therefore lie on the same ordinal scale as the paper but are calibrated by BERTScore rather than MOCHA.

RAGAS Faithfulness (Answer quality, reference-free). RAGAS Faithfulness (Es et al., 2024) measures whether the generated answer is grounded in the supplied context, that is, whether every factual claim in the answer can be inferred from c alone. The metric proceeds in two stages. First, the answer a is decomposed into a set of atomic factual claims $\mathcal{S}(a) = \{s_1, \ldots, s_m\}$; in the original paper this decomposition uses GPT-3.5/4. In our implementation, the decomposer is GPT-4o-mini when an OpenAI key is supplied, and otherwise falls back to NLTK sentence tokenization (each sentence is treated as one claim). Second, each claim s_i is checked against the context with a natural language inference (NLI) model: the pair (c, s_i) is fed to the entailment classifier and we record whether the predicted label is entailment (as opposed to neutral or contradiction). The faithfulness score is the fraction of claims that the NLI model judges entailed by the context:

$$\begin{aligned}&\text{Faithfulness}(a, c)\\ &\quad= \frac{1}{|\mathcal{S}(a)|} \sum_{s_i \in \mathcal{S}(a)} \mathbf{1}[\text{NLI}(c, s_i) = \text{ENT}].\end{aligned} \tag{9}$$

Range $[0, 1]$; higher means more answer content is supported by the context (fewer hallucinations). Our NLI model is cross-encoder/nli-deberta-v3-small (He et al., 2021).

QAFactEval (Answer quality, reference-free). QAFactEval (Fabbri et al., 2022) is the original QA-based factual-consistency metric for summarization: it generates QA pairs from the candidate text using a question-generation model, runs a question-answering model on the source document for each generated question, and measures answer-overlap with the candidate text via LERC (Chen et al., 2020). Our setting evaluates an answer rather than a summary, so we use a simplified UnifiedQA-based variant. For each sentence-level claim s_i in the answer a we form a yes/no probe "Is it true that: s_i?" and run UnifiedQA on (probe, c):

$$\begin{aligned}\text{QAFactEval}(a, c) &= \frac{1}{|\mathcal{S}(a)|} \sum_{s_i \in \mathcal{S}(a)} y_i,\\ y_i &= \mathbf{1}\big[\text{“yes”} \in \text{QA}(p_i, c)\big].\end{aligned} \tag{10}$$

Range $[0, 1]$; higher means more claims in the answer are confirmed by the context.

Round-Trip Consistency (Joint QA quality, reference-free). Round-Trip Consistency (RTC) (Alberti et al., 2019) tests whether the answer the candidate produced can be recovered by a QA model that is given only the candidate's question and the context. The original RTC uses exact-match between the QA model's output and the candidate answer:

$$\text{RTC}(q, a, c) = \text{F1}_{\text{tok}}\big(\text{QA}(q, c),\ a\big), \tag{11}$$

where QA is UnifiedQA-v2 and F1_{tok} is token-level F1 between the QA model's output and the candidate's generated answer. We replace the original exact-match with token-level F1 to obtain a continuous score that distinguishes near-correct answers from completely wrong ones. Range $[0, 1]$; higher means the candidate question and answer are mutually consistent (the question can recover the answer from the context).

RAGAS Answer Relevancy (Joint QA quality, reference-free). RAGAS Answer Relevancy (Es et al., 2024) measures whether the answer actually addresses the question that was asked. The metric reverse-generates n candidate questions from the answer a using an LLM (we use GPT-4o-mini with $n = 3$), embeds the original question q and each reverse-generated question $\hat{q}_i$ with a sentence embedding model (we use OpenAI `text-embedding-3-small`), and averages the cosine similarities:

$$\text{AR}(q, a) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{e}_q^\top \mathbf{e}_{\hat{q}_i}}{\|\mathbf{e}_q\| \, \|\mathbf{e}_{\hat{q}_i}\|}, \tag{12}$$

where $\mathbf{e}_x$ is the embedding of text x. The intuition: if the answer can stand in for a paraphrase of the original question, the original and reverse-generated questions will embed close together. Range typically $[0, 1]$; higher means the answer is relevant to the question that was asked.

Bloom's BERT Classifier (Pedagogical quality). A BERT-base classifier (`cip29/bert-blooms-taxonomy-classifier`), fine-tuned on Bloom's-labeled question data, takes the generated question q as input and predicts a probability over the six levels of Bloom's Revised Taxonomy (Anderson et al., 2001):

$$\hat{b}(q) = \underset{b\in\{1,\dots,6\}}{\arg\max}\ \mathrm{softmax}(W \cdot \mathrm{BERT}(q))_b, \tag{13}$$

where $\mathrm{BERT}(q)$ is the [CLS] representation of q and W is the classifier head. The predicted Bloom level is then normalized to a continuous *Educational Value Score* in $[0, 1]$:

$$\mathrm{EVS}(q) = \frac{\hat{b}(q) - 1}{5} \in [0, 1]. \tag{14}$$

Level 1 (Remember) maps to 0, Level 6 (Create) maps to 1. A higher EVS means the classifier judges the question to test deeper cognitive skills.

Bloom's LLM judge (Pedagogical quality). As an LLM-based complement to the BERT classifier we also query GPT-4o-mini with a structured prompt that defines each of the six Bloom levels with verb anchors (e.g. "Remember: define, list, state"; "Create: design, propose, formulate") and asks the judge to return, for the question q, the predicted Bloom level b together with a one-sentence justification. The predicted level is normalized to EVS exactly as in Eq. (14). We report the BERT classifier and LLM judge separately (columns $\mathrm{B}_{\mathrm{cls}}$ and $\mathrm{B}_{\mathrm{LLM}}$) so that the reader can see how strongly the two pedagogical-depth proxies agree.

Four-axis LLM judge (overall QA quality). For each (c, q, a) triple we also query GPT-4o-mini with a strict 1–5 rubric on four orthogonal axes: context grounding, educational value, answer correctness, answer relevance (full prompt in Appendix H). The per-axis scores $\{s_g, s_e, s_c, s_r\} \in \{1, \dots, 5\}^4$ are averaged and rescaled to $[0, 1]$:

$$\text{LLM-J}(c, q, a) = \frac{\bar{s} - 1}{4}, \quad \bar{s} = \frac{s_g + s_e + s_c + s_r}{4}, \tag{15}$$

where $\bar{s} = \frac{1}{4}(s_g + s_e + s_c + s_r)$. LLM-J is the headline "overall" column and is also the per sample quantity from which PES and KTI are computed.

N Public Testbed Reproducibility

This appendix documents the artifacts needed to reproduce the public testbed exactly as we use it. We will release four components: (i) the chunked plain-text source files derived from the MIT 18.657, Stanford CS229, and arXiv ML paper corpora; (ii) the QA annotations generated by GPT-4o, where each QA pair carries a topic label, a Bloom's-Taxonomy level (integer 1–6), a difficulty tag, and an answerability flag; (iii) the 3-fold cross-validation splits, produced deterministically from a fixed random seed (seed $= 42$); and (iv) an MD5 checksum for every split file, so that any party regenerating the splits can verify they match ours byte-for-byte.

Table 11 reports the full composition of the public testbed.

Table 11: Public-testbed composition. Chunk and QA-pair counts are the full per-institution totals; train and val are mean per fold QA counts derived from the chunk-level 3-fold CV on the 85% non-test split (seed $= 42$), and *local test* is the 15% chunk-level holdout. The global test pool is the union of the three local test sets. Sources: BART (Lewis et al., 2020), Flan-T5 (Chung et al., 2024), LED (Beltagy et al., 2020).

	Source	Content	Chunks	QA	Backbone	Tr/Val	Test
$\mathcal{I}_0^{\mathrm{pub}}$	MIT 18.657 (Massachusetts Institute of Technology, 2015)	Lecture notes: statistical learning theory and optimization	110	547	BART-base (Lewis et al., 2020)	311 / 156	80
$\mathcal{I}_1^{\mathrm{pub}}$	Stanford CS229 (Stanford University, 2024)	Undergraduate ML course notes: supervised/unsupervised learning, neural nets, RL	110	559	Flan-T5-base (Chung et al., 2024)	319 / 159	81
$\mathcal{I}_2^{\mathrm{pub}}$	arXiv ML papers	Curated recent ML preprints: vision, NLP, RL	981	5,037	LED-base (Beltagy et al., 2020)	2,855 / 1,427	755
Global test pool (union of local tests)			179	916			

The MD5 checksum manifest is released with the artifacts so that any third party can verify their regenerated splits bit-for-bit against ours.

O ARCHITECTURE GRAPHS

The architecture graph our constructor builds places one node per meaningful sub-module of the backbone (every attention projection, layer norm, dense layer, plus the embedding and language model head modules), so its size scales with the sub-module count of the SLM: $|\mathcal{V}_k| \in [228, 499]$ across the five SLM families (BART-base 236, Flan-T5-base 486, LED-base-16384 228, Pegasus-X-base 461, ProphetNet-large 499), with the chain-of-blocks edge topology described in Section 3. Node feature vectors are 16-dimensional, and edges connect consecutive sub-modules in forward order with self-loops.

Table 12 expands each feature index. Thirteen of the sixteen entries are *static*: they depend only on the SLM's architecture and are fixed at graph-construction time. The three *dynamic* entries (indices 6–8) re-compute the mean, std, and ℓ_2-magnitude of the node's own weight matrix at every refresh. The LoRA factor matrices $\mathbf{A}^{(k,\ell)}$ and $\mathbf{B}^{(k,\ell)}$ are nodes of their own, so their entries follow the adapters as they train; a projection node carries the statistics of its frozen pretrained weight. These are the only features that change during training. They enter the pooled graph embedding directly; a block's own embedding sees them only when the adapter nodes lie within the GAT's receptive field of that block's node, which the three-layer GAT does not reach on the sequential graphs of BART and LED, so the per-block modulation of those clients depends on the architecture graph and the federated GAT weights. Removing features 6–8 reduces the GAT to a purely architectural prior.

Table 12: Per-index breakdown of the 16-dimensional node feature vector $\mathbf{x}_\ell \in \mathbb{R}^{16}$ attached to each architecture-graph node (one per SLM sub-module). **Type** indicates whether the feature is fixed at graph-construction time (*static*) or refreshed every mini-batch from the node's current weight matrix (*dynamic*); for the LoRA factor nodes this matrix is the adapter factor itself. The three dynamic features are the only features that change during training.

Idx	Name	Type	Computation	Semantic
0	`type_id`	static	layer-class integer id (Linear, Attn, LN, Emb, …; vocab size 28)	module class identity
1	`n_params`	static	$\lvert\text{params}(\ell)\rvert/10^6$	per-block parameter count (M)
2	`feat_in`	static	`in_features`/1024 (Linear); `embed_dim`/1024 (Attn); $d_{\text{model}}/1024$ (LN/Emb)	input dimensionality
3	`feat_out`	static	`out_features`/1024 (Linear); `num_heads`/1024 (Attn); $d_{\text{ff}}/1024$ (FF)	output dim. or head count
4	`depth_ratio`	static	$\ell \,/\, \max(\lvert\mathcal{V}_k\rvert - 1,\ 1)$	normalized position $\in [0, 1]$
5	`is_encoder`	static	$\mathbb{1}[\text{“decoder”} \notin \text{name}(\ell)]$	encoder/decoder flag
6	`mean_w`	**dynamic**	mean of the node's weight matrix $\mathbf{W}_\ell$ ($\mathbf{A}$ or $\mathbf{B}$ for LoRA factor nodes)	weight mean
7	`std_w`	**dynamic**	standard deviation of $\mathbf{W}_\ell$	weight std
8	`norm_w`	**dynamic**	$\lVert\mathbf{W}_\ell\rVert_2 / \sqrt{\text{numel}(\mathbf{W}_\ell)}$	weight magnitude
9	`is_attn`	static	$\mathbb{1}$[attention-class]	module-type one-hot
10	`is_ff`	static	$\mathbb{1}$[feed-forward-class]	module-type one-hot
11	`is_norm`	static	$\mathbb{1}$[LayerNorm/T5LayerNorm]	module-type one-hot
12	`is_emb`	static	$\mathbb{1}$[Embedding]	module-type one-hot
13	`has_bias`	static	$\mathbb{1}$[module exposes a non-null `bias`]	bias-presence flag
14	`rel_size`	static	$\lvert\text{params}(\ell)\rvert \,/\, \sum_{\ell'} \lvert\text{params}(\ell')\rvert$	relative-size weight
15	`reserved`	static	0	placeholder (future use)

P COMPUTATIONAL COST

Table 8 reports per client compute under the two configurations our methods occupy: **LoRA-only** (used by Indiv, TC-Best, TC-Ens, and Centralized) and **LoRA + FiLM + GAT** (used by UniLearnLM and UniLearnLM+). All numbers were collected on a single accelerator with BF16 autocast, batch size 4, input/target sequence length 512→128, and beam size 4. Training time is wall-clock for one epoch over a fixed 313-sample slice of the public fold-1 train split, identical across rows. Inference time is per sample over a fixed 100-sample slice (beam-4, with two batches of warm-up before timing).

Trainable parameters. The GAT-conditioning channel adds a $+15\%$ to $+50\%$ trainable-parameter footprint relative to LoRA-only, depending on backbone d_{model}. The shared GAT itself has only 0.088M parameters and is the only component exchanged between clients and the server each round.

Per-mini-batch GAT overhead. UniLearnLM runs the shared GAT once per mini-batch to refresh the per-block conditioning. With $\sim 10^5$ parameters and a graph of at most 499 nodes, its forward pass is negligible relative to the SLM forward/backward over a 512-token batch.

Communication cost per round. Each client uploads only the 0.088M-parameter GAT and downloads the aggregated GAT of identical size, 7 to 27× smaller than the LoRA adapters a weight sharing federation would exchange (Table 8), and three to four orders of magnitude smaller than the backbones themselves.

Q Qualitative Examples

To make the per-cell numbers in the result tables concrete, we present three side-by-side QA-generation examples drawn from the public global test set (fold 1). Two of them (examples 703 and 245) originate from institution $\mathcal{I}_2^{\text{pub}}$ (arXiv ML-papers, LED backbone), the most technically demanding curriculum, on which UniLearnLM shows its largest gains (§5); the third (example 74) comes from institution $\mathcal{I}_0^{\text{pub}}$ (MIT OpenCourseWare, BART backbone), so the set spans two of the three public institutions and both backbones. For each context we show the gold reference QA pair alongside the QA pairs produced by local fine-tuning (**Indiv**), selective-consensus distillation (**TC-Best**), and UniLearnLM, together with the three per-example metrics: the LLM-judge overall score ($\text{J} \in [0, 1]$), educational value ($\text{EV} \in \{1, \ldots, 5\}$), and Bloom level ($\text{BL} \in \{1, \ldots, 6\}$).

Table 13 illustrates pedagogical depth. The context describes an interior optimum in bound width for uncertainty-adjusted sorting. **Indiv** and **TC-Best** both ask a shallow recall question (Bloom level 1) and copy a fragment of the source text, scoring 0.500 and 0.625. UniLearnLM poses an Analyze-level question (Bloom level 4) about how the optimum varies with model flexibility and answers it with the correct mechanism, earning the maximum educational value of 5 and a judge score of 1.000. Table 16 illustrates a coherence failure. The context names gradient descent as a common optimization method, but **Indiv** produces a circular, tautological answer ("gradient descent is used to optimize gradient descent") and scores 0.000, while **TC-Best** gives a vague non-answer that never names the method. UniLearnLM instead produces a coherent answer that correctly identifies the method present in the context ($\text{J} = 0.875$). Table 15 illustrates a factual-accuracy failure on a citation list. **Indiv** and **TC-Best** both hallucinate incorrect publication years, one for a paper that does not even appear in the excerpt, whereas UniLearnLM asks about the Shaw & Gopaladesikan citation and recovers the correct year (2020) directly from the listing ($\text{J} = 0.875$).

Failure-mode analysis. Categorizing the per sample LLM-judge justifications of the lowest-scoring generations reveals three recurrent failure classes: (i) *degenerate or echo answers* that restate the question without adding content; (ii) *ungrounded answers* that assert facts absent from, or contradicted by, the context; and (iii) *topic drift*, where the QA pair is well-formed but does not address the conditioning topic. All paradigms exhibit all three classes; the side-by-side examples above include representative instances.

R Equity-Metric Derivations and Implementation

We provide derivations and implementation notes for the two equity metrics in Section F.4: the *Pairwise Equity Score* $E_{kl}^{(m)}$ (Eq. 16) and the *Knowledge Transfer Index* $\text{KTI}(l \leftarrow k)$ (Eq. 17).

Pairwise Equity Score (PES). The Pairwise Equity Score asks the following question: *how good is a pair of institutions (k, l) at QA generation, relative to the gap in resources that the federation has to bridge between them?* A federation that lifts the QA quality of two institutions with very different curricula and very different model sizes earns a higher PES than the same numeric quality lift achieved between two institutions that already start from comparable resources.

For any learning paradigm $m \in \{\text{Indiv}, TC-Best, TC-Ens, UniLearnLM, UniLearnLM+\}$ and any two institutions k, l:

$$E_{kl}^{(m)} = \frac{Q_k^{(m)} + Q_l^{(m)}}{|\mathcal{C}_k - \mathcal{C}_l| + |\mathcal{R}_k - \mathcal{R}_l| + \varepsilon}. \tag{16}$$

The numerator is the summed QA quality of the two institutions under method m ($Q_k^{(m)}$ is institution k's aggregate QA-quality score; we use the LLM-judge overall mean by default, though any of the

	Question / Answer	J	EV	BL
Context (excerpt): "… These opposing forces generate an interior optimum in bound width rather than a monotonic relationship between adjustment strength and performance. The location of this optimum varies systematically with model flexibility: models with greater estimation uncertainty benefit more from moderate adjustment, whereas already stable models exhibit smaller gains and are more sensitive to over-adjustment. …"				
Gold	*Q:* What is the optimum bound width's relationship with model flexibility? *A:* The location of the optimum bound width varies systematically with model flexibility; models with greater estimation uncertainty benefit more from moderate adjustment.	–	–	–
Indiv	*Q:* What is the optimal bound width for effective uncertainty-adjusted sorting? *A:* The optimum bound width is defined as the location of the optimal optimum in bound width, which varies systematically with model flexibility.	0.500	3	1
TC-Best (Li & Wang, 2019)	*Q:* What is the optimal bound width for effective uncertainty-adjusted sorting? *A:* The optimum bound width is not so wide that the ranking signal itself is diluted.	0.625	3	1
UniLearnLM	*Q:* How does the optimal bound width in uncertainty-adjusted sorting vary with model flexibility? *A:* Models with greater estimation uncertainty benefit more from moderate adjustment, whereas already stable models exhibit smaller gains and are more sensitive to over-adjustment.	**1.000**	**5**	**4**

Table 13: Generated QA pairs for global-test example 703 (public testbed, fold 1, institution $\mathcal{I}_2^{\text{pub}}$). **Indiv** and **TC-Best** ask a shallow recall question (Bloom level 1), whereas UniLearnLM poses an Analyze-level question (Bloom level 4) and answers it with the correct mechanism. J = LLM-judge overall ($[0, 1]$), EV = educational value (1–5), BL = Bloom level (1–6); all are computed per example. The context is truncated for display. The TC-Best examples are from the run before the consensus-parser fix. BL is the level assigned by the BERT Bloom classifier (B_{cls}). Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019).

metrics in Tables 2–29 can be plugged in). The denominator is the absolute resource gap between the two institutions along two axes:

- $\mathcal{C}_k$ is the *curricular richness* of institution k, a proxy combining how many context chunks the institution holds and how many distinct question topics those chunks cover: $\mathcal{C}_k = \frac{1}{2}(\hat{N}_k + \hat{D}_k)$, where $\hat{N}_k$ is the chunk count and $\hat{D}_k$ is the distinct-topic count, each normalized to $[0, 1]$ by the federation maximum.
- $\mathcal{R}_k$ is the *model capacity* of institution k, taken as the total parameter count of its deployed SLM, normalized to $[0, 1]$ by dividing by the largest backbone in the federation.

The constant $\varepsilon = 10^{-6}$ prevents division by zero when two institutions happen to be matched on both axes. The per-institution mean PES is $\bar{E}_k^{(m)} = \frac{1}{m-1} \sum_{l \neq k} E_{kl}^{(m)}$, the average over institution k's pairs; the unweighted all pairs mean reported as $\bar{E}$ in Table 7 averages $E_{kl}^{(m)}$ over every institution pair.

PES is a presentational lens, not an independent measurement. For a fixed pair (k, l), the denominator $|\mathcal{C}_k - \mathcal{C}_l| + |\mathcal{R}_k - \mathcal{R}_l| + \varepsilon$ depends only on institution attributes and is therefore the same constant under every paradigm m. Hence $E_{kl}^{(m)}$ is a fixed positive rescaling of $Q_k^{(m)} + Q_l^{(m)}$, and ranking the methods by E_{kl} is identical to ranking them by pair-summed quality. PES adds no evidence beyond the per-method quality scores; what it does add is a resource-gap normalization that makes the same quality numbers comparable across pairs with different $(\mathcal{C}, \mathcal{R})$ gaps. It should be read as "how much

quality does method m buy per unit of resource gap", not as cross-validation of the underlying quality ranking.

Knowledge Transfer Index (KTI). The Knowledge Transfer Index asks the following question: *when institution l joins the federation, does it gain quality on topics that only its peer k has seen?* The metric isolates the cross institution transfer component of federated learning by restricting evaluation to a subset of held-out contexts that institution l could not have learned about locally, then comparing l's federated quality on that subset against its individual-training quality on the same subset.

Formally, for a directed pair $(l \leftarrow k)$ and any federated method m:

$$\mathrm{KTI}(l \leftarrow k;\, m) = Q_l^{(m)}(\mathcal{T}_{k\setminus l}) - Q_l^{(\mathrm{Indiv})}(\mathcal{T}_{k\setminus l}). \tag{17}$$

The two terms are evaluated on the same held-out subset $\mathcal{T}_{k\setminus l}$, defined as the union of context chunks whose topic label appears in institution k's training data but not in institution l's training data. These are precisely the topics that institution l never saw during its own local fine-tuning; if its federated SLM does better on them than its individual-training SLM, that improvement must have come from the federation channel rather than from l's own data. A positive KTI value therefore indicates that information from k's curriculum has flowed across the federation and is now lifting l's answers on k-exclusive topics, with magnitude reflecting how much. A KTI near zero indicates the federation did not transfer useful k-specific knowledge to l; a negative KTI indicates the federation hurt l on those topics relative to standalone training.

Failure modes. Each equity metric has a degenerate regime. PES is undefined when both $|\mathcal{C}_k - \mathcal{C}_l|$ and $|\mathcal{R}_k - \mathcal{R}_l|$ approach zero; the ε term keeps it finite but very large, so for nearly-matched pairs we report $\bar{E}$ but not the individual E_{kl}. KTI requires $\mathcal{T}_{k\setminus l}$ to be non-empty, and is reported as "–" when two institutions cover the same topics. KTI also loses resolution as the curricula converge. On the public testbed the three institutions all teach machine learning from overlapping source material, so $\mathcal{T}_{k\setminus l}$ is small and the resulting KTI magnitudes are correspondingly small; we therefore do not rest the equity claim on the public-testbed KTI values, and treat them as a consistency check on the sign rather than as an effect size. The private testbed is where KTI is informative, because the five curricula are genuinely disjoint in parts, and the full-participation run of Appendix AE is where it is measured cleanly, because every receiver–source pair is defined for every method there. A further caveat applies to both testbeds: KTI is computed from the LLM-judge per sample score, so it inherits that judge's limitations (Appendix AD), which is one more reason to read it as a directional rather than an absolute quantity.

S Distillation-Based Baseline Implementations

Distillation pool on the private testbed. The distillation pool of the live private federations is described in Appendix C.2.

We re-implement the two distillation-based *Hypothesis 1* baselines, TC-Best and TC-Ens, both as text space public pool variants in the FedMD (Li & Wang, 2019) family (one selective, one ensemble), within the same codebase as UniLearnLM, so they share tokenization, splitting, and evaluation. In both, each client's SLM is the same as in UniLearnLM: a frozen backbone with a trainable LoRA adapter on the attention projections.

Shared public dataset. Both baselines distill over an unlabeled public dataset. We use the public-testbed chunks (Section C.1) as that dataset: each client generates QA pairs on these chunks and exchanges only the generated text, never raw curricular material. The public testbed thus serves a dual role, as one of our experimental settings, and as the distillation surface for the two baselines.

TC-Best: selective text-consensus on a public pool. TC-Best is a selective variant of FedMD-family public pool distillation. Each round, every client generates one (q, a) pair on every public pool context with its current local SLM; the server picks a single consensus generation per context (the peer generation with the highest mean pairwise ROUGE-L F1 to its peers); each client then distils toward that selected consensus in the next round. Concretely, in each round t: (i) every client trains its local LoRA-adapted SLM for E=3 epochs on its private data (standard CE loss, weight

Table 14: Pedagogical and LLM-judge metrics on the **public testbed**, per institution (validation-selected model, mean over three folds). Column abbreviations follow Table 2: B_{cls} and B_{LLM} are EVS scores in $[0, 1]$ (Eq. 14); LLM-J is the overall mean across the four rubric dimensions, also in $[0, 1]$. Method rows: **Indiv** (Section C.5), **TC-Best** and **TC-Ens**, both public pool variants of FedMD (Li & Wang, 2019), and UniLearnLM. The TC-Best and TC-Ens rows are the rerun of Table 2; B_{LLM} and LLM-J were not computed for that rerun (−). Bold = best per (institution, metric) among the available cells; cells are shaded by rank within each institution group per column: darkest green = best, then 2nd, 3rd.

Institution	Method	B_{cls}	B_{LLM}	LLM-J
$\mathcal{I}_0^{pub}$	Indiv	**0.2384**	0.3079	0.6716
	TC-Best (Li & Wang, 2019)	0.1514	–	–
	TC-Ens (Li & Wang, 2019)	0.1539	–	–
	UniLearnLM	0.2183	**0.3111**	**0.6801**
$\mathcal{I}_1^{pub}$	Indiv	0.2718	**0.2397**	**0.7138**
	TC-Best	0.1464	–	–
	TC-Ens	0.1584	–	–
	UniLearnLM	**0.2768**	0.2361	0.7122
$\mathcal{I}_2^{pub}$	Indiv	0.1611	0.2787	0.6893
	TC-Best	0.1891	–	–
	TC-Ens	0.1841	–	–
	UniLearnLM	**0.1989**	**0.2880**	**0.7051**

	Question / Answer	J	EV	BL
Context (excerpt): "[25] Laurie Shaw and Sudarshan Gopaladesikan. Routine inspection: A playbook for corner kicks. In … MLSA 2020 … Springer, 2020. [26] Pascal Bauer, Gabriel Anzer, and Joshua Wyatt Smith. Individual role classification for players defending corners in football (soccer). Journal of Quantitative Analysis in Sports, 18(2):147–160, 2022. …"				
Gold	*Q:* In which year was the work by Pascal Bauer, Gabriel Anzer, and Joshua Wyatt Smith published? *A:* 2022.	–	–	–
Indiv	*Q:* What year was the paper by Laurie Shaw and Sudarshan Gopaladesikan published? *A:* The paper was published in 2017.	0.375	3	1
TC-Best (Li & Wang, 2019)	*Q:* In which year was the paper on modeling defensive dynamics in football published? *A:* The paper was published in 2023.	0.250	1	1
UniLearnLM	*Q:* What year was the paper 'Routine inspection: A playbook for corner kicks' published? *A:* The paper was published in 2020.	**0.875**	**3**	1

Table 15: Generated QA pairs for global-test example 245 (public testbed, fold 1, institution $\mathcal{I}_2^{pub}$, LED backbone). The context is a reference list; **Indiv** asks about the Shaw & Gopaladesikan citation but reports the wrong year (the listing clearly shows 2020), **TC-Best** asks about a paper that does not appear in the excerpt and gives an unsupported year, whereas UniLearnLM asks about the Shaw & Gopaladesikan paper and recovers the correct publication year (2020) directly from the citation. J = LLM-judge overall ($[0, 1]$), EV = educational value (1–5), BL = Bloom level (1–6); all are computed per example. The context is truncated for display. The TC-Best examples are from the run before the consensus-parser fix. BL is the level assigned by the BERT Bloom classifier (B_{cls}). Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019), LED (Beltagy et al., 2020).

	Question / Answer	J	EV	BL
Context (excerpt): "Gradient descent is a common optimization method. Thus, using Mirror descent would be most beneficial. Other Potential Functions: There are other potential functions which are strongly convex with respect to the ℓ_1 norm. . . ."				
Gold	*Q:* What optimization method is mentioned as being commonly used? *A:* Gradient descent is mentioned as a common optimization method.	–	–	–
Indiv	*Q:* What is the common optimization method used in the context of gradient descent? *A:* Gradient descent is used to optimize gradient descent.	0.000	1	1
TC-Best (Li & Wang, 2019)	*Q:* What is the optimization method mentioned in the context? *A:* The optimization method is used to optimize potential functions.	0.500	3	1
UniLearnLM	*Q:* What is the common optimization method used to optimize the potential functions? *A:* The technique used is Gradient descent.	**0.875**	**3**	1

Table 16: Generated QA pairs for global-test example 74 (public testbed, fold 1, institution $\mathcal{I}_0^{\text{pub}}$, BART backbone). **Indiv** produces a circular, tautological answer ("gradient descent is used to optimize gradient descent") and scores 0.000; **TC-Best** gives a vague non-answer that never names the method; whereas UniLearnLM produces a coherent answer that correctly names the method present in the context. J = LLM-judge overall ($[0, 1]$), EV = educational value (1–5), BL = Bloom level (1–6); all are computed per example. The context is truncated for display. The TC-Best examples are from the run before the consensus-parser fix. BL is the level assigned by the BERT Bloom classifier (B_{cls}). Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019), BART (Lewis et al., 2020).

1.0) and, for $t > 1$, on the previous round's consensus QAs over the public pool (CE loss, weight β=0.5); (ii) each client generates one (q, a) pair per public pool row (deterministic beam search, beam=4, max_new_tokens=128); (iii) the server picks per-row consensus as the candidate with the highest mean pairwise ROUGE-L F1 to peers (with K=3 clients, this is the mean of $K-1$=2 peer scores); (iv) the consensus QAs are broadcast back and used as the public-CE supervision in round t+1. Federation cadence (T=33 rounds, E=3 local epochs) and backbone hyperparameters match UniLearnLM. The public pool holds $\approx$ 180 rows per fold, 60 per source institution, peeled from the union of all clients' train and val rows whose context is in no client's test set. The selective vs ensemble contrast against TC-Ens below mirrors the codistillation literature's "single best peer" vs "averaged ensemble" axis; we cite FedKD (Wu et al., 2022) in Section 2 only as part of the broader distillation-based FL landscape and do not claim its mentor/mentee FedAvg mechanism as our ancestor.

TC-Ens (adapted from FedMD (Li & Wang, 2019)): ensemble distillation adaptation. TC-Ens shares the round structure and public pool of the TC-Best adaptation above, and differs only in the consensus rule. Where TC-Best selects the single per-context peer generation with the highest mean pairwise ROUGE-L F1, TC-Ens distils each client toward the *whole* ensemble: in round t every client generates one (q, a) pair per public pool context, and for $t > 1$ each peer generation enters the client's next-round objective as a cross-entropy term, weighted so the ensemble contributes a combined β=0.5 against the unit-weight private-data loss. This is the text-generation form of FedMD's client-side consensus distillation. The two baselines thus contrast the same public pool under *selective* (TC-Best) versus *ensemble* (TC-Ens) consensus, and reuse the same per fold public pool.

Implementation footprint. Both baselines share UniLearnLM's per fold local-training cost (Appendix P); their additional cost is one beam-search generation pass per client over the $\approx$ 180-row public pool each round, which is small relative to the LM training forward/backward passes.

T GENERIC BENCHMARK: SEEDS, CLIENTS AND SENSITIVITY TO MODEL SELECTION

This appendix tables the generic-benchmark numbers that Section 5.1 and the Limitations quote. Table 17 gives the three-seed means for local training and UniLearnLM and the per-client scores of the four-task federation. Table 18 reports the same trained weights of local training and Centralized evaluated as the final-epoch model instead of the model selected on validation data. All values are the `main` task metric on the global test split, mean over the four clients unless stated.

Table 17: Generic benchmark over three seeds (42, 1, 2; Dirichlet(0.3); validation-selected model): mean ± sample standard deviation over seeds for local training (Indiv) and UniLearnLM. The seed-42 XSum cells of Table 1 come from a rerun (Indiv 0.2529, UniLearnLM 0.2546, against 0.2533 and 0.2534 here). Bottom: per-client scores of the four-task federation, seed 42 (client i holds 40%, 30%, 20%, 10% of the rows for $i = 0, 1, 2, 3$). Sources: CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), SQuAD (Rajpurkar et al., 2016), XSum (Narayan et al., 2018), SAMSum (Gliwa et al., 2019). † marks a comparator that UniLearnLM beats in a paired two-sided t-test over the three seeds ($p < 0.05$).

Method	CoNLL (Tjong Kim Sang & De Meulder, 2003)	SQuAD (Rajpurkar et al., 2016)	AG News (Zhang et al., 2015)	XSum (Narayan et al., 2018)	SAMSum (Gliwa et al., 2019)	Four task
indiv	$0.7323 \pm 0.0099^{\dagger}$	0.7614 ± 0.0185	0.8118 ± 0.0446	0.2525 ± 0.0015	0.3897 ± 0.0021	0.3309 ± 0.0142
UniLearnLM	0.7668 ± 0.0025	0.7697 ± 0.0070	0.8664 ± 0.0065	0.2541 ± 0.0011	0.3873 ± 0.0060	0.2836 ± 0.0171
Four task, seed 42	**Client 0**	**Client 1**	**Client 2**	**Client 3**	**Mean**	
indiv	0.6267	0.1871	0.4101	0.1343	0.3396	
UniLearnLM	0.5620	0.1993	0.3352	0.0975	0.2985	

Table 18: Sensitivity of the generic benchmark to model selection: the same trained weights of local training (Indiv) and Centralized evaluated as the model selected on validation data (best, as in Table 1) and as the weights when training stopped (final-epoch model, final); seed 42. The federated methods select per round and are not affected. The Indiv XSum cells are from the original run and the Centralized XSum cells from the run of Table 1. Sources: CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), SQuAD (Rajpurkar et al., 2016), XSum (Narayan et al., 2018), SAMSum (Gliwa et al., 2019).

Method	Model	CoNLL (Tjong Kim Sang & De Meulder, 2003)	SQuAD (Rajpurkar et al., 2016)	AG News (Zhang et al., 2015)	XSum (Narayan et al., 2018)	SAMSum (Gliwa et al., 2019)	Four task
indiv	best	0.7237	0.7758	0.8374	0.2533	0.3909	0.3396
	final	0.7440	0.7751	0.8367	0.2472	0.3834	0.3061
	final − best	+0.0202	−0.0007	−0.0007	−0.0061	−0.0075	−0.0334
Centralized	best	0.8260	0.8031	0.9242	0.2705	0.4109	0.6372
	final	0.8267	0.7996	0.8914	0.2734	0.4095	0.6657
	final − best	+0.0006	−0.0035	−0.0328	+0.0028	−0.0014	+0.0285

Table 19 (plotted in Figure 5) repeats three of the single-task settings under two further partitions, Dirichlet(0.5) and IID, for local training and UniLearnLM over three seeds. Under Dirichlet(0.5) the difference between UniLearnLM and local training is less negative than under IID on SQuAD (+0.0028 against −0.0110) and SAMSum (−0.0033 against −0.0051), and equal on XSum (+0.0011 under both); UniLearnLM remains below local training on SQuAD IID and on both SAMSum settings.

Table 19: Local training (Indiv) and UniLearnLM under Dirichlet(0.5) and IID partitions: task metric on the global test split (ROUGE-L for XSum and SAMSum, answer F1 for SQuAD), mean over the four clients, then mean ± sample standard deviation over seeds 42, 1 and 2 (n=3); validation-selected model. The largest deviation of a single seed from the other two is SQuAD Dirichlet(0.5) Indiv, seed 2 (0.7616, 0.0167 below the mean of seeds 42 and 1). Sources: SQuAD (Rajpurkar et al., 2016), XSum (Narayan et al., 2018), SAMSum (Gliwa et al., 2019).

	XSum (Narayan et al., 2018)		SQuAD (Rajpurkar et al., 2016)		SAMSum (Gliwa et al., 2019)	
Method	**Dir(0.5)**	**IID**	**Dir(0.5)**	**IID**	**Dir(0.5)**	**IID**
indiv	0.2519 ± 0.0014	0.2479 ± 0.0014	0.7727 ± 0.0104	0.7795 ± 0.0055	0.3923 ± 0.0016	0.3897 ± 0.0038
UniLearnLM	0.2530 ± 0.0015	0.2490 ± 0.0004	0.7755 ± 0.0062	0.7685 ± 0.0031	0.3890 ± 0.0013	0.3846 ± 0.0009

Graph encoder ablation on the generic benchmark. Replacing the graph encoder by a learned client index changes CoNLL-2003 entity F1 by -0.0061 ± 0.0135 and AG News accuracy by -0.0142 ± 0.0208 (paired over three seeds; Table 20 and Figure 6). The client-index control also

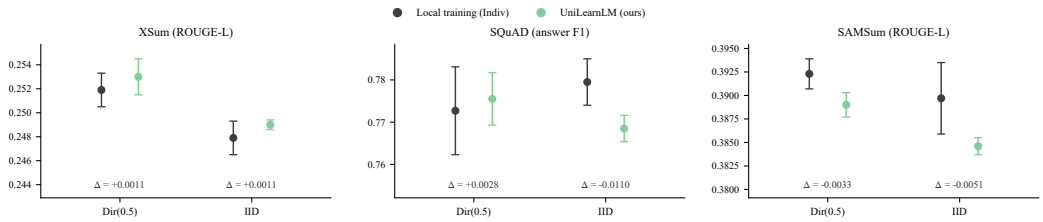


Figure 5: Local training (Indiv) and UniLearnLM under Dirichlet(0.5) and IID partitions (Table 19): task metric on the global test split, mean over the four clients, then mean ± 1 sample standard deviation over three seeds. Δ = UniLearnLM minus local training.

Table 20: Generic benchmark over three seeds in one series: ablations of the graph encoder and the strongest published heterogeneous federation (FedHeNN) (seeds 42, 1, 2; Dirichlet(0.3); validation-selected model): entity F1 on CoNLL-2003 and accuracy on AG News, global test split, mean over the four clients, then mean $\pm$ sample standard deviation over seeds. All cells come from one series of runs under the protocol of Section 4.1, separate from Table 17, and the difference is paired by seed. The client-index control replaces the graph encoder by a learned client embedding (Appendix U); the untyped graph keeps the architecture graph and sets its node-type features to zero. Sources: CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), AG News (Zhang et al., 2015).

Method	**CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003)**	**AG News (Zhang et al., 2015)**
FedHeNN (Makhija et al., 2022)	0.7693 ± 0.0087	0.8679 ± 0.0113
UniLearnLM	0.7695 ± 0.0059	0.8659 ± 0.0057
Client-index control	0.7634 ± 0.0076	0.8517 ± 0.0245
Untyped graph	0.7744 ± 0.0066	0.8652 ± 0.0214
FedHeNN $-$ UniLearnLM (paired)	-0.0003 ± 0.0074	$+0.0020 \pm 0.0068$
Client index $-$ UniLearnLM (paired)	-0.0061 ± 0.0135	-0.0142 ± 0.0208
Untyped $-$ UniLearnLM (paired)	$+0.0049 \pm 0.0074$	-0.0007 ± 0.0270

shows a wider seed spread on AG News (0.0245 against 0.0057). Setting the node-type features of the architecture graph to zero changes CoNLL-2003 entity F1 by $+0.0049 \pm 0.0074$ and AG News accuracy by -0.0007 ± 0.0270; the untyped control is nominally higher on CoNLL-2003. In the same series, FedHeNN differs from UniLearnLM by -0.0003 ± 0.0074 on CoNLL-2003 and $+0.0020 \pm 0.0068$ on AG News (paired over three seeds); the XSum and SAMSum comparisons of Table 1 remain single seed.

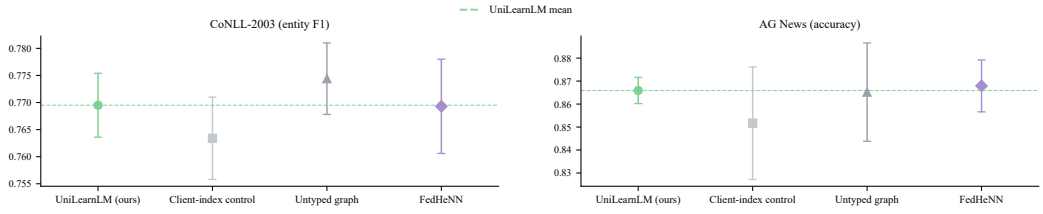


Figure 6: Generic benchmark over three seeds in one series (Table 20): UniLearnLM, the client-index control, the untyped graph and FedHeNN, mean ± 1 sample standard deviation over seeds; dashed line = UniLearnLM mean.

U Heterogeneous Federated Learning Baselines, Data Heterogeneity and Length Matched Comparison

This appendix reports the three published heterogeneous federated learning methods and the client-index control announced in Section 4, the heterogeneity index, all eleven metrics for the methods of Figure 2 (Figure 7), and the length matched comparison with the zero shot LLMs. Every run uses the protocol of UniLearnLM: the same three public testbed clients (BART-base, Flan-T5-base,

LED-base-16384), the same data splits and 3-fold cross-validation, LoRA rank 16, α=32, dropout 0.1 on the query and value projections, 33 rounds of 3 local epochs, batch size 4, and the same model-selection rule (best round by greedy validation ROUGE-L on the fixed 50 row subsample). Every method saves only its LoRA adapters and is scored by the same evaluation script on $\mathcal{T}^{\text{global}}$. Hyperparameters that the papers tune are tuned on the fold 1 validation set over the papers' own grids; the selected setting is then run on all folds. Cells marked – in Table 21 are not available, and * marks a mean over fewer than three folds.

Text consensus federations and reruns of our own systems. The TC-Best and TC-Ens rows of Table 2 are runs under this protocol. The earlier TC-Best implementation split every peer generation on a fixed answer marker and silently dropped the consensus rows won by the Flan-T5 client, whose generations use a different layout; the shared parser of the rerun keeps them, and the rerun also draws the public pool from training chunks only, so no validation row reaches any client through the pool. UniLearnLM in Table 2 is the original run, whose round was selected by the same greedy validation ROUGE-L rule computed on the LoRA adapters alone, with the FiLM modulation of the round inactive during the selection pass; a rerun in which the selection pass applies the modulation (UniLearnLM rerun rows of Table 21) selects rounds 5, 12 and 9 instead of later ones, moves ROUGE-L, BLEU-4 and BERTScore up by less than the fold spread and moves B_{cls}, QAFactEval, RQUGE and round trip consistency down, B_{cls} by about 0.03. We report the original run in the headline table and the rerun beside it so that the sensitivity of the depth result to the round selection is visible. UniLearnLM+ in Table 2 was trained with the earlier pool, which also drew on validation chunks; its rerun with the training only pool is listed in the same rows.

pFedHN (Shamsian et al., 2021). A hypernetwork $h(\mathbf{v}_k; \boldsymbol{\varphi})$ maps a learned client embedding $\mathbf{v}_k$ (dimension $\lfloor 1 + K/4 \rfloor$, as in the official code) through $\text{Linear}(d, 100)$ and three $\text{ReLU} \to \text{Linear}(100, 100)$ blocks to one linear head per LoRA tensor of the client; because the three clients have different tensor shapes, each architecture owns its head set, the variant the paper describes as emitting all architectures and slicing. The server visits one client per step: it generates $\boldsymbol{\theta}_k = h(\mathbf{v}_k)$, the client runs the inner loop (E=3 epochs, SGD with momentum 0.9, weight decay $5{\cdot}10^{-5}$, gradient clipping 50, optimizer state fresh per visit, inner learning rate on the paper's grid $\{5{\cdot}10^{-2}, 10^{-2}, 5{\cdot}10^{-3}\}$) and returns $\Delta\boldsymbol{\theta}_k = \boldsymbol{\theta}_k - \tilde{\boldsymbol{\theta}}_k$, and the server applies the vector Jacobian product $(\partial\boldsymbol{\theta}_k/\partial\boldsymbol{\varphi})^\top \Delta\boldsymbol{\theta}_k$ with SGD (learning rate 10^{-2}, momentum 0.9, weight decay 10^{-3}, clipping 50), line for line as in the official implementation. The evaluated model is the regenerated $h(\mathbf{v}_k; \boldsymbol{\varphi})$, not the client's fine-tuned copy. One deviation is documented: the heads that emit LoRA B matrices are zero initialized so that the generated adapter starts as the identity, as LoRA itself does. Fold 1 validation ROUGE-L selects the inner learning rate $5{\cdot}10^{-3}$ (0.483 against 0.382 and 0.259; an added 10^{-3} gives 0.472, and a larger outer learning rate lowers the statistic), and every swept setting is listed with its statistic in Table 21.

FedHeNN (Makhija et al., 2022). No weights are averaged. Each client keeps its own model and, following Algorithm 2 and Eq. (6) of the paper, adds a representation alignment penalty to its loss: on a representation alignment dataset (RAD) the client computes the linear kernel Gram matrix K_k of its final decoder hidden state at the first generation position, the server averages the clients' Gram matrices into $\bar{K}$, and the client minimizes $\mathcal{L}_k^{\text{CE}} + \eta(t)\,(1 - \text{CKA}(K_k, \bar{K}))$ with HSIC centered CKA and $\eta(t) = \eta_0\, t$. The RAD is the fold's 180 row public pool, the only shared material our setting allows. K_k is recomputed with gradient at every optimization step on a random 16 row sub-block of the RAD (the full 180 rows at every step would multiply the cost of a round by about 45). No official code exists; the implementation follows the paper. η_0 is tuned in $\{10^{-3}, 10^{-2}, 10^{-1}\}$; fold 1 validation ROUGE-L selects 10^{-3} (0.495 against 0.490 and 0.494), and all three settings are evaluated on fold 1 in Table 21.

FedMKT (Fan et al., 2025). Mutual knowledge transfer between the client SLMs and a server LLM through token aligned logits, following the FATE-LLM implementation: the server model is Qwen2.5-7B-Instruct (Qwen Team, 2024) with LoRA rank 8 (α=16, dropout 0.05, query, key, value and output projections, learning rate $3{\cdot}10^{-5}$, AdamW β_2=0.95, weight decay 0.1, cosine schedule, warmup 0.008, effective batch 4); vocabulary tables are built by minimum edit distance over the whole target vocabulary in both directions, token sequences are aligned by dynamic time warping, the top 128 logits are exchanged at temperature 1 with λ=0.9, and the dual minimum cross entropy

selection keeps the stale self candidate exactly as the code does. Each round runs one public epoch per client, as in the official code; a variant with three public epochs per round is listed as a sensitivity row in Table 21. Two adaptations are forced by our setting and documented in the code: FedMKT requires a labeled public set, so fixed pseudo targets for the unlabeled public pool are generated once by the un-adapted server LLM; and the clients are encoder-decoders, so alignment and distillation run on the target token sequence.

Client index and client vector controls. Both keep the FiLM head, the LoRA adapters, the schedule and the federated averaging of UniLearnLM and replace the graph encoder by a multilayer perceptron over a learned client embedding concatenated with a block position embedding (client-index control, which keeps the per block structure) or over the client embedding alone (client vector control). The client embeddings are local and never averaged; the MLP is the federated object. These controls answer the question whether the architecture graph carries information beyond a client identifier. Under the identical protocol, the rerun of UniLearnLM with FiLM aware round selection (UniLearnLM rerun rows of Table 21) leads the client-index control on the metrics that both rows report, on three folds, and both rows place the depth proxy below local training. The T5 block node row registers the T5 block class as a node type so that every hooked block of the Flan-T5 client resolves to its own node (Appendix AF.1); the seed row repeats UniLearnLM with a second random seed.

Judge columns. B_{LLM}, answer relevancy and LLM-J of Table 2 are computed with GPT-4o-mini. The holistic judge was not run for the published heterogeneous, text-consensus and control rows of this appendix, and no new judge calls were made for them. Under the Gemini-2.5-Flash replication protocol of Appendix AD (identical prompt and rubric), local training scores 0.4282 over three folds, against 0.4289 to 0.4687 for the three UniLearnLM clients in Table 44. Faithfulness for these rows (§) uses sentence level decomposition of the answer in place of the GPT-4o-mini claim decomposition, followed by the same DeBERTa-v3 entailment model, so it is comparable within the rows of Table 21 and Table 23 but not with the Faith column of the other tables.

Table 21: Published heterogeneous federated learning methods and controls under the UniLearnLM protocol on $\mathcal{T}^{\text{global}}$ (validation-selected model, mean over the institutions, mean±std over folds). *Folds* lists the folds in the mean; * = fewer than three folds; − = not available. Rows other than Indiv, TC-Best, TC-Ens, UniLearnLM and UniLearnLM+ are runs under the identical protocol, whose Faith uses sentence level claim decomposition (§, see text). Swept settings list their fold 1 validation ROUGE-L, the selection statistic, and the UniLearnLM rerun rows, followed by its ablations, are the reruns of our own systems described in the text. Column abbreviations follow Table 2.

Method	Folds	R-L	B-4	BERT	B_{cls}	Faith	QAFE	RQU	RTC
Indiv	1,2,3	0.4316±0.005	0.1612±0.003	0.8602±0.002	0.2238±0.011	0.2091±0.004	0.1832±0.006	4.0741±0.007	0.2049±0.005
TC-Best (Li & Wang, 2019)	1,2,3	0.4387±0.002	0.1714±0.002	0.8632±0.000	0.1623±0.013	0.1966±0.001	0.1724±0.008	4.0647±0.001	0.2045±0.003
TC-Ens (Li & Wang, 2019)	1,2,3	0.4364±0.002	0.1717±0.003	0.8628±0.001	0.1655±0.009	0.1893±0.004	0.1729±0.009	4.0652±0.003	0.2023±0.001
UniLearnLM	1,2,3	0.4385±0.005	0.1715±0.004	0.8640±0.002	0.2313±0.016	0.2092±0.005	0.1808±0.007	4.0826±0.006	0.2130±0.002
UniLearnLM+	1,2,3	0.4450±0.004	0.1760±0.003	0.8653±0.001	0.1746±0.020	0.2091±0.005	0.1784±0.005	4.0744±0.007	0.2069±0.004
pFedHN (Shamsian et al., 2021), inner lr $5{\cdot}10^{-2}$ (val R-L 0.259)	–	–	–	–	–	–	–	–	–
pFedHN, inner lr 10^{-2} (val R-L 0.382)	–	–	–	–	–	–	–	–	–
pFedHN, inner lr $5{\cdot}10^{-3}$ (val R-L 0.483, selected)	1,2,3	0.4313±0.001	0.1624±0.001	0.8614±0.000	0.1816±0.017	0.1798±0.005	0.1884±0.002	4.0576±0.007	0.2011±0.000
pFedHN, inner lr 10^{-3} (val R-L 0.472)	–	–	–	–	–	–	–	–	–
pFedHN, inner lr $5{\cdot}10^{-3}$, outer lr $5{\cdot}10^{-2}$ (val R-L 0.382)	–	–	–	–	–	–	–	–	–
pFedHN, inner lr $5{\cdot}10^{-3}$, outer lr 10^{-1} (val R-L 0.317)	–	–	–	–	–	–	–	–	–
FedHeNN (Makhija et al., 2022), $\eta_0{=}10^{-3}$ (val R-L 0.495, selected)	1,2,3	0.4372±0.001	0.1700±0.001	0.8636±0.000	0.2062±0.014	0.1933±0.007	0.1815±0.001	4.0666±0.005	0.2049±0.005
FedHeNN, $\eta_0{=}10^{-2}$ (val R-L 0.490)	1*	0.4388	0.1716	0.8639	0.2027	0.1980	0.1782	4.0663	0.2075
FedHeNN, $\eta_0{=}10^{-1}$ (val R-L 0.494)	1*	0.4374	0.1694	0.8633	0.1990	0.1838	0.1750	4.0802	0.2064
FedMKT (Fan et al., 2025), one public epoch per round (val R-L 0.451, selected)	1,2,3	0.3997±0.003	0.1494±0.002	0.8524±0.001	0.2243±0.012	0.1791±0.017	0.1762±0.006	4.0496±0.008	0.1949±0.007
FedMKT, three public epochs per round (val R-L 0.436)	1,2,3	0.3871±0.002	0.1373±0.001	0.8484±0.002	0.2216±0.023	0.1729±0.009	0.1565±0.020	4.0519±0.009	0.1978±0.007
UniLearnLM, seed 1	1*	0.4460	0.1768	0.8661	0.2001	0.1994	0.1852	4.0779	0.2092
UniLearnLM, rerun with FiLM aware round selection	1,2,3	0.4425±0.001	0.1743±0.002	0.8647±0.000	0.2028±0.004	0.1907±0.003	0.1762±0.011	4.0750±0.003	0.2085±0.003
UniLearnLM+, rerun with the training only pool	1,2,3	0.4440±0.007	0.1771±0.005	0.8656±0.001	0.1686±0.012	0.2010±0.001	0.1758±0.007	4.0683±0.007	0.2083±0.002
Client-index control	1,2,3	0.4402±0.007	0.1742±0.004	0.8645±0.002	0.2021±0.002	0.1905±0.004	0.1724±0.005	4.0647±0.014	0.2057±0.002
Client vector control	1*	0.4421	0.1733	0.8648	0.1951	0.1845	0.1794	4.0713	0.2078
UniLearnLM, T5 block nodes	1*	0.4404	0.1730	0.8643	0.2107	0.1996	0.1886	4.0784	0.2088

Data heterogeneity index. We embed every unique teaching chunk of each institution with the sentence encoder all-MiniLM-L6-v2 (Reimers & Gurevych, 2019; Wang et al., 2020) (mean pooled, unit normalized). The heterogeneity index H is the silhouette coefficient (Rousseeuw, 1987) of the institution labels in the original 384 dimensional embedding space under the cosine metric, computed on a balanced sample of up to 100 chunks per institution and averaged over 20 draws so that the 981 chunk arXiv client does not dominate: $H{=}0.067 \pm 0.003$ for the public testbed (1,201 chunks) and $H{=}0.038 \pm 0.004$ for the private testbed (2,670 chunks). Unbalanced silhouettes are 0.036 and 0.043, and the mean pairwise cosine distance between institution centroids is 0.343 and 0.412.

Table 22: Per institution results of the heterogeneous federated learning baselines and the client-index control on $\mathcal{T}^{\text{global}}$ (validation-selected model; *Folds* lists the folds in the mean, * = fewer than three folds, − = not available; § as in Table 21). The corresponding rows of the paper's methods are in Table 2. Sources: pFedHN (Shamsian et al., 2021), FedHeNN (Makhija et al., 2022), FedMKT (Fan et al., 2025), BART (Lewis et al., 2020), Flan-T5 (Chung et al., 2024), LED (Beltagy et al., 2020).

Method	Inst.	Folds	R-L	B-4	BERT	B_{cls}	Faith§	QAFE	RQU	RTC
pFedHN (Shamsian et al., 2021)	$\mathcal{I}_0^{\text{pub}}$ (BART (Lewis et al., 2020), MIT)	1,2,3	0.4274	0.1609	0.8592	0.2307	0.1801	0.1807	4.0619	0.1957
	$\mathcal{I}_1^{\text{pub}}$ (Flan-T5 (Chung et al., 2024), Stanford)	1,2,3	0.4425	0.1668	0.8663	0.1385	0.1861	0.2038	4.0747	0.2189
	$\mathcal{I}_2^{\text{pub}}$ (LED (Beltagy et al., 2020), ML papers)	1,2,3	0.4241	0.1596	0.8587	0.1755	0.1732	0.1809	4.0363	0.1888
FedHeNN (Makhija et al., 2022)	$\mathcal{I}_0^{\text{pub}}$ (BART, MIT)	1,2,3	0.4340	0.1657	0.8612	0.2242	0.2038	0.1754	4.0635	0.1997
	$\mathcal{I}_1^{\text{pub}}$ (Flan-T5, Stanford)	1,2,3	0.4461	0.1745	0.8673	0.2032	0.1903	0.1952	4.0904	0.2215
	$\mathcal{I}_2^{\text{pub}}$ (LED, ML papers)	1,2,3	0.4316	0.1699	0.8622	0.1911	0.1858	0.1739	4.0459	0.1935
FedMKT (Fan et al., 2025)	$\mathcal{I}_0^{\text{pub}}$ (BART, MIT)	1,2,3	0.3937	0.1441	0.8506	0.2123	0.1680	0.1865	4.0344	0.1830
	$\mathcal{I}_1^{\text{pub}}$ (Flan-T5, Stanford)	1,2,3	0.4049	0.1525	0.8539	0.2409	0.1903	0.1843	4.0813	0.2176
	$\mathcal{I}_2^{\text{pub}}$ (LED, ML papers)	1,2,3	0.4005	0.1516	0.8526	0.2196	0.1792	0.1578	4.0330	0.1842
Client index ctrl.	$\mathcal{I}_0^{\text{pub}}$ (BART, MIT)	1,2,3	0.4383	0.1688	0.8622	0.2132	0.1895	0.1747	4.0589	0.1947
	$\mathcal{I}_1^{\text{pub}}$ (Flan-T5, Stanford)	1,2,3	0.4455	0.1765	0.8672	0.1999	0.2046	0.1865	4.0940	0.2214
	$\mathcal{I}_2^{\text{pub}}$ (LED, ML papers)	1,2,3	0.4368	0.1772	0.8640	0.1932	0.1775	0.1559	4.0412	0.2010

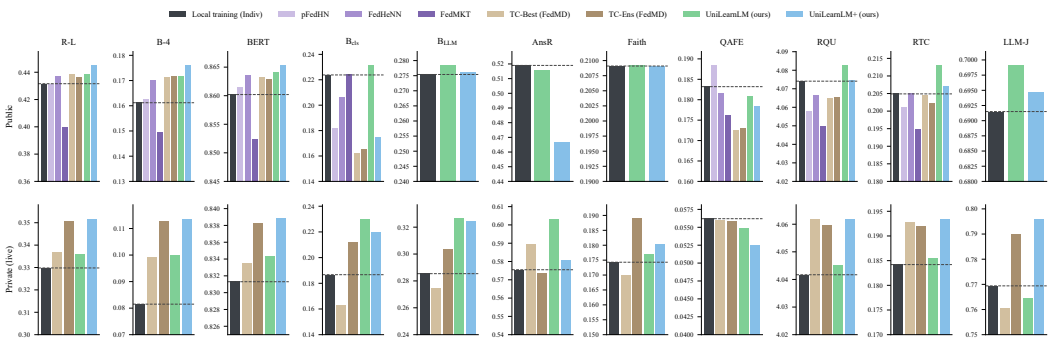


Figure 7: All eleven metrics for the methods of Figure 2, on the global test set of the public and private testbeds (mean over institutions; Mean rows of Tables 2 and 30). Dashed line = local training; each panel's y-axis starts just below its lowest value. Only methods with a value are drawn: the published heterogeneous federations were evaluated on the public testbed only, their B_{LLM}, answer relevancy and LLM judge cells are not available (Table 21), and Faithfulness values from sentence level claim decomposition (§ in Table 2) are omitted as not comparable. Private testbed: live runs.

$H > 0$ means chunks lie closer to their own institution than to the nearest other one; both testbeds are positive but small, because every curriculum is machine learning, so institutions differ in size, level and style rather than in subject, and the two course note clients of the public testbed form the tightest clusters.

Length matched comparison. Table 23 answers the confound raised in Section 5: the UniLearnLM mixture writes answers of 14.8 words against 19.5 to 47.1 for the LLMs (Table 23), and the two grounding metrics reward short extractive answers. Truncating every system to 15 words removes the length difference (all systems at 10.8 to 14.7 words); the mixture keeps the highest QAFactEval (0.148 against at most 0.138) and the highest Faithfulness (0.175 against at most 0.171), and the same ordering holds at 20 words (0.165 against at most 0.155; 0.176 against at most 0.173). The margins are small, under half a point on Faithfulness and about one point on QAFactEval, and truncation lifts the LLMs' lexical overlap because their first words track the reference more closely, so the accurate reading is that answer length explains most, not all, of the grounding gap: at a common budget the federated SLMs remain marginally more entailed by the source than the zero shot LLMs, and the LLMs remain ahead on ROUGE-L, BERTScore and depth.

Table 23: Length matched comparison on $\mathcal{T}^{\text{global}}$ (fold 1 mixture and LLM generations, re-scored in a separate truncation pass). Every answer is truncated to its first K words before scoring, so all systems are compared at a common answer budget; $K{=}\infty$ is the untruncated answer; these are this pass's untruncated fold-1 scores and differ from Table 34 by at most 0.008. Metrics are recomputed offline with the same implementation for every row (Faith with sentence level decomposition, §), so the columns are comparable within this table. Bold = best per column within each budget. Sources: Qwen2.5 (Qwen Team, 2024), GPT-4o (Hurst et al., 2024), Gemini (Comanici et al., 2025), Llama (Grattafiori et al., 2024), Mistral (Jiang et al., 2023).

System	K	words	R-L	BERT	RTC	Faith§	QAFE	RQU
GPT-4o (Hurst et al., 2024)	15	14.7	0.4762	0.8651	0.2095	0.1652	0.1219	4.0826
GPT-4o-mini	15	14.7	0.4842	0.8685	0.2151	0.1556	0.1179	4.0850
Llama-3.1-8B-Inst (Grattafiori et al., 2024)	15	10.8	0.4567	0.8657	0.1908	0.1386	0.1376	4.0610
Qwen-2.5-7B-Inst (Qwen Team, 2024)	15	14.5	0.4697	0.8633	0.2124	0.1498	0.1067	4.0732
Mistral-7B-Inst (Jiang et al., 2023)	15	14.3	0.3828	0.8360	0.1646	0.1714	0.1234	4.0045
Gemini-2.5-Flash (Comanici et al., 2025)	15	14.2	**0.4877**	**0.8733**	**0.2233**	0.1693	0.1359	**4.0893**
UniLearnLM (fed. mixture)	15	12.2	0.4508	0.8653	0.2107	**0.1747**	**0.1479**	4.0820
GPT-4o	20	19.1	0.4899	0.8725	0.2095	0.1606	0.1347	4.0826
GPT-4o-mini	20	19.3	0.4893	0.8725	0.2151	0.1359	0.1174	4.0850
Llama-3.1-8B-Inst	20	13.2	0.4612	0.8676	0.1908	0.1376	0.1550	4.0610
Qwen-2.5-7B-Inst	20	18.7	0.4776	0.8678	0.2124	0.1447	0.1004	4.0732
Mistral-7B-Inst	20	18.5	0.3853	0.8386	0.1646	0.1569	0.1148	4.0045
Gemini-2.5-Flash	20	18.1	**0.4985**	**0.8774**	**0.2233**	0.1728	0.1279	**4.0893**
UniLearnLM (fed. mixture)	20	13.8	0.4523	0.8670	0.2107	**0.1758**	**0.1654**	4.0820
GPT-4o	40	33.5	0.4747	0.8753	0.2095	0.1450	0.1276	4.0826
GPT-4o-mini	40	33.8	0.4645	0.8724	0.2151	0.1262	0.1084	4.0850
Llama-3.1-8B-Inst	40	18.0	0.4588	0.8682	0.1908	0.1370	0.1517	4.0610
Qwen-2.5-7B-Inst	40	30.4	0.4670	0.8693	0.2124	0.1445	0.0987	4.0732
Mistral-7B-Inst	40	31.6	0.3778	0.8417	0.1646	0.1479	0.1302	4.0045
Gemini-2.5-Flash	40	28.1	**0.4941**	**0.8802**	**0.2233**	0.1662	0.1298	**4.0893**
UniLearnLM (fed. mixture)	40	14.8	0.4535	0.8678	0.2107	**0.1774**	**0.1687**	4.0820
GPT-4o	∞	47.1	0.4487	0.8725	0.2095	0.1438	0.1335	4.0826
GPT-4o-mini	∞	42.2	0.4484	0.8707	0.2151	0.1223	0.1150	4.0850
Llama-3.1-8B-Inst	∞	19.5	0.4570	0.8679	0.1908	0.1359	0.1567	4.0610
Qwen-2.5-7B-Inst	∞	35.9	0.4586	0.8680	0.2124	0.1426	0.1004	4.0732
Mistral-7B-Inst	∞	42.4	0.3644	0.8406	0.1646	0.1450	0.1269	4.0045
Gemini-2.5-Flash	∞	31.9	**0.4883**	**0.8797**	**0.2233**	0.1610	0.1328	**4.0893**
UniLearnLM (fed. mixture)	∞	14.8	0.4534	0.8678	0.2107	**0.1774**	**0.1687**	4.0820

V Per-Fold Breakdown of the Headline Result

Table 24 reports the per fold values that underlie the mean-over-3-folds entries of the global-test, validation-selected-model headline (Table 2). Table 27 reports the final-epoch model counterpart, i.e. per fold values for the global-test, last-epoch / last-round state, with the in-table Mean rows giving the per-institution averages.

Per-fold standard deviation. The standard deviation over the three folds, computed from the per fold rows of Table 24, quantifies the per fold variation behind each Mean entry. On ROUGE-L it is small for every method on every institution (0.001–0.013), as it is on BERTScore (≤ 0.004); the larger spread sits on B_{cls}, the Bloom-classifier EVS, where the distillation baselines vary most (± 0.030 for TC-Best and ± 0.041 for TC-Ens on $\mathcal{I}_0^{\text{pub}}$ against ± 0.026 for UniLearnLM and ± 0.008 for Indiv).

W Supplementary Results Tables

This appendix collects the supplementary result tables referenced from Section 5.2: the in distribution local val/test breakdown (Table 28) and the pedagogical and LLM-judge metrics on the public testbed (Tables 14 and 3).

Table 24: Per-fold breakdown of the **global-test, validation-selected model** result for the public testbed, reporting the per fold values of Indiv, TC-Best, TC-Ens, UniLearnLM and the published heterogeneous federations that underlie the mean-over-3-folds headline (Table 2); Tables 25 and 26 continue it for $\mathcal{I}_1^{\text{pub}}$ and $\mathcal{I}_2^{\text{pub}}$; UniLearnLM+ and the client-index control are reported as fold means only (Tables 2 and 21). Each row reports a single fold, and the "Mean" rows are the per fold averages. Bold = best of $\{\text{indiv}, \text{TC-Best}, \text{TC-Ens}, \text{uni}\}$ per (institution, fold, metric). Cells are shaded by rank within each (institution, fold) group per column: darkest green = best, then 2nd, 3rd; the 4th-ranked cell is left unshaded. Column abbreviations follow Table 2. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019), pFedHN (Shamsian et al., 2021), FedHeNN (Makhija et al., 2022), FedMKT (Fan et al., 2025), BART (Lewis et al., 2020).

Inst.	Method	Fold	R-L	B-4	BERT	B_{cls}
$\mathcal{I}_0^{\text{pub}}$ (BART (Lewis et al., 2020), MIT)	indiv	1	0.4289	0.1603	0.8586	**0.2474**
	indiv	2	0.4210	0.1579	0.8569	0.2352
	indiv	3	0.4368	0.1637	0.8629	**0.2325**
	indiv	Mean	0.4289	0.1607	0.8595	**0.2384**
	pFedHN (Shamsian et al., 2021)	1	0.4255	0.1623	0.8586	0.2683
	pFedHN	2	0.4265	0.1579	0.8583	0.2384
	pFedHN	3	0.4303	0.1624	0.8607	0.1854
	pFedHN	Mean	0.4274	0.1609	0.8592	0.2307
	FedHeNN (Makhija et al., 2022)	1	0.4306	0.1651	0.8609	0.2406
	FedHeNN	2	0.4365	0.1675	0.8616	0.2406
	FedHeNN	3	0.4350	0.1646	0.8612	0.1915
	FedHeNN	Mean	0.4340	0.1657	0.8612	0.2242
	FedMKT (Fan et al., 2025)	1	0.3965	0.1482	0.8502	0.2290
	FedMKT	2	0.3918	0.1399	0.8505	0.2159
	FedMKT	3	0.3929	0.1442	0.8510	0.1919
	FedMKT	Mean	0.3937	0.1441	0.8506	0.2123
	TC-Best (Li & Wang, 2019)	1	**0.4366**	0.1641	0.8607	0.1664
	TC-Best	2	**0.4372**	**0.1654**	**0.8620**	0.1607
	TC-Best	3	**0.4406**	0.1724	**0.8633**	0.1271
	TC-Best	Mean	**0.4381**	**0.1673**	**0.8620**	0.1514
	TC-Ens (Li & Wang, 2019)	1	0.4293	0.1635	0.8592	0.1576
	TC-Ens	2	0.4276	0.1629	0.8598	0.1517
	TC-Ens	3	0.4370	**0.1751**	0.8629	0.1524
	TC-Ens	Mean	0.4313	0.1672	0.8606	0.1539
	UniLearnLM	1	0.4322	**0.1654**	**0.8613**	0.2207
	UniLearnLM	2	0.4254	0.1609	0.8590	**0.2430**
	UniLearnLM	3	0.4338	0.1683	0.8607	0.1913
	UniLearnLM	Mean	0.4305	0.1648	0.8604	0.2183

The table reports the full eleven-metric suite on the local test splits.

Decomposing the LLM-judge overall score into its four rubric dimensions (Table 3) shows that UniLearnLM's judged quality is carried by context grounding and answer relevance, both near $4/5$ on every institution, while educational value is the weakest dimension at roughly $3/5$, consistent with the generated questions concentrating at the lower Bloom levels.

The direction-resolved Knowledge Transfer Index on the private testbed is reported in Table 32 (Appendix X).

X Private-Testbed Results

This appendix reports the full result set for the five-institution private testbed described in Section C.2 (Tunisia, Germany, Nigeria, Colombia, Peru). Four tables follow. Table 29 reports the eleven-metric per-institution results on each institution's own local test split, providing a within distribution view of QA quality. Table 30 reports the same eleven metrics on the cross distribution global private test set, which is built from the materials contributed by the four dataset-only institutions $\mathcal{D}_0$–$\mathcal{D}_3$ and on which no institution trains. Table 31 reports the Pairwise Equity Score for every institution pair, with Q taken from either the global private test set (cross distribution) or each institution's own local test split (within distribution). Table 32 reports the Knowledge Transfer Index extended to all four federated methods (each compared against the individual-training baseline) on the global private test set, where $\mathcal{T}_{k\setminus l}$ (topics in institution k's curriculum but not in l's) is well defined. The same convention used in the public-testbed tables holds throughout: bold = best per row or column group,

Table 25: Per-fold breakdown of the global test, validation-selected model result, continued: $\mathcal{I}_1^{\text{pub}}$ (Flan-T5, Stanford). Conventions as in Table 24. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019), pFedHN (Shamsian et al., 2021), FedHeNN (Makhija et al., 2022), FedMKT (Fan et al., 2025), Flan-T5 (Chung et al., 2024).

Inst.	Method	Fold	R-L	B-4	BERT	B_{cls}
$\mathcal{I}_1^{\text{pub}}$ (Flan-T5 (Chung et al., 2024), Stanford)	indiv	1	0.4493	0.1693	0.8643	0.2489
	indiv	2	0.4416	0.1675	0.8619	0.2533
	indiv	3	0.4360	0.1566	0.8605	**0.3131**
	indiv	Mean	0.4423	0.1645	0.8623	0.2718
	pFedHN (Shamsian et al., 2021)	1	0.4436	0.1682	0.8667	0.1468
	pFedHN	2	0.4439	0.1659	0.8663	0.1354
	pFedHN	3	0.4399	0.1662	0.8659	0.1333
	pFedHN	Mean	0.4425	0.1668	0.8663	0.1385
	FedHeNN (Makhija et al., 2022)	1	0.4446	0.1742	0.8670	0.1976
	FedHeNN	2	0.4455	0.1733	0.8669	0.2122
	FedHeNN	3	0.4480	0.1760	0.8681	0.1998
	FedHeNN	Mean	0.4461	0.1745	0.8673	0.2032
	FedMKT (Fan et al., 2025)	1	0.4117	0.1557	0.8559	0.2408
	FedMKT	2	0.3971	0.1495	0.8513	0.2483
	FedMKT	3	0.4058	0.1523	0.8545	0.2336
	FedMKT	Mean	0.4049	0.1525	0.8539	0.2409
	TC-Best (Li & Wang, 2019)	1	0.4506	**0.1818**	0.8674	0.1314
	TC-Best	2	**0.4507**	**0.1805**	**0.8668**	0.1670
	TC-Best	3	0.4437	0.1737	0.8654	0.1406
	TC-Best	Mean	**0.4483**	0.1787	0.8665	0.1464
	TC-Ens (Li & Wang, 2019)	1	**0.4511**	0.1805	0.8676	0.1498
	TC-Ens	2	0.4461	0.1802	0.8663	0.1566
	TC-Ens	3	0.4439	**0.1774**	0.8661	0.1688
	TC-Ens	Mean	0.4470	**0.1794**	0.8667	0.1584
	UniLearnLM	1	0.4505	0.1765	**0.8682**	**0.2734**
	UniLearnLM	2	0.4430	0.1686	0.8665	**0.2806**
	UniLearnLM	3	**0.4470**	0.1746	**0.8668**	0.2764
	UniLearnLM	Mean	0.4468	0.1732	**0.8672**	**0.2768**

Table 26: Per-fold breakdown of the global test, validation-selected model result, continued: $\mathcal{I}_2^{\text{pub}}$ (LED, ML papers). Conventions as in Table 24. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019), pFedHN (Shamsian et al., 2021), FedHeNN (Makhija et al., 2022), FedMKT (Fan et al., 2025), LED (Beltagy et al., 2020).

Inst.	Method	Fold	R-L	B-4	BERT	B_{cls}
$\mathcal{I}_2^{\text{pub}}$ (LED (Beltagy et al., 2020), ML papers)	indiv	1	0.4265	0.1586	0.8599	0.1561
	indiv	2	0.4146	0.1481	0.8557	0.1624
	indiv	3	0.4301	0.1687	0.8613	0.1648
	indiv	Mean	0.4237	0.1585	0.8590	0.1611
	pFedHN (Shamsian et al., 2021)	1	0.4237	0.1584	0.8584	0.1738
	pFedHN	2	0.4259	0.1601	0.8597	0.1810
	pFedHN	3	0.4229	0.1601	0.8581	0.1716
	pFedHN	Mean	0.4241	0.1596	0.8587	0.1755
	FedHeNN (Makhija et al., 2022)	1	0.4355	0.1738	0.8639	0.1900
	FedHeNN	2	0.4286	0.1671	0.8613	0.2007
	FedHeNN	3	0.4308	0.1686	0.8614	0.1828
	FedHeNN	Mean	0.4316	0.1699	0.8622	0.1911
	FedMKT (Fan et al., 2025)	1	0.3959	0.1496	0.8516	0.2297
	FedMKT	2	0.3994	0.1502	0.8531	0.2225
	FedMKT	3	0.4061	0.1548	0.8532	0.2066
	FedMKT	Mean	0.4005	0.1516	0.8526	0.2196
	TC-Best (Li & Wang, 2019)	1	0.4268	0.1629	0.8601	0.1686
	TC-Best	2	0.4352	0.1741	0.8622	0.2044
	TC-Best	3	0.4274	0.1678	0.8604	0.1943
	TC-Best	Mean	0.4298	0.1683	0.8609	0.1891
	TC-Ens (Li & Wang, 2019)	1	0.4333	0.1657	0.8616	0.1598
	TC-Ens	2	0.4276	0.1654	0.8599	0.1976
	TC-Ens	3	**0.4321**	**0.1747**	**0.8619**	**0.1948**
	TC-Ens	Mean	0.4310	0.1686	0.8611	0.1841
	UniLearnLM	1	**0.4518**	**0.1862**	**0.8681**	**0.1845**
	UniLearnLM	2	**0.4379**	**0.1748**	**0.8644**	**0.2231**
	UniLearnLM	3	0.4250	0.1685	0.8613	0.1891
	UniLearnLM	Mean	**0.4383**	**0.1765**	**0.8646**	**0.1989**

Table 27: Per-fold breakdown of the **global-test, final-epoch model** result for the public testbed (sister to Table 24), available for Indiv and UniLearnLM only, because the fixed TC-Best rerun has no final-epoch model. Bold = best of {indiv, uni} per (institution, fold, metric). Cells are shaded by rank within each (institution, fold) group per column: darker green = best, lighter green = 2nd. Column abbreviations follow Table 2. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019), BART (Lewis et al., 2020), Flan-T5 (Chung et al., 2024), LED (Beltagy et al., 2020).

Inst.	Method	Fold	R-L	B-4	BERT	B_{cls}
$\mathcal{I}_0^{pub}$ (BART (Lewis et al., 2020), MIT)	indiv	1	**0.4240**	**0.1624**	**0.8584**	**0.2832**
	indiv	2	**0.4241**	**0.1584**	**0.8591**	**0.2736**
	indiv	3	**0.4372**	**0.1635**	**0.8622**	**0.1913**
	indiv	Mean	**0.4284**	**0.1614**	**0.8599**	**0.2493**
	UniLearnLM	1	0.4196	0.1566	**0.8584**	0.2483
	UniLearnLM	2	0.4152	0.1500	0.8558	0.2424
	UniLearnLM	3	0.4220	0.1594	0.8585	0.1786
	UniLearnLM	Mean	0.4189	0.1553	0.8576	0.2231
$\mathcal{I}_1^{pub}$ (Flan-T5 (Chung et al., 2024), Stanford)	indiv	1	0.4328	0.1611	0.8604	0.2784
	indiv	2	0.4411	0.1626	0.8623	0.2775
	indiv	3	0.4248	0.1506	0.8575	**0.3295**
	indiv	Mean	0.4329	0.1581	0.8601	0.2951
	UniLearnLM	1	**0.4354**	**0.1695**	**0.8647**	**0.3203**
	UniLearnLM	2	**0.4422**	**0.1662**	**0.8659**	**0.2806**
	UniLearnLM	3	**0.4291**	**0.1625**	**0.8621**	0.3266
	UniLearnLM	Mean	**0.4356**	**0.1661**	**0.8642**	**0.3092**
$\mathcal{I}_2^{pub}$ (LED (Beltagy et al., 2020), ML papers)	indiv	1	**0.4334**	0.1684	0.8621	0.1718
	indiv	2	0.4226	0.1568	0.8590	0.1539
	indiv	3	**0.4364**	**0.1706**	**0.8627**	**0.2258**
	indiv	Mean	**0.4308**	0.1653	0.8613	0.1838
	UniLearnLM	1	0.4298	**0.1705**	**0.8631**	**0.2172**
	UniLearnLM	2	**0.4277**	**0.1678**	**0.8625**	**0.2253**
	UniLearnLM	3	0.4127	0.1634	0.8585	0.2171
	UniLearnLM	Mean	0.4234	**0.1673**	**0.8614**	**0.2199**

cells are rank-shaded within each institution group, and − indicates an institution that did not connect a federation client to that particular live run (Section C.2).

Unlike the public testbed, where we report the mean over 3-fold cross-validation, the private-testbed federated methods are reported on a single fold (fold 1). Running three live federation rounds across five geographically distributed institutions requires coordinating five independent compute environments, time zones, and connectivity windows for every fold; doing this three times over was not logistically feasible within the project timeline. The individual-training baseline does not have this constraint and is therefore reported as the mean across all three folds, matching the public-testbed convention. Per-fold variance for the federated methods is therefore not estimated in the *live* private-testbed runs; the public-testbed per fold breakdown (Table 24, Appendix V) gives a calibration of how large that variance typically is for our federated training pipeline. Appendix AE additionally re-runs the same five private curricula as a controlled full-participation federation with all five clients present and full 3-fold cross-validation, which supplies the cross fold standard deviations the live runs cannot and removes the dropout confound discussed below; it is a re-execution on the same private data rather than a second live deployment.

Note that per-method KTI means aggregate over different pair sets: cells involving non-participating clients are undefined, and in particular the most resource-constrained institution ($\mathcal{I}_2^{priv}$) never appears as a receiver in the live UniLearnLM column, so cross-method comparison of the means should be made with care. Table 46 (Appendix AE) repeats this analysis under full participation, where all 20 directed pairs are defined for every method and the means are therefore directly comparable; the sign of the UniLearnLM mean reverses there, which we read as evidence that the negative live mean reflects the missing client rather than the method.

Table 28: In-distribution comparison on the public testbed: each client's **local test** set (validation-selected model, mean over 3 folds). Bold = best per (institution, metric). Cells are shaded by rank within each institution group per column: darkest green = best, then 2nd, 3rd. Column abbreviations are defined in the caption of Table 2. Red rows are the systems run under the identical protocol (Appendix U); their Faith is the sentence level variant (§, excluded from the ranking) and their GPT-4o-mini judge columns are not computed (−); * = fewer than three folds. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019), pFedHN (Shamsian et al., 2021), FedHeNN (Makhija et al., 2022), FedMKT (Fan et al., 2025), GPT-4o (Hurst et al., 2024).

Inst.	Method	Lightweight				LLM-based / reference-free						
		R-L	B-4	BERT	B_{cls}	B_{LLM}	AnsR	Faith	QAFE	RQU	RTC	LLM-J
$\mathcal{I}_0^{pub}$	indiv	**0.4803**	0.1999	0.8775	**0.2000**	0.3175	0.5086	0.2394	**0.3528**	4.1198	0.2322	0.6878
	pFedHN (Shamsian et al., 2021)	0.4775	0.1954	0.8793	0.1908	–	–	0.2368§	0.3375	4.1325	0.2359	–
	FedHeNN (Makhija et al., 2022)	0.4851	0.2120	0.8796	0.2042	–	–	0.2938§	0.3479	4.1766	0.2517	–
	FedMKT (Fan et al., 2025)	0.4373	0.1721	0.8708	0.1933	–	–	0.2715§	0.3792	4.1414	0.2336	–
	TC-Best (Li & Wang, 2019)	0.4755	0.2002	**0.8776**	0.1667	–	–	0.2687§	0.2819	**4.1734**	**0.2486**	–
	TC-Ens (Li & Wang, 2019)	0.4753	0.1960	0.8773	0.1417	–	–	0.2687§	0.3125	4.1229	0.2239	–
	UniLearnLM	0.4752	**0.2017**	0.8776	0.1475	**0.3275**	**0.5087**	**0.2831**	0.3229	4.1518	0.2342	**0.7143**
$\mathcal{I}_1^{pub}$	indiv	0.4793	0.1808	0.8681	0.2198	0.1975	0.5541	**0.2651**	**0.2757**	4.0613	0.2247	**0.6903**
	pFedHN	0.4556	0.1704	0.8714	0.1210	–	–	0.2572§	0.2510	4.1752	0.2849	–
	FedHeNN	0.4855	0.2038	0.8821	0.1901	–	–	0.2634§	0.2634	4.1527	0.2674	–
	FedMKT	0.4454	0.1724	0.8658	0.2189	–	–	0.2277§	0.2634	4.1398	0.2633	–
	TC-Best	0.4706	0.1927	0.8739	0.1556	–	–	0.2428§	0.2716	4.1609	0.2701	–
	TC-Ens	**0.4930**	**0.2094**	**0.8812**	0.1671	–	–	0.2387§	0.2469	**4.1632**	**0.2770**	–
	UniLearnLM	0.4682	0.1927	0.8767	**0.2535**	**0.2082**	**0.5558**	0.2536	0.2510	4.0992	0.2536	0.6620
$\mathcal{I}_2^{pub}$	indiv	0.4196	0.1556	0.8576	0.1706	0.2764	**0.4761**	0.2050	0.1424	4.0470	0.1855	0.6908
	pFedHN	0.4197	0.1562	0.8573	0.1879	–	–	0.1589§	0.1640	4.0201	0.1805	–
	FedHeNN	0.4290	0.1690	0.8610	0.1988	–	–	0.1777§	0.1563	4.0345	0.1892	–
	FedMKT	0.3984	0.1512	0.8516	0.2268	–	–	0.1623§	0.1348	4.0218	0.1785	–
	TC-Best	0.4252	0.1657	0.8595	**0.2013**	–	–	0.1642§	**0.1448**	4.0383	0.1856	–
	TC-Ens	0.4265	0.1658	0.8598	0.1978	–	–	0.1631§	0.1413	4.0432	0.1875	–
	UniLearnLM	**0.4369**	**0.1758**	**0.8638**	0.1947	**0.2831**	0.4724	**0.2078**	0.1433	**4.0507**	**0.1962**	**0.7065**
Mean	indiv	0.4597	0.1788	0.8677	0.1968	0.2638	**0.5129**	0.2365	**0.2570**	4.0760	0.2141	0.6896
	pFedHN	0.4509	0.1740	0.8693	0.1666	–	–	0.2176§	0.2508	4.1093	0.2337	–
	FedHeNN	0.4665	0.1949	0.8743	0.1977	–	–	0.2449§	0.2559	4.1212	0.2361	–
	FedMKT	0.4270	0.1652	0.8627	0.2130	–	–	0.2205§	0.2591	4.1010	0.2251	–
	TC-Best	0.4571	0.1862	0.8703	0.1745	–	–	0.2253§	0.2328	**4.1242**	**0.2348**	–
	TC-Ens	**0.4649**	**0.1904**	**0.8728**	0.1688	–	–	0.2235§	0.2336	4.1098	0.2294	–
	UniLearnLM	0.4601	0.1901	0.8727	**0.1986**	**0.2729**	0.5123	**0.2482**	0.2391	4.1006	0.2280	**0.6943**

Y Extended Results Discussion

This appendix discusses the results as numbered discoveries (Discovery 1 to Discovery 8), each a finding from the tables of the paper stated with its evidence and limits.

Discovery 7: the residual gap to the LLMs is on fluency and the holistic judge, and the judge column is an ordering only. GPT-4o, GPT-4o-mini and Gemini-2.5-Flash each lead the mixture on 7 of the 9 metrics of Table 34, lexical overlap and the holistic judge among them, and the weakest LLM leads it on 2. The judge rubric rewards fluent surface form and the judge shares a family with two scored systems (Panickssery et al., 2024; Zheng et al., 2023), and a Gemini-2.5-Flash replication ranks the nine systems almost identically (Appendix AD), so LLM-J is read as an ordering. On the depth proxy five of the six LLMs score above the mixture and the weakest scores highest, which says more about a proxy that rewards verb choice than about depth (Appendix Y).

The empirical core is the public testbed, evaluated on $\mathcal{T}^{global}$ as the mean over 3-fold cross-validation, every method as the model selected on validation data by ROUGE-L. The live private runs are reported in full in Appendix X. We compare the live runs institution by institution because their participant sets differ; Section 5.2 reports the counts. Discovery 5 draws on the full-participation re-run, which is participant matched and has folds; Appendix Y gives the long form of every discussion below.

Table 29: Equity comparison on the **private testbed** (Section C.2). Five institutions (Section C.2, Table 4), four methods per institution: the per-institution local baseline **Indiv** (Section C.5), the two FedMD-family public pool baselines **TC-Best** (selective consensus) and **TC-Ens** (ensemble consensus), both adapted from FedMD (Li & Wang, 2019), and our UniLearnLM. The eleven metric columns are identical to Table 2 (same abbreviations and Lightweight / LLM-based grouping); $\bar{E}$ is the mean Pairwise Equity Score (Section F.4). Indiv rows report each institution's local test split (topic conditioning, validation-selected model, mean over 3 folds). TC-Best, TC-Ens, UniLearnLM, and UniLearnLM+ rows are populated from the three live federations (fold 1 only). Where a site did not take part in a given live run we have no federation-trained model for that institution and the corresponding cells are left as −: institution $\mathcal{I}_2^{\text{priv}}$ (LED) was unable to connect a federation client to any of the live runs, so all four federated rows for $\mathcal{I}_2^{\text{priv}}$ are −; institution $\mathcal{I}_0^{\text{priv}}$ (BART) participated in TC-Best and UniLearnLM but was unable to connect to the TC-Ens and UniLearnLM+ live runs, so those two rows are − as well. Mean rows are averaged over each method's actual participant set (4 institutions for TC-Best/UniLearnLM, 3 for TC-Ens/UniLearnLM+). Sources: BART (Lewis et al., 2020), Flan-T5 (Chung et al., 2024), LED (Beltagy et al., 2020), Pegasus-X (Phang et al., 2023), ProphetNet (Qi et al., 2020).

Institution / SLM	Method	Lightweight				LLM-based / reference-free						
		R-L	B-4	BERT	B_{cls}	B_{LLM}	AnsR	Faith	QAFE	RQU	RTC	LLM-J
$\mathcal{I}_0^{\text{priv}}$ / BART-base (Lewis et al., 2020)	Indiv	0.3223	0.0657	0.8368	0.1633	0.2133	0.6920	0.2483	0.1722	4.0449	**0.1921**	0.7448
	TC-Best (Li & Wang, 2019)	0.3200	0.0631	0.8322	0.0700	0.1300	0.6299	**0.3367**	**0.2750**	4.0090	0.1755	0.7438
	TC-Ens (Li & Wang, 2019)	–	–	–	–	–	–	–	–	–	–	–
	UniLearnLM	**0.3341**	**0.0730**	**0.8376**	**0.3900**	**0.3700**	**0.7036**	0.2367	0.0750	**4.0970**	0.1879	**0.8125**
	UniLearnLM+	–	–	–	–	–	–	–	–	–	–	–
$\mathcal{I}_1^{\text{priv}}$ / Flan-T5-base (Chung et al., 2024)	Indiv	0.3571	0.1084	0.8398	0.2169	0.3109	0.5934	**0.1646**	0.0993	4.0613	0.1937	0.7954
	TC-Best	0.3612	0.1120	0.8421	0.1916	0.2919	0.5937	0.1621	0.1073	4.0661	0.1994	0.7962
	TC-Ens	0.3642	0.1137	**0.8441**	0.2348	0.3348	0.5913	0.1590	0.0895	4.0590	0.1907	**0.8163**
	UniLearnLM	0.3595	0.1120	0.8429	**0.2610**	**0.3532**	**0.5972**	0.1604	0.0949	**4.0706**	0.1980	0.8146
	UniLearnLM+	**0.3653**	**0.1205**	0.8425	0.2413	0.3345	0.5951	0.1640	**0.1091**	4.0682	**0.1997**	0.8048
$\mathcal{I}_2^{\text{priv}}$ / LED-base-16384 (Beltagy et al., 2020)	Indiv	**0.4138**	**0.1440**	**0.8557**	**0.0752**	**0.1292**	**0.5298**	**0.2213**	**0.1270**	**4.0467**	**0.2282**	**0.7313**
	TC-Best	–	–	–	–	–	–	–	–	–	–	–
	TC-Ens	–	–	–	–	–	–	–	–	–	–	–
	UniLearnLM	–	–	–	–	–	–	–	–	–	–	–
	UniLearnLM+	–	–	–	–	–	–	–	–	–	–	–
$\mathcal{I}_3^{\text{priv}}$ / Pegasus-X-base (Phang et al., 2023)	Indiv	**0.3508**	0.0948	0.8391	0.0841	0.2133	0.6802	**0.1835**	0.1276	4.0526	0.1846	0.8175
	TC-Best	0.3498	0.0959	**0.8394**	0.1210	0.2278	0.6751	0.1724	0.1207	4.0510	0.1827	0.8066
	TC-Ens	0.3497	0.0953	0.8386	0.1180	0.2273	0.6567	0.1735	**0.1329**	4.0352	0.1812	**0.8239**
	UniLearnLM	0.3462	0.0969	0.8383	0.1615	0.2595	**0.6833**	0.1599	0.1130	4.0513	0.1860	0.8189
	UniLearnLM+	0.3489	**0.0978**	0.8384	**0.1771**	**0.2620**	0.6652	0.1590	0.1301	**4.0587**	**0.1875**	0.8123
$\mathcal{I}_4^{\text{priv}}$ / ProphetNet-large (Qi et al., 2020)	Indiv	0.3344	0.0643	0.8332	0.1486	0.2781	0.6130	0.2130	0.0726	4.0249	0.1847	0.7548
	TC-Best	0.3398	0.1122	0.8358	**0.2800**	**0.3286**	**0.6258**	0.1881	0.0690	4.0966	0.2285	0.8045
	TC-Ens	0.3463	0.1082	0.8368	0.1800	0.2686	0.6149	0.2364	0.0786	4.0496	0.1981	0.8063
	UniLearnLM	0.3383	0.1050	**0.8374**	0.1143	0.2971	0.6156	0.1714	0.0643	**4.1096**	**0.2361**	**0.8107**
	UniLearnLM+	**0.3555**	**0.1148**	0.8374	0.1229	0.2743	0.6131	**0.2776**	**0.1214**	4.0260	0.1973	0.8098
Mean across $\mathcal{I}_{0..4}^{\text{priv}}$	Indiv	0.3557	0.0954	**0.8409**	0.1376	0.2290	0.6217	0.2061	0.1198	4.0461	0.1967	0.7687
	TC-Best	0.3427	0.0958	0.8374	0.1657	0.2446	0.6311	**0.2148**	**0.1430**	4.0557	0.1965	0.7878
	TC-Ens	0.3534	0.1057	0.8398	0.1776	0.2769	0.6210	0.1896	0.1003	4.0479	0.1900	**0.8155**
	UniLearnLM	0.3445	0.0967	0.8391	**0.2317**	**0.3200**	**0.6499**	0.1821	0.0868	**4.0821**	**0.2020**	0.8142
	UniLearnLM+	**0.3566**	**0.1110**	0.8394	0.1804	0.2903	0.6245	0.2002	0.1202	4.0510	0.1948	0.8090

Y.1 Reading the live private testbed runs

The empirical core is the public testbed, evaluated on a cross distribution global test pool $\mathcal{T}^{\text{global}}$ of 916 evaluation rows drawn from 179 held-out contexts, reporting the mean over 3-fold cross-validation. All methods are compared using the model selected by validation ROUGE-L (Indiv via early stopping, the federations via the best round). Results on the private testbed, the live five institution federation across Tunisia, Germany, Nigeria, Colombia, and Peru, mirror these discoveries on real institutional data and are reported in full in Tables 29–32 (Appendix X). We compare the live runs institution by institution because their participant sets differ; Section 5.2 reports the counts. Because the deployment lost clients to connectivity failures, the Mean rows of Tables 29–32 average over different participant sets, five institutions for local training against four or three for the federated methods, so they are not like for like. For a comparison that is participant matched and has folds, we re-ran the same five private curricula as a full participation 3-fold federation with no dropouts (Appendix AE), on which Discovery 5 draws. We discuss our discoveries in the following subsections.

Y.2 Discovery 1, extended discussion

Table 30: Cross-institution generalization on the **private global test set**: 996 QA pairs over 200 contexts, built by pooling the lecture and research material of the four dataset-only contributors $\mathcal{D}_0$–$\mathcal{D}_3$ (Table 4, Section C.2) and generating five GPT-4o QA pairs per context. No institution trains on this set, so it is the private analog of the public global mixed test (Table 2). Five institutions (Section C.2, Table 4), four methods per institution: the per-institution local baseline **Indiv** (Section C.5), the two FedMD-family public pool baselines **TC-Best** (selective consensus) and **TC-Ens** (ensemble consensus), both adapted from FedMD (Li & Wang, 2019), and our UniLearnLM. The eleven metric columns, abbreviations, and Lightweight / LLM-based grouping are identical to Table 2. Indiv rows report topic conditioning, validation-selected model, mean over 3 folds. TC-Best, TC-Ens, UniLearnLM, and UniLearnLM+ rows are populated from the three live federations (fold 1 only). $\mathcal{I}_2^{\text{priv}}$ (LED) did not participate in any online run and remains – for the federated methods; TC-Ens and UniLearnLM+ additionally have – for $\mathcal{I}_0^{\text{priv}}$ (BART) since only $\mathcal{I}_{1,3,4}^{\text{priv}}$ participated in those two runs. Mean rows are averaged over each method's actual participants: 4 for TC-Best/UniLearnLM ($\mathcal{I}_{0,1,3,4}^{\text{priv}}$), 3 for TC-Ens/UniLearnLM+ ($\mathcal{I}_{1,3,4}^{\text{priv}}$). Sources: BART (Lewis et al., 2020), Flan-T5 (Chung et al., 2024), LED (Beltagy et al., 2020), Pegasus-X (Phang et al., 2023), ProphetNet (Qi et al., 2020), GPT-4o (Hurst et al., 2024).

Institution / SLM	Method	Lightweight				LLM-based / reference-free						
		R-L	B-4	BERT	B_{cls}	B_{LLM}	AnsR	Faith	QAFE	RQU	RTC	LLM-J
$\mathcal{I}_0^{\text{priv}}$ / BART-base (Lewis et al., 2020)	Indiv	**0.3165**	**0.0776**	**0.8263**	0.2135	0.3161	**0.6257**	0.1538	0.0420	4.0321	0.1801	**0.7306**
	TC-Best (Li & Wang, 2019)	0.3042	0.0672	0.8194	0.0548	0.1962	0.5707	0.1503	**0.0614**	**4.0375**	**0.1836**	0.6840
	TC-Ens (Li & Wang, 2019)	–	–	–	–	–	–	–	–	–	–	–
	UniLearnLM	0.2991	0.0675	0.8214	**0.2502**	**0.3410**	0.6021	**0.1638**	0.0464	4.0127	0.1713	0.6942
	UniLearnLM+	–	–	–	–	–	–	–	–	–	–	–
$\mathcal{I}_1^{\text{priv}}$ / Flan-T5-base (Chung et al., 2024)	Indiv	0.3606	0.1111	0.8409	0.2102	0.2971	0.5947	0.1876	0.0570	4.0583	0.1955	0.8113
	TC-Best	**0.3667**	0.1152	0.8439	0.1972	0.3014	0.5992	0.1844	0.0589	4.0741	0.1993	0.8247
	TC-Ens	0.3613	0.1150	0.8440	0.2510	0.3273	0.5919	0.1772	0.0641	4.0722	0.2044	**0.8267**
	UniLearnLM	0.3653	**0.1174**	**0.8446**	**0.2512**	**0.3345**	**0.6037**	0.1765	**0.0649**	**4.0815**	**0.2084**	0.8259
	UniLearnLM+	0.3631	0.1137	0.8433	0.2492	0.3299	0.5984	**0.1886**	0.0568	4.0696	0.2021	0.8267
$\mathcal{I}_2^{\text{priv}}$ / LED-base-16384 (Beltagy et al., 2020)	Indiv	**0.2963**	**0.0525**	**0.8214**	**0.1286**	**0.1977**	**0.4602**	**0.1754**	**0.0763**	**4.0105**	**0.1675**	**0.7639**
	TC-Best	–	–	–	–	–	–	–	–	–	–	–
	TC-Ens	–	–	–	–	–	–	–	–	–	–	–
	UniLearnLM	–	–	–	–	–	–	–	–	–	–	–
	UniLearnLM+	–	–	–	–	–	–	–	–	–	–	–
$\mathcal{I}_3^{\text{priv}}$ / Pegasus-X-base (Phang et al., 2023)	Indiv	**0.3494**	0.1013	0.8355	0.1531	0.2644	0.6342	0.1861	0.0517	4.0679	0.1962	**0.8019**
	TC-Best	0.3443	0.1012	0.8355	0.2191	0.3076	0.6250	0.1735	**0.0562**	**4.0767**	**0.2025**	0.7900
	TC-Ens	0.3489	**0.1015**	0.8364	0.2088	0.2811	0.6084	**0.1894**	0.0507	4.0700	0.1989	0.7962
	UniLearnLM	0.3476	0.1015	**0.8376**	**0.2456**	**0.3131**	**0.6401**	0.1841	0.0527	4.0594	0.1912	0.7892
	UniLearnLM+	0.3448	0.0992	0.8353	0.2436	0.3084	0.6125	0.1852	0.0467	4.0620	0.1935	0.7826
$\mathcal{I}_4^{\text{priv}}$ / ProphetNet-large (Qi et al., 2020)	Indiv	0.3261	0.0649	0.8323	**0.2280**	**0.3518**	0.5626	0.1686	0.0542	4.0398	0.1816	0.7391
	TC-Best	0.3323	0.1132	0.8353	0.1825	0.2936	0.5631	0.1709	0.0474	**4.0597**	**0.1858**	0.7432
	TC-Ens	0.3419	0.1226	0.8341	0.1761	0.3014	0.5202	**0.2001**	0.0525	4.0370	0.1725	0.7472
	UniLearnLM	0.3313	0.1130	0.8336	0.1741	0.3165	**0.5678**	0.1829	**0.0556**	4.0274	0.1712	0.7482
	UniLearnLM+	**0.3474**	**0.1283**	**0.8379**	0.1679	0.3347	0.5314	0.1668	0.0537	4.0544	0.1849	**0.7791**
Mean across $\mathcal{I}_{0..4}^{\text{priv}}$	Indiv	0.3298	0.0815	0.8313	0.1867	0.2854	0.5755	0.1743	**0.0562**	4.0417	0.1842	0.7694
	TC-Best	0.3369	0.0992	0.8335	0.1634	0.2747	0.5895	0.1698	0.0560	**4.0620**	0.1928	0.7605
	TC-Ens	0.3507	0.1130	0.8382	0.2120	0.3033	0.5735	**0.1889**	0.0558	4.0597	0.1919	0.7900
	UniLearnLM	0.3358	0.0999	0.8343	**0.2303**	**0.3263**	**0.6034**	0.1768	0.0549	4.0452	0.1855	0.7644
	UniLearnLM+	**0.3518**	**0.1137**	**0.8388**	0.2202	0.3243	0.5808	0.1802	0.0524	4.0620	**0.1935**	**0.7961**

The per-institution table behind Discovery 1 is reproduced here.

Discovery 1: unified learning keeps question depth furthest above local training; the text-consensus routes fall below it. Pedagogical depth is where the methods separate. On the Mean row of Table 2, B_{cls} is 0.2238 for local training. The text-consensus federations lower it, TC-Best to 0.1623 (-27%) and TC-Ens to 0.1655 (-26%), and so does UniLearnLM+, which adds the public-pool channel to unified federation, to 0.1746 (-22%). UniLearnLM is highest, at 0.2313 ($+3\%$); the only other method above local training is FedMKT, at 0.2243. Per-institution values are in Table 35.

How much weight this result can carry. Everything above is measured with B_{cls}, and three properties of that measurement bound what it can support. We set them out here rather than in an appendix because they qualify the paper's main claim.

The effect is small in the metric's own units. B_{cls} is a Bloom level normalized by $\text{EVS}(q) = (\hat{b}(q) - 1)/5$, so one whole Bloom level is 0.200. The UniLearnLM advantage over local training on the Mean row, 0.0075, is under 4% of one level. These are shifts in the average level of a distribution of questions, not a change in what level the typical question sits at.

Table 31: **Pairwise Equity Score (PES) on the private testbed** for all five learning paradigms, with $Q_k^{(m)}$ taken as the LLM-judge overall on either (**top**) the global private test set $\mathcal{T}_{\text{priv}}^{\text{global}}$ (996 QA pairs, cross distribution generalization) or (**bottom**) each institution's own local private test split (within distribution). Federated methods are fold-1-only (live runs); Indiv is the 3-fold mean. Cells are – when either institution did not participate in the method's live run. The *common-pairs mean* row averages over the three institution pairs $(\mathcal{I}_1^{\text{priv}}, \mathcal{I}_3^{\text{priv}})$, $(\mathcal{I}_1^{\text{priv}}, \mathcal{I}_4^{\text{priv}})$, $(\mathcal{I}_3^{\text{priv}}, \mathcal{I}_4^{\text{priv}})$ on which every method has a value, enabling apples-to-apples comparison; no all pairs row is shown, because each method's coverage differs. Bold = best per row in each sub-block. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019).

PES pair $E_{kl}^{(m)}$	Indiv	TC-Best (Li & Wang, 2019)	TC-Ens (Li & Wang, 2019)	UniLearnLM	UniLearnLM+
*Q from global private test (**cross distribution**)*					
$(\mathcal{I}_0^{\text{priv}}, \mathcal{I}_1^{\text{priv}})$ Tunisia, Germany	**1.864**	1.824	–	1.838	–
$(\mathcal{I}_0^{\text{priv}}, \mathcal{I}_2^{\text{priv}})$ Tunisia, Nigeria	**1.445**	–	–	–	–
$(\mathcal{I}_0^{\text{priv}}, \mathcal{I}_3^{\text{priv}})$ Tunisia, Colombia	**2.245**	2.160	–	2.174	–
$(\mathcal{I}_0^{\text{priv}}, \mathcal{I}_4^{\text{priv}})$ Tunisia, Peru	**2.114**	2.052	–	2.074	–
$(\mathcal{I}_1^{\text{priv}}, \mathcal{I}_2^{\text{priv}})$ Germany, Nigeria	**2.439**	–	–	–	–
$(\mathcal{I}_1^{\text{priv}}, \mathcal{I}_3^{\text{priv}})$ Germany, Colombia	5.957	5.963	**5.993**	5.964	5.943
$(\mathcal{I}_1^{\text{priv}}, \mathcal{I}_4^{\text{priv}})$ Germany, Peru	1.790	1.810	1.817	1.817	**1.854**
$(\mathcal{I}_2^{\text{priv}}, \mathcal{I}_3^{\text{priv}})$ Nigeria, Colombia	**1.708**	–	–	–	–
$(\mathcal{I}_2^{\text{priv}}, \mathcal{I}_4^{\text{priv}})$ Nigeria, Peru	**0.994**	–	–	–	–
$(\mathcal{I}_3^{\text{priv}}, \mathcal{I}_4^{\text{priv}})$ Colombia, Peru	2.588	2.575	2.592	2.582	**2.623**
Common-pairs mean $(\mathcal{I}_1^{\text{priv}},\mathcal{I}_3^{\text{priv}})+(\mathcal{I}_1^{\text{priv}},\mathcal{I}_4^{\text{priv}})+(\mathcal{I}_3^{\text{priv}},\mathcal{I}_4^{\text{priv}})$	3.445	3.449	3.467	3.454	**3.473**
*Q from each institution's local private test (**within distribution**)*					
$(\mathcal{I}_0^{\text{priv}}, \mathcal{I}_1^{\text{priv}})$ Tunisia, Germany	1.862	1.862	–	**1.968**	–
$(\mathcal{I}_0^{\text{priv}}, \mathcal{I}_2^{\text{priv}})$ Tunisia, Nigeria	**1.427**	–	–	–	–
$(\mathcal{I}_0^{\text{priv}}, \mathcal{I}_3^{\text{priv}})$ Tunisia, Colombia	2.289	2.272	–	**2.390**	–
$(\mathcal{I}_0^{\text{priv}}, \mathcal{I}_4^{\text{priv}})$ Tunisia, Peru	2.156	2.226	–	**2.334**	–
$(\mathcal{I}_1^{\text{priv}}, \mathcal{I}_2^{\text{priv}})$ Germany, Nigeria	**2.364**	–	–	–	–
$(\mathcal{I}_1^{\text{priv}}, \mathcal{I}_3^{\text{priv}})$ Germany, Colombia	5.956	5.918	**6.057**	6.032	5.972
$(\mathcal{I}_1^{\text{priv}}, \mathcal{I}_4^{\text{priv}})$ Germany, Peru	1.790	1.848	1.873	**1.876**	1.864
$(\mathcal{I}_2^{\text{priv}}, \mathcal{I}_3^{\text{priv}})$ Nigeria, Colombia	**1.690**	–	–	–	–
$(\mathcal{I}_2^{\text{priv}}, \mathcal{I}_4^{\text{priv}})$ Nigeria, Peru	**0.983**	–	–	–	–
$(\mathcal{I}_3^{\text{priv}}, \mathcal{I}_4^{\text{priv}})$ Colombia, Peru	2.641	2.706	**2.738**	2.737	2.724
Common-pairs mean $(\mathcal{I}_1^{\text{priv}},\mathcal{I}_3^{\text{priv}})+(\mathcal{I}_1^{\text{priv}},\mathcal{I}_4^{\text{priv}})+(\mathcal{I}_3^{\text{priv}},\mathcal{I}_4^{\text{priv}})$	3.462	3.491	**3.556**	3.548	3.520

Table 32: **Knowledge Transfer Index (KTI) on the private testbed**, extended to all four federated methods (each vs the Indiv baseline). Computed on the global private test set ($\mathrm{T}_{k\setminus l}$ is only sensible on the cross institution pool). 16 directed pairs $(l \leftarrow k)$: the receiver l is one of the four institutions that ran a client in at least one live run ($\mathcal{I}_2^{\text{priv}}$ never connected, so it appears only as a topic source k, whose topic set is defined by its curriculum alone). Cells are – when the receiver l did not participate in that method's live run, so that $\mathrm{Q}_l^{(m)}$ is undefined. Positive = federated method gains over Indiv on topics in k's curriculum but not l's. Bold = best (most positive) per row. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019).

KTI $(l \leftarrow k;\ m)$	TC-Best (Li & Wang, 2019)	TC-Ens (Li & Wang, 2019)	UniLearnLM	UniLearnLM+
$\mathcal{I}_1^{\text{priv}}$ Germany ← $\mathcal{I}_0^{\text{priv}}$ Tunisia	−0.092	−0.029	−0.054	**+0.058**
$\mathcal{I}_3^{\text{priv}}$ Colombia ← $\mathcal{I}_0^{\text{priv}}$ Tunisia	−0.051	−0.060	**−0.006**	−0.015
$\mathcal{I}_4^{\text{priv}}$ Peru ← $\mathcal{I}_0^{\text{priv}}$ Tunisia	−0.047	**+0.047**	−0.036	+0.010
$\mathcal{I}_0^{\text{priv}}$ Tunisia ← $\mathcal{I}_1^{\text{priv}}$ Germany	−0.076	–	**−0.044**	–
$\mathcal{I}_3^{\text{priv}}$ Colombia ← $\mathcal{I}_1^{\text{priv}}$ Germany	−0.032	**−0.005**	−0.064	−0.010
$\mathcal{I}_4^{\text{priv}}$ Peru ← $\mathcal{I}_1^{\text{priv}}$ Germany	−0.021	−0.019	+0.014	**+0.025**
$\mathcal{I}_0^{\text{priv}}$ Tunisia ← $\mathcal{I}_2^{\text{priv}}$ Nigeria	**−0.012**	–	−0.038	–
$\mathcal{I}_1^{\text{priv}}$ Germany ← $\mathcal{I}_2^{\text{priv}}$ Nigeria	+0.058	+0.033	**+0.065**	+0.052
$\mathcal{I}_3^{\text{priv}}$ Colombia ← $\mathcal{I}_2^{\text{priv}}$ Nigeria	+0.014	+0.049	**+0.076**	+0.049
$\mathcal{I}_4^{\text{priv}}$ Peru ← $\mathcal{I}_2^{\text{priv}}$ Nigeria	**+0.103**	−0.030	+0.025	+0.076
$\mathcal{I}_0^{\text{priv}}$ Tunisia ← $\mathcal{I}_3^{\text{priv}}$ Colombia	−0.078	–	**−0.019**	–
$\mathcal{I}_1^{\text{priv}}$ Germany ← $\mathcal{I}_3^{\text{priv}}$ Colombia	+0.021	−0.009	+0.002	**+0.028**
$\mathcal{I}_4^{\text{priv}}$ Peru ← $\mathcal{I}_3^{\text{priv}}$ Colombia	−0.031	−0.002	**+0.022**	+0.015
$\mathcal{I}_0^{\text{priv}}$ Tunisia ← $\mathcal{I}_4^{\text{priv}}$ Peru	−0.051	–	**−0.026**	–
$\mathcal{I}_1^{\text{priv}}$ Germany ← $\mathcal{I}_4^{\text{priv}}$ Peru	−0.003	+0.000	−0.001	**+0.015**
$\mathcal{I}_3^{\text{priv}}$ Colombia ← $\mathcal{I}_4^{\text{priv}}$ Peru	+0.016	**+0.049**	−0.052	+0.029
Defined pairs (non-–)	16	12	16	12
Positive-pair count	5	5	6	10
Mean (over defined pairs)	−0.018	+0.002	−0.009	**+0.028**

The classifier is not validated on this domain. B_{cls} comes from a publicly released model (Appendix M) that we did not train and whose published card reports no accuracy figure on any Bloom benchmark. We did not hand label a sample of our own generated questions to estimate its agreement with human Bloom labels, and Bloom level annotation is known to be difficult even for trained annotators. Until that check exists, B_{cls} should be read as a consistent automatic proxy rather than as a validated measure of cognitive depth.

The second proxy does not agree with the first. We report two independent depth estimators and they disagree. We therefore state the depth result as holding *under B_{cls}*, note that B_{LLM} supports the same sign and a much smaller effect, and treat reconciling the two, by hand labeling a sample and reporting each estimator's agreement with human Bloom labels, as work this paper has not done. The same ordering appears on the private testbed under full participation (Table 45), where local training scores 0.1867 and ensemble distillation scores 0.1625. Two independent testbeds agreeing on a gap of 0.06 to 0.07 between unified learning and text consensus distillation is the most robust result in this paper, and it is a result about the baselines: distillation over a public pool buys surface fluency and pays for it in question depth.

The mechanism is not new even if the measurement in this domain is. Gudibande et al. (2024) show that models trained to imitate the outputs of a stronger system learn its style while closing little of the underlying capability gap, which is the same shape as what we observe: the distillation baselines gain on ROUGE-L, BLEU-4 and BERTScore and lose on the axis that measures what the question asks a student to do. Reading our result through theirs also sets its scope. It says that text level imitation of peer generations is a poor channel for transferring pedagogical depth. It does not say that distillation is a poor channel in general.

We are deliberately more cautious about the gap between UniLearnLM and local training itself. That gap is $+0.0075$ on B_{cls}, while the cross fold standard deviations for the same two systems are ± 0.0113 and ± 0.0156 (Appendix V), so with three folds and one seed we cannot separate it from run to run variation. On the holistic judge the picture is better behaved, $+0.0076$ against fold standard deviations of ± 0.0023 and ± 0.0010, and that is the one per metric effect here that clears fold level noise. Appendix AF returns to what would settle this. Counting first places, UniLearnLM+ leads on $3/11$ metrics (ROUGE-L, BLEU-4, BERTScore) and UniLearnLM on $3/11$ (B_{cls}, B_{LLM}, RQUGE), so the two unified federation methods together take $6/11$. Of the remaining five, local training with no federation at all takes two (Answer Relevancy 0.5190 and QAFactEval 0.1832) and the distillation baselines take three. We give no head to head win count, because several of those wins are margins below 0.001.

How much weight this result can carry, extended. Three properties of B_{cls} bound what it can support. *The effect is small in the metric's own units*: $\mathrm{EVS}(q) = (\hat{b}(q) - 1)/5$, so one Bloom level is 0.200, and UniLearnLM's 0.0075 advantage over local training is under 4% of one. *The classifier is not validated on this domain*: it is a publicly released model (Appendix M) with no reported accuracy on any Bloom benchmark, and we did not hand label a sample of our generated questions to check it, so it is a consistent automatic proxy, not a validated measure of cognitive depth. *The second proxy disagrees*: the two estimators are not measuring the same thing, so we state the depth result as holding *under B_{cls}*. The same ordering appears on the private testbed under full participation (Table 45: local training 0.1867, ensemble distillation 0.1625), and two testbeds agreeing on a gap of 0.06 to 0.07 between unified learning and text consensus distillation is the most robust result in this paper. It is a result about the baselines, and it has the shape Gudibande et al. (2024) describe: imitating a stronger system's outputs transfers style, not capability, and our distillation baselines gain on ROUGE-L, BLEU-4, and BERTScore while losing on the axis that measures what a question asks a student to do. The gap between UniLearnLM and local training itself, $+0.0075$ against cross fold standard deviations of ± 0.0113 and ± 0.0156 (Appendix V), is stated as a direction only; on the holistic judge $+0.0076$ against cross fold standard deviations of ± 0.0023 and ± 0.0010 has the same sign in all three folds (per fold client means 0.6911, 0.6895, 0.6940 for local training and 0.6981, 0.6993, 0.7000 for UniLearnLM), and a paired t-test across folds gives $p = 0.022$. Counting first places on the Mean row, UniLearnLM+ and UniLearnLM lead on $3/11$ metrics each, local training on two, and the distillation baselines on three; we drop our earlier $33/44$ head to head count because many of those margins are below 0.001.

Y.3 Discovery 2, extended discussion

Discovery 2: the benefit reaches every institution relative to the alternatives, and it is largest where the curriculum is largest. The Mean row hides who actually gains. Table 33 decomposes it. On B_{cls} the change from local training to UniLearnLM is -0.0201 at $\mathcal{I}_0^{pub}$, $+0.0050$ at $\mathcal{I}_1^{pub}$, and $+0.0378$ at $\mathcal{I}_2^{pub}$. Since $\mathcal{I}_2^{pub}$ holds 981 chunks and 5,037 QA pairs against roughly 110 chunks and 550 QA pairs at each of the other two, the institution with about nine times more material than either of the other two, and four and a half times their combined total, contributes more than the whole of the mean effect, and the smallest institution moves the other way. On the holistic judge the pattern is the same in direction if not in sign at $\mathcal{I}_1^{pub}$: $+0.0085$, -0.0016, and $+0.0158$.

This is the opposite of a uniform gain, which is why we state the equity result as depth preserved relative to the distillation alternatives rather than as a gain at every site. The leave one out study (Table 41, Appendix AA) points the same way. When $\mathcal{I}_2^{pub}$ is removed from the cohort, the two remaining clients reach a federated B_{cls} of 0.2053 in the two client federation and 0.2476 in the full three client federation, against 0.2551 for local training, so without the data rich client the federation sits below local training on depth.

We read this as bounding *Hypothesis 1*: the benefit is unevenly distributed and, on the depth proxy, negative for the smallest participant, so the equity claim holds relative to the distillation alternatives rather than as a uniform gain. Two things in our own numbers stop us from claiming that the benefit tracks contributed data. First, $\mathcal{I}_0^{pub}$ and $\mathcal{I}_1^{pub}$ hold the same number of chunks and differ by twelve QA pairs, which is 2.2%, yet their B_{cls} deltas have opposite signs. A monotone dependence on contributed data cannot be read off three points of which two are tied on the independent variable and disagree on the dependent one. Second, institution, curriculum and backbone are perfectly confounded here: $\mathcal{I}_0^{pub}$ is also the smallest backbone (BART-base, 140M) and $\mathcal{I}_2^{pub}$ is also the only LED client, so "negative at the smallest curriculum" and "negative at the smallest model" are observationally identical. The leave one out study does not separate them either, since removing $\mathcal{I}_2^{pub}$ removes the largest curriculum, the broadest curriculum and the LED backbone in one move. The experiment that would identify the cause is to subsample $\mathcal{I}_2^{pub}$'s curriculum to 110 chunks and re-run the three client federation; we have not run it, and Appendix AF lists it as the most valuable missing experiment.

A third caution applies to the whole of this subsection. The global pool is not balanced across institutions: the per-institution test splits are $|\mathcal{T}_0|$=80, $|\mathcal{T}_1|$=81 and $|\mathcal{T}_2|$=755, so $\mathcal{I}_2^{pub}$ supplies 82.4% of the 916 evaluation rows (Table 11). $\mathcal{I}_2^{pub}$'s model is therefore scored close to in distribution while the other two are scored largely out of distribution, and "the institution with the most data gains most" has an equally available alternative reading, namely "the institution whose distribution dominates the pool gains most". Recomputing the Mean row on an institution balanced subsample of the pool would separate these, uses generations we already have, and is not done here.

What does hold at $\mathcal{I}_0^{pub}$ is a weaker but still useful claim about the baselines. Distillation over a public pool drops B_{cls} there from 0.2384 to 0.1515, a relative fall of 36%, while UniLearnLM falls to 0.2183, a relative fall of 8%. So at the smallest client, unified federation does not improve depth but it does preserve most of it, whereas the standard route removes a third of it. On the reference-free axes at $\mathcal{I}_2^{pub}$, UniLearnLM ranks first on five of the seven columns that Table 2 groups under that heading; UniLearnLM+ leads on the remaining two, Faithfulness (0.2188) and the holistic judge (0.7098). Finally, the rank order is not stable across folds. In the per fold breakdown (Table 24, Appendix V), UniLearnLM's B_{cls} rank at $\mathcal{I}_2^{pub}$ moves between first and third across the three folds, and at $\mathcal{I}_0^{pub}$ the BERTScore gap over local training (0.0009) is smaller than the cross fold standard deviation of local training itself (≈ 0.0031).

Y.4 The in distribution ordering

The ordering reverses on each institution's own test split, and we report it. Everything above is measured on the cross distribution global pool, where each model is scored on material from all three curricula. On each institution's own in distribution test split the ranking is different, and the

Table 33: Per institution change from local training to UniLearnLM on the two pedagogical depth axes, computed from Table 2. The mean gain is carried by $\mathcal{I}_2^{\text{pub}}$, the institution with by far the largest curriculum, and is negative at $\mathcal{I}_0^{\text{pub}}$, the smallest. Corpus sizes are chunks / QA pairs from Table 11. Δ is computed from unrounded cells; the rounded headline difference on B_{cls} is 0.0075. Sources: BART (Lewis et al., 2020), Flan-T5 (Chung et al., 2024), LED (Beltagy et al., 2020).

Institution	Corpus	B_{cls}			LLM-J		
		indiv	UniLearnLM	Δ	indiv	UniLearnLM	Δ
$\mathcal{I}_0^{\text{pub}}$ (MIT, BART (Lewis et al., 2020))	110 / 547	0.2384	0.2183	**−0.0200**	0.6716	0.6801	+0.0085
$\mathcal{I}_1^{\text{pub}}$ (Stanford, Flan-T5 (Chung et al., 2024))	110 / 559	0.2718	0.2768	+0.0050	0.7138	0.7122	−0.0016
$\mathcal{I}_2^{\text{pub}}$ (arXiv, LED (Beltagy et al., 2020))	981 / 5,037	0.1611	0.1989	**+0.0378**	0.6893	0.7051	+0.0159
Mean		0.2238	0.2313	+0.0076	0.6915	0.6991	+0.0076

difference is not small. At $\mathcal{I}_0^{\text{pub}}$ specifically UniLearnLM scores 0.1475 against 0.2000 for local training.

This matters because the in distribution split is the setting our own motivating scenario describes (Appendix A): an instructor generating self assessment items from their own reading list. We read the two results together as a trade. Unified federation gives up in distribution surface fit and gains generalization to material the institution does not hold, which is consistent with the transfer picture in Table 46. We did not design an experiment to test that reading, so we offer it as an interpretation and not as a finding. What we do not do is leave the reversal in an appendix without comment, because a reader entitled to the global pool result is equally entitled to this one.

Y.5 Discovery 3, extended discussion

Discovery 3: pooling the data is not a uniform upper bound. The *Centralized* model leads the Mean row on the five surface and holistic metrics (LLM-J 0.7109 against 0.6991), while local training and UniLearnLM stay above it on the three reference-free answer quality metrics and UniLearnLM also on B_{cls} (0.2313 against 0.2122), RQUGE, and RTC: pooling averages out per-institution signal that local fine tuning keeps, and federation still pays on the holistic judge. The public testbed equity scalars (Table 7) do not discriminate between methods, since PES re-ranks methods identically to pair summed quality for a fixed pair and the three curricula overlap heavily; the equity evidence we rely on is the private testbed under full participation, where the curricula are disjoint.

The *Centralized* baseline pools all three institutions' curricula into a single model, giving up privacy for what should be an upper bound. It is not one. On the Mean row of Table 2 it leads on the five surface and holistic metrics (ROUGE-L, BLEU-4, BERTScore, B_{LLM}, and the judge at 0.7109 against 0.6991), while local training and UniLearnLM both stay above it on the three reference-free answer quality metrics (Answer Relevancy, Faithfulness, QAFactEval) and UniLearnLM stays above it on B_{cls} (0.2313 against 0.2122), RQUGE, and RTC. The reading we take from this is narrow: pooling averages out per-institution signal that local fine tuning keeps, so a federated method can match or beat it on answer grounding and question depth while still trailing it on fluency. We do not claim that federation is free of a performance cost, because on the holistic judge it is not.

Table 7 does *not* corroborate the claim above. Its all pairs mean PES ranks UniLearnLM (1.916) within 0.017 of UniLearnLM+ (1.902) and local training (1.899), and on the six directed KTI pairs UniLearnLM+ takes four and UniLearnLM two. The text-consensus reruns and the published heterogeneous federations have no LLM-judge scores, so they do not appear in that table. On the public testbed the equity scalars simply do not discriminate between methods. Two reasons. PES is a presentational lens rather than an independent measurement, since for a fixed institution pair it re-ranks methods identically to pair summed quality and adds only a resource gap normalization (Appendix R). And public testbed KTI magnitudes are small because the three curricula overlap heavily, so cross curriculum transfer has little room to show. The equity evidence we do rely on is the private testbed under full participation (Table 46), where the curricula are genuinely disjoint.

Y.6 Discovery 4, extended discussion

Discovery 4: removing the averaging step does not change question depth. UniLearnLM adds 15 to 50% trainable parameters over a LoRA only client, so a gain over local training could come from the extra capacity rather than from the federation, and the *no-FedAvg* ablation is the control that separates the two. It keeps the whole architecture at an identical parameter count and never averages the encoder, so the two arms differ in one step only. Run on three folds under the same settings as unified learning (Table 37), it does not separate from unified learning on pedagogical depth: the difference in B_{cls} is $+0.0057$, $+0.0057$ and -0.0060 on the three folds, a mean of $+0.0018$. Two folds favor the federation and one goes against it by about the same amount, so the sign is not stable, and the spread across folds, 0.0116, is four times the only same setup estimate of seed variation we have, 0.003 from a pair of runs. That estimate rests on two runs and is not a standard deviation, which is the second reason not to read a depth effect into these numbers. Surface overlap behaves the same way, with mean differences of -0.0009 on ROUGE-L and -0.0007 on BERTScore. Faithfulness is $+0.0070$, -0.0044 and $+0.0070$ across the folds, mean $+0.0032$; we report the number and make no claim from it.

The runs reported here give 0.2028 for unified learning and 0.2010 for no-FedAvg. The difference is not significant. Three further ablations locate the contribution of the encoder itself: a parameter matched MLP that ignores the graph edges loses 0.0278 LLM-J, freezing the dynamic LoRA statistic node features costs pedagogical depth while helping surface overlap, and replacing the encoder with a single averaged shared vector drops below local training on B_{cls} and AnsR. Depth is an interior optimum: the 3-layer GAT beats both the 2-layer and the 4-layer variant on ten of eleven metrics. The shared vector and GAT depth ablations run on fold 1 only and the encoder ablations on folds 2 and 3, so they carry less weight than the headline table.

Y.7 Discovery 5, extended discussion

Discovery 5: with no dropouts, the depth result reproduces on real institutional curricula. The live deployment lost its most resource-constrained client to connectivity failures, which both weakens and biases the live equity evidence (Appendix X). We therefore re-ran the same five private curricula with all five institutions present in every round and full 3-fold cross-validation (Appendix AE). We stress that this is a controlled re-execution from one coordination point and not a second live deployment, and the two should not be conflated: every claim in this paragraph comes from the re-run.

The unified pair is best on 5 of 11 metrics (BERTScore, B_{cls}, B_{LLM}, AnsR, QAFactEval), TC-Best on four (ROUGE-L, BLEU-4, RQUGE, RTC) and TC-Ens on the other two, the holistic judge and Faithfulness; four metrics were not computed for TC-Best. In Table 45, UniLearnLM (.0932) is above local training (.0815) on BLEU-4 and below TC-Best, TC-Ens and UniLearnLM+. The depth result reproduces cleanly: UniLearnLM reaches 0.2218 on B_{cls} against 0.1867 for local training, while TC-Ens falls to 0.1625, so the gap between unified federation and ensemble distillation is 0.059 on that axis and 0.038 on B_{LLM}. On the Knowledge Transfer Index, now defined on all 20 receiver and source pairs for every method, the two unified variants attain the highest means over the twenty cells, UniLearnLM+ $+0.0098$ and UniLearnLM $+0.0097$, a difference inside rounding; more usefully, because it does not depend on magnitudes, UniLearnLM has the most pairs with positive transfer ($14/20$ against $12/20$ for each of the two competitors). Its two largest positive cells, Germany←Nigeria at $+0.0813$ and Colombia←Nigeria at $+0.0602$, both have the most resource-constrained institution as the *source*, and Nigeria is the source in three of the five largest. This is suggestive of the federation propagating topics only the under resourced site teaches, but the cross fold standard deviation on the KTI mean (± 0.0179) is larger than the mean itself, so we present it as a direction rather than a measured effect.

Y.8 Discovery 6, extended discussion

Discovery 6: the federation leads every zero shot LLM on both grounding metrics, and most of that lead is explained by shorter answers. Against six single LLM baselines prompted zero-shot (Table 34; GPT-4o family, Llama/Qwen/Mistral 7–8B, Gemini-2.5-Flash; greedy decoding, fixed template), the UniLearnLM mixture scores highest of all seven systems on the two reference-free grounding metrics, RAGAS Faithfulness (0.2104) and QAFactEval (0.1788). Prompting the LLMs harder does not close it: re-prompting the strongest baseline with k=3 in-context examples from the

Table 34: *Hypothesis 2*: the UniLearnLM federated mixture against six single LLM baselines prompted zero-shot on $\mathcal{T}^{\text{global}}$ (916 rows; mean over 3 folds). The UniLearnLM row comes from the mixture pipeline, its lightweight metrics from the lightweight pass and the remaining metrics from the heavy pass, whose own lightweight values differ by at most 0.0005 (Appendix Y). Bold = best per metric; shading by rank per column. †: self judged; ‡: same family as the judge. Gemini-2.5-Flash is decoded under a 256-token budget (others 160) because the shared cap truncated most of its outputs. Answer lengths are in Table 23.

Method	R-L	B-4	BERT	RTC	Faith	QAFE	RQU	B_{cls}	LLM-J
GPT-4o (Hurst et al., 2024)	0.4487	0.1869	0.8725	0.2095	0.1764	0.1335	4.0826	0.2624	0.9523‡
GPT-4o-mini (Hurst et al., 2024)	0.4484	0.1812	0.8707	0.2155	0.1681	0.1150	4.0858	0.2585	**0.9542**†
Llama-3.1-8B-Inst (Grattafiori et al., 2024)	0.4571	0.1991	0.8678	0.1893	0.1768	0.1578	4.0607	0.1520	0.8985
Qwen-2.5-7B-Inst (Qwen Team, 2024)	0.4586	0.1917	0.8680	0.2124	0.1824	0.1004	4.0732	0.2605	0.9378
Mistral-7B-Inst (Jiang et al., 2023)	0.3724	0.1305	0.8437	0.1670	0.1778	0.1242	4.0165	**0.2686**	0.9308
Gemini-2.5-Flash (Comanici et al., 2025)	**0.4896**	**0.2069**	**0.8807**	**0.2220**	0.1844	0.1323	**4.0932**	0.2417	0.9395
UniLearnLM (fed. mixture)	0.4394	0.1688	0.8629	0.2094	**0.2104**	**0.1788**	4.0775	0.2296	0.6986

same institution buys GPT-4o a large jump in reference overlap (ROUGE-L 0.4487 → 0.5513) but leaves it below the mixture on both (Faithfulness 0.1881, QAFactEval 0.1463; Appendix AC).

We do not read this as evidence of better grounding. Both metrics ask whether the claims of an answer are entailed by, or recoverable from, the context, so both rise as an answer gets shorter and copies more. Ladhak et al. (2022) show for summarization that apparent faithfulness gains routinely come from increased extractiveness, and argue that systems should be compared on a faithfulness and abstractiveness curve rather than on a raw score. We can document the length part of that check. On fold 1, the UniLearnLM mixture answers in 14.8 words on average, against 19.5 (Llama-3.1-8B) to 47.1 (GPT-4o) words for the six LLMs (Table 23, rows $K{=}\infty$), so its answers are shorter than those of every LLM. Coverage, density and novel-bigram statistics are not reported, and we report no correlation between those statistics and the grounding metrics. The grounding columns are therefore confounded with answer length, and the length-matched comparison below is the control we rely on.

We therefore restate the claim. UniLearnLM produces answers that stay closer to the source than a zero-shot LLM does, and on context entailment metrics that is rewarded. That is a real and useful property for self assessment items, where an answer a student can check against the reading is often what an instructor wants. It is not a demonstration that a federation of 140M to 392M parameter models grounds text as well as a frontier LLM; the claim we make is approximation of the LLMs' grounding at a fraction of their size. *Hypothesis 2* is supported in that form. The length matched comparison that settles the underlying question is a decode and score pass over stored contexts rather than a retraining run: cap every system at a common answer budget and re-score Faithfulness and QAFactEval. Table 23 (Appendix U) reports it: at a common budget of 15 or 20 words the mixture keeps a lead on both metrics, under half a point on Faithfulness and about one point on QAFactEval.

Y.9 Discovery 7, extended discussion

UniLearnLM trails the stronger LLMs on lexical overlap (R-L, B-4, BERT) and on the holistic GPT-4o-mini judge, where the gap is large (≈ 0.70 against 0.90–0.95). Two factors compound. The holistic rubric rewards fluent surface form beyond its four explicit criteria, a margin SLM decoders trained on small curricula do not close even when they are factually correct. And the GPT-4o-mini judge shares a family with two of the scored systems, GPT-4o and itself, which inflates their scores; Panickssery et al. (2024) show that LLM evaluators recognize their own generations and prefer them, and Zheng et al. (2023) document the related position and verbosity biases. Tables 13–15 (Appendix Q) illustrate the tradeoff on three examples: UniLearnLM states facts that Indiv and TC-Best get wrong, but its wording diverges from the gold reference, which was itself LLM generated. Replicating the whole judging protocol with a non-OpenAI judge (Gemini-2.5-Flash) shows that the ordering is not a family artifact: absolute scores shift sharply, but the two judges rank the nine systems almost identically (Spearman $\rho{=}0.93$, Pearson $r{=}0.99$), and the new judge inflates its own family in the same way (Appendix AD). We therefore read the LLM-judge column only as a relative ordering. One caveat we cannot remove by replication: the claim decomposer inside RAGAS Faithfulness is

itself GPT-4o-mini (Appendix F), so the grounding column is not fully independent of the OpenAI family either.

Taken together with Discovery 6, the honest summary of this subsection is that a federation of small models does *not* reach LLM level quality on this task. It is behind on fluency and on holistic judgment, and its advantage on context entailment is explained by answer length and copying. What it offers instead is per-institution privacy, a compute footprint 18 to 57× smaller than the open 7 to 8B baselines, and question depth that no distillation baseline matches.

Y.10 Discovery 8, extended discussion

Adding the encoder and modulator raises training time per epoch by 2% (Pegasus-X) to 130% (LED) and peak memory by a 0.21 to 0.27M parameter FiLM MLP per client, while each round exchanges only the 0.088M parameter encoder, 7 to 27× smaller than the LoRA adapters a weight sharing federation would send (Table 8, Appendix P). The table's faster per sample inference for UniLearnLM cannot come from the extra forward pass; the likely cause is shorter outputs, and we claim no speedup.

Discovery 8: training cost rises by a model dependent amount; communication does not. Table 8 (Appendix P) compares the per client compute budget. Adding the encoder and the modulator raises training time per epoch by between 2% (Pegasus-X) and 130% (LED) depending on the backbone, which for LED is a real cost and not a rounding error, and it raises peak memory because each client stores a FiLM MLP of 0.21 to 0.27M parameters alongside its LoRA adapter. Communication is the part that is close to free: each round exchanges the 0.088M parameter encoder, which is 7 to 27× smaller than the LoRA adapters a weight sharing federation would send and three to four orders of magnitude smaller than the backbones, and it does not grow with the size of the largest participant.

One number in that table needs a caveat rather than an explanation. Per sample inference is reported as *faster* for UniLearnLM than for the LoRA only client on BART (484.3 against 613.8 ms), LED (520.9 against 671.7) and Pegasus-X (805.8 against 882.1), under a fixed beam width. Adding a forward pass cannot reduce compute, so the most likely explanation is that the federated models emit shorter sequences, which is consistent with the answer lengths in Table 23. We report the measurements as taken and do not claim an inference speedup; timings normalized per generated token, with repeats, are needed to settle it.

Y.11 Note on the UniLearnLM row of Table 34

The UniLearnLM values in Table 34 come from the mixture evaluation pipeline, which scores the pooled generations of the three backbones in one pass, whereas the Mean row of Table 2 averages the three per-institution scores. The two therefore differ in the third or fourth decimal on every shared metric (for example R-L 0.4394 against 0.4385, QAFE 0.1788 against 0.1808); every difference is below 0.01, but on RTC (0.2094 against 0.2130) the choice moves the mixture from fifth to third of seven in Table 34, so per-institution claims cite Table 2. We report both rather than silently harmonize them; Table 2 is the one to cite for per-institution claims, and the ordering conclusions of Table 34 are unchanged if its UniLearnLM row is replaced by the Mean row of Table 2, since 0.2092 and 0.1808 still exceed every LLM value in the Faith and QAFE columns.

Z Ablation and Significance Analysis

Depth: paired, context-clustered bootstrap. Table 36 gives the difference in B_{cls} between UniLearnLM and each comparator on the public global test set, for the three UniLearnLM runs: the run tabled in Table 2, a rerun with FiLM-aware model selection, and a second seed (fold 1 only). Resampling unit: the 179 test contexts (5,000 resamples); estimand: the evaluation contexts given the trained models, so fold and seed variance are not included. One Bloom level corresponds to 0.200. A second implementation with matched (fold, client, row) differences changes every value by at most 0.0005 and no significance flag.

Table 35: Per institution values behind Table 2, on $\mathcal{T}^{\text{global}}$ (validation-selected model, mean over 3 folds). Bold = best per (institution, metric); cells shaded by rank within each institution group (darkest green = best); *Centralized* is excluded from both. Conventions and abbreviations follow Table 2. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019), BART (Lewis et al., 2020), Flan-T5 (Chung et al., 2024), LED (Beltagy et al., 2020).

Inst.	Method	Lightweight				LLM-based / reference-free						
		R-L	B-4	BERT	B_{cls}	B_{LLM}	AnsR	Faith	QAFE	RQU	RTC	LLM-J
$\mathcal{I}_0^{\text{pub}}$ (BART (Lewis et al., 2020), MIT)	indiv	0.4289	0.1607	0.8595	**0.2384**	0.3079	**0.5201**	**0.2174**	0.1828	0.4631	0.1972	0.6716
	TC-Best (Li & Wang, 2019)	**0.4381**	**0.1673**	**0.8620**	0.1514	–	–	0.2058§	0.1536	4.0569	0.1987	–
	TC-Ens (Li & Wang, 2019)	0.4313	0.1672	0.8606	0.1539	–	–	0.1865§	0.1698	4.0568	0.1954	–
	UniLearnLM	0.4305	0.1648	0.8604	0.2183	**0.3111**	0.5103	0.2156	**0.1890**	4.0678	**0.2005**	0.6801
	UniLearnLM+	0.4356	0.1654	0.8617	0.1712	0.2818	0.4786	0.2141	0.1812	**4.0685**	0.1998	**0.6939**
	Centralized	0.4749	0.2079	0.8751	0.2279	0.3005	0.4920	0.2073	0.1823	4.0782	0.2108	0.7297
$\mathcal{I}_1^{\text{pub}}$ (Flan-T5 (Chung et al., 2024), Stanford)	indiv	0.4423	0.1645	0.8623	0.2718	0.2397	**0.5600**	0.2029	**0.2010**	4.0955	0.2242	**0.7138**
	TC-Best	0.4483	0.1787	0.8665	0.1464	–	–	0.2043§	0.1960	4.0864	0.2216	–
	TC-Ens	0.4470	0.1794	0.8667	0.1584	–	–	0.2012§	0.1840	4.0841	0.2167	–
	UniLearnLM	0.4468	0.1732	0.8672	**0.2768**	0.2361	0.5589	**0.2033**	0.1862	**4.1136**	**0.2365**	0.7122
	UniLearnLM+	**0.4517**	**0.1806**	**0.8680**	0.1746	**0.2701**	0.4441	0.1942	0.1873	4.0889	0.2204	0.6807
	Centralized	0.4871	0.2064	0.8750	0.2166	0.2774	0.4624	0.1984	0.1921	4.0908	0.2277	0.7122
$\mathcal{I}_2^{\text{pub}}$ (LED (Beltagy et al., 2020), ML papers)	indiv	0.4237	0.1585	0.8590	0.1611	0.2787	0.4769	0.2071	0.1659	4.0637	0.1932	0.6893
	TC-Best	0.4298	0.1683	0.8609	0.1891	–	–	0.1796§	**0.1678**	4.0507	0.1934	–
	TC-Ens	0.4310	0.1686	0.8611	0.1841	–	–	0.1803§	0.1650	4.0545	0.1947	–
	UniLearnLM	0.4383	0.1765	0.8646	**0.1989**	**0.2880**	**0.4793**	0.2088	0.1670	**4.0663**	**0.2020**	0.7051
	UniLearnLM+	**0.4476**	**0.1822**	**0.8663**	0.1782	0.2761	0.4767	**0.2188**	0.1666	4.0657	0.2006	**0.7098**
	Centralized	0.4269	0.1627	0.8605	0.1921	0.2864	0.4782	0.2065	0.1576	4.0526	0.1947	0.6909

Table 36: Difference in B_{cls} (UniLearnLM minus comparator) on the public global test set, paired context-clustered bootstrap, 95% intervals. * = interval excludes zero; n.s. = it does not. The comparison with local training is given with its interval. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019), pFedHN (Shamsian et al., 2021), FedHeNN (Makhija et al., 2022), FedMKT (Fan et al., 2025).

UniLearnLM run	vs Indiv	vs pFedHN (Shamsian et al., 2021)	vs FedHeNN (Makhija et al., 2022)	vs FedMKT (Fan et al., 2025)	vs TC-Best (Li & Wang, 2019)	vs TC-Ens (Li & Wang, 2019)	vs UniLearnLM+
Tabled run (3 folds)	$+0.0076$ $[+0.0007, +0.0147]^*$	$+0.0500^*$	$+0.0252^*$	$+0.0071$ n.s.	$+0.0691^*$	$+0.0659^*$	$+0.0567^*$
FiLM-aware rerun (3 folds)	-0.0210 $[-0.0285, -0.0135]^*$	$+0.0214^*$	-0.0034 n.s.	-0.0215^*	$+0.0405^*$	$+0.0373^*$	$+0.0281^*$
Second seed (fold 1)	-0.0173 $[-0.0296, -0.0055]^*$	$+0.0043$ n.s.	-0.0092 n.s.	-0.0330^*	$+0.0447^*$	$+0.0444^*$	$+0.0287^*$

Inside the encoder: the graph edges, the LoRA statistic features and the encoder depth each matter. A parameter matched MLP that ignores the graph edges loses judge score, freezing the LoRA statistic node features costs depth while helping surface overlap, a single shared vector drops below local training on depth and answer relevancy, and the three layer GAT beats the two and four layer variants on 10 of 11 metrics; a control that replaces the graph encoder by a learned client index trails a matched rerun of unified learning on the metrics both report. These ablations run on one or two folds (Appendices U and Z).

This appendix isolates the contribution of each design choice in UniLearnLM. Section Z.1 separates federation from added local capacity; Section Z.2 replaces the graph encoder with a parameter-matched MLP and with a single learned shared vector, and removes the dynamic node features; Section Z.3 varies the depth of the GAT; and Section Z.4 explains why we report no paired per sample significance tests. Unless stated otherwise, every variant is trained and evaluated under exactly the protocol of Section 4 on the public global test set $\mathcal{T}^{\text{global}}$, and differs from UniLearnLM only in the component named.

Z.1 Separating Federation from Added Capacity: the no-FedAvg Ablation

UniLearnLM adds a GAT unification encoder and a FiLM modulator on top of the LoRA adapters used by indiv, so a gain over indiv could in principle come from the extra local capacity rather than from federation. The *no-FedAvg* ablation is designed to tell those apart: every client keeps the complete UniLearnLM architecture (LoRA + GAT + FiLM, identical trainable-parameter count and identical optimizer settings), but the server never averages the GAT, so each client trains its own encoder in isolation. The two arms differ in the averaging step alone, which is what makes the comparison worth running.

Table 37: Unified learning against the no-FedAvg ablation on the public global test set, three folds, same settings on both arms. Arm columns are three-fold means of three client means with the cross fold sample standard deviation. The last four columns are the per fold difference and its mean.

Measure	UniLearnLM	**no-FedAvg**	**fold 1**	**fold 2**	**fold 3**	**mean**
B_{cls}	0.2028 ± 0.0039	0.2010 ± 0.0042	$+0.0057$	$+0.0057$	-0.0060	$+0.0018$
R-L	0.4425 ± 0.0006	0.4435 ± 0.0025	-0.0032	-0.0021	$+0.0024$	-0.0009
BERT	0.8647 ± 0.0005	0.8654 ± 0.0005	-0.0010	-0.0014	$+0.0002$	-0.0007
Faith	0.1907 ± 0.0029	0.1875 ± 0.0077	$+0.0070$	-0.0044	$+0.0070$	$+0.0032$

Table 37 reports the comparison on three folds. The two arms do not separate on pedagogical depth. The per fold differences in B_{cls} are $+0.0057$, $+0.0057$ and -0.0060, a mean of $+0.0018$, so two folds favor the federation and the third goes against it by a similar amount. The spread across folds is 0.0116. Surface overlap gives mean differences of -0.0009 on ROUGE-L and -0.0007 on BERTScore, also with the sign changing between folds. Faithfulness differs by $+0.0070$, -0.0044 and $+0.0070$, mean $+0.0032$, which we report without a claim.

What this control can and cannot support matters for how the rest of the paper should be read. It does not show that the averaging step produces the depth result, and we do not claim it does. Nor does it show the step is useless, because three folds at one seed cannot rule out an effect of this size. It removes one specific alternative explanation only in the weak sense that added capacity without averaging does not visibly help either. The gains in Table 2 therefore rest on the comparisons in that table and on the private testbed, not on this ablation.

Z.2 What the Graph Encoder Contributes

The unification encoder is the only federated object, so it is worth asking how much of its behavior depends on (i) reading a *graph* at all, (ii) the *dynamic* node features that track the running statistics of the LoRA-adapted weights, and (iii) the fact that the encoder produces *per-block* rather than global modulation. We run three ablations against these three questions, all under the identical protocol.

MLP-enc (is the graph structure needed?). The GAT is replaced by a parameter-matched multi-layer perceptron that consumes the same node-feature matrix but ignores the edges, so it cannot propagate information between transformer blocks. Parameter count and node features are held fixed; only message passing is removed.

static (are the dynamic node features needed?). Node features 6–8, which carry the running statistics of each node's weights (for the LoRA factor nodes, the adapters) and are therefore the only part of the graph that changes during training, are held at their initialization. The architecture graph then encodes structure alone. For the BART and LED clients, whose modulation reads per-block embeddings that the adapter nodes do not reach, this ablation changes the modulation only through training dynamics, not through the encoder's input.

SharedVector (is a graph needed at all?). The entire encoder is replaced by a single learned vector, shared and FedAvg-ed across clients, that is broadcast to every transformer block. This keeps a federated object of comparable size but removes both the graph and any architecture-conditioning: the modulation no longer depends on which model family the client runs.

Tables 38 and 39 report the results. Three observations hold consistently. First, removing message passing (*MLP-enc*) costs the most on exactly the axes UniLearnLM is designed to protect: it drops 0.0278 on LLM-J and 0.0328 on AnsR relative to UniLearnLM, and lands below indiv on seven of eleven metrics, so a parameter-matched encoder that cannot see the architecture as a graph does not reproduce the effect. Second, freezing the dynamic features (*static*) is the mildest of the three ablations and even helps the reference-based metrics (R-L 0.4427 vs 0.4354), but it costs pedagogical depth (B_{cls} 0.2065 vs 0.2339) and the holistic judge (0.6885 vs 0.6996). Third, collapsing the encoder to a single shared vector (*SharedVector*) recovers part of the reference-based gain over indiv but loses on ten of eleven metrics against the full graph encoder, and in particular falls *below* indiv on B_{cls} (0.2017 vs 0.2175) and AnsR (0.4873 vs 0.5195). Modulation conditioned on the architecture, per block, is doing real work and is not replaceable by a global bias.

Table 38: **Encoder ablations** on the public global test set (mean ± std over folds 2 and 3; these two folds were re-run for all five variants under an identical protocol). *MLP-enc* = parameter-matched MLP encoder that ignores graph edges; *static* = dynamic LoRA-statistic node features (features 6–8) frozen at initialization. Bold = best per row. Column abbreviations follow Table 2.

Metric	indiv	**no-FedAvg**	**MLP-enc**	**static**	UniLearnLM
R-L	.4300 ± .0061	.4355 ± .0019	.4285 ± .0048	**.4427 ± .0009**	.4354 ± .0001
B-4	.1604 ± .0037	.1660 ± .0056	.1620 ± .0031	**.1738 ± .0033**	.1693 ± .0017
BERT	.8599 ± .0024	.8624 ± .0003	.8615 ± .0009	**.8645 ± .0004**	.8631 ± .0003
B_{cls}	.2269 ± .0140	.1801 ± .0171	.2264 ± .0015	.2065 ± .0162	**.2339 ± .0212**
B_{LLM}	.2801 ± .0077	.2778 ± .0069	**.2943 ± .0079**	.2873 ± .0075	.2812 ± .0097
AnsR	**.5187 ± .0056**	.4802 ± .0054	.4840 ± .0047	.4824 ± .0025	.5168 ± .0006
Faith	.2109 ± .0023	.1952 ± .0064	**.2132 ± .0037**	.2072 ± .0051	.2073 ± .0060
QAFE	**.1852 ± .0065**	.1720 ± .0108	.1706 ± .0034	.1840 ± .0009	.1767 ± .0004
RQU	4.0750 ± .0093	4.0704 ± .0018	4.0578 ± .0082	4.0685 ± .0021	**4.0830 ± .0088**
RTC	.2048 ± .0069	.2048 ± .0014	.2016 ± .0061	.2050 ± .0001	**.2131 ± .0028**
LLM-J	.6918 ± .0032	.6726 ± .0090	.6718 ± .0013	.6885 ± .0012	**.6996 ± .0006**

Table 39: **Shared-vector ablation** on the public global test set (fold 1). *SharedVector* replaces the architecture-graph encoder with a single learned vector that is FedAvg-ed across clients and broadcast to every transformer block, removing both the graph and the architecture conditioning while keeping a federated object. Bold = best per row.

Metric	indiv	**SharedVector**	UniLearnLM (full)
R-L	0.4349	0.4405	**0.4448**
B-4	0.1627	0.1714	**0.1760**
BERT	0.8609	0.8639	**0.8659**
B_{cls}	0.2175	0.2017	**0.2262**
B_{LLM}	0.2661	**0.2887**	0.2729
AnsR	**0.5195**	0.4873	0.5148
Faith	0.2055	0.2054	**0.2130**
QAFE	0.1792	0.1797	**0.1889**
RQU	4.0724	4.0743	**4.0817**
RTC	0.2051	0.2120	**0.2127**
LLM-J	0.6911	0.6934	**0.6981**

Z.3 Depth of the Unification Encoder

Table 40 varies the number of GAT layers with every other hyperparameter held fixed. The 3-layer encoder used throughout the paper is best on ten of eleven metrics. The behavior is not monotone in depth: two layers under-propagate (the largest drop is on B_{cls}, 0.1710 vs 0.2262, and on LLM-J, 0.6594 vs 0.6981), while four layers begin to over-smooth the per-block embeddings, so the modulation applied to different transformer blocks becomes less distinguishable and the reference-based metrics fall back to roughly the 2-layer level. The choice of three layers is therefore an interior optimum rather than a compute-constrained default.

Z.4 Paired Significance Tests

Several gaps in Table 2 are small relative to the cross fold standard deviation, so we also tested them at the level of individual generations rather than of fold means. For every (fold, client, sample) triple of the public global test pool ($n = 8243$ paired generations) one could take the per sample metric under two systems and test the paired difference with a two-sided Wilcoxon signed-rank test, reporting alongside it a paired-bootstrap 95% confidence interval (10^4 resamples, seed 42). Pairing was exact: the two systems were scored on the same context and the same fold.

Why those p values overstate the evidence, and what to read instead. Those tests should not be read as establishing the smaller effects, for three reasons.

Table 40: **GAT depth** on the public global test set (fold 1). All other hyperparameters are held at the values in Appendix I. Bold = best per row; GAT-3 is the configuration used everywhere else in the paper.

Metric	GAT-2	GAT-3 (ours)	GAT-4
R-L	0.4375	**0.4448**	0.4329
B-4	0.1638	**0.1760**	0.1665
BERT	0.8630	**0.8659**	0.8629
B_{cls}	0.1710	**0.2262**	0.2257
B_{LLM}	0.2648	0.2729	**0.2913**
AnsR	0.4905	**0.5148**	0.4881
Faith	0.2104	**0.2130**	0.1966
QAFE	0.1817	**0.1889**	0.1756
RQU	4.0632	**4.0817**	4.0615
RTC	0.2013	**0.2127**	0.2024
LLM-J	0.6594	**0.6981**	0.6800

First, the 8243 observations are not independent. They are 916 evaluation rows drawn from only 179 held-out contexts, scored across three clients and three folds, so each context contributes roughly 45 correlated observations and the three folds share a single fixed test set and 85% of their training pool. A Wilcoxon signed-rank test assumes independent pairs. The correct analysis is a mixed model with random intercepts for context, client, and fold, which we have not run.

Second, the fold level spread from the same runs contradicts the smaller results. Appendix V gives indiv B_{cls} as 0.2238 ± 0.0113 and UniLearnLM as 0.2313 ± 0.0156 over the same three folds. The effect being tested is $+0.0075$, which is around half of one fold standard deviation. Two analyses of the same runs cannot both be right about that quantity, and the difference between them is entirely the choice of denominator.

Third, we apply no correction for multiplicity across the eleven metrics reported, and we run no test at all against TC-Best.

Our own reading of these runs is therefore the conservative one. The gaps we regard as established are those of UniLearnLM against the two distillation baselines on B_{cls}, significant in the paired tests of Table 2 and in every bootstrap run of Table 36, and the LLM-judge gap over indiv (paired t-test, $p = 0.022$). The B_{cls} gap between UniLearnLM and indiv is not among them, and Section 5 says so.

We do not report paired per sample tests: the 8,243 observations come from 179 contexts and are not independent, and the per sample scores were not retained, so a mixed model cannot be fitted without re scoring; fold level spread is reported instead.

AA Robustness to Federation Composition: Leave-One-Out Analysis

A federation that only works at one particular client roster is of little practical use, since institutions join and leave. We therefore test how UniLearnLM behaves when the cohort changes, by removing one client and re-running the federation from scratch with the remaining two. Each row of Table 41 isolates a different question, so the comparison must be read horizontally rather than against Table 2:

- **LOO 2-client fed**: the federation is retrained with only the two surviving clients, and the two surviving clients are then evaluated. This is the "the third institution left" case.
- **Full fed (same 2)**: the same two clients, but evaluated from the full 3-client federation of Table 2. This is the "the third institution is still here" reference.
- **Indiv (same 2)**: the same two clients trained without any federation, which bounds how much of the score is federation at all.

Two conclusions follow. First, *federation degrades gracefully*: dropping a client never sends the survivors below their own indiv baseline on ROUGE-L, and the 2-client federation retains most of the 3-client gain (dropping BART: 0.4414 vs 0.4425 full-fed and 0.4330 indiv). The method does

not require a specific cohort to be better than training alone. Second, *more heterogeneous partners help more, and the largest client helps most*: removing the arXiv/LED client, which contributes by far the largest and most diverse curriculum (5,037 QA pairs against 547 and 559), causes the biggest loss for the survivors on pedagogical depth (B_{cls} 0.2053 vs 0.2476 with LED present) and on the holistic judge (0.6780 vs 0.6962). Removing Flan T5 leaves the survivors essentially unchanged (R-L $+0.0024$, B_{cls} -0.0013). This is the equity story stated as a sensitivity: the direction of benefit runs from the data rich institution toward the data poor ones, so the federation is most valuable exactly when a resource-rich partner is present, and its value is not an artifact of the particular three-client roster.

Table 41: **Leave-one-out federation composition study** (public testbed, mean ± std over 3 folds). One client is removed and the federation is retrained from scratch on the remaining two; the reported scores are averaged over the *two surviving* clients only, so each block is internally comparable but blocks are not comparable to each other. Bold = best within each block. Sources: BART (Lewis et al., 2020), Flan-T5 (Chung et al., 2024), LED (Beltagy et al., 2020).

Dropped client	Setting	R-L	B_{cls}	LLM-J
BART (Lewis et al., 2020) (MIT)	LOO 2-client fed	.4414 ± .0013	.2180 ± .0098	.6851 ± .0106
	Full fed (same 2)	**.4425 ± .0078**	**.2378 ± .0123**	**.7087 ± .0065**
	Indiv (same 2)	.4330 ± .0049	.2164 ± .0197	.7015 ± .0027
Flan-T5 (Chung et al., 2024) (Stanford)	LOO 2-client fed	**.4368 ± .0100**	.2073 ± .0128	.6851 ± .0184
	Full fed (same 2)	.4344 ± .0067	**.2086 ± .0221**	**.6926 ± .0070**
	Indiv (same 2)	.4263 ± .0079	.1997 ± .0017	.6804 ± .0058
LED (Beltagy et al., 2020) (arXiv ML)	LOO 2-client fed	.4360 ± .0063	.2053 ± .0182	.6780 ± .0043
	Full fed (same 2)	**.4386 ± .0039**	.2476 ± .0140	**.6962 ± .0040**
	Indiv (same 2)	.4356 ± .0039	**.2551 ± .0155**	.6927 ± .0033

AB A Modern Decoder-Only Client

All three public-testbed backbones are encoder-decoder models released between 2019 and 2022, which raises a fair question: does the architecture-graph representation generalize to a modern decoder-only SLM, or is it an artifact of the encoder-decoder family? To test this we replace the BART client at $\mathcal{I}_0^{\text{pub}}$ (MIT) with Qwen2.5-1.5B-Instruct, a decoder-only model an order of magnitude larger than the other clients and from a different structural family, and re-run the federation without changing anything else: the same architecture-graph construction, the same node features, the same GAT, the same FiLM modulation, and the same federated schedule. The Qwen block class is not a registered node type, so, as for the T5 family, the client receives the pooled graph embedding rather than per-block modulation (Section 3).

Table 42 reports the result as the mean over the cohort's three clients. The cohort with the federated Qwen client is competitive with the encoder-decoder configuration and leads it on the two LLM-judged axes (B_{LLM} 0.2838, LLM-J 0.7159), while trailing on reference-overlap metrics, which is the expected signature of a much larger decoder that paraphrases the source material instead of reproducing the reference wording; the Qwen client itself reaches ROUGE-L 0.361 on the global test set. The graph encoder therefore handles a decoder-only client without any change; this single fold does not show whether it helps that client more than it helps the encoder-decoder ones.

AC Few-Shot LLM Control

The *Hypothesis 2* comparison in Table 34 prompts every LLM zero-shot, which is the operating point an institution actually gets when it has no annotated curriculum to draw examples from. It is nonetheless a weak control for the *format* of the reference corpus: the fine-tuned clients have seen thousands of GPT-4o-written reference pairs and so match their style, whereas a zero-shot LLM has seen none. To separate style matching from grounding we re-run the strongest baseline, GPT-4o, with $k = 3$ in-context examples drawn from the training split of the same institution (never from the test chunk), leaving decoding and the rest of the template unchanged.

Table 42: **A decoder-only client** (public global test set, fold 1). "Qwen2.5-1.5B on MIT" replaces the BART client at $\mathcal{I}_0^{\text{pub}}$ with a modern decoder-only SLM, leaving the rest of the cohort and the entire method unchanged. Values are means over the three clients. Bold = best per row. Sources: BART (Lewis et al., 2020), Qwen2.5 (Qwen Team, 2024).

Metric	BART (Lewis et al., 2020) on MIT (paper)	Qwen2.5-1.5B (Qwen Team, 2024) on MIT
R-L	**0.4448**	0.4123
B-4	**0.1760**	0.1445
BERT	**0.8659**	0.8536
B_{cls}	**0.2262**	0.2170
B_{LLM}	0.2729	**0.2838**
AnsR	**0.5148**	0.5121
Faith	**0.2130**	0.1877
QAFE	**0.1889**	0.1478
RQU	**4.0817**	4.0555
RTC	**0.2127**	0.2047
LLM-J	0.6981	**0.7159**

Table 43 shows the expected split. Few-shot prompting buys GPT-4o a very large gain on reference overlap (ROUGE-L $0.4487 \rightarrow 0.5513$), confirming that a substantial part of the overlap gap in Table 34 is format imitation rather than content quality, and we do not claim ROUGE-L parity with LLMs. It does *not* close the grounding gap: even with in-context examples, GPT-4o remains below the UniLearnLM mixture on both reference-free grounding metrics (Faithfulness 0.1881 vs 0.2104; QAFactEval 0.1463 vs 0.1788), and its holistic self-family judge score drops slightly. GPT-4o's deficit on the grounding metrics is therefore not an artifact of weak prompting. Whether that deficit reflects grounding or answer length is a separate question, and Section 5.3 answers it: the few-shot control shortens GPT-4o's answers from 43.6 to 27.6 tokens and moves its QAFactEval from 0.1335 to 0.1463 in the same step, which is the direction the extractiveness account predicts. We therefore do not read this appendix as supporting a claim that the federation exceeds frontier LLMs on grounding beyond the length matched margin of Discovery 6. The overlap result is correctly attributed to reference style matching.

Table 43: **Few-shot LLM control** on the public global test set. GPT-4o is re-prompted with $k = 3$ in-context examples drawn from the training split of the same institution. The UniLearnLM row is the federated mixture of Table 34 (mean $\pm$ std over 3 folds); the two GPT-4o rows are single deterministic decodes. Bold = best per column. Sources: GPT-4o (Hurst et al., 2024).

System	R-L	Faith	QAFE	LLM-J
GPT-4o (Hurst et al., 2024), zero-shot	0.4487	0.1764	0.1335	**0.9523**
GPT-4o, few-shot ($k = 3$)	**0.5513**	0.1881	0.1463	0.9284
UniLearnLM (fed. mixture, 3 folds)	$0.4394 \pm .0071$	$\mathbf{0.2104 \pm .0089}$	$\mathbf{0.1788 \pm .0122}$	$0.6986 \pm .0133$

AD JUDGE ROBUSTNESS: A NON-OPENAI REPLICATION

The evaluation pipeline uses OpenAI models in two places, GPT-4o wrote the reference QA pairs of the public testbed, and GPT-4o-mini is the holistic judge, while two of the seven systems being scored are themselves from the OpenAI family. This is a genuine circularity, and we measure it rather than argue it away. We re-run the entire holistic judging protocol with Gemini-2.5-Flash as the judge: identical rubric, identical prompt (Appendix H), identical temperature-0 decoding, identical generations. Only the judge model changes.

Table 44 reports both judges side by side. Three things are visible. First, the *absolute* scale shifts sharply: Gemini-2.5-Flash is a much harsher judge of small models, scoring the UniLearnLM clients around 0.43 where GPT-4o-mini scored them around 0.70, so absolute LLM-judge values are not comparable across judges and should never be read as calibrated quality. Second, the *ordering* is nearly preserved: the rank correlation between the two judges across the nine systems is Spearman $\rho = 0.93$ (Pearson $r = 0.99$ on the raw scores), and the identity of the top and bottom groups is unchanged.

Third, the same-family inflation the paper flags for GPT-4o-mini reappears symmetrically for the new judge, with Gemini-2.5-Flash placing itself first among the LLMs it scores. The conclusion we draw is deliberately narrow: the LLM-judge column supports *relative* statements between systems, which is how it is used in Sections 5.2–5.3, and the SLM–LLM gap on this axis is a real ordering rather than an artifact of the judge's family. It does not support absolute claims, and the grounding comparison of *Hypothesis 2* rests on RAGAS Faithfulness and QAFactEval, neither of which uses the holistic judge. Those two metrics are not independent of the OpenAI family either, since the claim decomposer inside RAGAS Faithfulness is GPT-4o-mini (Appendix M); replication with a second judge does not remove that dependence.

Table 44: **Judge replication with a non-OpenAI judge.** Holistic LLM-judge score under the paper's GPT-4o-mini judge and under a Gemini-2.5-Flash judge, on identical generations with an identical rubric and prompt. †: self judged; ‡: same family as the judge. Bold = best per column. ± is the sample standard deviation over folds. The per client values come from the mixture evaluation pipeline of Table 34; the per-institution pipeline of Table 2 gives 0.6801, 0.7122 and 0.7051. Sources: BART (Lewis et al., 2020), Flan-T5 (Chung et al., 2024), Qwen2.5 (Qwen Team, 2024), LED (Beltagy et al., 2020), GPT-4o (Hurst et al., 2024), Gemini (Comanici et al., 2025), Llama (Grattafiori et al., 2024), Mistral (Jiang et al., 2023).

System	GPT-4o-mini judge (mixture pipeline) (Hurst et al., 2024)	Gemini-2.5-Flash judge (Comanici et al., 2025)
GPT-4o	0.9523‡	0.8882
GPT-4o-mini	**0.9542**†	0.8729
Llama-3.1-8B-Inst (Grattafiori et al., 2024)	0.8985	0.8198
Mistral-7B-Inst (Jiang et al., 2023)	0.9308	0.8242
Qwen-2.5-7B-Inst (Qwen Team, 2024)	0.9378	0.8519
Gemini-2.5-Flash	0.9395	**0.9033**†
UniLearnLM BART (Lewis et al., 2020) client (3 folds)	0.6840 ± 0.0174	0.4289 ± 0.0222
UniLearnLM Flan-T5 (Chung et al., 2024) client (3 folds)	0.7099 ± 0.0135	0.4687 ± 0.0167
UniLearnLM LED (Beltagy et al., 2020) client (3 folds)	0.7018 ± 0.0042	0.4324 ± 0.0030
Cross-judge agreement over the nine systems: Spearman $\rho = 0.93$, Pearson $r = 0.99$.		

AE Full-Participation Private-Testbed Federation

The live private-testbed runs reported in Appendix X are honest about their dropouts: $\mathcal{I}_2^{\text{priv}}$ (Nigeria), the most resource-constrained site, was never able to connect a federation client, and $\mathcal{I}_0^{\text{priv}}$ (Tunisia) missed two of the four runs. That is a real property of deploying across three continents on donated hardware, and we keep it in the paper. It is, however, a confound for the equity claim: the institution the framework is most supposed to help is missing from the live UniLearnLM column, and the per-method means in Tables 29 and 32 are consequently taken over different participant sets.

To remove that confound we re-run the private testbed as a *full-participation* federation: the same five private curricula, the same five backbones, the same protocol and hyperparameters, but with all five clients guaranteed present in every round and with the full 3-fold cross-validation used on the public testbed. This is a controlled re-execution on the same private data, orchestrated from a single coordination point, rather than a second live deployment across the five countries; it therefore answers "what does the federation do when nobody drops out" and not "what happens in the field", and the two should not be conflated.

Quality under full participation. Table 45 (Figure 8) reports all eleven metrics with cross fold standard deviations, which the live single-fold runs could not provide. The unified federation family takes 5 of the 11 metrics (BERTScore, B_{cls}, B_{LLM}, AnsR, and QAFactEval), TC-Best takes four (ROUGE-L, BLEU-4, RQUGE, and RTC), and TC-Ens takes the other two, Faithfulness and the holistic judge; B_{LLM}, AnsR, Faithfulness and the judge were not computed for TC-Best. UniLearnLM (.0932) is above local training (.0815) on BLEU-4 and below TC-Best, TC-Ens and UniLearnLM+. So this is not a sweep, and we do not present it as one. The part of the pattern that does match the public testbed is the depth axis: UniLearnLM reaches B_{cls} 0.2218 and B_{LLM} 0.3056, 0.059

above TC-Ens on the first and 0.038 above it on the second, and it also leads on answer relevancy (0.5793). TC-Best reaches B_{cls} 0.148 ± 0.013 over three folds, against 0.163 ± 0.008 for TC-Ens and 0.222 ± 0.014 for UniLearnLM. Notably, TC-Ens falls *below* the non-federated indiv baseline on B_{cls}, B_{LLM} and AnsR, reproducing on real institutional data the finding of Section 5.2 that distillation over a public pool trades pedagogical depth for fluency. That reproduction across two independent testbeds is the result this appendix is here to support.

Table 45: **Full-participation private-testbed federation** (cross distribution global private test set, mean $\pm$ std over 3 folds, all five institutions present in every round). Same five private curricula, backbones, protocol and hyperparameters as the live runs of Appendix X, but with no dropouts and with full 3-fold cross-validation. Bold = best per row. Column abbreviations follow Table 2. "–" = not computed. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019).

Metric	indiv	TC-Best (Li & Wang, 2019)	TC-Ens (Li & Wang, 2019)	UniLearnLM	UniLearnLM+
R-L	$.3298 \pm .0021$	$\mathbf{.3416 \pm .0021}$	$.3380 \pm .0015$	$.3329 \pm .0015$	$.3403 \pm .0015$
B-4	$.0815 \pm .0004$	$\mathbf{.0972 \pm .0009}$	$.0964 \pm .0015$	$.0932 \pm .0008$	$.0967 \pm .0008$
BERT	$.8313 \pm .0013$	$.8338 \pm .0010$	$.8335 \pm .0005$	$.8335 \pm .0009$	$\mathbf{.8344 \pm .0010}$
B_{cls}	$.1867 \pm .0138$	$.1477 \pm .0131$	$.1625 \pm .0075$	$\mathbf{.2218 \pm .0141}$	$.1611 \pm .0161$
B_{LLM}	$.2854 \pm .0074$	–	$.2676 \pm .0062$	$\mathbf{.3056 \pm .0050}$	$.2603 \pm .0093$
AnsR	$.5755 \pm .0070$	–	$.5308 \pm .0072$	$\mathbf{.5793 \pm .0023}$	$.5277 \pm .0079$
Faith	$.1743 \pm .0055$	–	$\mathbf{.1867 \pm .0130}$	$.1742 \pm .0056$	$.1845 \pm .0079$
QAFE	$.0562 \pm .0042$	$.0591 \pm .0040$	$.0586 \pm .0011$	$.0548 \pm .0006$	$\mathbf{.0631 \pm .0033}$
RQU	$4.0417 \pm .0040$	$\mathbf{4.0510 \pm .0020}$	$4.0505 \pm .0044$	$4.0445 \pm .0052$	$4.0308 \pm .0347$
RTC	$.1842 \pm .0025$	$\mathbf{.1876 \pm .0017}$	$.1863 \pm .0024$	$.1866 \pm .0014$	$.1862 \pm .0016$
LLM-J	$.7694 \pm .0042$	–	$\mathbf{.7801 \pm .0017}$	$.7748 \pm .0048$	$.7787 \pm .0058$

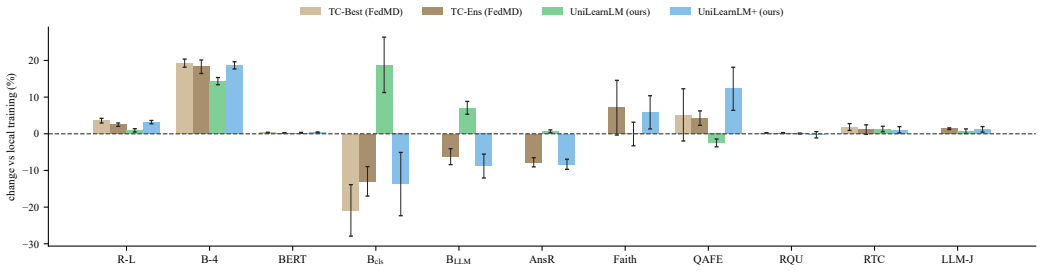


Figure 8: Change of each federated method relative to local training for all eleven metrics of Table 45 (full-participation federation, all five institutions, 3 folds), in percent of the local-training mean. Error bars = ± 1 standard deviation across folds, relative to that mean; TC-Best has no value on B_{LLM}, AnsR, Faith and LLM-J (not computed).

Knowledge transfer under full participation. The live KTI matrix (Table 32) is the weakest part of the equity evidence, both because its mean is negative for UniLearnLM and because, with $\mathcal{I}_2^{\text{priv}}$ absent, the per-method means are not taken over the same pair set. Under full-participation all 20 directed receiver–source pairs are defined for every method, so the comparison is finally like-for-like. Table 46 reports it. The two unified variants attain the highest mean transfer, UniLearnLM+ $+0.0098$ and UniLearnLM $+0.0097$, a difference inside rounding, and UniLearnLM has the largest number of pairs with *positive* transfer, 14 of 20 against 12 of 20 for both competitors. The strongest single direction is Germany $\leftarrow$ Nigeria ($+0.0813$). Counting from the table, Nigeria, the most resource-constrained site, is the *source* in two of the four largest positive cells for UniLearnLM ($+0.0813$ and $+0.0602$) and in three of the five largest (adding $+0.0233$). That is suggestive of the federation propagating topics only the under resourced institution teaches.

Two cautions belong with these numbers. Transfer is not uniformly positive: six of the twenty pairs are negative for UniLearnLM, and we report each of them. And the cross fold standard deviation on the UniLearnLM mean (± 0.0179) is larger than the mean itself ($+0.0097$; UniLearnLM+ reaches $+0.0098$, a difference inside rounding), so the KTI mean should be read as a direction rather than as a measured effect. What is more robust is the count of positive pairs, which does not depend on the

magnitudes. The sign of the average also reverses relative to the dropout-affected live run, which supports the reading that the negative live mean was driven by the missing client rather than by the method.

Table 46: **Knowledge Transfer Index under full participation** (private testbed, cross distribution global private test set, 3-fold mean). All 20 directed receiver ← source pairs are defined for every method because no client drops out, so the per-method means are taken over identical pair sets, unlike the live runs of Table 32. Positive = the federated method gains over indiv on topics present in the source's curriculum but absent from the receiver's. Bold = best per row. The ± on the mean row is the standard deviation across the three folds. Sources: TC-Best and TC-Ens (FedMD) (Li & Wang, 2019).

Receiver ← Source	TC-Ens (Li & Wang, 2019)	UniLearnLM	UniLearnLM+
Germany ← Tunisia	**+0.0042**	−0.0292	−0.0292
Nigeria ← Tunisia	−0.0521	**+0.0134**	−0.0283
Colombia ← Tunisia	−0.0208	**+0.0268**	+0.0208
Peru ← Tunisia	**+0.0486**	+0.0156	+0.0152
Tunisia ← Germany	**+0.0253**	−0.0096	+0.0058
Nigeria ← Germany	−0.0075	+0.0079	**+0.0204**
Colombia ← Germany	−0.0198	**−0.0128**	−0.0167
Peru ← Germany	+0.0126	**+0.0205**	−0.0096
Tunisia ← Nigeria	**+0.0392**	+0.0233	+0.0221
Germany ← Nigeria	+0.0146	**+0.0813**	+0.0729
Colombia ← Nigeria	+0.0255	+0.0602	**+0.0694**
Peru ← Nigeria	**+0.0260**	+0.0182	−0.0129
Tunisia ← Colombia	**+0.0290**	+0.0019	+0.0139
Germany ← Colombia	−0.0141	−0.0130	**+0.0079**
Nigeria ← Colombia	−0.0166	+0.0004	**+0.0080**
Peru ← Colombia	+0.0141	**+0.0305**	−0.0116
Tunisia ← Peru	**+0.0366**	+0.0017	+0.0289
Germany ← Peru	−0.0458	−0.0310	**−0.0123**
Nigeria ← Peru	+0.0185	+0.0200	**+0.0455**
Colombia ← Peru	**−0.0017**	−0.0312	−0.0139
Mean over all 20 pairs	+0.0058 ± .0273	+0.0097 ± .0179	**+0.0098 ± .0173**
Pairs with positive transfer	12/20	**14/20**	12/20

AF Extended Limitations

We group the limitations by theme; the Limitations paragraph of the main text summarizes them.

Seeds and statistical evidence. Seed variance is measured sparingly. Every headline run of the educational testbeds uses seed 42 (Tables 9–10); the one second seed run of unified learning, under the identical protocol of Appendix U, keeps the gap to the text consensus federations on B_{cls} but lands below local training. Three seeds exist only for the generic benchmark, its ablations and the partition controls. Elsewhere the variation we report is across three cross-validation folds that share a single fixed test set and 85% of their training pool, so the folds are not independent replicates of the measurement. The advantage of UniLearnLM over local training on B_{cls} in the tabled run is a case in point: its bootstrap interval excludes zero, yet the rerun and the second seed reverse its sign (Table 36), which is why Section 5 reports depth relative to local training as comparable rather than as a gain. A replication over at least three seeds at fixed fold is the single cheapest experiment that would settle which of our smaller effects are real, and we did not run it. The significance analysis in Appendix Z also overstates precision. Its paired tests treat 8,243 triples of (fold, client, sample) as independent, but those come from 179 distinct contexts scored across three clients and three folds, so each context contributes roughly 45 correlated observations. We report the tests alongside the fold level spread, but the correct analysis is a mixed model with random intercepts for context, client, and

fold, and we do not claim those p values as evidence. We apply no correction for multiplicity across the metrics compared, and that analysis includes no test against TC-Best.

Model selection. The results are sensitive to model selection, whose criterion is not shared across methods: local training and Centralized stop on greedy ROUGE-L over 50 validation samples, whereas the federated methods select per round. With the final-epoch model instead of the validation-selected one, most of the local-training lead on the four-task federation disappears (Table 18).

Heterogeneity across participants and test sets. The headline advantage is uneven across participants. As Table 33 shows, the depth change at $\mathcal{I}_0^{\text{pub}}$ is negative under the classifier proxy, although that institution improves on local training on most other metrics and keeps the depth that every distillation route removes. We report this rather than the mean alone: the equity claim rests on no participant being left worse off than under the alternatives, and the size of the gain depends on contributed data. The ordering on each institution's own in distribution test set is also different. On the local test splits (Table 28) the distillation baselines lead on most metrics and UniLearnLM is not first on B_{cls}; our reading, given in Section 5, is that unified federation trades in distribution surface fit for cross curriculum generalization, but that reading is post hoc and we did not design an experiment to test it. The global test pool is unbalanced as well: $\mathcal{I}_2^{\text{pub}}$ supplies 82.4% of the 916 global evaluation rows, so every Mean row in Table 2 is dominated by one institution's distribution. Finally, institution, curriculum, and backbone are perfectly confounded on the public testbed, with one backbone per institution and n=3. We cannot separate a curriculum effect from a backbone effect; Appendix AB swaps one backbone on one curriculum for one fold, which is not enough to break the confound.

The live deployment. The live deployment was incomplete. $\mathcal{I}_2^{\text{priv}}$ (Nigeria), the most resource-constrained site and the one the equity argument is about, was never able to connect a federation client to any live run, and $\mathcal{I}_0^{\text{priv}}$ missed two of four. The live federated rows in Tables 29–32 therefore cover four institutions for UniLearnLM and TC-Best and three for the other two methods, and their Mean rows average over different participant sets. The live private comparisons also set a 3-fold mean for local training against single fold federated runs, which is not a like for like comparison. The full-participation study of Appendix AE repairs this, but it is a controlled re-execution from one coordination point and not a second live deployment, and the two should not be conflated.

Metrics and evaluation. Three of the eleven metrics are modified re-implementations: RQUGE's trained span scorer is replaced by BERTScore, QAFactEval is a simplified UnifiedQA variant, and RTC uses token F1 in place of exact match (Appendix F documents each deviation). RQUGE in particular spans less than 0.08 of its 1–5 range across every system in Table 2, so it carries little discriminative weight despite counting as one metric of eleven, and we have not correlated our variants against the published implementations. The metric pipeline is not independent of the OpenAI family either. The holistic judge is GPT-4o-mini, and so is the claim decomposer inside RAGAS Faithfulness and the reverse question generator inside Answer Relevancy. Replication with Gemini-2.5-Flash (Appendix AD) shows the ordering is stable and that each judge favors its own family (Panickssery et al., 2024), but it does not remove the dependence from the grounding column. The depth proxy B_{cls} is a public classifier not validated on this domain, and the reported effects are a fraction of one Bloom level. No human evaluation with educators accompanies the automatic metrics, and the absolute educational value scores are low: Table 3 places the generated items at roughly 2.9 to 3.1 out of 5 on educational value, which is the rubric band for basic recall and surface level facts. The differences we report are movements inside a regime that our own judge rates as shallow, and instructor validation is needed before any claim about classroom value.

Baselines. Our strongest result is a negative result about a baseline, and a negative result is only as strong as the baseline's tuning. The public pool holds roughly 180 rows per fold against 313 to 2,855 private training rows per client, at distillation weight β=$0.5/K$, and TC-Best's "distil toward the single peer with the highest mean pairwise ROUGE-L" consensus rule is our own construction rather than FedMD's averaging. We ran no sweep over pool size or over β. If depth keeps falling as the pool grows, the finding is a property of the text space channel; if it recovers, the finding is a property of our pool. Six extra federated runs at one fold, over pool sizes of roughly 180, 540 and 1,800 rows, would settle it, and we did not run them. The baseline set is also bounded. The two text consensus

federations are our own text space adaptations of one method family (Li & Wang, 2019); the three published heterogeneous federations (pFedHN (Shamsian et al., 2021), FedHeNN (Makhija et al., 2022) and FedMKT (Fan et al., 2025)) run under our protocol and our tuning budget (Appendix U), not their authors' full search. Federated LoRA methods for a single shared backbone, such as FlexLoRA and HetLoRA, do not apply to clients of different families and are not compared.

What the architecture graph contributes. The client index and client vector controls, which swap the architecture graph for a learned client identifier in the same FiLM modulator, trail unified learning by small margins, so what the graph structure itself contributes remains open. Nor do we isolate the one design step we claim over the graph hypernetwork line: a hypernetwork that reads the same architecture graph and emits LoRA or adapter tensors for the frozen backbone, the closest published route (Litany et al., 2022) adapted to parameter efficient tuning, is not run, so whether emitting FiLM modulation instead of parameters matters is untested. The T5-family and Qwen clients also receive graph-level rather than per-block modulation (Section 3). Ablation coverage is uneven: the encoder ablations use folds 2 and 3, the shared vector and GAT depth studies use fold 1 only, and the decoder-only client study is a single fold at a single seed, so mechanism claims drawn from them are weaker than the headline table.

Language. Every curriculum in this study is in English, including all reference QA pairs, model inputs, judge prompts, and baseline outputs. The framework is language agnostic, since no tokens, vocabularies, or text representations cross institutions, so a multilingual cohort is structurally supported, but nothing here establishes it empirically. This is a live gap in our own cohort: two contributing sites list bilingual materials in Table 4 and we evaluate only in English.

AF.1 Flan T5 with block nodes

Registering the T5 block class as a node type gives every hooked block of the Flan-T5 client the embedding of its own node instead of the mean pooled graph embedding used by the released configuration (Section 3). Table 21 reports this variant as the "T5 block nodes" row, trained under the protocol and folds of UniLearnLM and scored by the same pipeline, so the effect of per block modulation on the one client that lacked it can be read directly against the UniLearnLM row.