FEDSOCRATIC: EXPLORING KNOWLEDGE TRANSFER THROUGH SOCRATIC QUESTIONING IN HETEROGENEOUS FEDERATED LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning enables collaborative improvement of large language models while keeping training data local. However, architectural differences hinder parameter aggregation, while output-based knowledge exchange often relies on shared public training corpora. We study how heterogeneous LLMs can identify and transfer complementary knowledge without a shared public training corpus or parameter alignment. Inspired by Socratic inquiry, we propose FEDSOCRATIC, which combines structured questioning, answer verification, and local adaptation. Base questions and answer-preserving and answer-changing follow-ups probe client capabilities; executable rules verify final answers, and receivers learn from accepted supervision while retaining their architectures. Experiments with four heterogeneous clients across four verifiable task domains demonstrate improved performance on unseen procedural states. Controlled studies show that complete question coverage improves pooled weak-domain accuracy over omitting question types at matched record and update budgets. With response text retained, capability-selected peers outperform verified self-training and random peers in mean cross-domain accuracy on matched accepted questions and record counts. These findings support knowledge transfer through structured questioning within verifiable task families and highlight question coverage and supervision-source selection as key design choices.

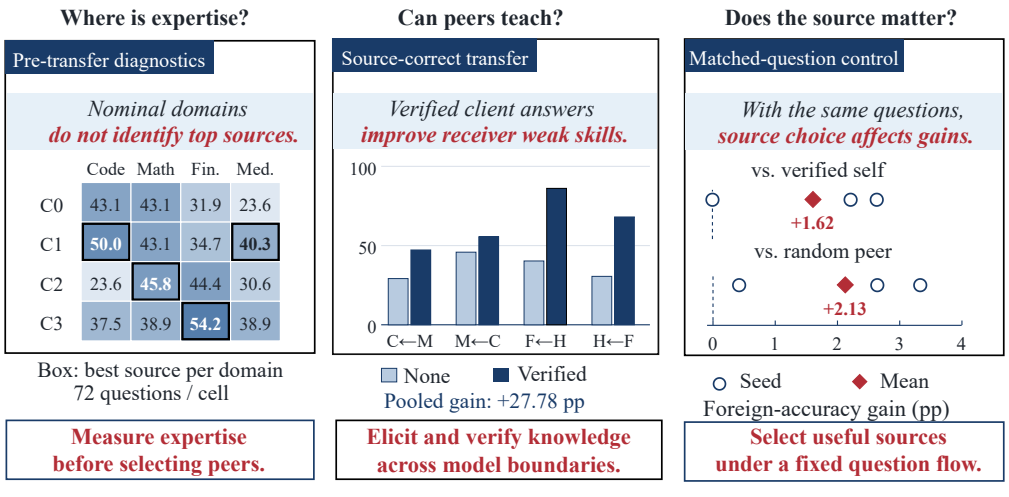


Figure 1: Three separate studies of knowledge exchange: capability diagnostics (left), transfer from verified client answers (middle), and capability-selected responses versus self and random peers on matched questions (right). Gains use within-study references.

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.

1 INTRODUCTION

"I cannot teach anybody anything. I can only make them think."
Attributed to Socrates

Federated learning improves language models collaboratively while keeping training data local (McMahan et al., 2017). OpenFedLLM supports decentralized instruction tuning and alignment (Ye et al., 2024). FedPepTAO uses efficient prompt tuning (Che et al., 2023), while FlexLoRA and FLoRA support heterogeneous adapter ranks (Bai et al., 2024; Wang et al., 2024). FedDQC improves instruction tuning through data-quality control (Du et al., 2025). Together, these advances improve learning quality, efficiency, and resource flexibility.

In practice, participants choose different model sizes, architectures, and tokenizers for their applications. OpenFedLLM requires aligned updates; FlexLoRA and FLoRA accommodate different ranks within compatible backbone parameter spaces. FedPepTAO's trainable prompts likewise remain model-dependent. Alternatively, FedMD, Fed-ET, and FedMKT exchange predictions on public or auxiliary inputs (Li & Wang, 2019; Cho et al., 2022; Fan et al., 2025). Suitable corpora may be unavailable under domain-coverage, licensing, or access constraints. FedGen and DENSE generate alternative supervision without public data (Zhu et al., 2021; Zhang et al., 2022). For heterogeneous LLMs, this raises a learning question: which inputs reveal complementary capabilities, and which responses help another client learn?

This motivates our central research question:

Research question

How can heterogeneous federated LLMs identify and transfer complementary knowledge without a shared public training corpus or parameter alignment?

Question coverage. Do answer-preserving and answer-changing follow-ups improve transfer at a fixed training budget?

Supervision selection. On matched questions, do capability-selected peer responses teach receivers more effectively than self-generated or random-peer responses?

Inspired by Socratic-style probing, we use follow-up questions to test model responses under answer-preserving and answer-changing conditions. Teaching tailored to demonstrated strengths and weaknesses further motivates capability-based exchange. Natural-language questions connect different LLMs, while structured variations diagnose capabilities beyond nominal specialization. Figure 1 traces capability diagnostics (left), verified-answer transfer (middle), and source selection (right).

We propose FEDSOCRATIC, connecting structured questioning, answer verification, and local adaptation (Figure 2). Known task-generation rules produce base, answer-preserving, and answer-changing questions to assess stability and sensitivity. Code execution or deterministic rules verify final answers; receivers learn from accepted records with local replay and checkpoint selection. Experiments across four heterogeneous clients cover code, mathematics, financial calculations, and medical rule execution. Verified client answers improve weak skills on unseen procedural states. The selected answer-only portfolio leads the controlled-task comparison against setting-adapted baselines in cross-domain and overall accuracy (Section 4.2).

Beyond feasibility, controlled comparisons examine which questioning and supervision choices support transfer. At matched record and update budgets, complete base, invariant, and flip coverage improves pooled weak-domain accuracy over omitting a phase and repeating retained questions (Section 5). Teacher choice depends on content: normalized correct answers can be identical across sources, whereas retained responses preserve source-specific text. On matched accepted questions and record counts, capability-selected responses yield higher mean cross-domain accuracy than verified self-training and random-peer supervision (Section 4.3). These controls distinguish coverage and source effects behind system-level gains, identifying useful decisions for knowledge exchange.

Our contributions are:

- We develop FEDSOCRATIC, connecting structured questions, verified supervision, and local updates without a shared public training corpus or parameter alignment. Experiments demonstrate transfer from verified client answers to unseen states in verifiable task families.
- We show that complete base, invariant, and flip coverage improves pooled weak-domain accuracy over phase deletions at matched record and update budgets. This comparison tests complementary question coverage against repetition, rather than additional training.
- We show that capability-selected responses yield higher mean cross-domain accuracy than verified self-training and random peers on matched accepted questions and record counts. This identifies useful response-source differences beyond supervision volume.

2 RELATED WORK

From shared parameters to heterogeneous models. Federated learning enables collaboration while keeping training data local, with federated averaging combining updates in a shared parameter space (McMahan et al., 2017). This setting has expanded to accommodate unequal client resources through nested subnetworks (Diao et al., 2021) and low-rank adapters of different sizes (Bai et al., 2024; Wang et al., 2024). For language models, prompt tuning and federated instruction tuning make adaptation more practical (Che et al., 2023; Ye et al., 2024), while data-quality control and representative sample selection improve the supervision used locally (Du et al., 2025; Qin et al., 2025). These developments accommodate client differences while retaining compatible components for aggregation. Here, model heterogeneity also includes differences in backbone architecture and tokenizer. Each client retains its model, requiring a learning signal that crosses parameter spaces.

Knowledge exchange without a shared public corpus. Distillation provides such a signal through model outputs (Hinton et al., 2015). Public or auxiliary inputs connect heterogeneous models in FedMD, FedDF, and Fed-ET (Li & Wang, 2019; Lin et al., 2020; Cho et al., 2022); FedMKT extends this exchange to large and small language models through token alignment (Fan et al., 2025). Data-free approaches construct alternative supervision. FedGen learns a generator that guides local training (Zhu et al., 2021), while DENSE and DFRD synthesize inputs for global distillation (Zhang et al., 2022; Luo et al., 2023). FedHPL instead exchanges class-conditioned logits for visual prompt learning without public data (Ma et al., 2024). We consider textual questions as the common interface. Server-known rules generate questions and check final answers, replacing a shared public training corpus with explicit procedural task knowledge.

From capability testing to useful supervision. Once questions provide the exchange interface, their design determines which capabilities clients reveal. CheckList shows how basic functionality, invariance, and directional expectations organize behavioral tests (Ribeiro et al., 2020). Generated responses can also become training material: Self-Instruct filters self-generated instruction data (Wang et al., 2023), STaR learns from rationales that yield correct answers (Zelikman et al., 2022), and Distilling Step-by-Step uses teacher rationales as additional supervision (Hsieh et al., 2023). This connects capability elicitation to learning, while distinguishing answer correctness from the information carried by a response. FEDSOCRATIC studies that connection across heterogeneous clients: structured question variations diagnose capabilities, capability-based selection identifies sources, and receivers learn from supervision whose final answers pass executable checks. Our focus is how question coverage and response selection shape transfer within verifiable task families.

3 QUESTION-BASED KNOWLEDGE EXCHANGE

FEDSOCRATIC connects structured questioning, supervision verification, and local adaptation (Figure 2). The answer-only configuration forms the FEDSOCRATIC *portfolio*; the response-retaining configuration evaluates source selection. Appendix B provides policies and pseudocode.

Table 1: Illustrative Socratic question chains. Blue marks answer-preserving changes; red marks answer-changing conditions. The answer column reports code-execution or fixed-cash-flow discounting results. Each follow-up is relative to its base question.

Phase	Question or follow-up	Verified answer
Code: when does a closure bind its variable?		
Base	Create `fs = [lambda: i for i in [1,2,3]]`. What does `[f() for f in fs]` return?	`[3,3,3]`
Invariant	Rename `i` to `j` in both the lambda body and the loop. Does the output change?	Unchanged: `[3,3,3]`
Flip	Replace `lambda: i` with `lambda i=i: i`, binding the value when each function is created. What is the output now?	Changed: `[1,2,3]`
Finance: how does yield affect a fixed cash flow?		
Base	Bond A pays 105 in one year. At an annual yield of 5%, what is its price, rounded to two decimals?	100.00
Invariant	Rename Bond A to Bond B, keeping the payment, maturity, and yield unchanged. Does its price change?	Unchanged: 100.00
Flip	Raise the yield from 5% to 10%, keeping the payment and maturity unchanged. What is its price now?	Changed: 95.45

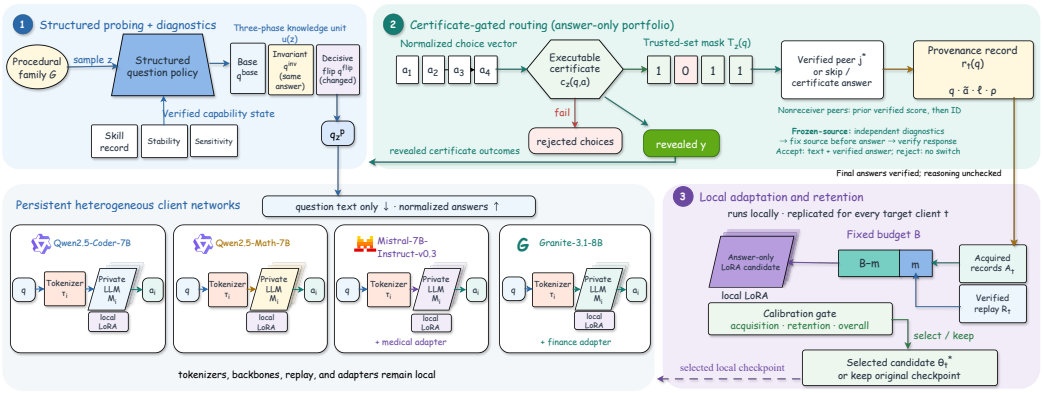


Figure 2: The FEDSOCRATIC answer-only portfolio: structured probing, certificate-gated peer selection, and local LoRA updates with verified replay. The inset distinguishes frozen-source response retention, where a source is fixed before its current answer is checked.

3.1 STRUCTURED QUESTIONS AND CAPABILITY DIAGNOSTICS

Related probes reveal whether a client remains stable under irrelevant changes and responds to decisive ones. We pair each base question with answer-preserving and answer-changing variations, using executable checks to characterize capabilities beyond nominal specialization.

Let $\mathcal{C} = \{1, \ldots, N\}$ index clients with models $M_i = (\alpha_i, \tau_i, \theta_i)$: architecture, tokenizer, and backbone-plus-adapter parameters. Without accessing client training examples or a shared public corpus, a coordinator samples a procedural state z and generates a knowledge unit

$$u(z) = \left(q_z^{\text{base}}, q_z^{\text{inv}}, q_z^{\text{flip}}, c_z\right), \tag{1}$$

where q_z^{base} is the base question, and q_z^{inv} and q_z^{flip} change answer-irrelevant and answer-determining fields, respectively. The executable certificate $c_z(q, a) \in \{0, 1\}$ checks final answers using code or deterministic rules. Clients receive questions without z or the checker (Table 1).

The question interface supports fixed and adaptive schedules. Adaptive acquisition prioritizes skills without verified teachers, close teacher rankings, and uncertain outcomes; target-specific chains adjust difficulty and skill after each certified outcome (Appendix B.2).

3.2 Verifying and retaining supervision

Prior capability guides source selection, while verification determines response acceptance. Separating these decisions allows source policies to be compared on the same question flow.

For a question q from state z, let $a_i(q)$ be client i's answer. The certificate defines the set of clients whose answers pass verification:

$$T_z(q) = \{i \in \mathcal{C} : c_z(q, a_i(q)) = 1\}. \tag{2}$$

The answer-only pipeline selects a passing non-receiver peer by prior verified domain performance, with deterministic tie breaking. An empty set triggers skipping or certificate-answer completion, depending on the condition. Records retain the question, verified answer, option label, and source identity (Eqs. 7–8). The frozen-source study instead ranks peers on separate diagnostic questions by skill accuracy, domain accuracy, and client identifier, then accepts or rejects the selected response without replacement (Appendix G).

Answer-only records share a final-answer format and option labels, so identical answers yield identical supervision. *Response-retaining* records keep source explanations and append verified answers. Verification applies to final answers, not intermediate reasoning. In the frozen-source study, peers provide response text and the server provides certificate answers, varying both source and content.

3.3 Local adaptation and retention

Local adapters avoid parameter alignment, while replay helps retain existing capabilities. Each receiver t trains a LoRA adapter (Hu et al., 2022) using LLaMA-Factory (Zheng et al., 2024), mixing acquired records $\mathcal{A}_t$ with replay records $\mathcal{R}_t$ it previously answered correctly. For question $q(r)$ and training output $s(r)$, the objective is

$$\mathcal{L}_t(\theta) = -\sum_{r \in \mathcal{A}_t \cup \mathcal{R}_t} \log p_\theta\big(s(r) \mid q(r)\big). \tag{3}$$

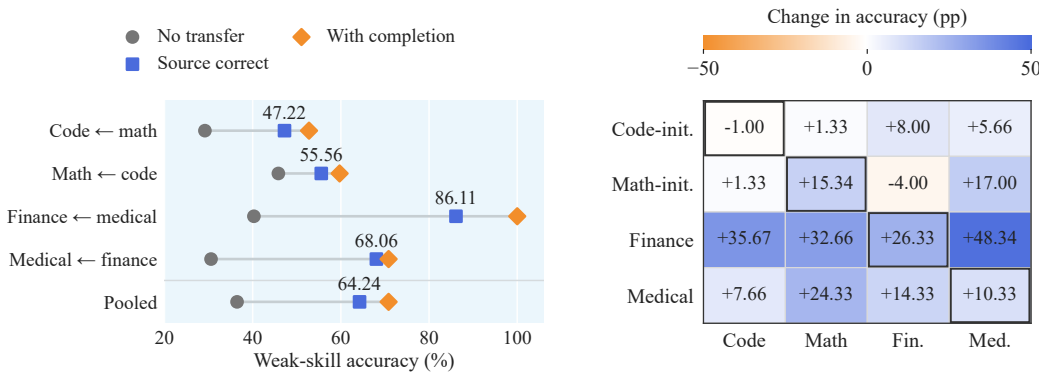
The loss covers output tokens: certified answers alone or answers with source text. Adapters and replay remain local. Separate calibration questions select acquisition–replay mixtures that improve weaker domains, retain stronger domains, and yield nonnegative overall change; otherwise, the pre-update checkpoint is retained (Eqs. 9–10). The common-selection study shares candidate ratios across acquisition policies (Appendix H); the frozen-source study fixes replay and optimizer steps.

4 Knowledge-transfer experiments

We describe the clients, tasks, and protocols (Section 4.1), then test transfer from verified client answers and compare complete system configurations (Section 4.2). We then examine capability-based response selection under a common candidate budget and on matched accepted questions (Section 4.3). Section 5 analyzes question coverage and supervision use.

Key Findings

- **Can heterogeneous peers teach each other?** Verified client answers improve weak-skill accuracy on unseen procedural states in all four tested directions (Section 4.2).
- **Does question coverage matter?** Complete base, invariant, and flip coverage improves pooled weak-domain accuracy over phase deletions at matched record and update budgets (Section 5).
- **Does the source matter?** With response text retained, capability-selected peers achieve higher mean cross-domain accuracy than verified self-training and random peers on matched accepted questions and record counts (Section 4.3).



(a) Source-correct transfer

(b) Portfolio gains

Figure 3: Transfer in two controlled studies. (a) Verified source answers improve weak-skill accuracy on new states. (b) Portfolio gains over no transfer on separate 300-question client–domain cells; rows are receivers, columns are domains, and outlines mark native cells.

4.1 Experimental setup

Clients and tasks. We evaluate four heterogeneous clients initialized from Qwen2.5-Coder-7B (Hui et al., 2024), Qwen2.5-Math-7B (Yang et al., 2024), Granite-3.1-8B (IBM Granite Team, 2024) with a finance adapter, and Mistral-7B-Instruct-v0.3 (Mistral AI, 2024) with a medical adapter. Their initialization domains identify provenance; observed capabilities determine source selection. The controlled evaluation spans 15 skills in code, mathematics, finance, and deterministic medical rule application. Each procedural state yields base, invariant, and decisive-flip questions. Appendix A describes the task families, the sources of the medical rules, and the generation checks. Figure 7 reports an initial capability catalog.

Protocols, metrics, and implementation. Table 4 lists each study's evaluation protocol and no-transfer reference. Foreign accuracy averages non-native client–domain pairs, Native averages nominal initialization-domain pairs, and Overall averages all pairs. Repeated training is summarized across seeds; paired latent-state intervals describe test-question uncertainty. Public benchmarks are reserved for evaluation. Appendix B specifies LoRA settings, replay, and checkpoint selection; Appendices G and H give study-specific splits and seeds. Supplementary calibration-based updates use a separate test protocol (Appendix I).

4.2 Transfer feasibility and system comparisons

Verified client answers teach heterogeneous receivers. We begin with supervision correctly produced by a source client, excluding certificate-completed records and evaluating receivers on separately generated states with zero overlap. Figure 3(a) shows improved weak-skill accuracy in all four tested directions, demonstrating transfer to new states within these task families. The pooled source-correct gain is 27.78 percentage points (paired latent-unit 95% interval: $[22.22, 33.33]$). Adding certificate completion further improves the pooled result, and equal-count resampling retains most of that benefit. This motivates examining which questions enter the training pool as well as which source supplies their supervision.

The selected portfolio leads the controlled-task system comparison. We compare the answer-only checkpoint portfolio with random transfer, majority voting, confidence selection, non-adaptive certificates, FedMD (Li & Wang, 2019), Fed-ET (Cho et al., 2022), and FedMKT (Fan et al., 2025). For the literature baselines, we adapt the shared-input procedures to generated questions and the semantic-answer interface. All configurations are evaluated on the fixed test set; trained conditions use 100 records and 1,000 final-update steps per client, with acquisition and candidate-search costs outside this budget. The FedSocratic portfolio has the highest foreign, native, and overall accuracy among these configurations (Table 2). Figure 3(b) shows gains in 14 of 16 client–domain

Table 2: System-level accuracy (%) on the controlled test set, with 300 questions per client–domain cell. Literature baselines use the setting adaptations described above. Each trained condition uses 1,000 final-update steps per client, excluding acquisition and selection costs. Best values are bold; domain-level results are in Table 5.

Method	Foreign	Native	Overall
No transfer	39.11	43.08	40.10
Random transfer	39.75	37.58	39.21
Majority vote	36.44	38.08	36.85
Confidence selection	39.06	40.08	39.31
Non-adaptive certificate	52.94	42.92	50.44
FedMD (Li & Wang, 2019)	40.83	41.50	41.00
Fed-ET (Cho et al., 2022)	38.94	43.33	40.04
FedMKT (Fan et al., 2025)	44.97	38.58	43.38
FEDSOCRATIC portfolio	**55.14**	**55.83**	**55.31**

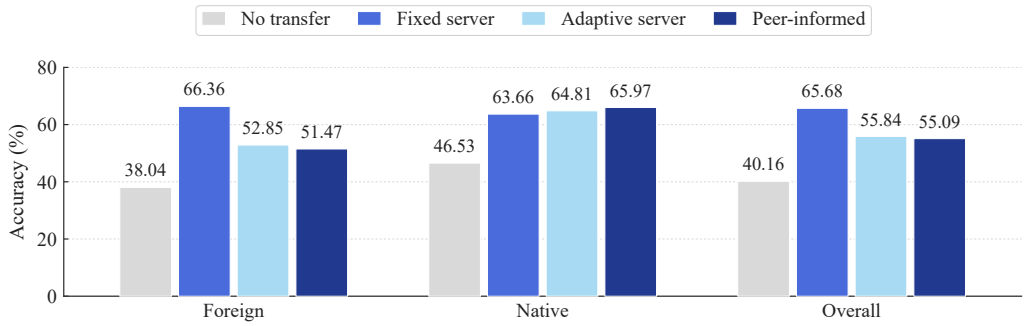


Figure 4: Acquisition under common answer-only update selection: mean accuracy (%) across three seeds. Candidate ratios and calibration gates are shared; pools remain policy-dependent. Evaluation is separate from Table 2; seed-level results are in Table 12.

cells, largest in the finance-adapted receiver; full reported accuracies appear in Figure 9. The code receiver's native score and the math receiver's finance score decrease, showing the variation behind the aggregate gain. Non-adaptive certificates remain strongest on code (Table 5).

Fixed server acquisition is strongest under common answer-only selection. We compare fixed server-certified acquisition, adaptive server acquisition without peer filtering, and peer-informed acquisition under a common answer-only selection rule. Each candidate uses 100 training records and 1,000 optimizer steps. The pipelines share four acquisition–replay ratios and calibration gates, with policy-dependent question pools. Candidate mixtures and evaluation states differ from Table 2; acquisition, calibration, and final states are disjoint for each seed (Appendix H). All trained pipelines improve over no transfer in Figure 4. Fixed server-certified acquisition achieves the highest foreign and overall accuracy, while peer-informed acquisition achieves the highest mean native accuracy.

4.3 CAPABILITY-BASED SELECTION OF RESPONSE SUPERVISION

We fix candidate flow to compare verified self, random peer, capability-selected peer, and direct server-certified supervision. An independent diagnostic set freezes source rankings before acquisition. The candidate-budget regime retains each source's accepted records; matched questions use their common accepted intersection, matching skills, difficulty, and record count. Both fix verified native replay and optimizer steps. Model-produced response text is retained, while the server supplies direct answers. Table 3 reports both regimes; Appendix G details selection and training.

Table 3: Frozen-source comparison: mean accuracy (%) over three training seeds. The common accepted intersection contains 8, 7, 9, and 12 unique questions for clients 0–3. Bold marks each regime's best trained value; no transfer is separate.

Regime	Supervision source	Foreign	Native	Overall
Reference	No transfer	38.06	42.08	39.06
Candidate budget	Verified self	35.97	36.25	36.04
	Random peer	35.19	**36.53**	35.52
	Capability-selected peer	**39.44**	35.69	**38.51**
	Server certified	35.09	35.14	35.10
Matched questions	Verified self	37.13	41.67	38.26
	Random peer	36.62	39.86	37.43
	Capability-selected peer	**38.75**	**43.06**	**39.83**
	Server certified	34.40	42.92	36.53

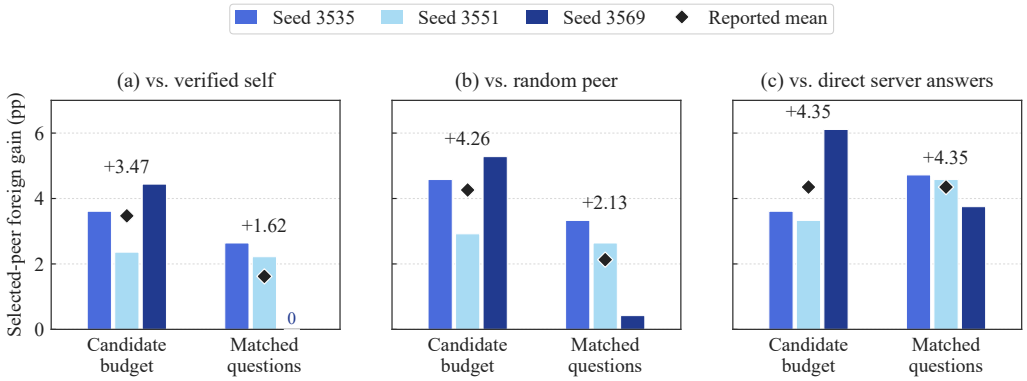


Figure 5: Foreign-accuracy gains of selected-peer responses over three supervision sources. Bars show training seeds sharing one response pool; diamonds and labels give reported means. Both candidate-budget and matched-question regimes are shown (Table 10).

Capability-based selection yields more verified records than random peers. Capability-based selection accepts more unique records than random-peer selection for every receiver, and more than self-generation for most receivers (Figure 13). Under the candidate budget, it exceeds the other trained sources in foreign accuracy in each seed. Server acceptance covers the entire flow but yields lower accuracy, separating supervision volume from learning outcomes.

Capability-selected responses outperform self and random peers on matched questions. On the common accepted questions, capability-selected responses achieve 38.75% foreign accuracy, exceeding verified self-training by 1.62 percentage points and random-peer supervision by 2.13 points. These comparisons match questions and record counts while allowing source response text to differ. Across the three training seeds, the foreign-accuracy contrast is positive against random peers and nonnegative against verified self (Figure 5). Overall accuracy reaches 39.83%. Relative to direct server answers, the selected responses are higher by 3.30 points overall and 4.35 points in foreign accuracy; this comparison also replaces direct answers with retained response text.

Cross-domain gains can accompany losses in native capabilities. The candidate-budget selected-peer condition improves foreign accuracy over no transfer while reducing native accuracy, leaving overall performance slightly lower. With matched questions, its net gains over no transfer are 0.69 percentage points in foreign accuracy and 0.77 points overall, smaller than its advantages over the trained source baselines. Native accuracy also increases on average, while client-level changes remain heterogeneous (Table 11).

5 QUESTION COVERAGE AND USE OF SUPERVISION

Complete phase coverage improves transfer at a fixed training budget. We compare complete base, invariant, and flip coverage with omitting one phase and repeating a retained question. Each condition uses 24 latent training states, 72 records, and 720 optimizer steps per target, with matched skill composition and source-correct counts. Figure 6 evaluates transfer on 24 new states per target. Complete coverage yields higher pooled weak-domain accuracy than every fixed deletion in each seed, with gains concentrated in the finance- and medical-adapted receivers. Within these task families, covering all three phases makes more effective use of the fixed budget than repeating questions from two phases. The comparison measures the complete coverage design, including the diversity supplied by the additional phase; Appendix H reports per-client and per-phase results.

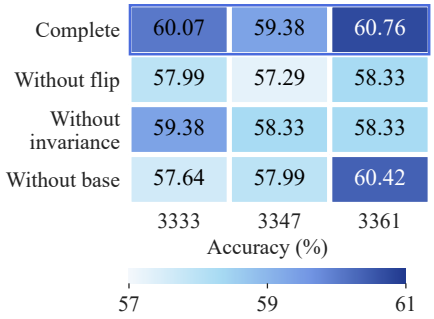


Figure 6: Pooled weak-domain accuracy (%) across seeds with matched training budgets. Color scale: 57–61%.

Response content gives source selection a role in transfer. Normalized correct answers can be token-identical across peer and server sources (Appendix F), while retained responses preserve differences in how answers are presented. In the frozen-source study, capability-selected responses improve mean foreign accuracy over self-generated and random-peer responses on the same accepted questions. Teacher choice matters in this response-retaining protocol, whereas provenance alone adds no information to identical answer-only records. At a fixed acquisition–replay mixture, generic replay outperforms personalized replay on average, while adaptive probing does not consistently outperform a fixed schedule (Appendix C.4).

Calibrated updates support retention with modest transfer gains. In supplementary experiments, two configurations using peer supervision improve internal overall and foreign accuracy over their respective no-transfer references across three training seeds each. One public evaluation seed per configuration gives overall gains of 0.59 and 0.25 percentage points, with near-preserved native performance. These follow-up experiments use a separate internal test with 180 questions per client–domain cell; protocols, seed-level scores, and public-task results appear in Appendix I.

6 DISCUSSION AND CONCLUSION

FEDSOCRATIC connects heterogeneous language models through structured questions, verified supervision, and local updates, without parameter alignment or a shared public training corpus. Verified client answers improve weak-skill accuracy on unseen states in all four tested directions. Complete three-phase coverage improves pooled transfer over fixed phase deletions at the same training budget, and the answer-only portfolio leads the controlled-task system comparison. Together, these findings establish question-based exchange within verifiable task families.

Teacher value depends on supervision content and local updates. Answer-only records erase source differences when certified answers coincide; fixed server-certified acquisition achieves the highest foreign and overall accuracy under common update selection. Retaining responses gives capability-based selection more accepted records than random peers and higher mean foreign accuracy than self-training or random-peer supervision, including on matched questions. These advantages exceed the net gain over no transfer, and the candidate-budget condition loses native accuracy.

The task families cover code execution, calculations, and deterministic rule application using server-known states; medical tasks assess rule execution rather than clinical judgment. On public-task subsets, the answer-only portfolio changes overall accuracy from 49.22% to 44.56%, while supplementary calibration-based updates yield small positive average changes on separately evaluated subsets (Appendices E and I). Structured coverage and response selection guide knowledge exchange, with update calibration helping receivers retain existing capabilities.

AI USE STATEMENT

A language-model coding assistant was used to help plan experimental controls, organize and interpret experiment records, revise prose, and restructure and check the LaTeX source. Reported measurements come from recorded experiment outputs. The authors remain responsible for the claims, citations, and final manuscript.

ETHICS STATEMENT

This study evaluates model-to-model knowledge transfer using procedurally generated tasks and public evaluation benchmarks. Medical tasks assess deterministic rule execution rather than clinical judgment, and the medical and financial results do not establish suitability for real-world decision-making. Client parameters and private training examples remain local, but exchanged responses carry no formal privacy guarantee. Applications involving sensitive information therefore require additional privacy and security safeguards. Executable checks establish final-answer correctness within the defined task rules; they do not assess the safety or fairness of retained response text. Reuse of pretrained models and evaluation datasets should follow their respective licenses and access conditions.

REPRODUCIBILITY STATEMENT

Our code is available at `https://anonymous.4open.science/r/FEDSOCRATIC-D93F/`. Appendix A specifies the procedural task families and generation checks. Appendix B provides the protocol map, acquisition and checkpoint-selection rules, local training settings, and Algorithm 1. Study-specific controls, evaluation sets, training seeds, and results are documented in Appendices C–I. Public-task metrics and evaluation procedures appear in Appendix E. These descriptions distinguish the system portfolio, common-selection acquisition study, and frozen-source study, allowing each comparison to be reproduced under its own protocol.

REFERENCES

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. arXiv preprint arXiv:2108.07732, 2021.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 14457–14483. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-0461.

Tianshi Che, Ji Liu, Yang Zhou, Jiaxiang Ren, Jiwen Zhou, Victor Sheng, Huaiyu Dai, and Dejing Dou. Federated learning of large language models with parameter-efficient prompt tuning and adaptive optimization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7871–7888. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.488.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R. Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.300.

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 Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI-2022, pp. 2881–2887. International Joint Conferences on Artificial Intelligence Organization, July 2022. doi: 10.24963/ijcai.2022/399.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

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

Yaxin Du, Rui Ye, Fengting Yuchi, Wanru Zhao, Jingjing Qu, Yanfeng Wang, and Siheng Chen. FedDQC: Data quality control in federated instruction-tuning of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15267–15291. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.791.

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 Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 243–255, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-Coder technical report. arXiv preprint arXiv:2409.12186, 2024.

IBM Granite Team. Granite 3.1 language models. Official model release, 2024.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259.

Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO clinical practice guideline for acute kidney injury. *Kidney International Supplements*, 2:1–138, 2012. URL https://kdigo.org/wp-content/uploads/2016/10/KDIGO-2012-AKI-Guideline-English.pdf.

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

W. S. Lim, M. M. van der Eerden, R. Laing, W. G. Boersma, N. Karalus, G. I. Town, S. A. Lewis, and J. T. Macfarlane. Defining community acquired pneumonia severity on presentation to hospital: An international derivation and validation study. *Thorax*, 58(5):377–382, 2003. doi: 10.1136/thorax.58.5.377.

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*, volume 33, pp. 2351–2363. Curran Associates, Inc., 2020.

Gregory Y. H. Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A. Lane, and Harry J. G. M. Crijns. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: The Euro Heart Survey on atrial fibrillation. *Chest*, 137 (2):263–272, 2010. doi: 10.1378/chest.09-1584.

Kangyang Luo, Shuai Wang, Yexuan Fu, Xiang Li, Yunshi Lan, and Ming Gao. DFRD: Data-free robustness distillation for heterogeneous federated learning. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 17854–17866. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-0784.

Yuting Ma, Lechao Cheng, Yaxiong Wang, Zhun Zhong, Xiaohua Xu, and Meng Wang. FedHPL: Efficient heterogeneous federated learning with prompt tuning and logit distillation. arXiv preprint arXiv:2405.17267, 2024.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017.

Mistral AI. Mistral-7B-Instruct-v0.3. Model card, 2024.

Zhen Qin, Zhaomin Wu, Bingsheng He, and Shuiguang Deng. Federated data-efficient instruction tuning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15550–15568. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.803.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442.

Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus. Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):762–774, 2016. doi: 10.1001/jama.2016.0288.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.754.

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 37*, NeurIPS 2024, pp. 22513–22533. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-0708.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122, 2024.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. OpenFedLLM: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 6137–6147. ACM, Aug 2024. doi: 10.1145/3637528.3671582.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. doi: 10.52202/068431-1126.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1472.

Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. DENSE: Data-free one-shot federated learning. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, pp. 21414–21428. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi: 10.52202/068431-1556.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 400–410, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-demos.38.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12878–12889. PMLR, 2021.

A TASK TAXONOMY AND GENERATION CHECKS

The controlled task family contains four code skills (Python closure binding, LRU state transitions, mutable-default traces, and SQL join row counts), four mathematics skills (linear recurrences, probability without replacement, modular exponentiation, and inclusion–exclusion), and three finance skills (bond yield–price direction, inventory cost flow, and covered-call profit). It also contains four medical rule-application skills: CURB-65 (Lim et al., 2003), qSOFA (Seymour et al., 2016), CHA_2DS_2-VASc (Lip et al., 2010), and KDIGO creatinine stages (Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group, 2012). The medical tasks assess deterministic rule execution rather than clinical judgment.

For every generated unit, the invariant probe changes only a nuisance field, the flip probe changes exactly one decisive field, all answer options are semantically unique, and the executable certificate agrees with the rendered correct option before any client call.

B IMPLEMENTATION AND INTEGRITY DETAILS

Table 4 links each study to its comparison and evaluation protocol. The system portfolio, common-selection acquisition policies, and frozen-source conditions are distinct question-based exchange configurations. Supplementary calibration-based updates are listed separately.

Table 4: Protocol map. Comparisons use within-study references. Question pools, supervision text, and end-to-end cost vary even when record or update budgets match.

Study	Comparison and controls	Evaluation and supported question
Source-correct transfer	Client-produced correct answers; certificate completion excluded from the source-only condition	New procedural states; can verified client answers improve a receiver's weak skill?
System portfolio	Exchange strategies with 100 records and 1,000 final-update steps per client; acquisition and search costs differ	300 questions per client–domain cell; which system configuration performs best here?
Common update selection	Acquisition policies with shared candidate ratios and calibration gates; 100 records and 1,000 steps per candidate	36 questions per client–domain cell, three seeds; how do policies behave under this selection rule?
Frozen sources	Same candidate flow, or common accepted questions; retained model responses, fixed replay and steps	New 240-question test, three seeds; does capability-based source selection help receivers?
Fixed phase deletions	24 training states, 72 records, and 720 steps per target; deletions repeat retained questions	24 new states per target, three seeds; does complete coverage use the fixed budget better?
Supplementary updates	Existing peer supervision with a short update selected per client; expanded-pool variant also selects data recipe and replay	180 questions per client–domain cell, three training seeds; public tasks use 200 questions per client–task cell (Appendix I)
Public update controls	Frozen 7–8B portfolio; separate 14B-composition rationale studies	Public benchmark subsets; how do content and update strength affect transfer?

The system-comparison and source-completion runs store frozen YAML configurations, SHA-256 code and configuration hashes, data manifests, raw JSONL responses, optimizer logs, adapter metadata, and machine-readable analyses. The shared baseline, final-update-matched portfolio, and source-completion ablation contain 48,000/48,000, 4,800/4,800, and 3,456/3,456 valid responses, respectively. The audited adapters match their expected optimizer-step counts, and the controlled-task bootstrap clusters probes by latent unit.

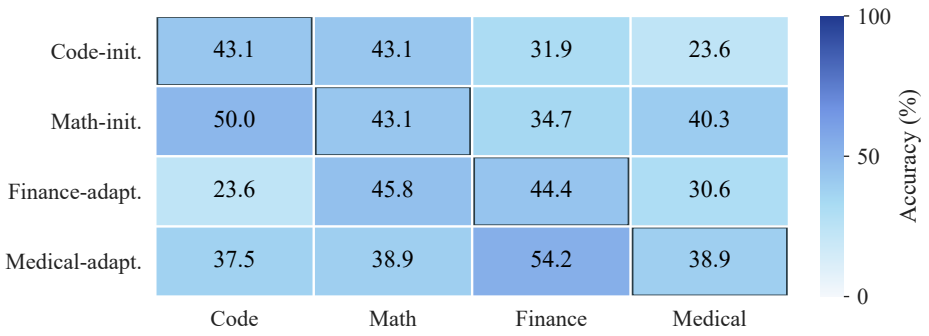


Figure 7: Pre-transfer diagnostic accuracy (%). Each cell contains 72 questions; outlined cells mark nominal initialization or adaptation domains. Transfer evaluations use separate questions.

The system-level comparison matches the final updates at 100 records and 1,000 steps per client, reusing the fixed system-comparison held-out probes. Its frozen records check the held-out hash and final step counts. Acquisition, calibration, and replay-candidate search are outside this matching budget. Paired intervals quantify test-question uncertainty conditional on the selected checkpoints.

The matched finance and medical updates use LoRA rank 8, alpha 16, dropout 0.05, `q_proj` and `v_proj`, cutoff length 1,024, an answer-only loss, batch size 1, and gradient accumulation 1. The recorded recipe uses ten epochs at 2×10^{-5}, cosine scheduling, warmup ratio 0.03, BF16, and seed 1212, starting from each domain adapter. The system comparison reuses trained client checkpoints where the final-update budget already matches.

The pre-transfer catalog in Figure 7 uses 72 questions per cell. System-comparison gains use paired evaluations on a separate 300-question-per-cell set (Table 5).

Algorithm 1 Certificate-gated Socratic acquisition and protected local update

Require: Clients $\{M_i\}$, generator $\mathcal{G}$, certificates $\{c_z\}$, interval K, budget B
Initialize chain states, Beta records, domain track records, and provenance logs
for each closed-loop acquisition block **do**
  Freeze all client checkpoints for the round
  for each target t and each of its K units **do**
    Choose a foreign-domain skill and difficulty using Eqs. 5–6
    Sample a novel state z, issue probe q, and collect normalized choices $\{a_i(q)\}$
    Verify choices, form $T_z(q)$, and select $j^*(q,t)$ if a peer passes
    Append $r_t(q)$ with peer or certificate-completion provenance
    Update only revealed track records and the target chain state
  end for
  Update each target on its certified answer-only block to form acquisition trajectories
end for
for each client t **do**
  Construct fixed-budget acquisition/replay candidates with Eq. 9
  Train candidates with Eq. 3 and select with Eq. 10
end for
return Personalized checkpoints and complete supervision provenance

Algorithm 1 separates question acquisition from final checkpoint selection. Within each acquisition block, client checkpoints remain fixed while verified responses update the capability records and question chains. The acquired answers then support local updates. After acquisition, each receiver trains acquisition–replay candidates and selects a checkpoint on separate calibration data. This algorithm describes the answer-only pipeline; Appendix G specifies pre-answer source selection.

B.1 Initial capability catalog

This pre-transfer catalog uses a diagnostic sample separate from transfer evaluation.

B.2 ANSWER-ONLY ACQUISITION AND CHECKPOINT SELECTION

The equations below specify the answer-only acquisition policy and per-receiver checkpoint selection for the system-comparison portfolio. The common-selection acquisition study uses the shared candidate ratios in Appendix H. The frozen-source study instead selects sources before observing their responses and fixes the update protocol (Appendix G).

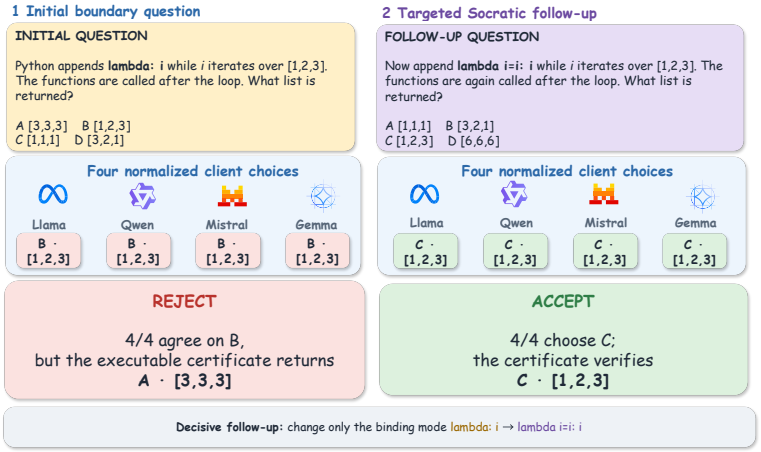


Figure 8: A structured questioning trace. The base Python closure probe elicits the same incorrect answer from all clients. Changing the binding mode to `lambda i=i: i` changes the correct result; the certificate checks each response independently of agreement.

Let $\mathcal{P} = \{\text{base}, \text{inv}, \text{flip}\}$ denote the question phases. The server maintains a phase-specific $\text{Beta}(1,1)$ record for each client–skill pair. If $n^+_{i,s,p}$ and $n^-_{i,s,p}$ are revealed successes and failures, the posterior mean is

$$\widehat{\pi}_{i,s,p} = \frac{1 + n^+_{i,s,p}}{2 + n^+_{i,s,p} + n^-_{i,s,p}}. \tag{4}$$

This estimate follows from Beta–Bernoulli conjugacy. Treating each observed correctness outcome as a Bernoulli trial, a $\text{Beta}(1,1)$ prior and counts $(n^+_{i,s,p}, n^-_{i,s,p})$ give a $\text{Beta}(1 + n^+_{i,s,p}, 1 + n^-_{i,s,p})$ posterior. Equation 4 gives its mean, scoring unobserved client–skill–phase combinations at $1/2$.

The phase-product capability score and lexicographic skill priority are

$$\begin{aligned} \widehat{\pi}_{i,s} &= \prod_{p \in \mathcal{P}} \widehat{\pi}_{i,s,p}, \\ R(s) &= \left(\mathbf{1}[s \notin \mathcal{H}],\ 1 - (\widehat{\pi}_{(1),s} - \widehat{\pi}_{(2),s}),\ \sum_{i,p} 4\widehat{\pi}_{i,s,p}(1 - \widehat{\pi}_{i,s,p}) \right), \end{aligned} \tag{5}$$

Here $\mathcal{H}$ contains skills with a verified teacher and $(1), (2)$ index the two largest client scores. The product combines phase-specific estimates into one ranking score. After initial skill coverage, lexicographic maximization prioritizes missing teachers, small top-two gaps, then uncertain correctness, measured by summing normalized Bernoulli variances $4\widehat{\pi}_{i,s,p}(1 - \widehat{\pi}_{i,s,p})$ over clients and phases. Only revealed discovery outcomes update this catalog; sealed outcomes remain evaluation-only.

For target-specific transfer, each foreign domain has a chain state $x = (k, h, w)$: skill index k, difficulty $h \in \{0, 1, 2\}$, and consecutive-error count w. Given the latest certified target outcome $y = c_z(q, a_t(q))$, the next state is

$$\Gamma(x, y) = \begin{cases} ((k+1) \bmod |\mathcal{S}_d|, \min\{h+1, 2\}, 0), & y = 1, \\ (k, \max\{h-1, 0\}, 1), & y = 0,\ w = 0, \\ ((k+1) \bmod |\mathcal{S}_d|, 0, 0), & y = 0,\ w \geq 1. \end{cases} \tag{6}$$

A correct answer advances and raises difficulty; the first error keeps the skill at an easier novel state; a second error advances at easy difficulty. The loop collects K units per local update, so K changes the number of cycles under fixed exposure.

If several non-target peers pass, the server uses only revealed domain history to choose a source. With $b_{i,d} = (n^{+}_{i,d} + 1)/(n_{i,d} + 2)$ and frozen domain owner $o(d)$, it selects

$$j^*(q,t) = \operatorname*{lexmax}_{i \in T_z(q) \setminus \{t\}} \left(b_{i,d}, \mathbf{1}[i = o(d)], -\operatorname{rank}(i)\right). \tag{7}$$

The final term is a deterministic client-identifier tie break. If no peer passes, the server either abstains or uses the certificate's semantic answer as a curriculum item. The normalized record is

$$r_t(q) = \big(q, \widetilde{a}_z(q), \ell_z(q), \rho_t(q)\big), \quad \rho_t(q) = \begin{cases} \text{peer} : j^*(q,t), & T_z(q) \setminus \{t\} \neq \varnothing, \\ \text{certificate completion}, & \text{otherwise}, \end{cases} \tag{8}$$

Here $\widetilde{a}_z(q)$ and $\ell_z(q)$ are the certified semantic answer and option label. Completion can broaden phase coverage, but its provenance prevents attribution to client knowledge. Unrestricted reasoning is excluded because an erroneous rationale can accompany a correct independently mapped label.

Let $\mathcal{A}_t$ contain acquired records and $\mathcal{R}_t$ contain client-correct, certificate-verified replay records from protected domains. For a fixed budget $B = 100$ and replay count m, a candidate dataset is

$$\mathcal{S}_t^{(m)} = \operatorname{Sample}(\mathcal{A}_t, B - m) \cup \operatorname{Sample}(\mathcal{R}_t, m), \qquad m \in \{0, 25, 50, 75\}, \tag{9}$$

Targeted domain allocations provide additional replay when a protected skill requires it.

On the disjoint calibration catalog, the two strongest no-transfer domains define protected set $\mathcal{P}_t$ and the remaining domains define acquisition set $\mathcal{Q}_t$. For candidate m, let $\Delta_{t,d}^{(m)}$ be its accuracy change from no transfer. Eligibility and selection are

$$\begin{aligned} E_t^{(m)} &= \mathbf{1}\left[\min_{d \in \mathcal{P}_t} \Delta_{t,d}^{(m)} \geq -0.02,\ \frac{1}{|\mathcal{Q}_t|} \sum_{d \in \mathcal{Q}_t} \Delta_{t,d}^{(m)} \geq 0.03,\ \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \Delta_{t,d}^{(m)} \geq 0\right], \\ m_t^* &= \operatorname*{lexmax}_{m : E_t^{(m)} = 1} \left(\overline{\Delta}_{t,\mathcal{Q}}^{(m)}, \overline{\Delta}_{t,\mathcal{D}}^{(m)}, \operatorname{tie}(m)\right), \end{aligned} \tag{10}$$

If no candidate is eligible, the client retains its pre-update checkpoint. A frozen tie order is deterministic. The gate checks calibration retention; held-out retention is evaluated separately. This rule defines checkpoint selection for the answer-only system portfolio. Appendix C.4 compares generic and personalized replay at one fixed mixture, without the portfolio's candidate search.

C Controlled-task component audits

These studies use generated, certifiable tasks with study-specific training and evaluation protocols.

C.1 System-level comparison

The system-level study compares the answer-only FedSocratic portfolio against no transfer, random transfer, majority vote, self-reported confidence, non-adaptive certificates, and semantic-interface adaptations of FedMD, Fed-ET, and FedMKT. The literature-derived conditions adapt the published shared-input procedures to generated questions and semantic answers. Final-update steps are matched, while acquisition and candidate search differ. Figure 4 reports the separate acquisition-policy study with common update selection and new evaluation states.

The selected portfolio has the highest overall score in this comparison, while the non-adaptive certificate control leads on code. Paired latent-unit 95% intervals for the portfolio minus non-adaptive certificates are $[3.35, 6.40]$ overall, $[0.44, 3.92]$ foreign, and $[9.83, 16.00]$ native, in percentage points. These intervals quantify test-question uncertainty for the selected checkpoints.

Table 5: Final-update-matched system comparison: accuracy (%) on the fixed controlled-task test set. Domain columns pool four clients. Existing methods are setting-adapted; each trained row uses 4,000 final-update steps, excluding acquisition and selection costs. Best values are bold.

Method	Code	Math	Finance	Medical	Foreign	Native	Overall
No transfer	38.75	42.50	48.00	31.17	39.11	43.08	40.10
Random transfer	36.92	39.67	47.50	32.75	39.75	37.58	39.21
Majority vote	37.83	36.83	39.17	33.58	36.44	38.08	36.85
Confidence selection	37.08	40.25	45.92	34.00	39.06	40.08	39.31
Non-adaptive certificate	**54.00**	47.08	58.83	41.83	52.94	42.92	50.44
FedMD (Li & Wang, 2019)	37.50	40.83	49.17	36.50	40.83	41.50	41.00
Fed-ET (Cho et al., 2022)	38.50	37.17	47.92	36.58	38.94	43.33	40.04
FedMKT (Fan et al., 2025)	44.58	43.25	49.67	36.00	44.97	38.58	43.38
FEDSOCRATIC portfolio	49.67	**60.92**	**59.17**	**51.50**	**55.14**	**55.83**	**55.31**

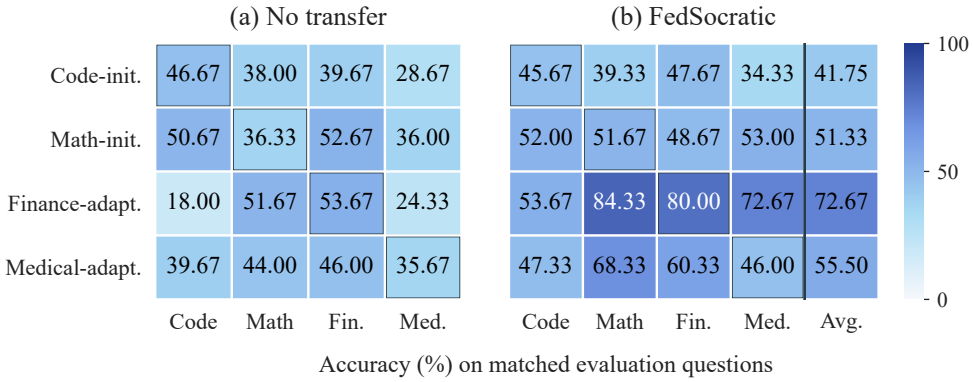


Figure 9: Paired client capability profiles before and after answer-only FEDSOCRATIC transfer on the same system-comparison questions (300 per client–domain cell). Both panels share an accuracy scale (%); the last column gives post-transfer client averages. Outlines mark nominal domains.

C.2 CLIENT-LEVEL PORTFOLIO DIAGNOSTICS

Complementary knowledge under client-specific transfer. Figure 9 reports paired pre-transfer and post-transfer accuracy for every client–domain cell in the selected portfolio. The finance-adapted client has the strongest post-transfer mathematics and finance scores, while the math-initialized and medical-adapted clients also perform well outside their nominal domains. These receiver profiles motivate measuring capabilities directly across domains.

C.3 UPDATE-INTERVAL SWEEP

All update intervals use 100 supervision records and 1,000 optimizer steps per client. Changing the interval also changes intermediate models and the subsequent acquisition trajectory. Figure 10 and Table 6 report the resulting sensitivity on a separate evaluation stream.

C.4 PROBING AND REPLAY FACTORIAL CONTROL

The factorial control crosses fixed versus target-adaptive acquisition with generic versus personalized verified-self replay. Each client receives 50 acquisition records and 50 replay records, 1,000 optimizer steps, and one candidate per cell. Templated training-token counts are matched across the four cells for each client. The generic control preserves the replay budget and samples without targeting the client's protected domains. Fixed and adaptive acquisition records come from the frozen non-adaptive and $K = 100$ trajectories, respectively. Each cell evaluates a single replay candidate.

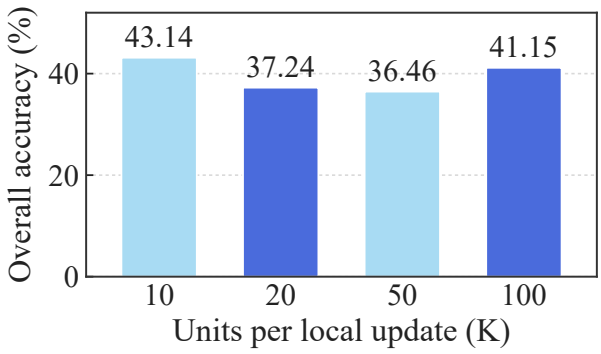


Figure 10: Overall accuracy by update interval on the isolated evaluation stream. The no-transfer reference is 38.98%; full domain scores appear in Table 6.

Table 6: Sensitivity to the number of knowledge units per local update on the separate isolated evaluation stream. Values are pooled accuracy (%) across all four clients; each client–domain cell contains 72 questions. All K conditions use 100 supervision records and 1,000 optimizer steps per client. Bold marks the best value in each column.

Update interval	Code	Math	Finance	Medical	Overall
Initial (no transfer)	38.54	42.71	41.32	33.33	38.98
$K = 10$	**38.89**	**46.88**	**43.40**	**43.40**	**43.14**
$K = 20$	34.03	42.36	38.54	34.03	37.24
$K = 50$	36.11	36.81	40.28	32.64	36.46
$K = 100$	38.19	45.83	42.01	38.54	41.15

Figure 11 shows the seed-level results and their means.

Personalized-minus-generic replay averages -3.26 points with fixed acquisition and -4.53 with adaptive acquisition. Adaptive-minus-fixed acquisition averages -2.59 points with generic replay. The interaction, defined as the personalized replay effect under adaptive probing minus that under fixed probing, is -4.15, $+0.71$, and -0.35 points across seeds: mean -1.26, sample standard deviation 2.55. The sample standard deviation measures training-seed variation.

C.5 Matched phase coverage

The source-conditioned phase study compares two-phase and three-phase coverage while matching rows, latent states, task composition, and source proportions. The two-phase pair varies by state according to source correctness, and one retained question is repeated. It is therefore distinct from the fixed-deletion study in Figure 6. Evaluation contains 72 weak-domain questions per target (288 in total, grouped into 96 latent units). Three-minus-two-phase gains are 2.43, 1.74, and 3.13 points for seeds 1717, 1729, and 1741, respectively, averaging 2.43 points (sample standard deviation 0.69).

The reported weak-domain grouping uses a documented post-freeze analysis repair: client 2 is evaluated on medical rather than mathematics, and client 3 on finance rather than code, matching the actual held-out domains. The repair selects the sole non-native domain in each target's held-out records, preserving predictions and training. Both the pre-correction analysis and the corrected grouping record are retained. The contrast supports broader phase coverage in these tasks.

C.6 Source completion and retention diagnostics

The completion-expanded pool reaches 70.83% weak-skill accuracy; three frozen resamples matched to the source-only row count average 70.72%. The paired interval for full completion minus the equal-count mean is $[-2.20, 2.55]$ percentage points. Source-correct transfer changes

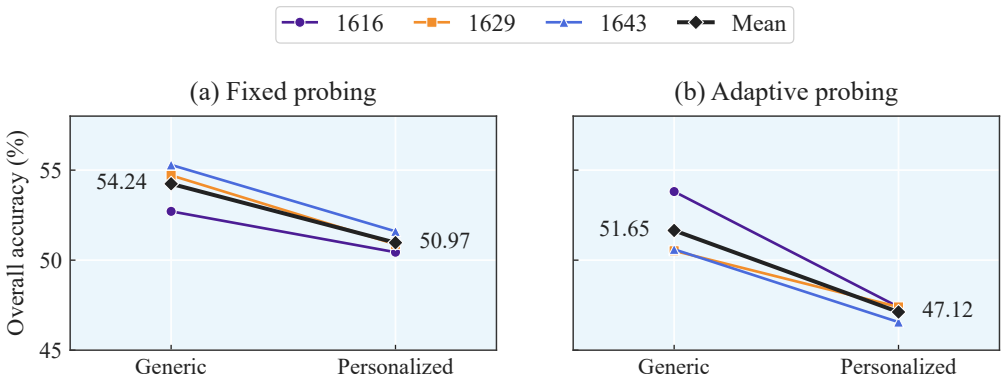


Figure 11: Probing–replay factorial accuracy (%). Thin lines show individual seeds; thick lines show three-seed means. Panels share an axis scale; lines connect categorical replay conditions.

the code-, math-, finance-, and medical-side strong-domain scores by $+2.78$, -4.17, -16.67, and -9.72 points, respectively. These outcomes distinguish answer acquisition from retention.

HellaSwag (Zellers et al., 2019) uses identical documents across compared checkpoints. Changes for the four clients are $+1.4$, -0.8, -2.7, and $+0.3$ points. The pooled mean is -0.45 with a document-clustered interval of $[-1.225, 0.325]$; the finance-adapted client's interval excludes zero.

D Model-composition and supervision-content extensions

These studies replace the code-initialized 7B client with Qwen2.5-14B-Instruct while retaining the other three clients. This composition test jointly changes size and specialization and uses controlled generated tasks; public-task evaluation appears in Appendix E.

D.1 Answer-only transfer with one 14B client

Table 7: Controlled-task accuracy (%) with the client composition including Qwen2.5-14B-Instruct.

Method	Overall	Cross-domain	Native
No transfer	46.88	45.14	52.08
Server-only certified	51.91	53.47	47.22
Fixed certificate + matched replay	53.99	55.56	49.31
FedSocratic, 14B composition	49.65	50.93	45.83
Peer-only without completion	44.62	45.14	43.06

The FedSocratic configuration with the 14B composition improves overall accuracy over its no-transfer reference but falls below both server-certified controls. Replacing one client with a larger model changes the comparison and the portfolio's ranking relative to the 7–8B composition.

D.2 Rationale-content controls

The initial exploration adds free-form rationales whose final answers pass verification to answer-only supervision. Overall accuracy is 49.83% with server rationales and generic replay, 53.47% with peer rationales and generic replay, and 48.44% with peer rationales and personalized replay; no transfer is 46.88%. Without replay, server and peer rationale conditions obtain 42.71% and 47.74%, respectively. Verification applies to final answers; intermediate reasoning remains unchecked.

A token-matched rationale study adds matched controls and training seeds on the internal multiple-choice protocol, with 144 evaluation questions per client (36 per domain). It audits 1,000 optimizer

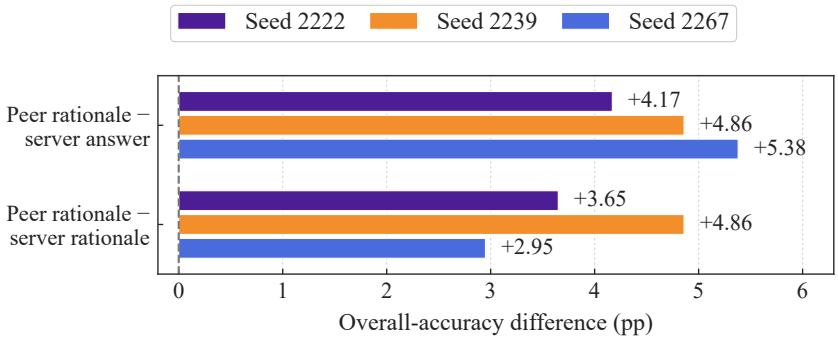


Figure 12: Token-matched rationale-study overall-accuracy differences (percentage points) for three training seeds. Both contrasts use peer rationales with generic replay, minus the named server supervision. Zero denotes no difference; bars show seed-level point estimates.

steps and exact integrated token budgets for the compared conditions. Figure 12 reports within-seed source contrasts; absolute scores use this study's evaluation protocol.

The point estimates favor peer rationales in all three seeds, but the last seed's peer-minus-server-rationale interval is $[-0.35, 6.25]$ points. The result supports a content-dependent effect on the internal protocol. Personalized-minus-generic replay is -2.78 and -3.99 points in the two additional seeds, consistent with the replay limitation in Appendix C.4.

E PUBLIC-TASK EVALUATION OF EARLIER UPDATE CONFIGURATIONS

To test whether gains on generated task families carry over to public tasks, we evaluate frozen checkpoints on 200 held-out questions each from MBPP (Austin et al., 2021), GSM8K (Cobbe et al., 2021), FinQA (Chen et al., 2021), and PubMedQA (Jin et al., 2019). Metrics are MBPP pass@1, GSM8K final-answer exact match, FinQA normalized final-answer exact match, and PubMedQA accuracy. Overall is the equally weighted client–task macro-average; native averages the four nominal client–task matches and cross-domain averages the other 12. The benchmark tasks are distinct from the procedural rule families used for training.

Table 8: Public-task accuracy (%). The 7–8B and 14B-composition groups use separate no-transfer references. The latter replaces Qwen2.5-Coder-7B with Qwen2.5-14B-Instruct (Appendix D).

Group / checkpoint	Overall	Native	Cross-domain
7–8B: no transfer	49.22	64.88	44.00
7–8B: $K = 10$	43.97	62.50	37.79
7–8B: $K = 100$	43.91	62.62	37.67
7–8B: FedSocratic portfolio	44.56	63.75	38.17
With 14B: no transfer	50.56	62.62	46.54
With 14B: server rationale + generic	50.37	62.87	46.21
With 14B: peer rationale + generic	44.59	57.63	40.25
With 14B: peer rationale + personalized	43.28	54.25	39.63
With 14B: FedMKT	48.38	57.25	45.42

The answer-only portfolio changes overall accuracy by -4.66 points. The 14B-composition peer-rationale updates also decrease accuracy. These results motivate the update-intensity controls in Appendix F and the calibrated peer configurations in Appendix I.

Evaluation lineage. Exploratory known-domain routing used separate 200-question benchmark subsets. It obtained 50.66% overall and 47.17% foreign accuracy with server-rationale modular

updates, versus 49.78% and 46.00% for its reference. Its global comparator used a higher learning rate and more epochs. Appendix F gives matched-source controls and independent confirmation.

F SUPERVISION CONTENT AND UPDATE INTENSITY

F.1 MATCHED QUESTIONS AND UPDATE PROFILES

This control study uses the composition with Qwen2.5-14B-Instruct (Appendix D). Server-generated foreign-domain pools retain questions whose peer and server final answers pass the same certificate: 74, 70, 82, and 80 unique questions for clients 0–3. On answer-only records, question order, clean-data hashes, tokenized-data hashes, and training settings coincide across sources, giving one token-equivalent condition. Retained rationales provide source-dependent text and token counts.

All conditions use LoRA rank 8, alpha 16, dropout 0.05, `q_proj` and `v_proj`, cutoff length 1,024, batch size and gradient accumulation 1, cosine scheduling, warmup ratio 0.03, and BF16. Low intensity uses three epochs at 5×10^{-6}; high intensity uses ten epochs at 2×10^{-5}. Global training updates one adapter per client on all foreign-domain records. With known task domains, modular training updates three domain adapters and routes native questions to the frozen model.

Lower update intensity recovers much of the peer-supervision loss. On the 200-question-per-benchmark development subsets, global low-minus-high intensity improves foreign accuracy by 9.17 points for peer rationales, with a question-clustered 95% interval of $[7.38, 10.96]$. Peer-minus-server rationale differences at low intensity are -0.46 points for modular and -0.29 for global updates; their intervals, $[-1.00, 0.08]$ and $[-1.04, 0.46]$, span zero. These controls connect performance to the supplied text and the update recipe.

F.2 CONFIRMATION AND FACTORIAL FOLLOW-UP

Before development training, 100 questions per benchmark were frozen with subset seed 30300001, disjoint from the preceding public evaluations. A predeclared rule selected server-rationale modular low intensity from the two modular rationale candidates, together with a source-matched global comparator. Three training seeds (3030, 3041, 3052) gave modular-minus-no-transfer foreign changes of -0.50, 0.00, and -0.33 points, averaging -0.28; the overall mean change was -0.21. Native accuracy remained unchanged through the frozen native route.

The factorial follow-up completes the source, update structure, and intensity combinations using the same seeds and already evaluated confirmation questions (Table 9).

Table 9: Full factorial follow-up, averaged across three seeds. Values are accuracy (%); the no-transfer reference is 50.50 overall, 62.25 native, and 46.58 foreign.

Source	Structure	Intensity	Overall	Native	Foreign
Server rationale	Global	Low	50.46	62.75	46.36
	Global	High	50.75	61.42	47.19
	Modular	Low	50.29	62.25	46.31
	Modular	High	49.50	62.25	45.25
Peer rationale	Global	Low	50.62	64.00	46.17
	Global	High	44.12	61.83	38.22
	Modular	Low	50.60	62.25	46.72
	Modular	High	46.54	62.25	41.31

Supervision and update intensity interact. Peer and server rationales have similar foreign scores at low intensity, while peer rationales perform worse in every seed at high intensity under both structures. Modular updates reduce the high-intensity peer loss. Rationale lengths vary by source. This motivates update-intensity control; Appendix I reports separate calibrated 7–8B peer tests.

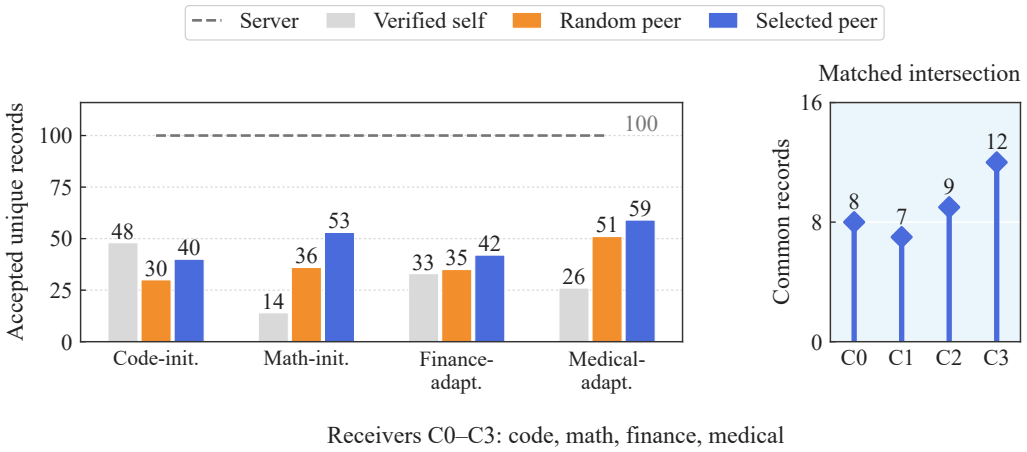


Figure 13: Accepted unique supervision records from 100 candidate questions per receiver. Left: model-generated sources, with server acceptance of all 100 questions as a dashed reference. Right: the common accepted intersection across all four sources.

Computational record. The preceding bridge and independent confirmation runs completed 84 training jobs and 240 evaluation units, using 1.21 cumulative GPU hours for training and 12.96 for evaluation. These totals exclude earlier model selection, generation, and source queries.

G Frozen-source supervision study

G.1 Acquisition, diagnostics, and updates

The frozen-source study uses the four-client 7–8B composition described in Section 4 and three training seeds: 3535, 3551, and 3569. Each receiver has 100 candidate questions, reusing the saved same-question raw responses from the common-selection acquisition study (seed 3232). A newly generated diagnostic set contains 48 latent states and 144 questions; the final test contains 80 states and 240 questions. Both diagnostic and final states have zero overlap with states used in the preceding controlled-task studies, and the final set is excluded from source selection and training.

Verified self uses the receiver's own response when it passes the verifier. Random peer selects one other client using a frozen random rule before its answer is inspected. Capability-selected peer ranks non-target clients by diagnostic skill accuracy, then domain accuracy, then client identifier; the ranking is frozen before acquisition decisions. Server-certified supervision uses the certificate's correct answer. The same verifier decides acceptance for model-generated responses. It checks final answers, not every step of a retained rationale.

The candidate-budget regime applies the same candidate-question flow and source-call rule to each model source, retaining the unique accepted records. Candidate and call budgets are matched; response lengths and total inference compute remain source-dependent. The matched-question regime takes the intersection of questions accepted under all four sources for a given receiver. It preserves question identity, skill and difficulty composition, record counts, native replay, and optimization steps. The intersection contains the unique-record counts in Figure 13, which vary across receivers. Response text and training-token counts remain source-dependent on matched questions.

All conditions use 32 verified native replay records and 500 LoRA optimizer steps per receiver. Before new training or final evaluation, the replay quota was amended from 40 to 32 for all clients because the medical-adapted client had 38 eligible replay records. The amendment was recorded before observing final results; all training uses generated questions.

G.2 WITHIN-SEED SOURCE CONTRASTS

Table 10 gives the prespecified foreign-domain contrasts. The means and sample standard deviations summarize variation across the three training seeds. Absolute mean accuracies are reported in Table 3. The shared-question comparison uses each receiver's small all-source verified intersection.

Table 10: Frozen-source foreign-accuracy differences in percentage points. Selected peer is the capability-based source. Values are reported from the unrounded experiment aggregates.

Regime	Selected peer minus	3535	3551	3569	Mean	SD
Candidate budget	Verified self	+3.61	+2.36	+4.44	+3.47	1.05
	Random peer	+4.58	+2.92	+5.28	+4.26	1.21
	Server	+3.61	+3.33	+6.11	+4.35	1.53
Matched questions	Verified self	+2.64	+2.22	+0.00	+1.62	1.42
	Random peer	+3.33	+2.64	+0.42	+2.13	1.52
	Server	+4.72	+4.58	+3.75	+4.35	0.53

Table 11: Capability-selected peer performance by receiver. Foreign accuracy is averaged across three training seeds; changes are relative to the receiver's no-transfer reference.

Regime	Receiver	Foreign (%)	Change (pp)
Candidate budget	Code-init.	32.59	+0.93
	Math-init.	39.26	−1.85
	Finance-adapt.	37.78	+4.44
	Medical-adapt.	48.15	+2.04
Matched questions	Code-init.	34.81	+3.15
	Math-init.	38.70	−2.41
	Finance-adapt.	36.85	+3.52
	Medical-adapt.	44.63	−1.48

The experiment records include the training-data manifest, frozen source policy, acquisition decisions, protocol amendment, and per-seed analysis. This study compares frozen source policies with response-retaining supervision and fixed replay. It differs from both the system-comparison portfolio and the answer-only acquisition policies evaluated under common update selection.

H COMMON UPDATE SELECTION AND PHASE-COVERAGE DETAILS

H.1 ACQUISITION UNDER COMMON UPDATE SELECTION

The common-selection study compares acquisition policies using the four 7–8B clients in Section 4 and seeds 3232, 3251, and 3271. Fixed server-certified acquisition, adaptive server acquisition without peer filtering, and peer-informed acquisition share four acquisition/replay ratios: 75/25, 50/50, 25/75, and 0/100. Each candidate has 100 training records, ten epochs, and 1,000 optimizer steps. The pure-replay checkpoint is shared across policies. This gives ten unique candidates per receiver and seed, for 120 training jobs.

Acquisition/discovery, calibration, and final states are disjoint per seed. Final evaluation uses 12 latent states per domain with three phases each (144 questions per client). Policies share candidate ratios and calibration gates, with policy-dependent acquisition pools. A deterministic token-length matching search records counts without enforcing equality. Table 12 reports seed-level scores.

This protocol uses a different candidate set and evaluation scale from the system-comparison portfolio. The portfolio includes acquisition/replay ratios 100/0, 75/25, 50/50, and 25/75, together with targeted domain allocations (Appendix B.2), and is evaluated on 300 questions per client–domain cell. Here, the shared set includes pure replay and evaluation uses 36 questions per cell. This study compares acquisition policies under a common selection rule.

Table 12: Held-out accuracy (%) under common answer-only update selection, by training seed. Policies share candidate ratios and calibration gates. Conditions correspond to Figure 4; Table 2 uses the separate portfolio protocol.

Seed	Pipeline	Foreign	Native	Overall
3232	No transfer	36.34	45.14	38.54
	Fixed server-certified	66.67	63.19	65.80
	Adaptive server, no peer	53.01	63.19	55.56
	Peer-informed	48.84	64.58	52.78
3251	No transfer	39.35	46.53	41.15
	Fixed server-certified	65.97	61.81	64.93
	Adaptive server, no peer	52.55	63.19	55.21
	Peer-informed	53.70	65.97	56.77
3271	No transfer	38.43	47.92	40.80
	Fixed server-certified	66.44	65.97	66.32
	Adaptive server, no peer	53.01	68.06	56.77
	Peer-informed	51.85	67.36	55.73

H.2 Fixed phase deletions

The fixed phase-deletion study compares complete three-phase coverage with fixed removal of flip, invariance, or base questions. Each target has 24 training states and 72 rows, with the same skill composition, source-correct row counts, and LoRA configuration. Each condition trains for 720 optimizer steps. Deletions use 48 distinct retained questions and 24 controlled repeated rows; complete coverage uses 72 distinct phase questions. Provenance records preserve repetition choices.

Each receiver is evaluated on 24 new states and 72 questions, with zero overlap against 5,925 states used in the preceding controlled-task studies. Training seeds are 3333, 3347, and 3361. The source-conditioned phase comparison in Appendix C selected phase pairs according to source correctness; here, the removed phase is fixed across states. These are distinct coverage contrasts.

Figure 14 reports the paired coverage effects; Table 13 gives the receiver and phase breakdown.

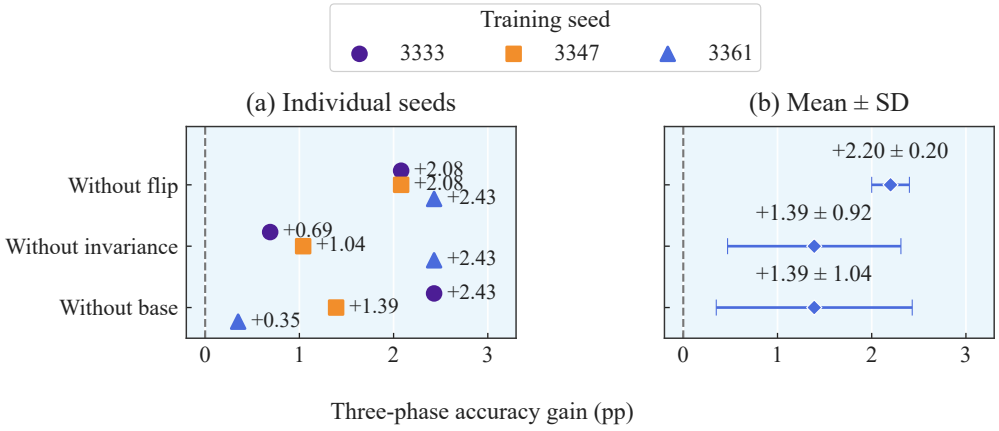


Figure 14: Complete three-phase coverage minus each fixed deletion (percentage points). Left: individual training-seed effects. Right: mean ± sample standard deviation across training seeds.

Table 13: Mean accuracy (%) by receiver and test phase across three training seeds.

Coverage	Client 0	Client 1	Client 2	Client 3	Base	Inv.	Flip
Complete	50.00	47.22	75.93	67.13	61.11	71.18	47.92
Without flip	50.93	48.15	69.44	62.96	58.33	68.06	47.22
Without inv.	50.46	46.76	72.69	64.81	60.42	68.40	47.22
Without base	50.93	48.15	75.00	60.65	59.72	69.79	46.53

I SUPPLEMENTARY CALIBRATION-BASED UPDATES

I.1 CONFIGURATIONS AND EVALUATION

These follow-up experiments examine whether adjusting local update strength helps receivers use the peer supervision from the frozen-source study (Appendix G). The *original peer pool* reuses capability-selected responses with verified final answers, mixed with 32 native replay records per receiver. Short local updates are selected separately for each client on calibration questions. The *expanded peer pool* additionally explores response cleaning, data recipes, and replay ratios. Both retain the initial checkpoint when no update meets the calibration criteria.

Selection requires improved foreign accuracy, non-decreasing overall accuracy, and at most one lost native answer on the calibration set. Final evaluation follows checkpoint selection. Both publicly evaluated configurations retain the original medical client.

Internal evaluation uses 180 questions per client–domain cell, or 720 per client. Public evaluation uses 200 questions per client–task cell on MBPP, GSM8K, FinQA, and PubMedQA, with the task metrics defined in Appendix E. These are supplementary protocols with their own no-transfer references; the system comparison in Table 2 uses 300 controlled questions per client–domain cell.

I.2 INTERNAL TRANSFER ACROSS TRAINING SEEDS

Table 14: Supplementary internal evaluation: accuracy (%) on 180 questions per client–domain cell. Each configuration has its own no-transfer reference; training seeds share acquisition data.

Peer pool	Seed / reference	Overall	Foreign	Native
Original peer pool	No transfer	38.65	37.78	41.25
	3801	39.69	38.38	43.61
	3811	39.65	38.43	43.33
	3821	39.76	38.94	42.22
Expanded peer pool	No transfer	38.61	37.82	40.97
	3901	39.13	38.70	40.42
	3911	39.72	39.17	41.39
	3921	39.58	38.75	42.08

Overall and foreign gains recur across training repeats. All six updates improve both metrics over their no-transfer references (Table 14), with overall gains of approximately 0.5–1.1 percentage points. Native changes vary across runs. Repeats assess training variability with acquisition fixed.

I.3 PUBLIC-TASK PERFORMANCE

Average public performance is largely retained, with small positive changes. The original-pool configuration increases overall accuracy from 49.13% to 49.72%, with a 0.79-point foreign gain and unchanged native average. The expanded-pool configuration increases overall accuracy from 48.84% to 49.09%, with a 0.38-point foreign gain and a 0.13-point native decrease. These aggregates summarize one complete four-client, four-task evaluation for each configuration.

Table 15: Public-task macro-averages (%), initial → updated, for one seed per row. Overall equally weights 16 cells; Foreign averages 12 off-diagonal cells, Native four diagonal cells.

Peer pool / seed	Overall	Foreign	Native
Original peer pool / 3801	49.13 → 49.72	43.83 → 44.63	65.00 → 65.00
Expanded peer pool / 3901	48.84 → 49.09	43.46 → 43.83	65.00 → 64.88

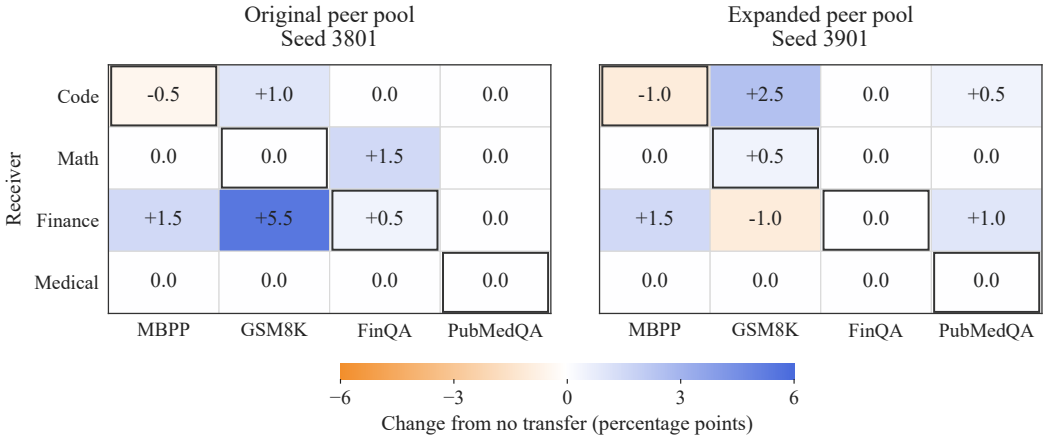


Figure 15: Public-task changes in percentage points for the two supplementary configurations. All 16 cells are shown in each panel; outlined diagonal cells mark native tasks. Panels share a zero-centered scale. Cells show single-seed before–after changes.

Changes are concentrated in particular receiver–task pairs. The original-pool finance receiver improves GSM8K by 5.5 points, while most other cells change by 0–1.5 points. The expanded-pool configuration has smaller mixed changes, including −1.0 on the code receiver's MBPP and the finance receiver's GSM8K. Tables 16 and 17 give every initial and updated score. Different pools and no-transfer references require within-panel interpretation.

Table 16: Original peer pool, seed 3801: accuracy (%) before/after updates; 200 questions/cell.

Receiver	MBPP	GSM8K	FinQA	PubMedQA
Code-initialized	74.0 → 73.5	78.0 → 79.0	18.0 → 18.0	44.5 → 44.5
Math-initialized	42.0 → 42.0	92.5 → 92.5	9.5 → 11.0	66.0 → 66.0
Finance-adapted	53.5 → 55.0	55.5 → 61.0	30.0 → 30.5	65.5 → 65.5
Medical-adapted	41.5 → 41.5	49.0 → 49.0	3.0 → 3.0	63.5 → 63.5

Table 17: Expanded peer pool, seed 3901: accuracy (%) before/after updates; 200 questions/cell.

Receiver	MBPP	GSM8K	FinQA	PubMedQA
Code-initialized	74.0 → 73.0	76.5 → 79.0	17.5 → 17.5	44.5 → 45.0
Math-initialized	42.0 → 42.0	92.0 → 92.5	10.0 → 10.0	65.0 → 65.0
Finance-adapted	54.0 → 55.5	59.0 → 58.0	30.5 → 30.5	64.5 → 65.5
Medical-adapted	41.5 → 41.5	44.0 → 44.0	3.0 → 3.0	63.5 → 63.5

The finance receiver's GSM8K gain persists across training seeds. Three training seeds of the original-pool configuration raise this score from 55.5% to 61.0%, 57.5%, and 58.5%. This repeat covers the finance–GSM8K pair; the complete public evaluation is reported for seed 3801.