D-ICL: DECENTRALIZED IN-CONTEXT LEARNING WITH TABULAR FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributed tabular learning presents a fundamental challenge: data are distributed across silos that cannot share raw records, while each silo alone lacks sufficient coverage for accurate modeling. We introduce Decentralized In-Context Learning (D-ICL), a framework that enables effective collaboration by leveraging the in-context learning capability of frozen tabular foundation models, without sharing either data or model parameters. Each of K agents maintains a private context and exchanges only predictive distributions over a shared query set (K is the number of agents). These predictions are aggregated to construct high-confidence pseudo-labels, which iteratively augment local contexts and drive collective improvement. We provide three theoretical guarantees: (i) a K-fold variance reduction via consensus aggregation, (ii) monotonic performance improvement under a calibrated high-confidence pseudo-labeling scheme, and (iii) an $\mathcal{O}(1/T)$ convergence rate of the consensus loss over T iterations. Empirically, across diverse tabular benchmarks under both IID and non-IID partitions, D-ICL consistently narrows the gap between isolated agents and a centralized oracle, often approaching centralized performance while strictly preserving data locality. The code base is available here.

1 INTRODUCTION

Distributed tabular learning (Kotal et al., 2022) is increasingly central to high-stakes domains such as healthcare (Nik et al., 2023), finance (Clements et al., 2020), and scientific discovery (Attari & Arroyave, 2025), where data are inherently fragmented across institutional silos that cannot share raw records due to legal, regulatory, or proprietary constraints. This fragmentation creates a fundamental tension: modern predictive performance relies on aggregating diverse and large-scale datasets, yet each individual silo often contains only limited samples, leading to unreliable models when trained in isolation. Bridging this gap of achieving the benefits of collaboration without violating data privacy remains a core challenge in modern machine learning.

Federated learning (FL) (Li et al., 2020) has emerged as the dominant paradigm here, training a shared model via iterative gradient aggregation across agents. However, FL faces critical limitations: gradients can leak sensitive information through reconstruction attacks (Yang et al., 2023a; Wei et al., 2020), undermining its own privacy guarantees. Communication overhead (Chen et al., 2021) scales with model size, requiring repeated transmission of high-dimensional updates and optimizer states, particularly acute for large foundation models and resource-constrained agents. Server failure can halt the entire training process (Mansouri et al., 2022), disrupting synchronization across clients.

Recent advances in Tabular Foundation Models (TFMs) (Hollmann et al., 2025) suggest a fundamentally different direction. Trained offline on large and diverse collections of tabular datasets, TFMs are deployed as frozen predictors that perform inference via in-context learning (ICL) (Dong et al., 2024), conditioning directly on a provided set of labeled examples at query time. This paradigm eliminates the need for gradient-based training during deployment and shifts the locus of learning from parameter updates to context construction. As a result, TFMs naturally decouple knowledge acquisition (during pretraining) from task adaptation (during inference), opening up new possibilities for distributed collaboration. These observations motivate the central question: *Can multiple agents collaborate to improve tabular prediction performance using only in-context inference without sharing raw data, gradients, or model parameters, while remaining communication-efficient and theoretically grounded?* This setting is not hypothetical. Institutions increasingly adopt a shared,

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.

vendor-distributed frozen model rather than training bespoke ones, e.g., hospitals piloting a common risk-scoring model, or banks using a single licensed credit-scoring model for auditability, so that every participant holds the identical frozen model. Gradient-based FL is a poor fit here as it jointly updates a shared trainable model, but once agents already hold the same frozen one, there is nothing left to jointly train, and fine-tuning it is often exactly what regulatory or resource constraints prohibit.

Contributions. To address this challenge, we propose **Decentralized In-Context Learning (D-ICL)** shown in Figure 1, a four-phase protocol in which each agent uses a frozen tabular foundation model with its private local context to generate predictions over a shared unlabeled query pool, broadcasts only these predictions to its neighbors, aggregates them via consensus, and uses the resulting high-confidence pseudo-labels to iteratively enrich its local context, all without sharing raw data or model parameters. D-ICL targets consortia that adopt a common, audited, frozen TFM and a shared unlabeled query pool. Within this scope, it offers a narrower direct-exposure surface than gradient sharing, orders-of-magnitude lower communication cost, and elimination of gradient-based optimization and parameter synchronization. This departs from ensembles (Dong et al., 2020) (identical frozen models, diversity from data contexts, not model training), self-training (Kumar et al., 2020) (multi-agent consensus, not a single model's own outputs), and FL (Lu et al., 2024) (prediction-space collaboration, not parameter synchronization). Our contributions are threefold. First, we formalize D-ICL and its four-phase protocol, i.e., local inference, broadcast, consensus, and context enrichment, the first framework to cast decentralized collaboration for tabular foundation models purely in prediction space. Second, we establish theoretical guarantees showing that consensus aggregation reduces variance, high-confidence pseudo-labeling yields monotonic improvement, and the overall procedure achieves sublinear convergence. Finally, we comprehensively evaluate two backbones (TabICL and TabPFN) within a unified decentralized framework across multiple tabular benchmarks, partition strategies, and communication topologies, including ablations on aggregation strategy, confidence threshold, and data heterogeneity, with D-ICL consistently narrowing the gap between isolated agents and a centralized oracle across all settings.

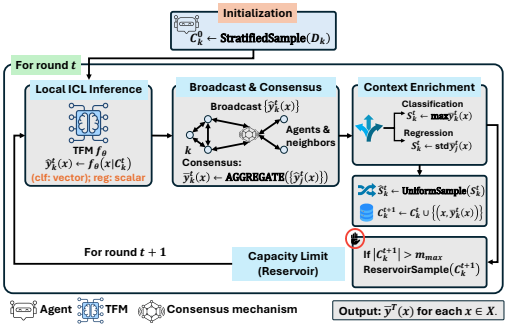


Figure 1: A systematic schematic illustrating the workflow of the decentralized In-Context Learning (D-ICL) method (single agent): The process begins with initialization through stratified sampling, followed by repeated rounds of local ICL inference, neighbor broadcast and consensus aggregation, and context enrichment based on classification max-predictions or regression standard deviation before applying reservoir-based capacity limits and outputting final predictions.

2 Related Work

Tabular Foundation Models. Tabular Foundation Models (TFMs) (Van Breugel & Van Der Schaar, 2024) have emerged as a strong alternative to classical methods for structured data. TabPFN (Hollmann et al., 2022) introduced a transformer trained on synthetic priors to approximate Bayesian inference in a single forward pass, with TabPFNv2 (Ye et al., 2025) and later versions improving robustness and accuracy. In contrast, TabICL (Qu et al., 2025) and TabICLv2 (Qu et al., 2026) leverage large real-world datasets and structured attention mechanisms, achieving state-of-the-art results on benchmarks such as TabArena (Erickson et al., 2025) and TALENT (Liu et al., 2025). Other transformer-based models (e.g., FT-Transformer (Dai et al., 2025), TabTransformer (Huang et al., 2020)) explore alternative inductive biases, but a common trend is the shift toward pretrain-then-infer paradigms, where frozen models adapt via context rather than gradients. Our work builds on this paradigm, using TFMs for decentralized collaboration without parameter sharing.

Federated and Decentralized Learning. Federated Learning (FL) enables distributed training under data privacy constraints by exchanging and aggregating gradients, with extensions addressing communication efficiency (Malekijoo et al., 2021), non-IID data (Lu et al., 2024), robustness (Xu et al., 2024), and personalization (Tan et al., 2022). However, FL faces key limitations: gradients can leak sensitive data via inversion attacks (Zhu et al., 2019), and communication costs scale with model size due to repeated transmission of high-dimensional updates. Decentralized variants (Saa-

dati et al., 2024; Yuan et al., 2024) improve scalability by removing the central server but introduce additional challenges in consensus and convergence. Crucially, both settings operate in parameter space and rely on iterative optimization. In contrast, D-ICL departs from this paradigm by operating entirely in prediction space, where agents exchange only low-dimensional predictive summaries, eliminating gradient sharing, parameter synchronization, and training overhead.

Ensemble Methods and Knowledge Distillation. Our approach is closely related to ensemble learning (Yang et al., 2023b), where combining predictors reduces variance with simple averaging offering robustness, and methods like Product of Experts (PoE) (Hinton, 2002) producing sharper but less stable predictions. D-ICL's consensus mechanisms can be viewed through this lens, with geometric aggregation corresponding to PoE across agents. It is also connected to knowledge distillation (Gou et al., 2021) and semi-supervised learning (Van Engelen & Hoos, 2020), where models leverage soft predictions or pseudo-labels to improve performance. However, D-ICL differs in two key ways. Pseudo-labels are derived from multi-agent consensus rather than a single model, improving robustness, and all agents use frozen foundation models, with learning occurring via context augmentation instead of parameter updates.

3 Problem Formulation and Preliminaries

Let $\mathcal{F} = \{(x_i, y_i)\}_{i=1}^N$ be a tabular classification dataset with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{P}_C := \{0, \dots, C-1\}$; for regression, $y_i \in \mathbb{R}$. Agent k holds private shard $D_k = \{(x_i, y_i)\}_{i \in \mathcal{I}_k}$, where $\mathcal{I}_1, \dots, \mathcal{I}_K$ partition $\{1, \dots, N\}$. Agents share a common unlabelled query pool $\mathcal{Q} = \{x_j\}_{j=1}^M$ drawn i.i.d. from marginal P_X, the distribution over inputs induced by the full data-generating process, and a held-out test set $X_\bullet$ used only for evaluation. A Tabular Foundation Model is a function

$$f_\theta : \mathbb{R}^d \times \big((\mathbb{R}^d \times \mathbb{P}_C)\big) \longrightarrow \Delta^{C-1},\ \Delta^{C-1} := \big\{p \in \mathbb{R}^C : p_c \geq 0,\ \textstyle\sum_c p_c = 1\big\}. \tag{1}$$

with frozen parameters θ. For regression, it becomes

$$f_\theta : \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R}) \longrightarrow \mathbb{R} \tag{2}$$

For a labelled context $\mathcal{C} = \{(x_i, y_i)\}_{i=1}^m$ and a query $x^* \in \mathbb{R}^d$, the TFM returns a posterior predictive distribution:

$$p = f_\theta(x^* \mid \mathcal{C}) \in \Delta^{C-1}(\text{classification}), \quad \mu = f_\theta(x^* | \mathcal{C}) \in \mathbb{R}\ (\text{regression}). \tag{3}$$

The computation is a single forward pass *without* parameter or gradient updates.

We model non-IID heterogeneity via the Dirichlet partition (Osting & Reeb, 2017). For each class $c \in \mathbb{P}_C$, the per-agent allocation vector is drawn as $(q_{1c}, \dots, q_{Kc}) \sim \mathrm{Dir}(\alpha \cdot \mathbf{1}_K)$, where $\alpha > 0$ controls homogeneity. As $\alpha \to 0$ each agent holds samples from a single class; as $\alpha \to \infty$ the partition converges to uniform IID. For regression, we bin y into K quantiles and apply the same mechanism. For agent k with context $\mathcal{C}_k$, define the local expected cross-entropy loss for classification: $L_k(\mathcal{C}_k) := -\mathbb{E}_{(x,y)\sim P_k}\big[\log f_\theta(x \mid \mathcal{C}_k)_y\big]$. While for regression, $L_k(\mathcal{C}_k) := \mathbb{E}_{(x,y)\sim P_k}\big[(f_\theta(x|\mathcal{C}_k) - y)^2\big]$. The global consensus objective is $\bar{L} := \frac{1}{K}\sum_{k=1}^K L_k(\mathcal{C}_k)$, subject to the hard constraint that agents communicate only output vectors (probabilities or scalars) and never elements of D_k. The communication graph $\mathcal{G} = (V, E)$ is a key design degree of freedom, where $V = \{1, ..., N\}$ and $E \subseteq V \times V$. Define a mixing matrix $W \in \mathbb{R}^{K \times K}$ such that 1) each element $w_{kj} > 0$ only if $j \in \mathcal{N}_k := \{j : (k, j) \in E\}$; 2) $W\mathbf{1} = \mathbf{1}$ (row-stochastic) 3) $\mathbf{1}^\top W = \mathbf{1}^\top$ (column stochastic); 4) $1 = \lambda_1(W) > \lambda_2(W) \geq ... \geq \lambda_K(W) > -1$. Define the spectral gap such that $\gamma := 1 - \lambda_2(W)$.

4 The D-ICL Algorithm

Initialization. At round $t = 0$, each agent k initializes its in-context dataset $\mathcal{C}_k^0$ as a *stratified* subsample of D_k of size n_0 (stratified by class label for classification; by quantile for regression). Stratification is critical for non-IID partitions as without it, an agent's context may be dominated by a single class or target range, causing the TFM to produce degenerate predictions for unseen outputs. We also denote by $n_{\max}$ and $s_{\max}$ the maximal size of the context and the effective number of informative pseudo-labels.

Formal Algorithm Description. Algorithm 1 presents the complete D-ICL protocol, applicable to both classification and regression. The four phases, including local inference, broadcast, consensus, and context enrichment, are executed synchronously each round. We assume honest-but-curious agents such that every agent executes Algorithm 1 faithfully but may attempt to infer information

from what it legitimately observes. Under this model, agent k's neighbors $\mathcal{N}_k$ directly observe its pre-consensus prediction $\hat{y}_k^t(x)$ for every $x \in \mathcal{Q} \cup X_\bullet$ (Phase 2), before aggregation. The query pool $\mathcal{Q}$ in our experiments is sampled from the pooled real training distribution, not synthetic data. What is exposed is the frozen TFM's prediction conditioned on agent k's private context, which is not a true label and not a direct assertion of shard membership, but this is nonetheless a real exposure channel we do not currently mitigate with formal guarantees.

Algorithm 1: D-ICL: Decentralized In-Context Learning (Classification & Regression)

Input: K agents with private shards $\{D_k\}$; frozen TFM f_θ; query pool $\mathcal{Q}$; test set $X_\bullet$; graph $\mathcal{G} = (V, E)$; rounds T; threshold τ; capacities n_0, $n_{\max}$, $s_{\max}$.

Initialize: $\mathcal{C}_k^0 \leftarrow \text{StratifiedSample}(D_k, n_0)$ for $k = 1, \ldots, K$

for $t = 0, 1, \ldots, T-1$ **do**

`// Phase 1: Local ICL Inference (no weight update)`

  forall k, $x \in \mathcal{Q} \cup X_\bullet$ **do**

    $\hat{y}_k^t(x) \leftarrow f_\theta(x \mid \mathcal{C}_k^t)$ *(probability vector for clf; scalar for reg)*

  end

`// Phase 2: Broadcast`

  Agent k sends $\{\hat{y}_k^t(x)\}_{x \in \mathcal{Q} \cup X_\bullet}$ to neighbors $\mathcal{N}_k$

`// Phase 3: Consensus`

  forall k **do**

    $\bar{y}_k^t(x) \leftarrow \text{AGGREGATE}(\{\hat{y}_j^t(x) : j \in \mathcal{N}_k\}) \quad \forall x$

  end

`// Phase 4: Context Enrichment`

  forall k **do**

    Classification: $S_k^t \leftarrow \{x \in \mathcal{Q} : \max_c \bar{y}_k^t(x)_c \geq \tau\}$

    Regression: $S_k^t \leftarrow \{x \in \mathcal{Q} : \text{std}_j(\hat{y}_j^t(x)) \leq \text{median}_x(\text{std}_j)\}$

    $\hat{S}_k^t \leftarrow \text{UniformSample}(S_k^t, \min(|S_k^t|, s_{\max}))$

    $\mathcal{C}_k^{t+1} \leftarrow \mathcal{C}_k^t \cup \{(x, \arg\max_c \bar{y}_k^t(x)_c) : x \in \hat{S}_k^t\}$

    if $|\mathcal{C}_k^{t+1}| > n_{\max}$ **then** $\mathcal{C}_k^{t+1} \leftarrow \text{ReservoirSample}(\mathcal{C}_k^{t+1}, n_{\max})$

  end

end

Output: $\bar{y}^T(x)$ for each $x \in X_\bullet$.

Phase 1: Local ICL Inference. The TFM performs a single frozen forward pass, where parameters θ are unchanged from pre-training. For example, if we use TABICL, the column-then-row attention mechanism encodes the context $\mathcal{C}_k^t$ into KV projections during in-context learning, which are then reused across all query calls within the same round via KV caching. This amortizes the context encoding cost. Only the query side x is processed in each inference call. For other popular TFMs, e.g., TABPFN, no explicit KV caching is available in the public API, but the model similarly performs a single forward pass over the concatenated context and query tokens. The computational cost of this phase is determined by the transformer attention complexity over the context-plus-query sequence, not by the number of model parameters, which enables effective cost saving.

Phase 2: Broadcast. Agent k transmits prediction vectors to each neighbor $j \in \mathcal{N}_k$. For classification the payload per query point is C floats (the probability vector); for regression it is one float (the scalar prediction). The total bandwidth per edge per round is $(M + |X_\bullet|) \times C$ floats for classification and $(M + |X_\bullet|)$ floats for regression. For $C = 10$, $M = 200$, and $|X_\bullet| = 200$, this amounts to 4 000 floats (16 KB) per edge per round for classification, which approximately implies 10^4 times less than transmitting the gradient of a 90M-parameter TFM in FP32. Crucially, the payload size is independent of $|\theta|$ and of the feature dimension d. **Note on broadcasting $\mathcal{Q}$ and $X_\bullet$.** Agents run inference on both the query pool $\mathcal{Q}$ and the test set $X_\bullet$ in the same forward pass and broadcast predictions for both. This is purely for computational efficiency (one forward pass per point). Only predictions on $\mathcal{Q}$ are used for consensus and pseudo-label enrichment; predictions on $X_\bullet$ are collected solely for final evaluation and never influence learning. The test set $X_\bullet$ is held out from all private shards to ensure unbiased benchmarking.

Phase 3: Consensus. Each agent independently aggregates predictions (from query pool $\mathcal{Q}$) from its neighborhood. We adopt the typical aggregation rule for the communication protocol (the AGGREGATE() function): $\bar{p}_k^t = \sum_{j \in \mathcal{N}_k} w_{kj}\, p_j^t, \sum_j w_{kj} = 1,\ w_{kj} \geq 0$, which is known for the weighted consensus in traditional decentralized learning algorithms (Jiang et al., 2017; 2015; Saadati et al., 2024; Kong et al., 2021). Its protocol preserves $\bar{p}_k^t \in \Delta^{C-1}$ for classification, while for regression, the weighted mean of scalar predictions is used correspondingly. In this work, the weighted mean is the default protocol for our theoretical guarantees analysis, the other two (arithmetic and geometric) will empirically be investigated. Please note that in the latter analysis, for the sake of convenience, we rewrite $\bar{p}_k^t$ as $\sum_{j=1}^K w_{kj}\, p_j^t$.

Phase 4: Context Enrichment. For classification, a query point x is pseudo-labeled if its maximum consensus probability exceeds the threshold τ. For regression, we use inter-agent prediction variance as a confidence proxy so that query points whose standard deviation across agents falls below the median standard deviation over all query points are selected. This is motivated by the observation that low inter-agent disagreement corresponds to high collective certainty, analogous to the committee disagreement measure in query-by-committee active learning (Seung et al., 1992). Reservoir sampling (Vitter, 1985) enforces the context cap $n_{\max}$, maintaining an exactly unbiased uniform sample of all pseudo-labeled examples. For classification, the context is enriched with the hard label $\arg\max_c \bar{p}_k^t(x)_c$, consistent with how the TFM backbones expect discrete class labels as in-context examples, not the soft probability vector.

5 Theoretical Analysis

In this section, we study how multiple agents enable variance reduction in the in-context learning setting, how context improvement monotonically facilitates performance enhancement, and what the convergence rate is and whether it retains the similar one as in traditional decentralized optimization. We provide the discussion of assumptions for each theoretical statement in Section C.1.

Proposition 1. *Let $Z_k := p_k^t(x)_c$ denote the prediction of agent $k \in \{1, \ldots, K\}$ on coordinate c for query x, and let the consensus estimate be $\bar{Z}_k = \sum_{j=1}^K w_{kj} Z_j$. Assume that for all k, $\mathrm{Var}(Z_k) \leq \sigma^2$, and for all $i \neq j$, $|\mathrm{Cov}(Z_i, Z_j)| \leq \sigma_{\mathrm{het}}^2$, where $\sigma, \sigma_{het} > 0$ are the self-variance and cross-variance. Then the average variance of the consensus estimates satisfies $\frac{1}{K}\sum_{k=1}^K \mathrm{Var}[\bar{Z}_k] \leq \frac{\sigma^2}{K_{\mathit{eff}}} + \sigma_{\mathrm{het}}^2\big(1 - \frac{1}{K_{\mathit{eff}}}\big)$, where $K_{\mathit{eff}} := \frac{1}{\|W\|_F^2}$, $\|W\|_F^2 := \frac{1}{K}\sum_{k=1}^K \sum_{j=1}^K w_{kj}^2$.*

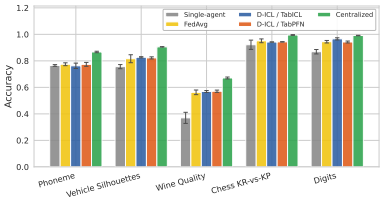
Impact of topology and interpretation. Proposition 1 shows that variance reduction in D-ICL is governed by the Frobenius norm $\|W\|_F^2$, which reflects how uniformly agents aggregate neighbor information. Fully connected graphs with uniform weights achieve optimal $O(\sigma^2/K)$ reduction, while sparse topologies (e.g., ring or star) yield $\|W\|_F^2 = \Theta(1)$ and thus limited gains independent of K. Random graphs interpolate between these regimes based on connectivity. Intuitively, decentralized aggregation acts as prediction-space ensembling, and averaging cancels independent noise but cannot remove correlated errors induced by data heterogeneity. As connectivity increases, variance reduction improves but correlation effects become more pronounced, leading to a fundamental trade-off between noise reduction and shared bias. In the IID case, this reduces to the classical $1/K_{\mathrm{eff}}$ variance scaling, presented as follows.

Corollary 1.1. *When the K shards are IID, predictions Z_k are independent, so $\sigma_{\mathrm{het}}^2 = 0$ and $\mathrm{Var}[\bar{p}_c^t] \leq \sigma^2/K_{\mathit{eff}}$. Hence, the variance is reduced by exactly K_{eff} relative to any single agent.*

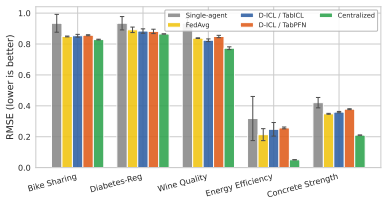
Proposition 2. *Let $L_k(\mathcal{C})$ denote the expected cross-entropy of agent k with context $\mathcal{C}$ on the marginal P_k. Assuming (i) TFM calibration: $P(\hat{y}^* = y^* | \kappa) \approx \kappa$ for consensus confidence κ, and (ii) approximately symmetric loss changes: $|\varepsilon_+| \approx |\varepsilon_-| \approx \delta$ for some $\delta > 0$, where $\varepsilon_+, \varepsilon_-$ are the loss changes from correct and incorrect pseudo-labels respectively. Then for a pseudo-labeled point $(x^*, \hat{y}^*)$ with confidence $\kappa = \max_c \bar{p}^t(x^*)_c \geq \tau$, there exists a threshold $\tau^* \in (1/2,\, 1)$ such that $\varepsilon(\kappa) < 0$ for all $\kappa > \tau^*$, where $\varepsilon(\kappa) := \mathbb{E}_{x^*, \hat{y}^*}\big[L_k\big(\mathcal{C} \cup \{(x^*, \hat{y}^*)\}\big) - L_k(\mathcal{C})\big]$.*

This guarantee is conditional since it requires the TFM to be approximately calibrated ($P(\hat{y}^* = y^* | \kappa) \approx \kappa$) and loss changes to be approximately symmetric (Section C.3). These are standard conditions in the self-training literature but need not hold uniformly. We discuss this further below and revisit it empirically in Section 6. Proposition 2 shows that under a sufficiently high confidence

threshold, iterative pseudo-labeling monotonically improves expected loss, enabling genuine learning without gradient updates. Consensus aggregation reduces variance and increases confidence, producing reliable pseudo-labels that expand local contexts, reduce bias, and improve generalization. The confidence threshold trades off label quality against data augmentation, which is high enough to ensure correctness, but not so restrictive as to limit learning. Thus D-ICL operates as decentralized self-training, where learning emerges from agent interaction rather than explicit optimization. A key insight: better consensus yields better pseudo-labels, which further improve predictions. However, in heterogeneous settings, biased consensus can propagate errors. Next, we analyze convergence.



(a) Classification performance comparison among different methods for different data under IID conditions with 4 agents



(b) Final RMSE results on regression task among different methods for different data under IID scenario with 4 agents

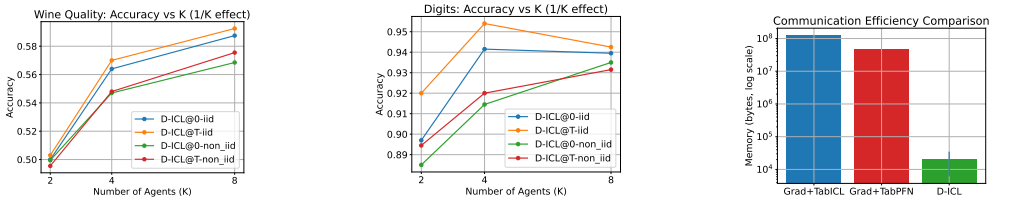
Figure 2: Classification and regression performance.

Proposition 3. *Let K agents communicate over a connected graph with symmetric doubly stochastic mixing matrix $W \in \mathbb{R}^{K\times K}$ having spectral gap $\gamma = 1 - \lambda_2 > 0$. Each agent k maintains a context set $\mathcal{C}_k^t$ of size at most $n_{\max}$. Let $L(p)$ be a loss function (e.g., cross-entropy) and L^* the optimal loss. Define $\bar{L}^t = \frac{1}{K}\sum_{k=1}^{K} L(\bar{y}_k^t)$, where $\bar{y}_k^t$ is agent k's consensus prediction after averaging. Suppose: (1) **Expected improvement from high-confidence pseudo-labels:** Based on Proposition 2, when adding a set $\mathcal{S}$ of size s high-confidence pseudo-labeled points (with $s \le s_{\max}$), we have the linear approximation $\mathbb{E}[L_k(\mathcal{C}\cup\mathcal{S})] \le (1-\rho)L_k(\mathcal{C}) + \rho\sigma_{het}^2$, where $\rho = \alpha s_{\max}/n_{\max}$ for some $\alpha \in (0,1]$, and σ_{het}^2 measures data heterogeneity. (2) **Lipschitz continuity of loss:** There exists $M > 0$ such that $|L(p) - L(q)| \le M\|p-q\|$ for all predictions p, q. Then after T rounds of Algorithm 1 (D-ICL), the expected average suboptimality satisfies $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\bar{L}^t - L^*\right] \le \frac{C_1}{T} + C_2\sigma_{het}$ with constants $C_1 = \frac{\bar{L}^0 - L^*}{\rho}$, $C_2 = \frac{2M(2-\rho)}{\rho\sqrt{\gamma}} + \sigma_{het}$.*

Proposition 3 yields an $\mathcal{O}(1/T)$ which is faster than $\mathcal{O}(1/\sqrt{T})$ under the assumption that high-confidence pseudo-labeling induces a consistent per-round contraction, which holds because consensus aggregation reduces variance while iterative enrichment steadily improves context quality. This is an idealized but meaningful characterization of prediction-space collaboration, governed by three factors: the contraction rate ρ (set by τ) the mixing speed τ (set by topology), and heterogeneity (σ_{het}). These expose a trade-off: higher τ improves pseudo-label quality but reduces coverage, while limited connectivity or heterogeneity slows convergence and raises the residual error. In short, D-ICL trades parameter-space optimization for prediction-space aggregation and context evolution as consensus reduces variance, pseudo-labeling reduces bias, and heterogeneity bounds what is achievable. Agents exchange predictive summaries rather than raw data, gradients, or parameters, narrowing direct exposure relative to gradient sharing. But a neighbor still observes an agent's raw pre-consensus predictions on the shared query pool (Algorithm 1, Phase 2), which is a real channel we do not currently mitigate. Topology bounds this exposure directly with sparser graphs (ring, sparse random) letting fewer parties observe any given agent's raw predictions than a fully-connected graph, at the cost of slower mixing (Proposition 3). This is a trade-off along with formal differential-privacy guarantees for prediction-space communication, we leave to future work.

6 Numerical Results

Datasets and Backbones. To validate our proposed method, we evaluate on standard benchmarks across domains, including classification datasets (`Phoneme`, `Vehicle Silhouettes`, `Wine Quality Red`, `Chess King-Rook vs King-Pawn endgame`, `NIST Handwritten Digits`) and regression datasets (`Bike Sharing`, `Diabetes`, `Wine Quality`, `Energy Efficiency`, `Concrete Strength`). Please see Section E for



(a) `Wine Quality` (b) `Digits` (c) Communication efficiency

Figure 3: K vs. accuracy and communication efficiency.

more detail. We use two tabular foundation models: TabICL and TabPFN, both kept fully frozen. TabICL uses 16 ensemble members, batch size 8, KV caching, logit averaging, and temperature 0.9 (with TabICLRegressor for regression). TabPFN uses the default Apache-2.0 v2 configuration (Zhang et al., 2025) (classifier and regressor variants). Key parameters include $n_0 = 64, n_{\max} = 1024, s_{\max} = 32$. Please check Section D for the full list of hyperparameters. For reference, TabICL and TabPFN have approximately 30M and 11M parameters respectively (derived from the FP32 gradient payload sizes reported below: 120MB and 44MB).

Baselines. We evaluate three main baselines. Single-agent (SA): Agent 0 queries the TFM using its initial context $\mathcal{C}_0^0$ (size m_0) with no communication, representing an isolated institution. Centralized: A single TFM uses a reservoir-sampled context from the full training set (size $\min(m_{\max}, N_{\text{train}})$), serving as a privacy-violating upper bound. D-ICL: Our method with $K \in \{2, 4, 8\}$ agents, $T = 5$ rounds, and default fully connected topology. Classical federated methods (e.g., FedAvg) are not directly comparable since they require parameter updates, whereas D-ICL uses frozen models. We therefore include **FedAvg+TFM**, a flat ensemble of initial predictions without iterative enrichment. All results are averaged over three random seeds. We emphasize that FedAvg+TFM is itself a gradient-free, prediction-space baseline using the same frozen backbone as D-ICL. It isolates the value of iterative consensus-driven context enrichment specifically, not a comparison against a stronger training-based method.

Table 1: Ablation on communication topology with 4 agents, IID setting (`Vehicle`).

BB	Topology	D-ICL
TabICL	fully_connected	**0.748**
TabPFN	fully_connected	**0.781**
TabICL	ring	0.646
TabPFN	ring	0.748
TabICL	sparse_random	0.742
TabPFN	sparse_random	0.760

Classification Results. Figures 2a shows the classification performance under the IID setting across three datasets. D-ICL consistently outperforms isolated agents across all datasets and partition settings (e.g., 0.75 vs. 0.83 on `Vehicle Silhouettes`; 0.87 vs. 0.97 on `Digits`), confirming that consensus-based pseudo-label sharing effectively expands each agent's context and reduces variance. FedAvg (flat ensemble) already improves over SA by averaging predictions across agents, confirming that aggregation alone reduces variance. D-ICL further exceeds FedAvg, demonstrating that iterative pseudo-label enrichment (in Proposition 2) yields monotonic improvement beyond simple averaging. At the same time, D-ICL closely approaches the centralized oracle, demonstrating strong collaboration without data sharing. Overall, D-ICL bridges the gap between SA and the centralized oracle. In Section E, we present extensive results on more datasets and non-IID settings with various numbers of agents, which also consistently validate our findings.

Regression Results. Figure 2b compares final RMSE under IID settings to three baselines. Similar conclusions have been drawn here as in classification results. Both FedAvg and D-ICL consistently improve over single-agent baselines, showing that aggregating predictive distributions enhances accuracy without data sharing, with larger gains on data-scarce or harder tasks (e.g., `Bike Sharing`, `Concrete`). While not fully matching the centralized oracle, D-ICL substantially narrows the gap, reflecting effective collaboration under decentralized constraints. Notably, gains are smaller on simpler datasets where single-agent performance is already strong. Overall, the results support the variance-reduction effect of consensus aggregation and demonstrate that D-ICL achieves stable performance relatively close to centralized models across different cases. More thorough results over different datasets and non-IID settings are included in Section E.

Topology. Table 1 shows that communication topology critically affects D-ICL performance. Fully connected graphs consistently achieve the best or near-best results, as they enable global averaging

each round and maximize variance reduction implied by Proposition 3. In contrast, sparse topologies (e.g., ring) restrict information flow to local neighborhoods, slowing propagation, delaying consensus, and degrading pseudo-label quality in early rounds. Interestingly, random sparse graphs can almost match fully connected performance, suggesting a balance between connectivity and information diversity. Overall, D-ICL is robust to topology choice, but denser connectivity significantly accelerates convergence, consistent with the theory of consensus-driven variance reduction.

Threshold τ. The temperature parameter τ, which controls the confidence threshold for pseudo-label selection, induces a clear quality–coverage trade-off. As shown in Table 2, intermediate values (around 0.7–0.8) achieve the best performance, while both extremes degrade results. Small τ admits low-confidence pseudo-labels, introducing noise that propagates across agents and can violate the conditions of Proposition 2, leading to error amplification. In contrast, large τ overly restricts selection to a few high-confidence samples, reducing data augmentation and slowing convergence, effectively behaving like a near-static ensemble.

Table 2: Ablation on confidence threshold τ with TabPFN, 4 agents, IID setting (`Vehicle`).

τ	D-ICL@0	D-ICL@T
0.6	0.614	0.633
0.7	0.722	0.724
0.8	**0.799**	**0.796**
0.9	0.770	0.770

Agent count K. The number of agents K directly affects aggregation strength in D-ICL. As shown in Table 3, performance improves consistently with larger K for both backbones, but with diminishing gains beyond a moderate scale. This matches Proposition 1, where aggregation variance decreases at rate $1/K$, as increasing K effectively enlarges the ensemble and stabilizes consensus predictions. However, due to the heterogeneity term, residual error persists and eventually dominates, leading to performance saturation once variance is sufficiently reduced. Figure 3 further supports this result by confirming variance reduction with increasing K, and also validates Corollary 1.1, showing that IID settings achieve lower variance due to the absence of σ^2_{het}.

Communication Efficiency and the ICL Advantage. The communication advantage of D-ICL over gradient-based FL is substantial and grows with model size as shown in Figure 3c. For TABICL with approximately 30 M parameters, gradient transmission in FP32 costs 120 MB per edge per round and TABPFN 44 MB. D-ICL transmits $C \times (M + |X_\bullet|)$ floats per edge per round: for `Digits`, $C = 10$ classes and 855 total query and test points, this is 8 550 floats or ∼34 KB, which is a reduction by a factor of approximately 3509. For regression with scalar outputs, the reduction is even larger: 3.4 KB per edge per round. A significant reduction on communication bandwidth by 3–4 orders of magnitude applies analogously to other datasets. Moreover, D-ICL is a zero-gradient protocol: each agent maintains only its context of size at most $m_{\max} = 1024$ samples, rather than gradient buffers and optimizer states. This makes D-ICL feasible on devices with limited memory and bandwidth.

Table 3: Ablation on number of agents K with IID setting (`Vehicle`).

BB	K	D-ICL@0	D-ICL@T
TabICL	2	0.733	0.704
TabPFN	2	0.705	0.727
TabICL	4	0.770	0.774
TabPFN	4	0.791	0.782
TabICL	8	**0.802**	**0.793**
TabPFN	8	0.785	0.785
TabICL	16	0.788	0.776
TabPFN	16	0.793	0.783

Self-training vs. D-ICL: isolating the consensus contribution. Figure 4 and Tables 13–14 (in Section E) ablate the consensus contribution. A single agent iteratively pseudo-labels its own predictions under identical query pool, threshold, budget, and rounds as D-ICL, but without cross-agent aggregation. D-ICL outperforms self-training on 4/5 classification datasets, with the gap scaling with difficulty, most notably on Wine Quality (Red), where self-training barely moves off round-0 ($0.358 \rightarrow 0.377$ TabICL; $0.414 \rightarrow 0.393$ TabPFN, a slight regression) while D-ICL reaches $0.544 - 0.552$ (Table 13). This supports Proposition 2. Without consensus-driven variance reduction, a single agent's predictions are not reliable enough as pseudo-labels and iterating on them compounds rather than corrects error. Phoneme is the exception (self-training matches D-ICL), consistent with this low-dimensional, near-linearly-separable task already saturating from a single agent's context. Regression (Table 14) mirrors this pattern, with one instructive instability. Energy Efficiency's self-training splits across seeds (Table 15) as two converge to RMSE $\sim 0.11 - 0.12$ (matching D-ICL), and one collapses to 0.61 and never recovers. Since self-training has no mech-

anism to correct an unlucky initial context, unlike D-ICL's cross-agent averaging (Proposition 1), this instability itself supports the paper's central claim rather than being a result to average away.

Membership-inference attack: quantifying privacy amplification. The theoretical privacy-amplification argument (Section E) predicts reduced exposure but does not itself establish how much protection consensus actually provides against a concrete adversary. We therefore evaluate a standard membership-inference attack to quantify this empirically. Figure 5 and Table 16 report the results. Aggregating to the consensus output reduces attack AUC below the isolated baseline in 9 of 10 conditions often sharply. Wine Quality (Red)/TabICL drops from 0.982 ± 0.007 (near-perfect membership inference against an isolated agent) to 0.516 ± 0.002 (indistinguishable from random guessing); Digits/TabICL shows a similar collapse ($0.948 \rightarrow 0.450$). Raw broadcast is identical to isolated exposure in every row of Table 16 by construction, since a neighbor observing agent k's Phase-2 broadcast sees exactly the same prediction an isolated agent would produce, confirming the privacy benefit comes entirely from consensus, not from the broadcast mechanism itself. One exception, Chess KR-vs-KP/TabPFN, shows consensus AUC (0.604) nominally above isolated (0.597), but the two are within one combined standard deviation, so we attribute this to sampling noise rather than a genuine reversal. Attack implementation details are in Section E.

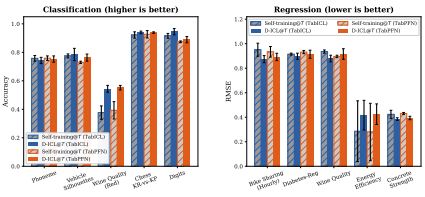


Figure 4: Self-training (no consensus) vs. D-ICL@T, classification (left) and regression (right), mean $\pm$ std over 3 seeds. Full per-dataset numbers in Table 13 and Table 14 in Section E.

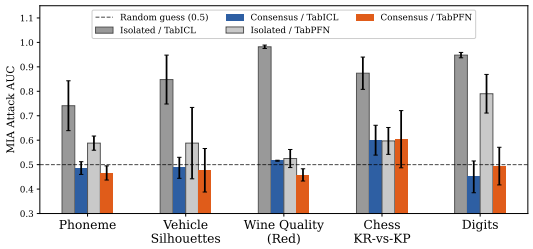


Figure 5: Membership-inference attack AUC: Isolated vs. Consensus, both backbones, mean $\pm$ std over 3 seeds. Raw Broadcast is identical to Isolated by construction and is omitted from the figure; full numbers in Table 16.

Limitations. While D-ICL is rigorous and validated, it may amplify errors under severe non-IID settings. Biased consensus can introduce incorrect pseudo-labels that propagate across rounds. It also relies on reasonably calibrated models and near-symmetric loss. Miscalibration or class imbalance can degrade performance. The convergence analysis assumes a constant contraction rate, whereas in practice gains diminish as contexts grow, leading to slower sublinear behavior. Our theory further uses worst-case bounds and does not fully capture data-dependent effects or the impact of reservoir sampling. Empirically, evaluation is limited to small- and medium-scale tabular benchmarks and excludes stronger baselines such as federated or advanced semi-supervised methods. Extending both theoretical analysis and large-scale empirical validation remains important future work.

7 Conclusion

We presented D-ICL, a decentralized framework where frozen tabular foundation models enable collaborative learning without sharing data, gradients, or parameters. Agents exchange predictive distributions over a shared query set, aggregate via consensus, and iteratively enrich contexts with high-confidence pseudo-labels. Key results include variance reduction at rate $1/K$, monotonic improvement under a confidence threshold, and $O(1/T)$ convergence to a heterogeneity-limited residual. Experiments across ten benchmarks confirm D-ICL bridges the gap between isolated agents and a centralized oracle at orders-of-magnitude lower communication cost than federated learning. D-ICL shifts learning from parameter optimization to prediction-space collaboration, ideal for data-siloed settings. Performance remains limited by data heterogeneity, motivating future work on adaptive aggregation and pseudo-labeling. More broadly, foundation models can serve as building blocks for decentralized systems that learn through interaction.

AI USE STATEMENT

AI-based tools were used solely as writing and development aids during the preparation of this manuscript, including assistance with prose refinement, LaTeX formatting, and code review. All experiments, analyses, numerical results, and theoretical proofs were independently developed, implemented, executed, and verified by the authors. AI tools did not generate or alter experimental data or scientific claims. The authors are solely responsible for the accuracy, integrity, and content of the work.

ETHIC STATEMENT

This work uses only publicly available benchmark datasets (OpenML-CC18, UCI Machine Learning Repository, and standard sklearn datasets); no new data was collected, and no human subjects were involved. D-ICL is motivated by data privacy-constrained collaborative learning, but we are careful not to overclaim. As our membership-inference results show, isolated-agent predictions can leak substantial information (AUC up to 0.98 on some datasets), and while consensus aggregation measurably reduces this exposure, D-ICL provides no formal differential-privacy guarantee. Practitioners deploying this framework in genuinely sensitive settings (e.g., healthcare or financial consortia) should not treat it as a certified privacy mechanism and should layer additional safeguards (e.g., calibrated noise, secure aggregation). We do not foresee direct harms from this work as the primary risk is misplaced trust in an informal exposure-reduction claim as if it were a formal guarantee, which this statement and the paper's own limitations discussion are intended to preempt.

REPRODUCIBILITY STATEMENT

All experiments use publicly available datasets (OpenML-CC18, UCI Machine Learning Repository, and standard sklearn datasets). Every table can be regenerated from the archived evaluation records using the codebase accompanying the paper, and every figure can be regenerated from the archived calibration statistics. Hyperparameters are listed in Appendix D.

REFERENCES

Vahid Attari and Raymundo Arroyave. Decoding non-linearity and complexity: deep tabular learning approaches for materials science. *Digital Discovery*, 4(10):2765–2780, 2025.

Mingzhe Chen, Nir Shlezinger, H Vincent Poor, Yonina C Eldar, and Shuguang Cui. Communication-efficient federated learning. *Proceedings of the National Academy of Sciences*, 118(17):e2024789118, 2021.

Jillian M Clements, Di Xu, Nooshin Yousefi, and Dmitry Efimov. Sequential deep learning for credit risk monitoring with tabular financial data. *arXiv preprint arXiv:2012.15330*, 2020.

Huangliang Dai, Shixun Wu, Jiajun Huang, Zizhe Jian, Yue Zhu, Haiyang Hu, and Zizhong Chen. Ft-transformer: Resilient and reliable transformer with end-to-end fault tolerant attention. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1085–1098, 2025.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14(2):241–258, 2020.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmuller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data. *arXiv preprint arXiv:2506.16791*, 2025.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

Zhanhong Jiang, Soumik Sarkar, and Kushal Mukherjee. On distributed optimization using generalized gossip. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 2667–2672. IEEE, 2015.

Zhanhong Jiang, Aditya Balu, Chinmay Hegde, and Soumik Sarkar. Collaborative deep learning in fixed topology networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Lingjing Kong, Tao Lin, Anastasia Koloskova, Martin Jaggi, and Sebastian Stich. Consensus control for decentralized deep learning. In *International Conference on Machine Learning*, pp. 5686–5696. PMLR, 2021.

Anantaa Kotal, Aritran Piplai, Sai Sree Laya Chukkapalli, and Anupam Joshi. Privetab: Secure and privacy-preserving sharing of tabular data. In *Proceedings of the 2022 ACM on international workshop on security and privacy analytics*, pp. 35–45, 2022.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International conference on machine learning*, pp. 5468–5479. PMLR, 2020.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020.

Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, Huai-Hong Yin, Tao Zhou, Jun-Peng Jiang, and Han-Jia Ye. Talent: A tabular analytics and learning toolbox. *Journal of Machine Learning Research*, 26 (226):1–16, 2025.

Zili Lu, Heng Pan, Yueyue Dai, Xueming Si, and Yan Zhang. Federated learning with non-iid data: A survey. *IEEE Internet of Things Journal*, 11(11):19188–19209, 2024.

Amirhossein Malekijoo, Mohammad Javad Fadaeieslam, Hanieh Malekijou, Morteza Homayounfar, Farshid Alizadeh-Shabdiz, and Reza Rawassizadeh. Fedzip: A compression framework for communication-efficient federated learning. *arXiv preprint arXiv:2102.01593*, 2021.

Mohamad Mansouri, Melek Önen, and Wafa Ben Jaballah. Learning from failures: Secure and fault-tolerant aggregation for federated learning. In *Proceedings of the 38th annual computer security applications conference*, pp. 146–158, 2022.

Alireza Hossein Zadeh Nik, Michael A Riegler, Pål Halvorsen, and Andrea M Storås. Generation of synthetic tabular healthcare data using generative adversarial networks. In *International Conference on Multimedia Modeling*, pp. 434–446. Springer, 2023.

Braxton Osting and Todd Harry Reeb. Consistency of dirichlet partitions. *SIAM Journal on Mathematical Analysis*, 49(5):4251–4274, 2017.

Jingang Qu, David HolzmAvzller, GaAFl Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. *arXiv preprint arXiv:2502.05564*, 2025.

Jingang Qu, David Holzmuller, Gael Varoquaux, and Marine Le Morvan. Tabiclv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Nastaran Saadati, Minh Pham, Nasla Saleem, Joshua R Waite, Aditya Balu, Zhanong Jiang, Chinmay Hegde, and Soumik Sarkar. Dimat: Decentralized iterative merging-and-training for deep learning models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27517–27527, 2024.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294, 1992.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022.

Boris Van Breugel and Mihaela Van Der Schaar. Why tabular foundation models should be a research priority. *arXiv preprint arXiv:2405.01147*, 2024.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1):37–57, 1985.

Wenqi Wei, Ling Liu, Margaret Loper, Ka-Ho Chow, Mehmet Emre Gursoy, Stacey Truex, and Yanzhao Wu. A framework for evaluating gradient leakage attacks in federated learning. *arXiv preprint arXiv:2004.10397*, 2020.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, and Radha Poovendran. Brave: Byzantine-resilient and privacy-preserving peer-to-peer federated learning. *arXiv preprint arXiv:2401.05562*, 2024.

Haomiao Yang, Mengyu Ge, Dongyun Xue, Kunlan Xiang, Hongwei Li, and Rongxing Lu. Gradient leakage attacks in federated learning: Research frontiers, taxonomy, and future directions. *IEEE Network*, 38(2):247–254, 2023a.

Yongquan Yang, Haijun Lv, and Ning Chen. A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review*, 56(6):5545–5589, 2023b.

Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A closer look at tabpfn v2: Understanding its strengths and extending its capabilities. *arXiv preprint arXiv:2502.17361*, 2025.

Liangqi Yuan, Ziran Wang, Lichao Sun, Philip S Yu, and Christopher G Brinton. Decentralized federated learning: A survey and perspective. *IEEE Internet of Things Journal*, 11(21):34617–34638, 2024.

Xingxuan Zhang, Gang Ren, Han Yu, Hao Yuan, Hui Wang, Jiansheng Li, Jiayun Wu, Lang Mo, Li Mao, Mingchao Hao, et al. Limix: Unleashing structured-data modeling capability for generalist intelligence. *arXiv preprint arXiv:2509.03505*, 2025.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

A METHODOLOGICAL COMPARISON

In this section, we present a compact table to show the difference among federated learning, decentralized learning, and D-ICL in multiple aspects.

Table 4: Comparison of different paradigms (Var. red.: variance reduction; con.upd: context update)

Property	**Decentralized Learning**	**Federated Learning**	**D-ICL (Ours)**
Model/params sharing	✓(peer-to-peer)	✓(via server)	✗
Communication	Model updates/gradients	Model updates/gradients	Predictive distributions
Model type	Trainable local models	Shared global model	Frozen foundation model
Personalization	Limited	Moderate (fine-tuning)	High (context-specific)
Handling non-IID data	Challenging	Challenging	Robust
Theoretical guarantees	Convergence	Convergence	Var. red. + convergence
Computation per agent	High (training)	High (training)	Low (inference + con. upd.)

B MORE DETAILS ON COMMUNICATION TOPOLOGIES

In this work, we focus primarily on three different topologies in this work.

- **Fully-connected (FC):** $E = \{(i, j) : i \neq j\}$. Every agent aggregates all K predictions, so it leads to $O(K^2)$ communication edges. This also marks maximum information exchange.
- **Ring:** Agent k communicates with $k-1 \pmod K$ and $k+1 \pmod K$ plus itself, having degree 3 (or 2 for $K = 2$). This results in total bandwidth $O(K)$. In practice, ring topology can be used to model wireless sensor networks with physical locality constraints.
- **Sparse Erdős–Rényi:** Each undirected edge (i, j), $i < j$, exists independently with a probability p. All self-loops are added, but isolated nodes are connected to agent 0. The graph has an expected degree $1 + (K-1)p$, particularly modeling ad-hoc peer-to-peer networks.

C MISSING PROOF AND ADDITIONAL ANALYSIS

C.1 DISCUSSION OF ASSUMPTIONS

Our theoretical results rely on assumptions that are standard in ensemble learning and semi-supervised learning, but may not hold uniformly across all practical settings. In particular, Proposition 2 assumes that high-confidence pseudo-labels are sufficiently accurate to yield expected improvement, which implicitly requires a degree of calibration in the underlying tabular foundation model. Similarly, Proposition 3 assumes a constant expected contraction rate induced by pseudo-labeling, which may weaken as the context grows and marginal gains diminish. Therefore, our theoretical guarantees should be interpreted as characterizing the regime in which D-ICL is most effective, rather than as worst-case guarantees. Empirically, we observe that these assumptions are approximately satisfied across a wide range of datasets (see Section 6), supporting their practical relevance.

C.2 PROOF OF PROPOSITION 1

Proof. We proceed in several steps.

We first decompose the variance. Fix an agent k. Since $\bar{Z}_k = \sum_{j=1}^{K} w_{kj} Z_j$, the variance of a weighted sum gives

$$\mathrm{Var}[\bar{Z}_k] = \sum_{i=1}^{K} w_{ki}^2 \mathrm{Var}(Z_i) + \sum_{i \neq j} w_{ki} w_{kj} \mathrm{Cov}(Z_i, Z_j). \tag{4}$$

We refer to the first term as the *self-variance* and the second as the *cross-covariance*. Taking average over k yields

$$\frac{1}{K} \sum_{k=1}^{K} \mathrm{Var}[\bar{Z}_k] = (A) + (B), \tag{5}$$

where

$$(A) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{K} w_{ki}^2 \mathrm{Var}(Z_i), \tag{6}$$

$$(B) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i\neq j} w_{ki} w_{kj} \mathrm{Cov}(Z_i, Z_j). \tag{7}$$

We then bound the self-variance term. Using $\mathrm{Var}(Z_i) \leq \sigma^2$, we obtain

$$(A) \leq \frac{1}{K}\sum_{k,i} w_{ki}^2 \sigma^2 = \sigma^2 \cdot \frac{1}{K}\sum_{k,i} w_{ki}^2 = \sigma^2 \|W\|_F^2. \tag{8}$$

Reordering the summation in (B) gives

$$(B) = \frac{1}{K}\sum_{i\neq j}\left(\sum_{k=1}^{K} w_{ki} w_{kj}\right)\mathrm{Cov}(Z_i, Z_j). \tag{9}$$

Define the interaction weights such that

$$\mathcal{A}_{ij} := \sum_{k=1}^{K} w_{ki} w_{kj}.$$

Then we obtain

$$(B) = \frac{1}{K}\sum_{i\neq j} \mathcal{A}_{ij} \mathrm{Cov}(Z_i, Z_j).$$

We now bound the cross-covariance term. Using $|\mathrm{Cov}(Z_i, Z_j)| \leq \sigma_{\mathrm{cov}}^2$, we obtain

$$(B) \leq \sigma_{\mathrm{cov}}^2 \cdot \frac{1}{K}\sum_{i\neq j} \mathcal{A}_{ij}. \tag{10}$$

We simplify the weight interaction by computing

$$\sum_{i,j} \mathcal{A}_{ij} = \sum_{i,j}\sum_{k=1}^{K} w_{ki} w_{kj} = \sum_{k=1}^{K}\left(\sum_{i=1}^{K} w_{ki}\right)^2. \tag{11}$$

Since W is row-stochastic, $\sum_i w_{ki} = 1$, hence

$$\sum_{i,j} \mathcal{A}_{ij} = K.$$

Separating diagonal and off-diagonal terms yields:

$$\sum_{i\neq j} \mathcal{A}_{ij} = K - \sum_{i=1}^{K}\sum_{k=1}^{K} w_{ki}^2.$$

Dividing by K, we obtain

$$\frac{1}{K}\sum_{i\neq j} \mathcal{A}_{ij} = 1 - \|W\|_F^2.$$

Substituting the above back into (B) gives

$$(B) \leq \sigma_{\mathrm{cov}}^2\big(1 - \|W\|_F^2\big).$$

Combining the bounds for (A) and (B) and the definition of K_{eff}, we conclude

$$\frac{1}{K}\sum_{k=1}^{K} \mathrm{Var}[\bar{Z}_k] \;\leq\; \frac{\sigma^2}{K_{\mathrm{eff}}} + \sigma_{\mathrm{cov}}^2\big(1 - \frac{1}{K_{\mathrm{eff}}}\big),$$

which completes the proof. □

C.3 Proof of Proposition 2

Proof. Let $y^* = \arg\max_c P(y \mid x^*)$ be the Bayes-optimal label.

Decomposing by correctness, based on the above definition, we have

$$\varepsilon(\kappa) = P(\hat{y}^* = y^*) \cdot \varepsilon_+ + P(\hat{y}^* \neq y^*) \cdot \varepsilon_-,$$

where $\varepsilon_+ < 0$ when the pseudo-label is correct and $\varepsilon_- > 0$ when it is incorrect.

Under sufficient context size and approximate TFM calibration, $P(\hat{y}^* = y^* \mid \kappa) \approx \kappa$ (Qu et al., 2025; Hollmann et al., 2022). Let $\delta > 0$ be such that $|\varepsilon_+| \approx |\varepsilon_-| \approx \delta$ (symmetric loss changes). Then we can obtain:

$$\varepsilon(\kappa) = \kappa(-\delta) + (1-\kappa)\delta = \delta(1-2\kappa).$$

We then derive the threshold. The condition $\varepsilon(\kappa) < 0$ becomes $\delta(1-2\kappa) < 0 \Leftrightarrow \kappa > \frac{1}{2}$, giving $\tau^* = \frac{1}{2}$ under symmetric loss changes. In practice $|\varepsilon_-| > |\varepsilon_+|$ (incorrect pseudo-labels propagate errors through subsequent rounds), which shifts τ^* strictly above $\frac{1}{2}$, motivating our empirical choice $\tau = 0.80$. □

C.4 Proof of Proposition 3

Proof. We denote by $\mathcal{F}_t$ the history of all random events up to round t (including the context sets, predictions, and pseudo-labels). All expectations are conditional on $\mathcal{F}_t$ unless otherwise stated. Define the consensus error at round t as

$$\Phi_t = \frac{1}{K}\sum_{k=1}^{K} \|\bar{y}_k^t - \bar{y}^t\|^2, \qquad \bar{y}^t = \frac{1}{K}\sum_{k=1}^{K} \bar{y}_k^t.$$

Because W is doubly stochastic, $\bar{y}^t = \frac{1}{K}\sum_{k=1}^{K} \hat{y}_k^t$, where $\hat{y}_k^t$ are the local predictions before averaging. From the spectral gap property (see, e.g., [1]), we have

$$\Phi_t \le (1-\gamma)\frac{1}{K}\sum_{k=1}^{K} \|\hat{y}_k^t - \bar{y}^t\|^2.$$

The pseudo-labeling step can increase the disagreement by at most σ_{het}^2 (Assumption 1, second part), so

$$\frac{1}{K}\sum_{k=1}^{K} \|\hat{y}_k^{t+1} - \bar{y}^{t+1}\|^2 \le \Phi_t + \sigma_{\text{het}}^2.$$

Combining these two inequalities yields the deterministic recursion

$$\Phi_{t+1} \le (1-\gamma)(\Phi_t + \sigma_{\text{het}}^2).$$

Solving this recurrence gives

$$\Phi_t \le (1-\gamma)^t \Phi_0 + \frac{\sigma_{\text{het}}^2}{\gamma}.$$

Consequently,

$$\sqrt{\Phi_t} \le \sqrt{(1-\gamma)^t \Phi_0} + \frac{\sigma_{\text{het}}}{\sqrt{\gamma}}.$$

All quantities here are deterministic given the initial conditions and the graph structure. We then let $\hat{L}^{t+1} = \frac{1}{K}\sum_{k=1}^{K} L(\hat{y}_k^{t+1})$ be the average loss of the local predictions after updating contexts (before averaging). From the expected improvement guarantee in Assumption 1, each agent adds up to $\Delta_{\max}$ pseudo-labeled points, each satisfying $\mathbb{E}[L_k(\mathcal{C}_k^{t+1}) \mid \mathcal{F}_t] \le (1-\rho)L_k(\mathcal{C}_k^t) + \rho\sigma_{\text{het}}^2$ (the linear approximation). Averaging over k gives

$$\mathbb{E}[\hat{L}^{t+1} \mid \mathcal{F}_t] \le (1-\rho)\hat{L}^t + \rho\sigma_{\text{het}}^2. \tag{1}$$

We now relate consensus loss to local loss. The consensus loss $\bar{L}^{t+1}$ and the local loss $\hat{L}^{t+1}$ differ because the averaging step changes the predictions. By Lipschitz continuity (Assumption 2) and the bound on the distance between $\bar{y}_k^{t+1}$ and $\hat{y}_k^{t+1}$ (see the mixing property), we have

$$|\bar{L}^{t+1} - \hat{L}^{t+1}| \le M\sqrt{\frac{1}{K}\sum_{k=1}^{K} \|\bar{y}_k^{t+1} - \hat{y}_k^{t+1}\|^2} \le 2M\sqrt{\Phi_{t+1}}.$$

Thus, deterministically,

$$\bar{L}^{t+1} \le \hat{L}^{t+1} + 2M\sqrt{\Phi_{t+1}}, \qquad \hat{L}^t \le \bar{L}^t + 2M\sqrt{\Phi_t}. \tag{2}$$

Take conditional expectation in the first inequality of (2):

$$\mathbb{E}[\bar{L}^{t+1} \mid \mathcal{F}_t] \le \mathbb{E}[\hat{L}^{t+1} \mid \mathcal{F}_t] + 2M\sqrt{\Phi_{t+1}}.$$

Substituting the improvement bound (1) into the above relationship, we acquire:

$$\mathbb{E}[\bar{L}^{t+1} \mid \mathcal{F}_t] \le (1-\rho)\hat{L}^t + \rho\sigma_{\text{het}}^2 + 2M\sqrt{\Phi_{t+1}}.$$

Now we use the second inequality of (2) (which is deterministic, so the expectation does not affect the bound) to replace $\hat{L}^t$ by $\bar{L}^t + 2M\sqrt{\Phi_t}$ and obtain:

$$\mathbb{E}[\bar{L}^{t+1} \mid \mathcal{F}_t] \leq (1-\rho)(\bar{L}^t + 2M\sqrt{\Phi_t}) + \rho\sigma_{\text{het}}^2 + 2M\sqrt{\Phi_{t+1}}.$$

Hence, the following relationship can be obtained:

$$\mathbb{E}[\bar{L}^{t+1} \mid \mathcal{F}_t] \leq (1-\rho)\bar{L}^t + 2M(1-\rho)\sqrt{\Phi_t} + 2M\sqrt{\Phi_{t+1}} + \rho\sigma_{\text{het}}^2. \tag{3}$$

Define $\Delta_t = \bar{L}^t - L^*$. Subtract L^* from both sides of (3) and take total expectation (i.e., $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}[\cdot \mid \mathcal{F}_t]]$):

$$\mathbb{E}[\Delta_{t+1}] \leq (1-\rho)\mathbb{E}[\Delta_t] + 2M(1-\rho)\sqrt{\Phi_t} + 2M\sqrt{\Phi_{t+1}} + \rho\sigma_{\text{het}}^2.$$

Summing this inequality for $t = 0, 1, \ldots, T-1$ gives

$$\sum_{t=0}^{T-1} \mathbb{E}[\Delta_{t+1}] \leq (1-\rho)\sum_{t=0}^{T-1} \mathbb{E}[\Delta_t] + 2M(1-\rho)\sum_{t=0}^{T-1}\sqrt{\Phi_t} + 2M\sum_{t=0}^{T-1}\sqrt{\Phi_{t+1}} + \rho T\sigma_{\text{het}}^2.$$

Let $S_T = \sum_{t=0}^{T-1} \mathbb{E}[\Delta_t]$. Then $\sum_{t=0}^{T-1} \mathbb{E}[\Delta_{t+1}] = S_T - \mathbb{E}[\Delta_0] + \mathbb{E}[\Delta_T]$. Since $\Delta_T \geq 0$, we have

$$S_T - \mathbb{E}[\Delta_0] \leq (1-\rho)S_T + 2M(1-\rho)\sum_{t=0}^{T-1}\sqrt{\Phi_t} + 2M\sum_{t=0}^{T-1}\sqrt{\Phi_{t+1}} + \rho T\sigma_{\text{het}}^2.$$

Rearranging the above relationship yields,

$$\rho S_T \leq \mathbb{E}[\Delta_0] + 2M(1-\rho)\sum_{t=0}^{T-1}\sqrt{\Phi_t} + 2M\sum_{t=0}^{T-1}\sqrt{\Phi_{t+1}} + \rho T\sigma_{\text{het}}^2. \tag{4}$$

We now bound the sums of $\sqrt{\Phi_t}$. Based on the previous derivation, $\sqrt{\Phi_t} \leq \sqrt{(1-\gamma)^t\Phi_0} + \frac{\sigma_{\text{het}}}{\sqrt{\gamma}}$. Therefore

$$\sum_{t=0}^{T-1}\sqrt{\Phi_t} \leq \sqrt{\Phi_0}\sum_{t=0}^{T-1}(1-\gamma)^{t/2} + \frac{T\sigma_{\text{het}}}{\sqrt{\gamma}}.$$

The geometric series satisfies $\sum_{t=0}^{\infty}(1-\gamma)^{t/2} = \frac{1}{1-\sqrt{1-\gamma}} \leq \frac{2}{\gamma}$ for $\gamma \in (0, 1]$. Hence

$$\sum_{t=0}^{T-1}\sqrt{\Phi_t} \leq \frac{2\sqrt{\Phi_0}}{\gamma} + \frac{T\sigma_{\text{het}}}{\sqrt{\gamma}}.$$

The same bound holds for $\sum_{t=0}^{T-1}\sqrt{\Phi_{t+1}}$ (the shift does not affect the constant). Plugging these into (4) yields

$$\rho S_T \leq \mathbb{E}[\Delta_0] + 2M(1-\rho)\left(\frac{2\sqrt{\Phi_0}}{\gamma} + \frac{T\sigma_{\text{het}}}{\sqrt{\gamma}}\right) + 2M\left(\frac{2\sqrt{\Phi_0}}{\gamma} + \frac{T\sigma_{\text{het}}}{\sqrt{\gamma}}\right) + \rho T\sigma_{\text{het}}^2.$$

By grouping terms, we have

$$\rho S_T \leq \mathbb{E}[\Delta_0] + \frac{4M(2-\rho)\sqrt{\Phi_0}}{\gamma} + \frac{2M(2-\rho)T\sigma_{\text{het}}}{\sqrt{\gamma}} + \rho T\sigma_{\text{het}}^2.$$

Lastly, we divide both sides by ρT:

$$\frac{S_T}{T} \leq \frac{\mathbb{E}[\Delta_0]}{\rho T} + \frac{4M(2-\rho)\sqrt{\Phi_0}}{\rho\gamma T} + \frac{2M(2-\rho)\sigma_{\text{het}}}{\rho\sqrt{\gamma}} + \sigma_{\text{het}}^2.$$

For large T, the first two terms are $O(1/T)$ and the remaining terms are $O(\sigma_{\text{het}})$. Absorbing the constant factors, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\bar{L}^t - L^*] \leq \frac{C_1}{T} + C_2\sigma_{\text{het}},$$

with $C_1 = \frac{\bar{L}^0 - L^*}{\rho}$ and $C_2 = \frac{2M(2-\rho)}{\rho\sqrt{\gamma}} + \sigma_{\text{het}}$. This completes the proof. □

D Detail About Experimental Setup

Table 5 summarizes the detail for all benchmarking datasets we have used in the empirical implementation for the classification and regression tasks. Table 6 shows the default D-ICL hyperparameters.

Table 5: Dataset summary. CLF = classification; REG = regression. *: main datasets for regression

Dataset	Task	N	d	C/range	Source
Phoneme	CLF-binary	5405	5	2	OpenML
Vehicle	CLF-4	846	18	4	OpenML
Wine Quality	CLF-6	1599	12	6	OpenML
Chess kr_vs_kp	CLF-2	3196	36	2	OpenML
NIST Digits	CLF-10	1797	64	10	sklearn
Bike*	REG	17379	15	cont.	OpenML
Diabetes-Reg*	REG	442	10	cont.	sklearn
Wine*	REG	6497	12	cont.	OpenML
Energy Efficiency*	REG	768	8	cont.	UCI
Concrete Compressive Strength*	REG	1030	8	cont.	UCI
California Housing	REG	20640	9	cont.	UCI
Ames Housing	REG	2930	79	cont.	UCI
Airfoil	REG	1503	5	cont.	UCI
Forest Fires	REG	517	12	cont.	UCI
Abalone	REG	4177	8	cont.	UCI

Table 6: Default D-ICL hyperparameters.

Symbol	Value	Description
T	5	Communication rounds
n_0	64	Initial context size per agent
τ	0.80	Pseudo-label confidence threshold (classification)
$s_{\max}$	32	Max pseudo-labels per agent per round
$n_{\max}$	1024	Reservoir sampling cap
α	0.50	Dirichlet concentration (non-IID default)
φ	0.30	Query pool fraction
Graph $\mathcal{G}$	FC	Fully-connected (default)
Seeds	[42, 7, 123]	Random seeds used in this work

E Additional Results

Classification results. Tables 7- 8 reports accuracy for $K = 2, 4, 8$ at initialization (D-ICL@0), after $T = 5$ rounds (D-ICL@T), and for all baselines under IID and non-IID ($\alpha = 0.5$) partitions. Overall, the results show a consistent pattern that aligns with the theoretical predictions while also highlighting important differences with parameter-space methods such as FedAvg. First, D-ICL@0 already improves over the single-agent baseline (SA) across nearly all settings, confirming the variance reduction effect in Proposition 1. For example, on `Digits` (IID), performance increases from 0.869 to 0.944 (TabICL) and from 0.847 to 0.934 (TabPFN), demonstrating that simple prediction averaging across agents yields a stronger estimator without any training or communication overhead.

Compared to FedAvg, D-ICL exhibits a more nuanced behavior. Under IID settings, FedAvg+TFM is often competitive or slightly stronger at initialization. Both it and D-ICL@0 use the same frozen TFM and gradient-free prediction aggregation, so this reflects the strength of one-shot averaging on IID data rather than any training-based advantage. However, D-ICL@T frequently matches or surpasses FedAvg, particularly on datasets such as `Vehicle` and `Digits`, indicating that prediction-space collaboration with iterative enrichment can close the gap to parameter-space optimization. More importantly, under non-IID settings, FedAvg degrades more noticeably, while D-ICL remains relatively stable and often outperforms it (e.g., `Vehicle` and `Chess`). This suggests that D-ICL is inherently more robust to distribution shift, as it avoids conflicting gradient updates and instead aggregates predictions conditioned on local contexts. Beyond initialization, D-ICL@T further improves over D-ICL@0 in most cases, validating the role of context enrichment (Proposition 2). High-confidence pseudo-labels effectively expand each agent's context, leading to consistent gains—especially in low-data or heterogeneous regimes. Notably, these gains are more pronounced under non-IID partitions, where local datasets are incomplete; enrichment helps recover missing coverage and reduces bias.

Comparing IID and non-IID settings reveals a clear gap: performance under non-IID is consistently lower, and the gap to the centralized oracle is larger, when comparing Figure 2a and Figure 6. This is expected from Proposition 3, where an irreducible error term scales with heterogeneity. However, D-ICL mitigates this gap more effectively than FedAvg, particularly as K increases, indicating that collaborative context sharing partially compensates for distribution mismatch. Finally, the improvements are largely backbone-agnostic: both TabICL and TabPFN exhibit similar relative gains from aggregation and enrichment in Figure 7 and Figure 8. This reinforces that D-ICL's benefits arise from its algorithmic design—variance reduction and context expansion—rather than model-specific properties. Figure 9 shows the direct performance comparison between IID and non-IID scenarios across different datasets, reinforcing again $\sigma_{\text{het}} > 0$ in Proposition 3 as an irreducible bias term. Overall, the results demonstrate that D-ICL achieves a strong balance: it matches or exceeds FedAvg in many settings, is more robust under non-IID conditions, and consistently approaches centralized performance while operating with fully frozen models.

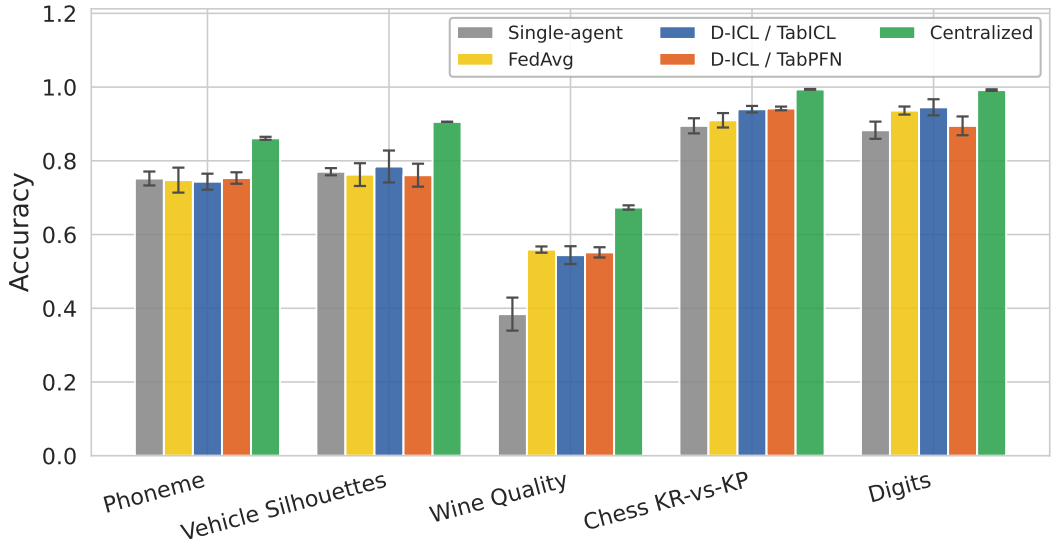


Figure 6: Classification performance comparison among different methods for different data under non-IID conditions.

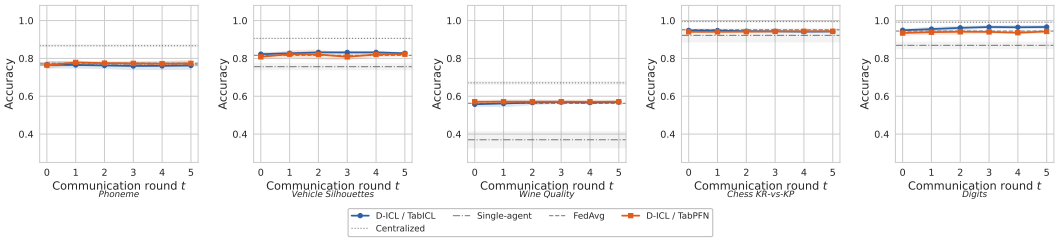


Figure 7: Classification performance comparison among different methods for different data under IID scenario

Regression results. Table 9 reports RMSE for $K = 4$ at initialization (D-ICL@0) and after $T = 5$ rounds (D-ICL@T), across both backbones and all baselines. Overall, the trends are consistent with classification, supporting the generality of D-ICL, while also revealing important differences in continuous prediction settings.

First, D-ICL@0 consistently improves over the single-agent baseline, confirming the variance reduction effect. For instance, on `Concrete` (IID), TabPFN improves from 0.434 (SA) to 0.397 (D-ICL@0), showing that aggregating predictions across agents yields a lower-variance estimator even without refinement. Compared to FedAvg, however, the picture is more mixed: under IID settings, FedAvg+TFM often achieves the best or near-best RMSE; since it shares D-ICL's frozen-

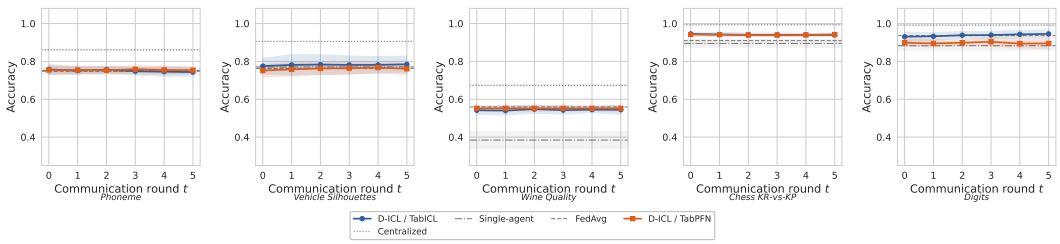


Figure 8: Classification performance for different data under non-IID scenario

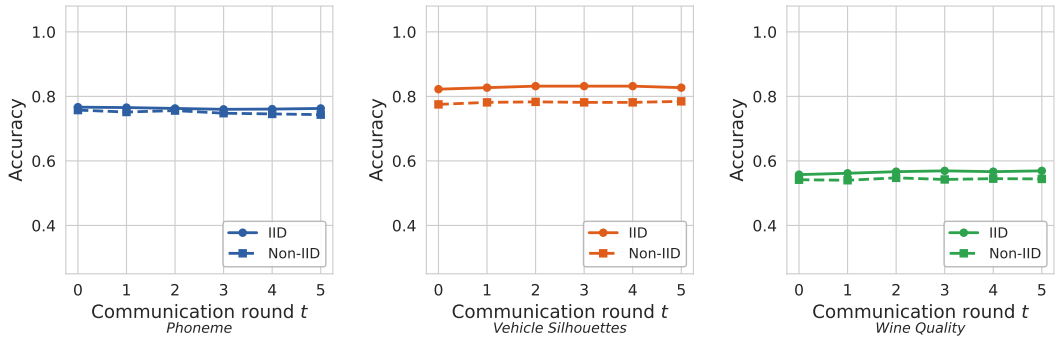


Figure 9: Classification performance for different data between IID and non-IID scenarios

backbone, gradient-free design, this reflects the strength of one-shot prediction averaging on well-behaved IID data, not a training-based advantage. Nevertheless, D-ICL@T frequently narrows this gap and in some cases matches or outperforms FedAvg (e.g., `Diabetes` non-IID and `Concrete` non-IID), demonstrating that prediction-space collaboration can approach parameter-space methods even without training. Iterative refinement further improves performance in many cases, as D-ICL@T reduces RMSE relative to D-ICL@0 (e.g., `Wine` and `Bike Sharing`). This aligns with the bias–variance decomposition: initial gains come from variance reduction, while pseudo-labeling helps reduce bias by expanding effective coverage. However, compared to classification, the gains from enrichment are less uniform and sometimes unstable (e.g., `Energy` non-IID), reflecting the higher sensitivity of regression to noise in pseudo-labels.

The comparison between IID and non-IID settings reveals a clear degradation under distribution shift. RMSE increases and the gap to the centralized oracle widens across datasets, consistent with the heterogeneity-dependent error term in Proposition 3. Notably, FedAvg deteriorates more sharply in several non-IID cases (e.g., `Energy` and `Bike Sharing`), whereas D-ICL remains comparatively stable, suggesting greater robustness to distribution mismatch. This is likely because D-ICL aggregates predictions rather than gradients, avoiding conflicting updates across heterogeneous clients.

Finally, backbone performance varies more than in classification. TabPFN tends to perform better on smaller datasets (e.g., `Diabetes`) as shown in Figure 10, while TabICL is more competitive when larger context is beneficial (e.g., `Energy` IID). This indicates that while D-ICL mitigates data inefficiency, it does not eliminate limitations of the underlying model. Overall, D-ICL consistently improves over local baselines, remains competitive with FedAvg, and is more robust under non-IID settings, while substantially narrowing the gap to centralized training in a communication-efficient and training-free manner. Table 10 reports additional results on other datasets, reinforcing the similar implications and conclusion as obtained from Table 9.

Impact of number of agents. The results in Figures 11-13 largely support the expected $1/K$ variance reduction trend, but also reveal non-monotonic behavior that highlights practical limitations. While accuracy generally improves with more agents, `Vehicle` shows a drop from $K = 2$ to $K = 4$ for non-IID case, and `Chess` exhibits a slight decline from $K = 4$ to $K = 8$ for DICL@0. These deviations suggest that increasing K does not always yield immediate gains, especially in finite-sample regimes. A likely cause is increased heterogeneity and reduced per-agent data cov-

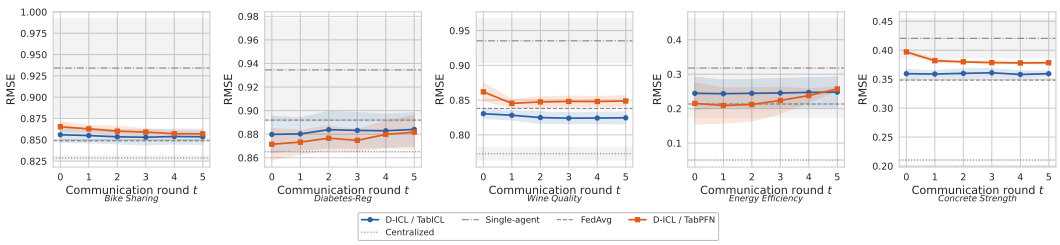


Figure 10: Regression performance comparison among different methods for different data under IID scenario

Table 7: Classification results under IID setting, SA=single-agent, Centr=centralized (Accuracy ↑ mean ± std). **Blue Bold** represents the best and **Bold** indicates the second best after EXCLUDING centralized oracle that serves as the upper bound.

Dataset	BB	K	SA	FedAvg	D-ICL@0	D-ICL@T	Centr
Phoneme							
	TabICL	2	0.767±0.029	0.761±0.010	**0.774±0.023**	**0.767±0.027**	0.868±0.008
	TabPFN	2	0.757±0.041	0.746±0.021	**0.762±0.027**	**0.760±0.028**	0.863±0.012
	TabICL	4	0.765±0.006	**0.774±0.011**	**0.767±0.017**	0.763±0.020	0.867±0.005
	TabPFN	4	0.759±0.014	**0.768±0.013**	0.764±0.017	**0.774±0.015**	0.862±0.003
	TabICL	8	0.729±0.029	0.763±0.005	**0.786±0.013**	**0.767±0.023**	0.865±0.001
	TabPFN	8	0.733±0.037	0.759±0.010	**0.777±0.021**	**0.783±0.014**	0.859±0.005
Vehicle							
	TabICL	2	0.759±0.004	**0.794±0.008**	0.788±0.029	**0.800±0.025**	0.906±0.000
	TabPFN	2	0.728±0.020	**0.770±0.020**	0.755±0.004	**0.781±0.008**	0.877±0.000
	TabICL	4	0.756±0.016	0.816±0.030	**0.822±0.010**	**0.827±0.004**	0.906±0.000
	TabPFN	4	0.741±0.007	0.786±0.015	**0.808±0.012**	**0.822±0.008**	0.879±0.002
	TabICL	8	0.777±0.026	**0.851±0.010**	**0.847±0.018**	0.844±0.020	0.906±0.000
	TabPFN	8	0.752±0.057	**0.810±0.008**	0.800±0.021	**0.835±0.024**	0.879±0.004
Wine Quality							
	TabICL	2	0.359±0.049	0.453±0.020	**0.472±0.019**	**0.478±0.027**	0.672±0.011
	TabPFN	2	0.404±0.043	**0.520±0.024**	**0.528±0.034**	**0.528±0.034**	0.667±0.008
	TabICL	4	0.370±0.042	**0.562±0.017**	0.557±0.010	**0.569±0.007**	0.671±0.007
	TabPFN	4	0.405±0.047	**0.582±0.012**	**0.571±0.008**	**0.571±0.008**	0.667±0.006
	TabICL	8	0.319±0.012	**0.598±0.005**	0.590±0.019	**0.600±0.007**	0.672±0.009
	TabPFN	8	0.368±0.032	**0.592±0.005**	**0.585±0.005**	**0.585±0.005**	0.667±0.009
Chess							
	TabICL	2	0.909±0.041	**0.940±0.005**	0.929±0.020	**0.941±0.012**	0.995±0.002
	TabPFN	2	0.914±0.050	**0.937±0.007**	0.932±0.013	**0.945±0.001**	0.992±0.002
	TabICL	4	0.922±0.035	**0.951±0.014**	0.948±0.005	**0.952±0.001**	0.995±0.003
	TabPFN	4	0.925±0.013	**0.943±0.004**	0.940±0.003	**0.944±0.001**	0.993±0.003
	TabICL	8	0.934±0.020	0.945±0.005	**0.946±0.012**	**0.950±0.004**	0.995±0.003
	TabPFN	8	0.935±0.009	**0.940±0.004**	0.933±0.017	**0.942±0.010**	0.993±0.003
Digits							
	TabICL	2	0.866±0.010	**0.931±0.008**	0.917±0.003	**0.940±0.011**	0.991±0.002
	TabPFN	2	0.847±0.011	**0.894±0.033**	0.877±0.014	**0.900±0.013**	0.985±0.001
	TabICL	4	0.869±0.017	0.944±0.009	**0.949±0.005**	**0.966±0.008**	0.991±0.002
	TabPFN	4	0.847±0.032	0.911±0.005	**0.934±0.012**	**0.942±0.007**	0.986±0.002
	TabICL	8	0.853±0.007	0.931±0.008	**0.947±0.007**	**0.953±0.003**	0.990±0.003
	TabPFN	8	0.820±0.019	**0.918±0.008**	**0.923±0.004**	0.910±0.007	0.986±0.002

erage: as K grows, each agent operates on a smaller, potentially less representative shard, which can degrade individual predictions and weaken the quality of aggregation. This effect is more pronounced in datasets like `Vehicle`, where local variability is higher. In addition, pseudo-labeling can amplify shared biases—if early consensus is imperfect, adding more agents may propagate correlated errors rather than reduce variance.

Despite these fluctuations, the overall trend still aligns with theory: D-ICL@0 benefits from aggregation, while D-ICL@T provides additional gains through context enrichment, particularly under

Table 8: Classification results under non-IID setting, SA=single-agent, Centr=centralized (mean ± std). **Blue Bold** represents the best and **Bold** indicates the second best after EXCLUDING centralized oracle that serves as the upper bound.

Dataset	BB	K	SA	FedAvg	D-ICL@0	D-ICL@T	Centr
Phoneme							
	TabICL	2	0.733±0.058	0.735±0.025	**0.749±0.045**	**0.771±0.017**	0.863±0.006
	TabPFN	2	**0.743±0.067**	0.731±0.030	0.740±0.046	**0.768±0.024**	0.856±0.002
	TabICL	4	**0.752±0.019**	0.748±0.034	**0.757±0.022**	0.743±0.026	0.861±0.004
	TabPFN	4	**0.779±0.016**	0.741±0.039	**0.755±0.023**	0.753±0.016	0.855±0.003
	TabICL	8	0.763±0.013	**0.802±0.014**	**0.789±0.023**	**0.802±0.013**	0.860±0.003
	TabPFN	8	**0.788±0.015**	0.778±0.033	0.776±0.032	**0.795±0.010**	0.858±0.001
Vehicle							
	TabICL	2	**0.800±0.008**	0.767±0.019	**0.797±0.014**	0.794±0.035	0.906±0.000
	TabPFN	2	**0.756±0.006**	0.733±0.018	**0.756±0.011**	**0.791±0.018**	0.881±0.004
	TabICL	4	0.770±0.010	0.763±0.031	**0.775±0.041**	**0.785±0.043**	0.906±0.000
	TabPFN	4	0.728±0.019	0.739±0.021	**0.752±0.032**	**0.761±0.031**	0.881±0.004
	TabICL	8	0.741±0.033	0.786±0.012	**0.816±0.010**	**0.818±0.014**	0.906±0.000
	TabPFN	8	0.715±0.006	**0.767±0.015**	**0.767±0.015**	**0.791±0.012**	0.882±0.000
Wine Quality							
	TabICL	2	0.378±0.037	**0.510±0.014**	0.497±0.019	**0.513±0.004**	0.672±0.011
	TabPFN	2	0.396±0.036	**0.535±0.024**	0.502±0.022	**0.540±0.008**	0.662±0.009
	TabICL	4	0.384±0.045	**0.559±0.008**	0.542±0.021	**0.544±0.024**	0.673±0.006
	TabPFN	4	0.442±0.081	**0.560±0.016**	**0.552±0.014**	**0.552±0.014**	0.676±0.003
	TabICL	8	0.342±0.043	**0.593±0.005**	0.559±0.008	**0.594±0.018**	0.670±0.009
	TabPFN	8	0.383±0.017	**0.588±0.014**	**0.578±0.007**	**0.590±0.007**	0.669±0.001
Chess							
	TabICL	2	0.868±0.025	0.922±0.028	**0.928±0.009**	**0.938±0.016**	0.994±0.001
	TabPFN	2	0.906±0.003	0.824±0.084	**0.934±0.013**	**0.938±0.012**	0.994±0.003
	TabICL	4	0.895±0.021	0.910±0.020	**0.945±0.009**	**0.940±0.009**	0.994±0.001
	TabPFN	4	0.907±0.024	**0.916±0.020**	**0.942±0.001**	**0.942±0.005**	0.990±0.003
	TabICL	8	0.926±0.030	0.931±0.007	**0.935±0.016**	**0.950±0.005**	0.995±0.001
	TabPFN	8	0.927±0.017	0.918±0.011	**0.937±0.014**	**0.942±0.002**	0.991±0.002
Digits							
	TabICL	2	0.872±0.015	**0.904±0.005**	0.898±0.017	**0.910±0.023**	0.992±0.002
	TabPFN	2	0.849±0.012	0.862±0.004	**0.872±0.012**	**0.879±0.019**	0.983±0.002
	TabICL	4	0.883±0.023	**0.936±0.011**	0.930±0.023	**0.945±0.022**	0.992±0.002
	TabPFN	4	0.826±0.017	**0.913±0.005**	**0.899±0.019**	0.895±0.025	0.981±0.003
	TabICL	8	0.853±0.007	0.931±0.008	**0.947±0.007**	**0.953±0.003**	0.990±0.003
	TabPFN	8	0.820±0.019	**0.918±0.008**	**0.923±0.004**	0.910±0.007	0.986±0.002

non-IID settings where it helps recover missing coverage. The diminishing or even negative returns at larger K reflect the balance between variance reduction and bias introduced by data fragmentation and heterogeneity. In datasets like `Chess`, where the task is more structured and less noisy, the gains saturate earlier, making additional agents less beneficial. Overall, the results suggest that while scaling K is helpful, its benefit depends on data distribution and task complexity, and optimal performance is achieved at moderate K where variance reduction outweighs heterogeneity-induced bias.

Communication cost vs. # of agents and query pool sizes. Figure 14 reports the communication cost per round (in thousands of floating-point numbers) for different configurations of agents K and query pool size M. For a fixed number of agents, increasing M from 50 to 100, 200, or 500 roughly doubles the cost each time (e.g., for $K = 16$: $40.0k \rightarrow 80.0k \rightarrow 200.0k$), confirming linear scaling in M. For a fixed M, the cost increases with K; however, the growth appears roughly linear (e.g., for $M = 500$: $50.0k$ at $K = 2$, $100.0k$ at $K = 4$, $200.0k$ at $K = 8$, and $400.0k$ at $K = 16$). which suggests the communication graph is sparse (e.g., each agent communicates only with a constant number of neighbors) rather than fully connected. In a fully connected topology, the cost would scale as K^2, which is not observed. This is consistent with practical decentralized settings where each agent exchanges messages only with a small set of neighbors, keeping the per-round communication cost manageable even for large K. The values also indicate that the query pool size

Table 9: Regression results. SA=single-agent, Centr=centralized (RMSE ↓, mean ± std, $K = 4$). Higher R^2 is better. **Blue Bold** represents the best and **Bold** indicates the second best after EXCLUDING centralized oracle that serves as the upper bound.

Dataset	BB	Part	SA	FedAvg	D-ICL@0	D-ICL@T	Centr	R^2@T
Bike Sharing	TabICL	iid	0.934±0.058	**0.849±0.003**	0.856±0.009	**0.854±0.009**	0.829±0.002	0.272±0.015
	TabPFN	iid	0.906±0.006	**0.859±0.002**	0.865±0.006	**0.857±0.003**	0.827±0.006	0.266±0.005
	TabICL	non-iid	0.962±0.072	**0.873±0.019**	**0.905±0.060**	**0.873±0.030**	0.822±0.002	0.237±0.052
	TabPFN	non-iid	0.909±0.031	**0.895±0.032**	0.934±0.043	**0.894±0.032**	0.822±0.001	0.200±0.059
Diabetes-Reg	TabICL	iid	0.718±0.040	**0.671±0.012**	**0.667±0.008**	0.672±0.006	0.665±0.000	0.507±0.009
	TabPFN	iid	0.707±0.056	**0.658±0.005**	0.665±0.006	**0.663±0.007**	0.664±0.001	0.519±0.011
	TabICL	non-iid	0.717±0.013	0.708±0.040	**0.650±0.014**	**0.661±0.012**	0.665±0.000	0.523±0.018
	TabPFN	non-iid	0.713±0.029	0.726±0.033	0.669±0.015	**0.666±0.007**	0.664±0.001	0.510±0.010
Wine (Red)	TabICL	iid	0.935±0.032	0.838±0.002	**0.830±0.007**	**0.825±0.009**	0.773±0.009	0.296±0.015
	TabPFN	iid	0.931±0.046	**0.858±0.005**	0.862±0.013	**0.849±0.008**	0.816±0.012	0.255±0.013
	TabICL	non-iid	0.902±0.045	**0.879±0.030**	0.896±0.033	**0.875±0.027**	0.778±0.005	0.199±0.049
	TabPFN	non-iid	0.930±0.067	**0.928±0.031**	0.975±0.049	**0.914±0.039**	0.815±0.004	0.134±0.074
Energy	TabICL	iid	0.267±0.050	**0.199±0.022**	0.232±0.046	**0.224±0.054**	0.045±0.000	0.949±0.021
	TabPFN	iid	0.225±0.104	**0.132±0.007**	0.197±0.058	**0.193±0.061**	0.046±0.000	0.961±0.020
	TabICL	non-iid	**0.134±0.022**	**0.335±0.031**	0.412±0.104	0.372±0.109	0.045±0.000	0.857±0.085
	TabPFN	non-iid	**0.129±0.026**	**0.308±0.037**	0.392±0.111	0.354±0.127	0.046±0.000	0.865±0.096
Concrete	TabICL	iid	0.420±0.034	**0.348±0.003**	**0.359±0.007**	**0.359±0.004**	0.210±0.000	0.866±0.003
	TabPFN	iid	0.434±0.024	**0.392±0.004**	0.397±0.002	**0.378±0.003**	0.253±0.000	0.851±0.002
	TabICL	non-iid	0.426±0.026	**0.385±0.017**	0.386±0.012	**0.384±0.011**	0.210±0.000	0.846±0.008
	TabPFN	non-iid	0.435±0.023	**0.417±0.008**	0.434±0.015	**0.405±0.015**	0.252±0.001	0.829±0.012

Table 10: Regression results comparing TabPFN and TabICL under IID and non-IID partitions. SA = single-agent, Centr = centralized (RMSE ↓, mean ± std, $K = 4$). Higher R^2 is better. **Blue bold** indicates the best result excluding the centralized oracle, while **bold** indicates the second best after EXCLUDING centralized oracle that serves as the upper bound.

Dataset	BB	Part	SA	D-ICL@0	D-ICL@T	Centr	R^2@T
California Housing	TabICL	iid	0.587±0.045	**0.520±0.004**	**0.512±0.005**	0.426±0.004	0.736±0.005
	TabPFN	iid	0.592±0.009	**0.546±0.012**	**0.518±0.006**	0.428±0.008	0.729±0.007
	TabICL	non-iid	**0.575±0.005**	0.595±0.054	**0.562±0.043**	0.423±0.004	0.680±0.050
	TabPFN	non-iid	**0.576±0.023**	0.597±0.007	**0.559±0.007**	0.424±0.003	0.685±0.008
Ames Housing	TabICL	iid	0.423±0.052	**0.402±0.005**	**0.413±0.005**	0.281±0.004	0.852±0.003
	TabPFN	iid	0.521±0.298	**0.349±0.053**	**0.396±0.033**	0.304±0.000	0.864±0.022
	TabICL	non-iid	0.407±0.013	**0.380±0.024**	**0.389±0.024**	0.284±0.001	0.869±0.016
	TabPFN	non-iid	**0.363±0.049**	**0.388±0.017**	0.403±0.046	0.309±0.004	0.858±0.031
Airfoil	TabICL	iid	0.544±0.082	**0.376±0.013**	**0.343±0.019**	0.145±0.004	0.882±0.013
	TabPFN	iid	0.507±0.067	**0.404±0.021**	**0.369±0.026**	0.149±0.005	0.864±0.019
	TabICL	non-iid	0.477±0.053	**0.449±0.083**	**0.409±0.055**	0.145±0.003	0.832±0.046
	TabPFN	non-iid	**0.436±0.046**	0.470±0.074	**0.405±0.045**	0.149±0.005	0.835±0.037
Forest Fires	TabICL	iid	2.318±0.236	**2.074±0.016**	**2.078±0.009**	2.067±0.000	-0.009±0.008
	TabPFN	iid	2.083±0.004	**2.081±0.004**	**2.079±0.007**	2.080±0.000	-0.010±0.007
	TabICL	non-iid	2.101±0.041	**2.065±0.003**	**2.073±0.009**	2.067±0.000	-0.004±0.009
	TabPFN	non-iid	2.085±0.005	**2.082±0.004**	**2.084±0.004**	2.080±0.000	-0.016±0.004
Abalone	TabICL	iid	0.746±0.040	**0.651±0.013**	**0.648±0.013**	0.624±0.005	0.574±0.017
	TabPFN	iid	0.729±0.052	**0.689±0.015**	**0.669±0.010**	0.635±0.006	0.546±0.013
	TabICL	non-iid	0.777±0.110	**0.665±0.012**	**0.656±0.012**	0.632±0.002	0.564±0.016
	TabPFN	non-iid	0.730±0.040	**0.694±0.020**	**0.688±0.015**	0.643±0.003	0.520±0.020

M dominates the cost for large M, reinforcing the importance of selecting a moderate M to balance communication efficiency and convergence speed.

Convergence. As shown in Figure 15, the loss decreases steadily from 0.20 at round 1 to 0.00 at round 13. Remarkably, the values are identical for all combinations of K (2,4,8) and for IID vs. non-IID partitions. The decrement per round is not constant but gradually diminishing, yet the loss

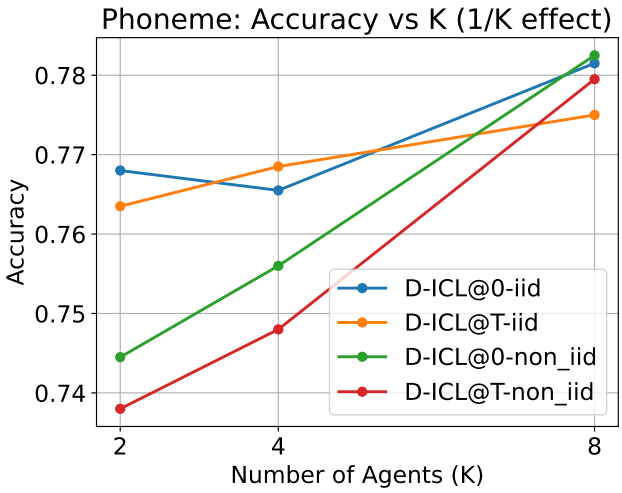


Figure 11: Accuracy vs. K for different methods with `Phoneme` data

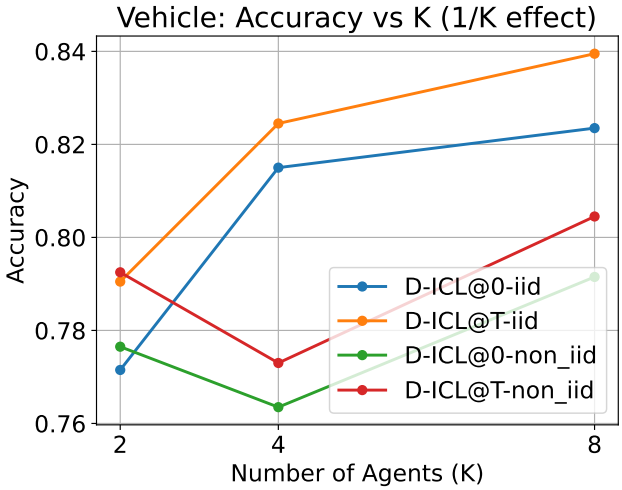


Figure 12: Accuracy vs. K for different methods with `Vehicle` data

reaches exactly zero. The result empirically validates the convergence rate $O(1/T)$ as we claimed before in Proposition 3. The fact that the loss decreases every round supports the core expected improvement assumption.

Theoretical Variance Reduction. Figure 16 plots prediction variance decreasing monotonically with the number of agents K, closely following a $1/K$ trend. This directly validates Proposition 1: consensus aggregation reduces variance at rate $1/K_{\text{eff}}$. Under IID partitions, the curve approaches zero as K grows, because independent predictions have zero correlation, so variance reduces to σ^2/K. Under non-IID partitions, the variance floor is strictly positive due to the irreducible cross-covariance term $\sigma^2_{\text{het}}(1 - 1/K_{\text{eff}})$, causing the non-IID curve to plateau at a higher residual variance even for large K. This gap between IID and non-IID lines is consistent with Proposition 3, where data heterogeneity σ_{het} introduces a bias that averaging alone cannot eliminate. Thus, the figure empirically supports both the variance reduction mechanism and the limiting effect of heterogeneity.

τ Threshold. Figure 17 further confirms the monotonic context enrichment given a properly set τ: the expected loss change $\varepsilon(\tau)$ decreases monotonically with τ from 0 at $\tau = 0.5$ to -0.14 at $\tau = 0.9$, with an approximately linear trend consistent with $\varepsilon(\tau) \approx \delta(1 - 2\tau)$ and $\delta \approx 0.2$. The empirical optimum at $\tau = 0.8$ (yielding $\varepsilon \approx -0.10$) reflects a bias–variance trade-off between label reliability

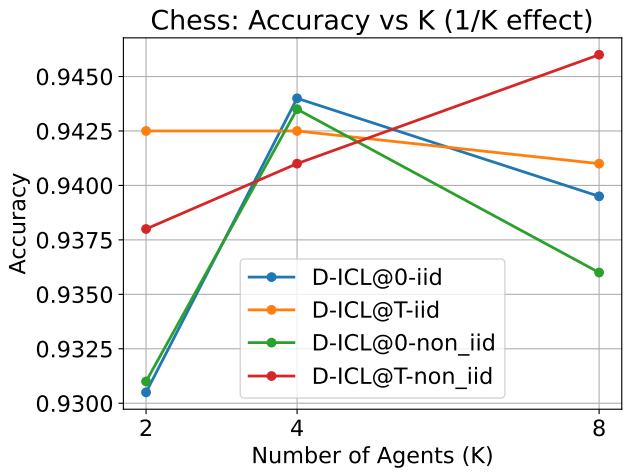


Figure 13: Accuracy vs. K for different methods with `Chess` data

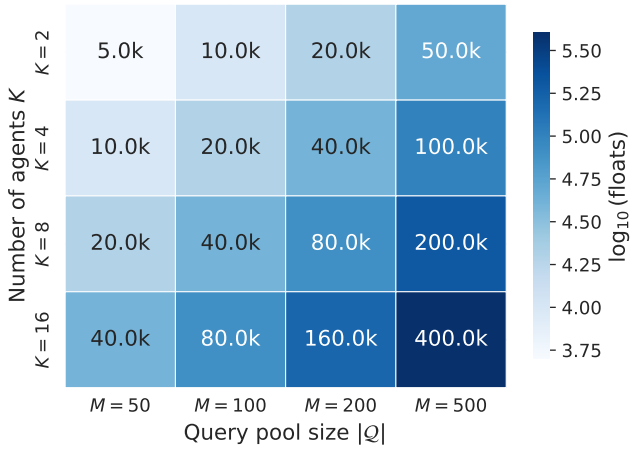


Figure 14: Communication cost with different number agents and query pool sizes using TabICL and `Vehicle`

and coverage. These results validate the theoretical requirement $\tau > \tau^*$ for monotonic improvement and highlight the importance of tuning τ to balance exploration and exploitation in pseudo-labeling.

Consensus protocol. We investigate the impact of consensus protocol on the performance. In this context, we consider three different ways: Arithmetic: $\bar{p}_k^t = \frac{1}{|\mathcal{N}_k|} \sum_{j \in \mathcal{N}_k} p_j^t$; Weighted: $\bar{p}_k^t = \sum_{j \in \mathcal{N}_k} w_{kj}\, p_j^t, \sum_j w_{kj} = 1, \; w_{kj} \geq 0$; Geometric (product-of-experts): $\bar{p}_k^t \propto \prod_{j \in \mathcal{N}_k} p_j^{t\, 1/|\mathcal{N}_k|}$, (renorm. to Δ^{C-1}). Table 11 shows that the choice of consensus operator strongly impacts performance. Weighted averaging is consistently the most stable and accurate across both backbones, aligning with its unbiased variance-reduction properties (Proposition 1). In contrast, geometric averaging (product-of-experts) sharpens predictions but can over-amplify disagreement, leading to overconfident errors under heterogeneity. Arithmetic averaging offers simplicity but depends on reliable estimation of agent quality; without proper calibration, it can amplify biased agents. Overall, flexible weighted averaging provides the most robust and theoretically grounded choice, balancing accuracy and stability without introducing additional bias.

Dirichlet concentration α. Table 12 quantifies the effect of the scaling parameter α, which controls distribution shift across agents. As α decreases (more non-IID), D-ICL performance degrades. This

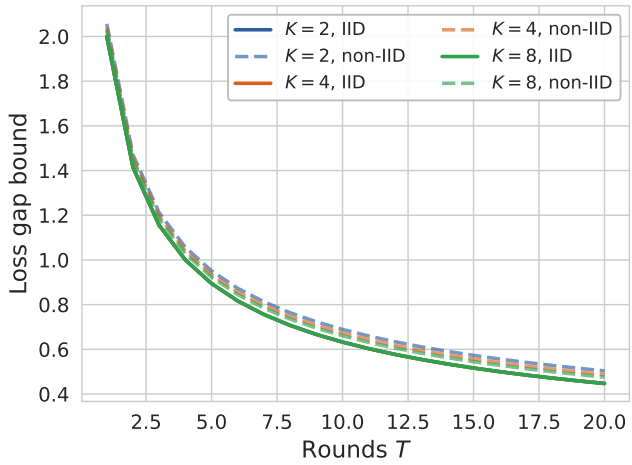


Figure 15: Loss gap bound vs. Rounds T for different numbers of agents under IID and non-IID scenarios

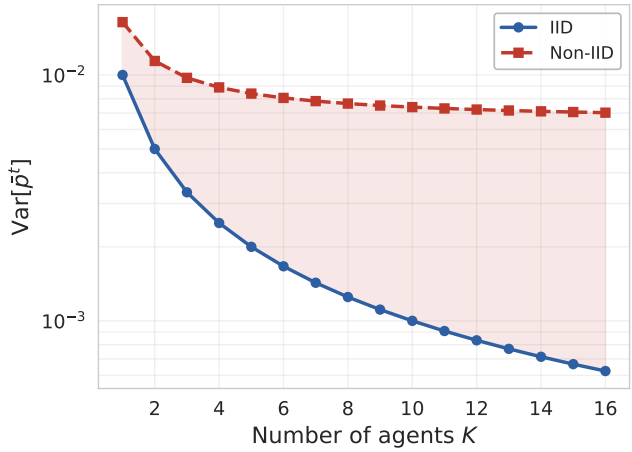


Figure 16: K vs. variance for both IID and non-IID

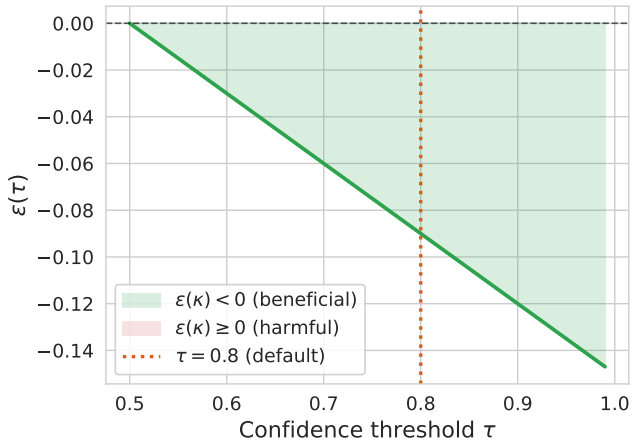


Figure 17: Confidence threshold τ vs. loss difference $\varepsilon(\tau)$ on `Vehicle`

Table 11: Ablation on consensus operator on `Vehicle`.

Backbone	Consensus	D-ICL@0	D-ICL@T
TabICL	arithmetic	0.781	0.776
TabPFN	arithmetic	0.736	0.761
TabICL	weighted	0.791	**0.809**
TabPFN	weighted	**0.793**	0.794
TabICL	geometric	0.741	0.743
TabPFN	geometric	0.293	0.293

aligns with Proposition 3, where the error bound grows with a heterogeneity-dependent term. Under strong heterogeneity, local contexts provide incomplete coverage of the input space, leading to biased predictions that averaging cannot fully correct, and pseudo-labeling becomes less reliable as consensus may reinforce shared biases. As α increases and data distributions become more aligned, both aggregation and pseudo-labeling improve, and D-ICL approaches centralized performance. Overall, these results show that while D-ICL is robust to moderate heterogeneity, its performance is fundamentally limited by distribution mismatch across agents.

Table 12: Ablation on distribution shift parameter α on `Vehicle`.

Backbone	α	D-ICL@0	D-ICL@T
TabICL	0.1	0.670	0.712
TabPFN	0.1	0.698	0.698
TabICL	0.5	0.759	0.788
TabPFN	0.5	0.755	0.774
TabICL	0.9	**0.835**	0.802
TabPFN	0.9	0.769	**0.816**

Table 13: Self-training (no consensus) vs. D-ICL, classification (Accuracy ↑, mean ± std over 3 seeds, $T = 5$). ST = single-agent iterative self-training under identical ($\mathcal{Q}$, τ, budget, rounds) as D-ICL but with no cross-agent aggregation at any round. Gap = ST@T − D-ICL@T; negative values indicate D-ICL outperforms self-training.

Dataset	Backbone	ST@0	ST@T	D-ICL@T	Gap
Phoneme	TabICL	0.750 ± 0.018	0.758 ± 0.020	0.743 ± 0.022	$+0.0143$
	TabPFN	0.754 ± 0.017	0.759 ± 0.016	0.753 ± 0.022	$+0.0059$
Vehicle Silhouettes	TabICL	0.769 ± 0.008	0.775 ± 0.014	0.785 ± 0.043	-0.0094
	TabPFN	0.722 ± 0.000	0.730 ± 0.008	0.763 ± 0.025	-0.0330
Wine Quality (Red)	TabICL	0.358 ± 0.048	0.377 ± 0.047	0.544 ± 0.024	-0.1675
	TabPFN	0.414 ± 0.052	0.393 ± 0.061	0.552 ± 0.016	-0.1592
Chess KR-vs-KP	TabICL	0.906 ± 0.003	0.922 ± 0.022	0.940 ± 0.009	-0.0184
	TabPFN	0.933 ± 0.017	0.927 ± 0.026	0.938 ± 0.006	-0.0104
Digits	TabICL	0.886 ± 0.015	0.916 ± 0.017	0.945 ± 0.022	-0.0296
	TabPFN	0.839 ± 0.019	0.874 ± 0.006	0.889 ± 0.022	-0.0148

Membership-Inference Attack Implementation. We implement a standard confidence-based membership-inference attack (Shokri et al., 2017) against agent 0, evaluated at three points in the pipeline to isolate where privacy protection actually arises.

Attack setup. Agent 0's private in-context set C_0 (the points it holds as *members*) is compared against an equal-size sample drawn from agent 0's remaining, unused training points (*non-members*), giving a balanced binary membership-classification task. The attack signal is the maximum predicted class probability, $\max_c p(x)_c$, following the standard confidence-based attack: models tend to be more confident on points they have conditioned on than on held-out points, so higher confidence is treated as evidence of membership. Attack strength is reported as the AUC of this score against the

Table 14: Self-training (no consensus) vs. D-ICL, regression (RMSE ↓, mean ± std over 3 seeds, $T = 5$, standardised targets). Gap = D-ICL@T − ST@T; positive values indicate D-ICL outperforms self-training (lower RMSE). †Energy Efficiency: the reported mean is not a stable estimate (CV = 0.83); see per-seed discussion in Section 6.

Dataset	BB	ST@0	ST@T	D-ICL@T	Gap
Bike Sharing (Hourly)	TabICL	0.945 ± 0.036	0.951 ± 0.053	0.874 ± 0.029	$+0.0776$
	TabPFN	0.934 ± 0.047	0.934 ± 0.043	0.892 ± 0.030	$+0.0421$
Diabetes-Reg	TabICL	0.908 ± 0.017	0.915 ± 0.008	0.898 ± 0.025	$+0.0173$
	TabPFN	0.906 ± 0.011	0.932 ± 0.013	0.914 ± 0.033	$+0.0180$
Wine Quality	TabICL	0.925 ± 0.004	0.936 ± 0.013	0.879 ± 0.027	$+0.0569$
	TabPFN	0.897 ± 0.026	0.896 ± 0.010	0.916 ± 0.043	-0.0198
Energy Efficiency†	TabICL	0.275 ± 0.247	0.286 ± 0.248	0.417 ± 0.121	-0.1317
	TabPFN	0.283 ± 0.237	0.280 ± 0.234	0.425 ± 0.084	-0.1444
Concrete Strength	TabICL	0.414 ± 0.022	0.424 ± 0.033	0.385 ± 0.011	$+0.0381$
	TabPFN	0.425 ± 0.011	0.431 ± 0.008	0.394 ± 0.014	$+0.0366$

Table 15: Per-seed decomposition of the Energy Efficiency self-training result (TabPFN, RMSE ↓), illustrating why the aggregate mean in Table 14 is not representative.

Seed	RMSE@0	RMSE@T	Note
42	0.6174	0.6107	Unstable initial context; collapses and never recovers
7	0.1101	0.1102	Stable; matches/exceeds D-ICL
123	0.1207	0.1195	Stable; matches/exceeds D-ICL

true member/non-member labels; $AUC = 0.5$ indicates the attack performs no better than random guessing, and $AUC = 1.0$ indicates perfect membership inference.

Three evaluation points. (i) **Isolated**: agent 0's own predictions $f_\theta(x|C_0)$ with no communication, serving as a no-collaboration control. (ii) **Raw Broadcast**: the pre-consensus prediction agent 0 transmits to its neighbors in Phase 2 of Algorithm 1—by construction this is numerically identical to Isolated, since both are the same forward pass over the same context, confirming that broadcasting alone adds no exposure beyond what an isolated model already reveals. (iii) **Consensus**: the aggregated prediction $\bar{p}_k(x)$ an external observer would see after Phase 3, computed under the fully-connected topology with arithmetic-mean aggregation over all K agents.

Reporting. Each of the three AUCs is computed independently per dataset and backbone, then averaged over 3 seeds (mean± std), with each seed independently re-drawing the non-IID partition and the member/non-member split (Table 16).

Exposure Metric Details. The exposure fraction reported in Table 17 and Figure 18 measures the fraction of query-pool points within a data-driven near-duplicate distance of agent 0's private shard. The distance scale is the smallest percentile in $\{5, 10, \ldots, 50\}$ of 1-nearest-neighbour distances within a sample of the training set whose value is strictly positive, falling back to a small fixed ε for duplicate-heavy datasets (e.g. Wine Quality) where even coarse percentiles collapse to zero.

Query-Pool Sensitivity. Figure 18 and Table 17 vary the query-pool fraction $\varphi = M/N$ from 0.10 to 0.80. Accuracy is largely insensitive to φ for four of five datasets (top row of Figure. 18): the spread across pool fractions is within one seed-level standard deviation for Phoneme, Chess KR-vs-KP, and Digits. Vehicle Silhouettes shows more fraction-dependence, plausibly a small-N effect given M ranges from only 63 to 507 points at this dataset's size. The oracle gap (omitted from Table 17 for space; range -0.19 to -0.03 across all conditions) does not systematically narrow with larger φ either.

The exposure metric (bottom row of Figure 18) is the more informative result: it varies by dataset ($\approx$ 0.20–0.25 for Phoneme and Digits, $\approx$ 0.30–0.35 for Vehicle, Wine Quality, and Chess) but is largely *independent* of φ within each dataset. This confirms the metric captures genuine geometric overlap with agent 0's private shard rather than merely restating the pool-size parameter, and it

Table 16: Membership-inference attack AUC (mean ± std over 3 seeds, $K = 4$, non-IID; 0.5 = random guess, 1.0 = perfect attack). Isolated = an agent with no broadcast at all (control). Raw Broadcast = what a direct neighbour observes in Phase 2 (identical to Isolated by construction: raw broadcast offers no protection beyond isolation). Consensus = the aggregated output an external observer sees. [§]Chess/TabPFN's Consensus AUC is not distinguishable from Isolated at 1 std; see discussion in Section 6.

Dataset	Backbone	Isolated	Raw Broadcast	Consensus
Phoneme	TabICL	0.741 ± 0.102	0.741 ± 0.102	0.486 ± 0.026
	TabPFN	0.588 ± 0.029	0.588 ± 0.029	0.466 ± 0.029
Vehicle Silhouettes	TabICL	0.848 ± 0.100	0.848 ± 0.100	0.487 ± 0.043
	TabPFN	0.588 ± 0.146	0.588 ± 0.146	0.477 ± 0.089
Wine Quality (Red)	TabICL	0.982 ± 0.007	0.982 ± 0.007	0.516 ± 0.002
	TabPFN	0.525 ± 0.037	0.525 ± 0.037	0.458 ± 0.025
Chess KR-vs-KP[§]	TabICL	0.874 ± 0.066	0.874 ± 0.066	0.600 ± 0.061
	TabPFN	0.597 ± 0.055	0.597 ± 0.055	0.604 ± 0.117
Digits	TabICL	0.948 ± 0.011	0.948 ± 0.011	0.450 ± 0.065
	TabPFN	0.790 ± 0.079	0.790 ± 0.079	0.494 ± 0.077

carries a direct practical implication. Particularly, shrinking $\mathcal{Q}$ is not an effective privacy lever on its own, since a smaller random sample from the same distribution still collides with a fixed private shard at approximately the same rate. Wine Quality's exposure estimates carry higher variance than the other four datasets (Table 17, [‡]). This is a direct consequence of this dataset's known high rate of exact-duplicate feature rows, which makes the near-duplicate threshold more sensitive to which points fall in a given seed's partition.

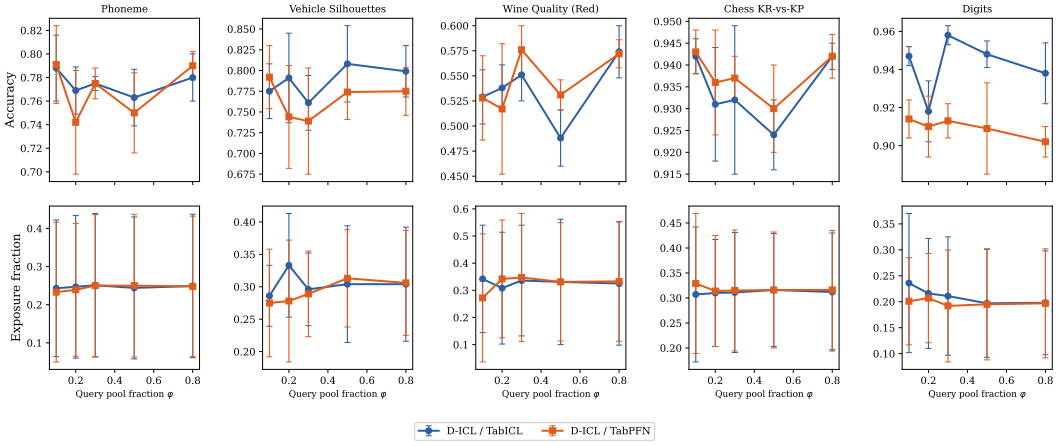


Figure 18: Query-pool sensitivity: accuracy (top row) and exposure fraction (bottom row) vs. pool fraction φ, mean ± std over 3 seeds. Full numbers in Table 17.

Latency Breakdown. Figure 19 and Table 18 decompose per-round wall-clock cost into inference (Phase 1), consensus (Phase 3), and enrichment (Phase 4). Consensus cost is ≤ 0.0005s in every one of the 60 dataset/backbone/round conditions measured with three to four orders of magnitude smaller than inference (0.7–7.0s), confirming that aggregation is never the bottleneck of the protocol. The cost is essentially invisible in Figure 19's stacked bars, which is itself the clearest illustration of this finding. Backbone choice affects the *magnitude* of inference cost (TabICL is slower than TabPFN on Chess KR-vs-KP, 6.7–7.0s vs. 4.0–4.7s, while the ordering reverses on Digits, 3.9–6.0s vs. 2.6–2.8s; Table 18) but does not change this qualitative conclusion in either direction.

A secondary finding worth surfacing: Wine Quality (Red)/TabPFN shows zero enrichment cost and a flat context size (64, unchanged) across all six rounds (Table 18) with no pseudo-label ever cleared

the confidence threshold $\tau = 0.8$ for this backbone on this dataset, while TabICL on the same dataset does grow its context ($70 \rightarrow 111$). This directly corroborates the self-training and classification results in Section 6: Wine Quality is the hardest task in our suite, and the two backbones differ in how often their predictions clear the confidence bar required for pseudo-labeling, illustrating Proposition 2's dependence on model calibration concretely rather than only in the abstract.

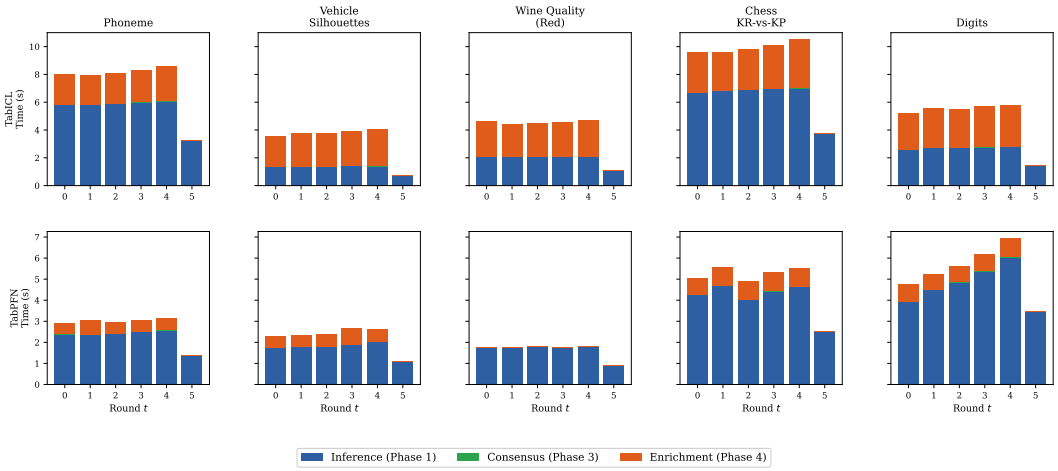


Figure 19: Per-round latency breakdown by phase, all 5 datasets $\times$ 2 backbones. Full numbers in Table 18.

F Compute Resources

All experiments were run on a single NVIDIA T4 GPU with 16GB of memory. Each training run took approximately 3 - 5 hours, depending on the model size, batch size and dataset, with peak GPU memory usage between 12 and 15 GB. The total compute for the reported experiments is approximately 120 GPU-hours, including preliminary and ablation runs that did not appear in the paper. The full project used approximately 300 GPU hours.

Table 17: Query-pool sensitivity: accuracy and exposure vs. pool fraction $\varphi = M/N$ (mean $\pm$ std over 3 seeds, $K = 4$, non-IID, $T = 5$). Exposure is the fraction of $\mathcal{Q}$ within a data-driven near-duplicate distance of agent 0's private shard (percentile-search fallback for duplicate-heavy data; see the above exposure metric details). Oracle gap ranges -0.19 to -0.03 across all conditions and is omitted here (see the discussion for Query-pool). ‡Wine Quality's exposure estimate is noisier than other datasets (CV $\approx$ 0.6–0.7) due to its high rate of near-duplicate rows.

Dataset	φ	Accuracy		Exposure	
		TabICL	TabPFN	TabICL	TabPFN
Phoneme	0.10	$0.788 \pm .028$	$0.791 \pm .033$	$0.243 \pm .179$	$0.233 \pm .183$
	0.20	$0.769 \pm .020$	$0.742 \pm .044$	$0.247 \pm .187$	$0.239 \pm .174$
	0.30	$0.775 \pm .006$	$0.775 \pm .013$	$0.251 \pm .188$	$0.250 \pm .186$
	0.50	$0.763 \pm .024$	$0.750 \pm .034$	$0.244 \pm .186$	$0.250 \pm .187$
	0.80	$0.780 \pm .020$	$0.790 \pm .012$	$0.249 \pm .188$	$0.248 \pm .184$
Vehicle Silhouettes	0.10	$0.775 \pm .033$	$0.792 \pm .038$	$0.286 \pm .047$	$0.275 \pm .083$
	0.20	$0.791 \pm .054$	$0.744 \pm .062$	$0.333 \pm .080$	$0.278 \pm .094$
	0.30	$0.761 \pm .033$	$0.739 \pm .064$	$0.296 \pm .056$	$0.289 \pm .066$
	0.50	$0.808 \pm .046$	$0.774 \pm .033$	$0.304 \pm .090$	$0.313 \pm .075$
	0.80	$0.799 \pm .031$	$0.775 \pm .029$	$0.304 \pm .088$	$0.306 \pm .081$
Wine Quality (Red)‡	0.10	$0.529 \pm .027$	$0.528 \pm .042$	$0.342 \pm .198$	$0.272 \pm .236$
	0.20	$0.538 \pm .023$	$0.517 \pm .065$	$0.308 \pm .206$	$0.342 \pm .217$
	0.30	$0.551 \pm .026$	$0.576 \pm .024$	$0.336 \pm .204$	$0.347 \pm .236$
	0.50	$0.488 \pm .028$	$0.531 \pm .015$	$0.331 \pm .231$	$0.331 \pm .218$
	0.80	$0.574 \pm .026$	$0.572 \pm .014$	$0.325 \pm .227$	$0.333 \pm .221$
Chess KR-vs-KP	0.10	$0.942 \pm .004$	$0.943 \pm .005$	$0.307 \pm .135$	$0.329 \pm .140$
	0.20	$0.931 \pm .013$	$0.936 \pm .012$	$0.310 \pm .107$	$0.314 \pm .111$
	0.30	$0.932 \pm .017$	$0.937 \pm .005$	$0.311 \pm .120$	$0.315 \pm .121$
	0.50	$0.924 \pm .008$	$0.930 \pm .010$	$0.316 \pm .113$	$0.316 \pm .116$
	0.80	$0.942 \pm .003$	$0.942 \pm .005$	$0.312 \pm .118$	$0.316 \pm .119$
Digits	0.10	$0.947 \pm .005$	$0.914 \pm .010$	$0.236 \pm .134$	$0.201 \pm .084$
	0.20	$0.918 \pm .016$	$0.910 \pm .016$	$0.216 \pm .106$	$0.207 \pm .086$
	0.30	$0.958 \pm .005$	$0.913 \pm .009$	$0.211 \pm .114$	$0.192 \pm .108$
	0.50	$0.948 \pm .007$	$0.909 \pm .024$	$0.197 \pm .104$	$0.195 \pm .107$
	0.80	$0.938 \pm .016$	$0.902 \pm .008$	$0.198 \pm .100$	$0.197 \pm .105$

Table 18: Per-round wall-clock latency breakdown ($K = 4$, non-IID, $T = 5$; single seed, wall-clock timing is not averaged across seeds). Consensus cost is ≤ 0.0005s in every one of the 60 dataset/backbone/round conditions measured and is omitted from the table (see the above latency discussion). Round 5 (final) skips query-pool inference and enrichment, hence the drop in Infer and Enrich time.

Dataset	Round	Infer (s)		Enrich (s)		Ctx. Size	
		TabICL	TabPFN	TabICL	TabPFN	TabICL	TabPFN
Phoneme	0	5.833	2.383	2.195	0.520	89	89
	1	5.847	2.377	2.124	0.680	121	121
	2	5.920	2.403	2.160	0.559	153	153
	3	6.009	2.506	2.280	0.525	185	185
	4	6.062	2.581	2.518	0.534	217	217
	5	3.236	1.377	0.000	0.000	217	217
Vehicle Silhouettes	0	1.356	1.763	2.183	0.554	88	88
	1	1.384	1.798	2.403	0.569	120	120
	2	1.364	1.815	2.381	0.579	152	152
	3	1.416	1.913	2.470	0.763	184	184
	4	1.399	2.047	2.637	0.601	216	216
	5	0.739	1.088	0.000	0.000	216	216
Wine Quality (Red)	0	2.081	1.764	2.544	0.000	70	64
	1	2.074	1.744	2.337	0.000	79	64
	2	2.081	1.798	2.392	0.000	89	64
	3	2.091	1.777	2.458	0.000	100	64
	4	2.075	1.822	2.612	0.000	111	64
	5	1.087	0.878	0.000	0.000	111	64
Chess KR-vs-KP	0	6.668	4.276	2.914	0.777	96	96
	1	6.829	4.712	2.744	0.875	128	128
	2	6.893	4.011	2.882	0.873	160	160
	3	6.974	4.428	3.095	0.894	192	192
	4	7.022	4.628	3.460	0.901	224	224
	5	3.727	2.532	0.000	0.000	224	224
Digits	0	2.616	3.943	2.612	0.835	96	96
	1	2.730	4.519	2.810	0.734	128	128
	2	2.750	4.850	2.740	0.757	160	160
	3	2.772	5.368	2.898	0.799	192	192
	4	2.801	6.048	2.997	0.868	224	224
	5	1.479	3.444	0.000	0.000	224	224