[SHORT] FEW-SHOT CROSS-TABLE DATA MIXTURE IN TABULAR IN-CONTEXT LEARNING: BENEFITS, FAILURE MODES, AND ALIGNMENT

Jia-Wei Liao[1*] **Kuan-Yu Chen**[2†] **Yu-Chen Den**[2] **Darby Tien-Hao Chang**[2]
[1]National Taiwan University [2]SinoPac Holdings
d11922016@csie.ntu.edu.tw, {lavamore,abnerden,darby}@sinopac.com

ABSTRACT

Tabular foundation models show promise for structured data prediction, but unlike text and images, tabular datasets exhibit heterogeneous schemas and label semantics. This raises a critical question: Does mixing tables during few-shot training improve in-context learning (ICL)? We systematically investigate cross-table training under controlled few-shot protocols, comparing single-table training versus augmentation with auxiliary datasets. We identify severe negative transfer under naive mixing and propose two alignment strategies: feature-level matching via optimal transport (OT) and label semantics alignment via pseudo-labeling. Our key finding reveals an architectural divide: TabPFN-v2 and MITRA fail to benefit from cross-table augmentation, while representation-based models (TabICL) achieve +1.02% average improvement. This indicates that cross-table learning requires learned embedding spaces where semantic correspondences can be preserved across heterogeneous schemas.

1 INTRODUCTION

Foundation models have revolutionized machine learning through their ability to transfer knowledge across diverse datasets, achieving remarkable success in text (Singh et al., 2025; Comanici et al., 2025; Grattafiori et al., 2024; Guo et al., 2025) and vision (Inc, 2023; Esser et al., 2024; Labs, 2024) domains. However, developing similar foundation models for tabular data, which underlies critical decision-making systems in finance, healthcare, and online platforms, remains challenging.

Recently, in-context learning (ICL) approaches (Hollmann et al., 2023; 2025; Zhang et al., 2025b; Jingang et al., 2025; Grinsztajn et al., 2025; Zhang et al., 2025a) have emerged as a promising paradigm for tabular foundation models, enabling strong generalization by learning to predict from few-shot examples without task-specific fine-tuning. TabPFN (Hollmann et al., 2023; 2025) pioneered this direction by formulating tabular prediction as in-context inference: a Transformer conditions on both training and query samples to perform single-pass amortized Bayesian inference. Pretrained on synthetic tabular tasks sampled from a prior over data-generating processes, it learns a general mapping from few-shot datasets to predictions, achieving strong performance in low-data regimes without gradient-based fine-tuning. Building on this paradigm, MITRA (Zhang et al., 2025b) enhances robustness through a principled mixture of synthetic priors, while TabICL (Jingang et al., 2025) explores representation-space ICL by encoding tabular samples into a shared embedding space where support and query instances interact through attention.

Unlike conventional gradient boosting tree models (Chen & Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018) and several deep tabular models (Gorishniy et al., 2021; Chen et al., 2024; Gorishniy et al.), which learns a global mapping from features to labels, ICL-based tabular foundation models operate in a fundamentally different manner. Their predictions are formed by dynamically weighting and aggregating relevant support samples through attention, effectively performing similarity-based inference within each task. This mechanism enables flexible adaptation to

[*]Work done during an internship at SinoPac Holdings.
[†]Project leader.

new datasets under limited supervision, but also makes performance highly dependent on the quality of similarity relationships among support samples.

In practice, real-world tabular learning frequently encounters data scarcity, making the use of auxiliary tables an attractive augmentation strategy. However, cross-table augmentation is inherently difficult due to the heterogeneity of schemas and feature semantics across datasets. Because ICL models such as TabPFN and MITRA rely on attention to weight and aggregate relevant support samples, naive mixing forces the model to compute similarities across misaligned feature dimensions. This distortion of the latent similarity space can lead the attention mechanism to select misleading support samples, ultimately resulting in negative transfer where auxiliary data degrades rather than enhances performance (Pan & Yang, 2009). This raises a fundamental research question: *Does cross-table data augmentation help or hurt ICL tabular foundation models in few-shot settings?*

To answer this question, we study cross-table training under few-shot regimes and identify key failure modes of naive table mixing. To mitigate these issues, we investigate two complementary alignment strategies: (i) feature-level alignment via optimal transport (OT) (Villani et al., 2008) column matching, which leverages distributional statistics to align heterogeneous feature spaces, and (ii) label semantics alignment via pseudo-labeling, which projects auxiliary data into the target label space. Our empirical findings reveal that naive cross-table mixing provides inconsistent benefits and can substantially degrade performance under high schema heterogeneity. In contrast, proper alignment enables representation-based ICL to substantially improve few-shot accuracy with statistical significance. Notably, TabICL, a representation-based ICL model, achieves the largest improvements from cross-table alignment, gaining 1.02% average accuracy across 11 datasets, demonstrating that learned embeddings can effectively bridge heterogeneous table schemas when proper correspondences are established.

2 METHOD

We study cross-table data mixture for ICL-based tabular foundation models in a few-shot setting. Let A denote the target table with d_A feature columns and B denote an auxiliary table with d_B feature columns. Our objective is to enhance the model's predictive performance on A by leveraging the information present in B during training. We define $(X_A^{\text{tr}}, y_A^{\text{tr}})$, $(X_B^{\text{tr}}, y_B^{\text{tr}})$ as the respective training sets for A and B. where the j-th feature column of a dataset is denoted as $X^{\text{tr},(j)}$. Under this framework, we evaluate two primary training paradigms. Our cross-table training objective is to incorporate auxiliary information from B while maintaining the label semantics of A. We consider the following variants: (i) Baseline: The model is trained exclusively on $(X_A^{\text{tr}}, y_A^{\text{tr}})$. (ii) Mixing $A + B$: The model is jointly trained on both A and B, both variants are evaluated on the target set X_A^{ts}. In practice, naive mixing is often suboptimal due to schema heterogeneity, which feature columns between A and B lack semantic alignment, which can distort the similarity relationships essential for ICL and induce negative transfer. To mitigate these effects, we introduce feature-level and label-level alignment strategies designed to preserve the target task's semantic integrity while maximizing auxiliary gain.

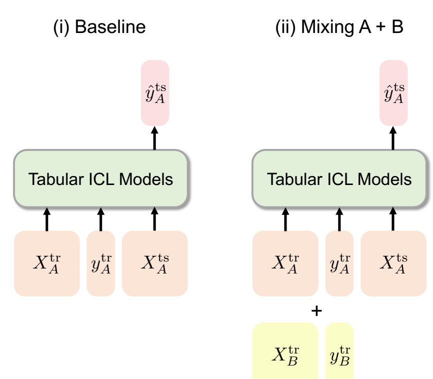


Figure 1: Comparison between single-table (Baseline) and cross-table training (Mixing $A+B$) for tabular ICL models. Left: Standard setting using only target table A. Right: Cross-table setting incorporating auxiliary table B.

2.1 FEATURE-LEVEL ALIGNMENT

To investigate the effect of feature correspondence across heterogeneous tables, we consider two feature-matching strategies. As a naive comparison method, we first introduce random feature matching, where the column order of the auxiliary table B is randomly permuted to match the dimensionality of the target table A. This produces arbitrary column-to-column correspondences without enforcing any semantic or statistical alignment, serving as a control setting to evaluate whether cross-table gains can arise from data mixture alone.

To obtain a principled schema alignment, we further propose an optimal transport (OT) feature matching (Villani et al., 2008). For each column $X_A^{\text{tr},(j)}$, we compute a distributional descriptor vector $m_A^{(j)} = \psi(X_A^{\text{tr},(j)})$, where $\psi(\cdot)$ extracts the first $p = 4$ moments $\mathbb{E}[(X_A^{\text{tr},(j)})^k]$ for $k = 1, ..., p$. Similarly, for B we compute $\{m_B^{(k)}\}_{k=1}^{d_B}$. We define an OT cost matrix $C \in \mathbb{R}^{d_A \times d_B}$ by $C_{jk} = \left\| m_A^{(j)} - m_B^{(k)} \right\|_2^2$. We solve the OT problem

$$\Gamma^\star = \underset{\Gamma \in \Pi(p_A, p_B)}{\operatorname{argmin}} \ \operatorname{trace}(\Gamma^\top C),$$

where $\Gamma \in \mathbb{R}_+^{d_A \times d_B}$ is the transport plan, $\Pi(p_A, p_B)$ denotes the set of couplings with uniform marginals $p_A = \mathbf{1}_{d_A}/d_A$ and $p_B = \mathbf{1}_{d_B}/d_B$. Given $\Gamma^\star$, we derive a hard schema alignment by assigning each column of A to the most strongly matched column of B: $\pi(j) = \operatorname{argmax}_k \Gamma^\star_{jk}$. We then construct a schema-aligned auxiliary table $\tilde{X}_B^{\text{tr}}$ by permuting the columns of X_B^{tr} according to the induced matching $\pi(\cdot)$, such that $\tilde{X}_B^{\text{tr}}$ follows the column ordering of A. This procedure aligns heterogeneous schemas by matching each column in A with its closest counterpart in B under the OT coupling, resulting in a discrete column-to-column correspondence.

2.2 Label Semantics Alignment

Besides schema mismatch, cross-table training also suffers from label semantics mismatch: the label space of B may not correspond to the target task defined by A. To address this, we propose to a method to align the supervision of B to A by pseudo-labeling. Specifically, we first train a surrogate model f_θ on dataset A only, and then generate pseudo labels $\hat{y}_B^{\text{tr}} = f_\theta(X_B^{\text{tr}})$ for dataset B. We then substitute the original labels y_B^{tr} with $\hat{y}_B^{\text{tr}}$ during cross-table training. This effectively projects the auxiliary dataset into the target label semantics, trading label noise for improved semantic consistency. Combining both feature and label alignment, we train on the mixed dataset $([X_A^{\text{tr}}, \tilde{X}_B^{\text{tr}}], [y_A^{\text{tr}}, \hat{y}_B^{\text{tr}}])$.

3 Experiment

We consider the following training variants in our experiments: (i) Baseline: train on table A only; (ii) Mixing $A + B$: mix training data from A and B directly. We also compare random vs OT and original label vs pseudo label.

3.1 Experimental Setup

We conduct experiments on 11 tabular classification tasks from the OpenML benchmark (Bischl et al., 2021). For each task, we consider a target table A and optionally incorporate an auxiliary table B for cross-table training. In cases where the number of columns in tables A and B differ, we apply padding with zeros to ensure uniformity in column dimensions during training. We evaluate three representative tabular foundation models: TabPFNv2 (Hollmann et al., 2025), MITRA (Zhang et al., 2025b), and TabICL (Jingang et al., 2025). All results are averaged over 30 random seeds to account the performance variability induced by few-shot sampling. In the few-shot setting, we randomly sample 100 labeled training instances from dataset A. When cross-table training is enabled, we additionally sample 100 labeled training instances from dataset B. For evaluation, we use the original number of query instances provided by OpenML. For datasets with more than 10,000 query instances, we subsample 10,000 evaluation samples to reduce computational cost.

3.2 Evaluation Protocol

To quantify the effect of cross-table training, we evaluate performance on all ordered dataset pairs. Specifically, given a collection of n tabular datasets, we treat each dataset $i \in \{1, \ldots, n\}$ as the target dataset and each dataset $j \in \{1, \ldots, n\}, j \neq i$ as an auxiliary dataset. We denote by $\text{Acc}_{i,j}$ the test accuracy on target dataset i when the model is trained using the cross-table setting with

Metrics	TabPFN-v2		MITRA		TabICL	
	Random	OT	Random	OT	Random	OT
Improved Accuracy	-0.15% ± 0.31%	-0.17% ± 0.34%	-0.29% ± 0.66%	0.08% ± 0.62%	1.02% ± 0.43%	0.80% ± 0.31%
Improved Ratio	-0.08% ± 0.52%	-0.10% ± 0.56%	-0.10% ± 1.15%	0.42% ± 1.04%	1.66% ± 0.66%	1.33% ± 0.48%

Table 1: Comparison of performance improvement under **random feature matching** and **OT feature matching** across different methods when using pseudo labels. Values are reported as mean ± standard deviation ($\mu \pm \sigma$) over 30 random seeds. Orange boxes highlight cases where $\mu - \sigma > 0$.

Metrics	TabPFN-v2		MITRA		TabICL	
	Original	Pseudo	Original	Pseudo	Original	Pseudo
Improved Accuracy	-0.93% ± 0.40%	-0.15% ± 0.31%	-1.03% ± 0.64%	-0.29% ± 0.66%	0.78% ± 0.45%	1.02% ± 0.43%
Improved Ratio	-1.26% ± 0.62%	-0.08% ± 0.52%	-1.25% ± 1.05%	-0.10% ± 1.15%	1.34% ± 0.69%	1.66% ± 0.66%

Table 2: Comparison of performance improvement under **original labels** and **pseudo labels** across different methods when using random matching. Values are reported as mean ± standard deviation ($\mu \pm \sigma$) over 30 random seeds. Orange boxes highlight cases where $\mu - \sigma > 0$.

auxiliary dataset j (i.e., A+B training). We further denote by Baseline_i the test accuracy on target dataset i under the baseline setting trained using only dataset i (A-only). Using these quantities, we define the **Improved Accuracy** and **Improved Ratio** in Appendix A.1.

3.3 RESULT

Finding 1

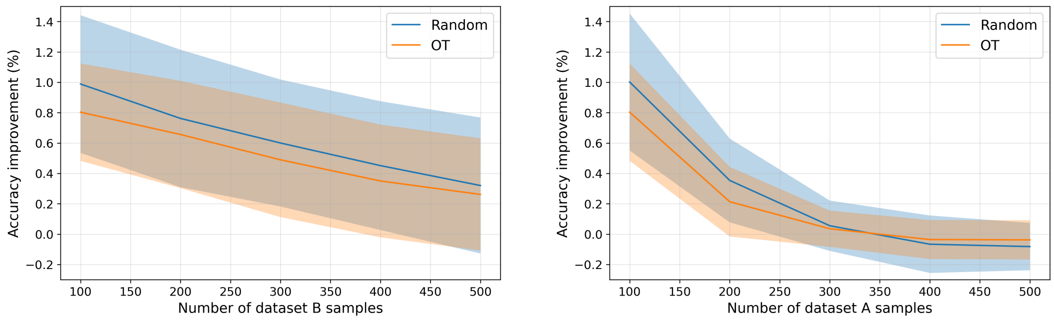
Representation-based ICL can effectively leverage auxiliary tables by learning shared semantic embeddings across heterogeneous schemas.

Table 1 shows that TabPFN-v2 and MITRA both degrade when augmented with auxiliary tables, with accuracy drops of 0.15%-0.29% under random matching and near-zero or negative gains under OT-based alignment. This consistent negative transfer across both models indicates that they learn largely invariant to cross-table information, rendering feature alignment ineffective. In contrast, TabICL achieves significant performance gains under both matching strategies (random: 1.02%, OT: 0.80%; $p < 0.001$, t-test). More than 75% of cross-table pairs exhibit improvements, indicating robust generalization. Furthermore, although TabICL with OT-based feature matching achieves slightly lower average performance gains than random matching, it yields substantially lower standard deviation, indicating more stable and reliable cross-table transfer across datasets. This contrast reveals a fundamental architectural distinction: *representation-based in-context learning enables effective cross-table knowledge transfer by operating in a learned embedding space where semantic correspondences across heterogeneous tables can be preserved and leveraged*. Comprehensive cross-table performance results are provided in Appendix Tables 4 and 5.

Finding 2

Pseudo-labeling effectively mitigates label semantics heterogeneity for representation-based ICL models.

As shown in Table 2, pseudo-labeling reduces negative transfer from cross-table augmentation, its effectiveness varies substantially across model architectures. TabPFN-v2 and MITRA experience persistent performance degradation even with pseudo-labeling, indicating that non representation-based ICL models struggle with heterogeneous auxiliary data. However, TabICL, a representation-based ICL model, demonstrate exceptional capacity to leverage alignment: pseudo-labeling further increases average accuracy gains from 0.78% to 1.02%, benefiting 88 out of 110 dataset pairs. These findings reveal that pseudo-labeling effectively bridges label semantics heterogeneity, but only representation-based models possess the inductive bias to fully exploit this alignment.



(a) Effect of B sample size on performance improvement on A.

(b) Effect of A sample size on performance improvement on A.

Figure 2: Effect of different sample sizes on performance improvement on A. Shaded regions indicate one standard deviation.

Finding 3

Cross-table augmentation exhibits diminishing marginal benefit for representation-based ICL models as data availability increases, limiting its applicability to few-shot learning scenarios.

We further investigate the impact of varying sample sizes on performance, as shown in Figure 2. Increasing target table A samples dramatically attenuates TabICL's cross-table augmentation gains, with improvements declining from 1.02% to near-zero or slightly negative as target sample size grows. This decline suggests that auxiliary tables provide critical signal primarily when target data is scarce; once sufficient in-domain examples are available, auxiliary information becomes redundant and may introduce heterogeneity-induced noise. In contrast, increasing auxiliary table samples moderates the decline in improvement, with gains remaining positive, suggesting that auxiliary data heterogeneity becomes less detrimental when more target data are available. These findings highlight that, for TabICL cross-table augmentation is most effective in few-shot settings, where data scarcity necessitates external information to improve model generalization.

4 Conclusion

This work empirically studies cross-table augmentation for tabular foundation models under in-context learning (ICL). Our key findings identify three critical principles: (i) representation-based ICL uniquely leverages auxiliary tables through learned embeddings; (ii) pseudo-labeling effectively bridges label semantics heterogeneity for representation-based models, achieving 1.02% average improvement; and (iii) cross-table augmentation exhibits diminishing returns as in-domain data increases, establishing that this strategy is optimally suited for few-shot learning. Together, these findings establish representation-based ICL as a powerful paradigm for cross-table learning which underscores the substantial untapped potential of representation-based ICL for tabular learning.

References

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael Gomes Mantovani, Jan N van Rijn, and Joaquin Vanschoren. Openml benchmarking suites. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

KuanYu Chen, Ping-Han Chiang, Hsin-Rung Chou, Chih-Sheng Chen, and Tien-Hao Chang. Dofen: Deep oblivious forest ensemble. *Advances in Neural Information Processing Systems*, 37:44624–44677, 2024.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. Tabm: Advancing tabular deep learning with parameter-efficient ensembling. In *The Thirteenth International Conference on Learning Representations*.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint*, 2024.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, et al. Tabpfn-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 2025.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*, 2023.

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*, 2025.

Midjourney Inc. Midjourney, 2023. URL `https://www.midjourney.com/`.

QU Jingang, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. In *International Conference on Machine Learning (ICML)*, 2025.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NIPS)*. 2017.

Black Forest Labs. Official weights of FLUX.1 dev, 2024. URL `https://huggingface.co/black-forest-labs/FLUX.1-dev`.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 2009.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint*, 2025.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

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*, 2025a.

Xiyuan Zhang, Danielle C Maddix, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael W Mahoney, et al. Mitra: Mixed synthetic priors for enhancing tabular foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

A APPENDIX

A.1 EVALUATION METRICS

We define the metrics as the follows:

Improved Accuracy. To measure the average absolute performance gain brought by cross-table training, we compute:

$$\text{Improved Accuracy} = \frac{1}{n(n-1)} \sum_{i \neq j} (\text{Acc}_{i,j} - \text{Baseline}_i)$$

This metric captures the magnitude of transfer, and it penalizes negative transfer when auxiliary datasets degrade performance.

Improved Ratio. Since different datasets may have different baseline difficulty levels, absolute improvements are not always comparable. We therefore additionally report the relative improvement normalized by the baseline performance:

$$\text{Improved Ratio} = \frac{1}{n(n-1)} \sum_{i \neq j} \left(\frac{\text{Acc}_{i,j} - \text{Baseline}_i}{\text{Baseline}_i} \right)$$

This metric measures the average percentage gain over the baseline and allows a fairer comparison across target datasets with varying baseline accuracies.

A.2 DATASET INFORMATION

Task ID	Dataset	#Column	#Testing Data	Test Label Ratio
361055	credit	10	4700	0.508
361060	electricity	7	10000	0.512
361061	covertype	10	10000	0.507
361062	pol	26	2118	0.512
361063	house_16H	16	2833	0.511
361065	MagicTelescope	10	2810	0.512
361066	bank-marketing	7	2222	0.510
361069	Higgs	24	10000	0.521
361070	eye_movements	20	1599	0.518
361273	Diabetes130US	7	10000	0.515
361275	default-of-credit-card-clients	20	2788	0.514

Table 3: Dataset information.

A.3 BASELINE SCORE OF TABICL AND CROSS-TABLE IMPROVEMENT

B DISCUSSION AND FUTURE WORK

Our results indicate that representation-based in-context learning (ICL) shows strong potential for few-shot cross-table learning. Models that operate in a learned representation space, such as TabICL, are better able to leverage auxiliary tables through alignment and achieve consistent and statistically significant improvements. These findings suggest that similarity in a learned representation space plays a key role in enabling effective knowledge transfer across heterogeneous tabular datasets.

For future work, an important direction is to identify which auxiliary data are most beneficial for a given target table. Not all auxiliary tables contribute equally, and indiscriminate data mixture may introduce noise or negative transfer. Designing methods to select or weight auxiliary samples based on their relevance to the target task, such as representation-space feature matching or data valuation strategies, may further improve the effectiveness and scalability of cross-table tabular foundation models.

B \ A	361055	361060	361061	361062	361063	361065	361066	361069	361070	361273	361275
361055	0.6997	0.7374	0.6776	0.8464	0.8246	0.7665	0.7407	0.5659	0.5208	0.5602	0.6655
361060	0.7333	0.7241	0.6795	0.8414	0.8238	0.7710	0.7410	0.5691	0.5247	0.5619	0.6672
361061	0.7335	0.7364	0.6588	0.8453	0.8234	0.7708	0.7415	0.5671	0.5227	0.5615	0.6655
361062	0.7305	0.7410	0.6819	0.8739	0.8241	0.7710	0.7421	0.5704	0.5247	0.5630	0.6715
361063	0.7345	0.7418	0.6793	0.8520	0.8227	0.7693	0.7410	0.5720	0.5248	0.5644	0.6652
361065	0.7338	0.7372	0.6834	0.8448	0.8235	0.7677	0.7403	0.5698	0.5235	0.5629	0.6677
361066	0.7326	0.7382	0.6774	0.8426	0.8255	0.7701	0.7128	0.5654	0.5253	0.5604	0.6664
361069	0.7345	0.7461	0.6896	0.8566	0.8246	0.7757	0.7443	0.5726	0.5285	0.5706	0.6708
361070	0.7327	0.7426	0.6844	0.8544	0.8230	0.7700	0.7409	0.5722	0.5220	0.5638	0.6675
361273	0.7328	0.7378	0.6798	0.8407	0.8246	0.7697	0.7397	0.5668	0.5261	0.5366	0.6681
361275	0.7315	0.7395	0.6817	0.8498	0.8212	0.7691	0.7393	0.5766	0.5255	0.5639	0.6593

Table 4: TabICL cross-table performance with random feature matching. Rows denote different auxiliary tables B and columns denote different target tables A. The diagonal entries represent the baseline (training on A only). Green cells indicate cases where cross-table training improves performance over the baseline.

B \ A	361055	361060	361061	361062	361063	361065	361066	361069	361070	361273	361275
361055	0.6997	0.7407	0.6606	0.8385	0.8272	0.7688	0.7358	0.5680	0.5207	0.5505	0.6651
361060	0.7210	0.7241	0.6764	0.8415	0.8237	0.7667	0.7309	0.5677	0.5266	0.5544	0.6670
361061	0.7338	0.7408	0.6588	0.8441	0.8259	0.7697	0.7434	0.5675	0.5255	0.5589	0.6661
361062	0.7276	0.7456	0.6789	0.8739	0.8246	0.7675	0.7428	0.5702	0.5213	0.5590	0.6657
361063	0.7332	0.7417	0.6815	0.8470	0.8227	0.7704	0.7453	0.5705	0.5238	0.5642	0.6655
361065	0.7361	0.7404	0.6765	0.8403	0.8209	0.7677	0.7419	0.5645	0.5244	0.5610	0.6664
361066	0.7201	0.7341	0.6680	0.8356	0.8252	0.7662	0.7128	0.5611	0.5233	0.5419	0.6637
361069	0.7358	0.7459	0.6905	0.8570	0.8259	0.7760	0.7478	0.5726	0.5284	0.5732	0.6669
361070	0.7322	0.7434	0.6820	0.8471	0.8252	0.7718	0.7457	0.5698	0.5220	0.5614	0.6621
361273	0.7179	0.7349	0.6687	0.8396	0.8253	0.7692	0.7166	0.5669	0.5235	0.5366	0.6642
361275	0.7327	0.7417	0.6746	0.8463	0.8228	0.7651	0.7398	0.5742	0.5222	0.5647	0.6593

Table 5: TabICL cross-table performance with OT matching. Rows denote different auxiliary tables B and columns denote different target tables A. The diagonal entries represent the baseline (training on A only). Green cells indicate cases where cross-table training improves performance over the baseline.