Block Redundancy in Tabular Foundation Models for Clinical Applications

Irene Cannistraci [1] **Julia E. Vogt** [1]

Abstract

In-context-learning Tabular Foundation Models (TFMs) are uniquely suited to clinical use because they require no training on patient data, however their architectures still demand GPU inference and discourage the on-premise deployment that hospitals need. Simpler alternatives (e.g. Gradient-Boosted Decision Tree) run on CPU and often match their performance, but at the cost of manual feature engineering, preprocessing, and hyperparameter tuning that transformer models avoid by design. In this paper we show that it is possible to simplify TFMs, by substituting up to $\sim$94% of the blocks using a closed-form linear translator, while preserving downstream performance and requiring minimal compute. We show results on eight clinical benchmarks of different sizes (including MIMIC-III and multi-center eICU-CRD) and three TabZilla controls, for classification (binary and multi-class) and regression.

1. Introduction

Tabular foundation models (e.g., TabPFNv2 (Hollmann et al., 2025), TabICL (QU et al., 2025), TabDPT (Ma et al., 2026), Mitra (Zhang et al., 2026)) are attractive for clinical use: the training set is passed as context at inference time, so a single pre-trained model serves every downstream task without gradient updates on the target dataset. The inference footprint, however, blocks their adoption: 12–16 transformer blocks, with diverse pre-training and attention mechanisms, demand GPUs that on-premise clinical environments typically do not have. Simpler alternatives such as Light Gradient-Boosting Machine (LightGBM) and Gradient-Boosted Decision Trees (GBDTs) run on CPU and often match or surpass deep tabular models on performance (Grinsztajn et al., 2022; McElfresh et al., 2023), but at a different cost: manual feature engineering, preprocessing, and per-dataset hyperparameter tuning that transformer models avoid by design. This similar performance suggests that TFM models could be over-parametrized, and that the redundancy observed in vision and language transformers (Dalvi et al., 2020; Cannistraci et al., 2024; Jacobs et al., 2026) might be present in this scenario as well. Tabular transformers, however, differ fundamentally from those settings: short sequences (8–74 tokens), heterogeneous feature types, no sequential grammar, making it non-obvious whether the same redundancy holds. In this paper we investigate whether TFMs are vastly overparametrized.

We evaluate four TFM across eleven datasets: eight clinical (including MIMIC-III (Johnson et al., 2016) and eICU-CRD (Pollard et al., 2018)) and three TabZilla (McElfresh et al., 2023)) controls for structural generality, spanning binary classification, multi-class classification, and regression. We first establish the existence of this redundancy by measuring per-block sensitivity and pairwise representation similarity. Next, we empirically demonstrate that this redundancy can be exploited: replacing contiguous blocks with a closed-form linear translator (Cannistraci et al., 2024) fit on just 500 calibration samples recovers the downstream Area Under the Receiver Operating Characteristic Curve (AUROC). Remarkably, this approach allows for extreme compression; as shown in Table 1, we can drop all but a single transformer block (e.g., approximating 15 of 16 layers) and still match or even exceed the full model's performance, whereas naive layer dropping results in catastrophic failure.

2. Method

Linear approximation error. Let $\mathbf{H}_l \in \mathbb{R}^{N\times d}$ denote the representation at depth $l \in \{0, 1, \ldots, L\}$: $\mathbf{H}_0$ is the tokenizer output and $\mathbf{H}_l$ for $l \geq 1$ is the output of block l (equivalently, the input to block $l{+}1$); L is the model's number of blocks. Following Cannistraci et al. (2024), given a start depth s and an end depth e with $0 \leq s < e \leq L$, we ask how well the contiguous stack of blocks $s{+}1, \ldots, e$ (taking $\mathbf{H}_s$ as input and producing $\mathbf{H}_e$ as output) can be replaced by a single linear map. We define $\varepsilon(s, e)$ as the relative residual of the best-fit linear map from

[1]Department of Computer Science, ETH Zurich. Correspondence to: Irene Cannistraci <irene.cannistraci@inf.ethz.ch>.

Accepted at the ICML 2026 Workshop on Structured Data for Health (SD4H), Seoul, South Korea. Copyright 2026 by the author(s).

$\mathbf{H}_s$ to $\mathbf{H}_e$:

$$\varepsilon(s,e) = \frac{\|\mathbf{H}_e - (\mathbf{H}_s\mathbf{W}^* + \mathbf{b}^*)\|_F}{\|\mathbf{H}_e\|_F}, \quad (\mathbf{W}^*, \mathbf{b}^*) = \arg\min_{\mathbf{W},\mathbf{b}} \|\mathbf{H}_e - (\mathbf{H}_s\mathbf{W} + \mathbf{b})\|_F^2, \tag{1}$$

where $\mathbf{b}^*$ is broadcast across the N rows. $(\mathbf{W}^*, \mathbf{b}^*)$ is computed in closed form via least squares on calibration representations extracted by forward hooks. A small $\varepsilon(s,e)$ means a single affine map closely reproduces what blocks $s{+}1, \ldots, e$ collectively compute, i.e. those blocks are linearly redundant.

Per-architecture translators. Equation 1 fits a single $(\mathbf{W}, \mathbf{b})$ that is shared across all positions, which suits single-axis self-attention (TabICL, TabDPT). Dual-axis attention models (TabPFNv2, Mitra) attend along both items and features, so a single shared map mixes the two axes and destroys the per-feature correspondence between $\mathbf{H}_s$ and $\mathbf{H}_e$. For these models we instead fit a separate $(\mathbf{W}_f, \mathbf{b}_f)$ per feature index f on tensors of shape $(\cdot, n_{\text{features}}, d)$, applied as $\mathbf{H}'[\cdot, f, \cdot] = \mathbf{H}[\cdot, f, \cdot]\,\mathbf{W}_f + \mathbf{b}_f$; for Mitra, whose blocks process the in-context support and query examples as a tuple of two such tensors, we additionally fit one per-feature translator per stream (Appendix A.5). With $n_{\text{features}}{=}1$ the per-feature parameterisation reduces exactly to the single-$(\mathbf{W}, \mathbf{b})$ case.

Block approximation. At inference time the trained affine translator $(\mathbf{W}^*, \mathbf{b}^*)$ replaces blocks $s{+}1, \ldots, e$: input $\mathbf{H}_s$ is mapped directly to $\mathbf{H}_e$, and the remaining blocks $e{+}1, \ldots, L$ together with the pre-trained classification head run unchanged. End-to-end extraction and the least-squares solve take 33–109 ms on a single GPU (calibration size $N{=}500$ items) and require no gradients, backpropagation, or parameter updates. This operation has two uses: with $w = e - s = 1$ we replace a single block at a time to measure each block's individual importance (per-block sensitivity); with $w \geq 2$ we replace a contiguous stack of blocks at once for higher compression (multi-block approximation).

3. Experiments

Setup. We study four pre-trained tabular foundation models with frozen weights: TabPFNv2, TabICL, TabDPT, and Mitra. We evaluate on eight clinical datasets, six binary (Diabetes (Kahn), Heart (Janosi et al., 1989), Breast Cancer (Zwitter & Soklic, 1988), SUPPORT2 (Investigators et al., 1995), MIMIC-III (Johnson et al., 2016), eICU-CRD (Pollard et al., 2018)), and two three-class (Cardiotocography (Campos & Bernardes, 2000), Thyroid (Quinlan, 1986)), and three non-medical TabZilla controls (McElfresh et al., 2023) (Jannis, Higgs, Covertype, subsampled to 30K each). Each dataset is split into train, calibration and test, where the calibration split is used to fit the linear translator and to select the skip range, never to evaluate. We report AUROC, macro-averaged one-vs-rest for the multi-class datasets. Regression on MIMIC-III length-of-stay is reported in Table 2. Dataset details and GBDT hyperparameters, as well as other implementation details are in Section A.

3.1. Analysis

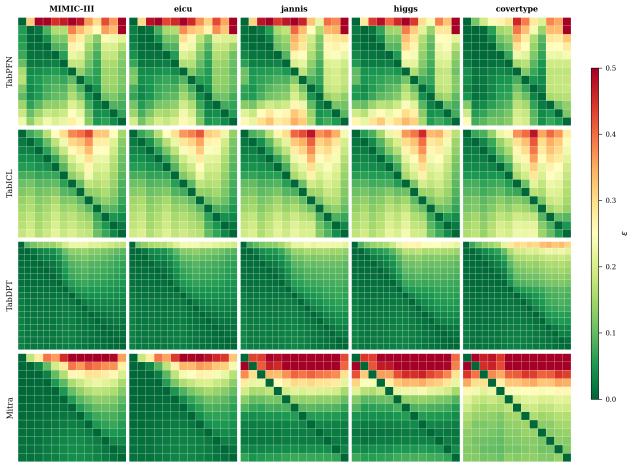

Figure 1. Pairwise linear approximation error $\varepsilon(s,e)$ (Eq. 1) on the largest clinical (MIMIC-III, eICU-CRD) and TabZilla (Jannis, Higgs, Covertype) datasets, and the four TFMs.

Figure 1 reports $\varepsilon(s,e)$ for all block pairs, where rows are the TFMs and columns the datasets. The figure shows that for some models (e.g., TabDPT and TabICL) representations are more similar to each others, while for example for Mitra, they are more difficult to approximate. In general, results are model dependent rather than dataset dependent, and the earlier blocks do most of the work. After the first blocks, TabDPT and TabICL saturate completely with an $\varepsilon < 0.06$.

We then evaluate whether this behaviour reflects on the downstream AUROC. Therefore we select the skip window (s,e) using the AUROC on the held-out calibration split: 500 items, the same validation split already used for hyperparameter and model selection. The procedure consumes no additional patient data and never touches test labels.

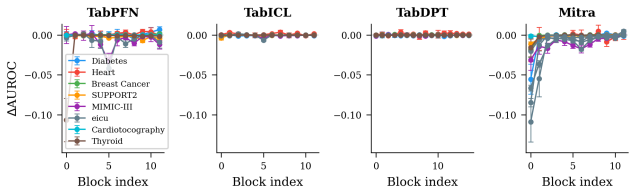

Figure 2. Per-block sensitivity (ΔAUROC when one block is replaced by a linear translator). Uniform y-axis across panels.

Figure 2 replaces each individual block with a linear translator and measures ΔAUROC. TabICL and TabDPT are entirely flat ($|\Delta| < 0.005$): no individual block is critical. TabPFNv2 shows scattered sensitivity, while Mitra shows higher sensitivity in the first blocks.

Figure 3 shows the best approximation window distribution for each model when considering 3 or 4 blocks. Block 0 is preserved across nearly every dataset (yellow color), and TabDPT shows no highly preferred approximation position over blocks 1 to $L-1$, indicating that for this model it is possible to pick *any* contiguous later window without per-dataset tuning.

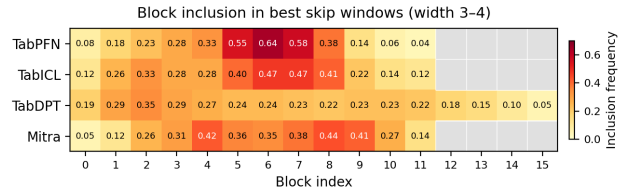

Figure 3. Block inclusion frequency in the best approximation windows (width 3-4) across datasets.

Attention NLP transformer heads' attention distributions tend to differentiate across layers (Dalvi et al., 2020), whereas we show that tabular foundation model heads look comparatively flat across depth (Section B.5). Figure 4 reports TabICL per-block mean attention entropy on surviving blocks before and after replacing a contiguous block window with a linear translator; the curves overlap closely highlighting that our approximation procedure does not meaningful change the attention distribution (Section B.6).

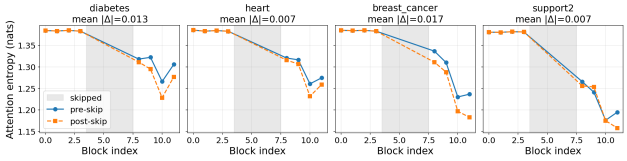

Figure 4. TabICL per-block mean attention entropy on surviving blocks before vs. after replacing blocks from 4 to 8 with a fitted linear translator (skip window shaded). Entropy in nats.

3.2. Results

Table 1 reports all regimes side-by-side. We tested approximating 3 or 4 blocks, 8 blocks and maximum-compression. Additionally we tested whether a simple identity function is enough instead of the linear approximator.

Approximating 3 or 4 contiguous blocks (Best Approx.) preserves accuracy in most cases. The best window is selected with the calibration-set AUROC as shown in Figure 3. On smaller clinical datasets (Diabetes, Heart, Breast Cancer) the approximation is essentially lossless or slightly improves AUROC (e.g., TabPFN on Diabetes improves from .841 to .844, and Mitra on Breast Cancer from .994 to .995). TFMs outperform GBDTs on Heart (Mitra .906 vs. XGBoost .879) and Breast Cancer; GBDTs remain best on the largest tabular benchmark (Covertype, XGBoost .979 vs. TabICL Best Approx. .961), consistent with the established large-data picture (Grinsztajn et al., 2022).

Approx. max scales aggressively to a single retained block. The degradation follows a clear complexity gradient: smaller datasets remain nearly lossless at any compression level, while MIMIC-III, SUPPORT2, eICU and the larger TabZilla benchmarks degrade gently at extreme widths. For TabDPT the retained block is block 0 on $10/11$ datasets, confirming that block 0 carries the dominant computation under single-axis self-attention; for TabICL it is block 0 on $7/11$. Under the per-feature translator that respects dual-axis attention, TabPFNv2 and Mitra select later blocks more often (block 0 retained on $1/11$ and $3/11$ respectively): the translator equalises the contribution of different blocks, so any contiguous window can be replaced.

The approximation is essential. Naive layer dropping (i.e., Identity) at the same maximal compression collapses every model AUROC. For example, dropping 11 blocks with Identity on TabPFN for the Heart dataset plummets the AUROC to .428, whereas our closed-form linear translator (Approx. max) preserves it at .861 (Table 1).

Ablations. We additionally test whether compressed models remain robust under random feature corruption and structured clinical missingness (all-vitals or all-labs dropped) (Sections C.2 and C.3), and if per-patient ε is uniform across outcome subgroups (Section C.1), confirming that compression does not introduce subpopulation-specific degradation. Furthermore, we show that *early exit* is uniformly worse (Section C.4). Indeed, a natural alternative is to discard the pre-trained head and refit a logistic regression on intermediate representations. Across all configurations, block approximation, which retains the pre-trained head and bridges $\mathbf{H}_L \to \mathbf{H}_N$ with a linear translator, matches or beats early-exit.

Regression results. To verify that block redundancy is not an artefact of binary classification, we evaluate block approximation on a regression task: predicting ICU length-of-stay (LOS) in hours from the same MIMIC-III cohort. We use the regressor variants of all four ICL foundation models with the same protocol: linear translator, best skip position per width selected by R^2. Block redundancy persists under regression: TabICL is near-lossless at widths 3 and 4 ($\Delta R^2 = -0.004$), TabDPT degrades by at most -0.013, Mitra by at most -0.005, and TabPFNv2 degrades more visibly. The finding that block redundancy extends beyond classification confirms it is a structural property of TFMs, not an artefact of binary decision boundaries.

Table 1. Downstream AUROC results for GBDT baselines and four foundation models. Per model: **Full** = pre-trained model unchanged; **Best Approx.** = best contiguous approximation window of width 3–4 (mild compression); **Linear** (w/L) = linear translator at intermediate compression; **Approx. max** = linear translator at maximum compression (11 of 12 blocks approximated, or 15 of 16 for TabDPT); **Identity** = naive layer dropping at the same maximum compression (no translator). All approximation windows are selected by calibration-set AUROC. **Bold**: best per column. Underline: approximate variant $\geq$ Full. *Italic*: degrades $> 1\%$ from Full.

Model		Diabetes	Heart	Br. Cancer	SUPPORT2	MIMIC	eICU	CTG†	Thyroid†	Jannis†	Higgs	Covertype‡
XGBoost		.832±.026	.879±.031	.994±.003	.981±.002	**.805±.007**	**.792±.005**	.997±.002	**1.00±.000**	.860±.005	.788±.007	**.979±.001**
LightGBM		.838±.016	.876±.026	.994±.002	.981±.002	.804±.007	.791±.005	.998±.002	**1.00±.000**	.858±.004	.789±.006	.976±.002
TabPFN	Full	.841	.897	.994	.982	.758	.758	.998	1.00	.846	.780	.957
	Best Approx.	.844	.900	.994	.980	.764	.763	.998	.999	.838	.778	.954
	Linear (8/12)	*.824*	*.883*	.994	.980	.764	.760	.997	.999	.842	.778	.952
	Approx. max (11/12)	*.831*	*.861*	.994	*.948*	*.710*	*.728*	*.983*	*.912*	*.756*	*.621*	*.807*
	Identity (11/12)	*.632*	*.428*	*.229*	*.492*	*.441*	*.441*	*.460*	*.490*	*.474*	*.463*	*.447*
TabICL	Full	**.845**	.904	.994	**.983**	.800	.786	.999	1.00	**.866**	**.790**	.962
	Best Approx.	.843	.904	.995	.981	.793	.783	**.999**	1.00	.863	.787	.961
	Linear (8/12)	.845	**.907**	.995	.979	.796	.781	.999	.999	.858	.785	.959
	Approx. max (11/12)	.838	.903	.994	.974	*.785*	.777	.999	.999	*.843*	*.768*	*.943*
	Identity (11/12)	*.725*	*.888*	.990	*.926*	*.669*	*.695*	*.897*	*.853*	*.697*	*.619*	*.659*
TabDPT	Full	.841	.898	.994	.979	.761	.757	.996	.998	.813	.734	.942
	Best Approx.	.841	.894	.993	.978	.760	.757	.998	.998	.812	.734	.940
	Linear (11/16)	.842	.891	.994	.975	.757	.749	.998	.997	.807	*.720*	.935
	Approx. max (15/16)	.842	.905	.994	.971	*.751*	*.747*	.996	.990	*.792*	*.692*	*.918*
	Identity (15/16)	*.432*	*.716*	*.623*	*.453*	*.527*	*.466*	*.578*	*.370*	*.528*	*.487*	*.460*
Mitra	Full	.839	**.906**	.994	.980	.781	.772	.998	.999	.828	.773	.927
	Best Approx.	*.819*	.904	**.995**	.979	.780	.770	.997	.999	.822	.771	.927
	Linear (8/12)	.834	.899	.993	.975	*.770*	*.759*	.996	.999	.801	*.741*	.922
	Approx. max (11/12)	*.739*	*.816*	.988	*.937*	*.696*	*.708*	*.984*	*.945*	*.694*	*.584*	*.837*
	Identity (11/12)	*.440*	*.378*	*.157*	*.493*	*.433*	*.409*	*.266*	*.555*	*.509*	*.500*	*.301*

†Macro-averaged OVR AUROC (CTG, Thyroid 3-class; Jannis 4-class). ‡Covertype 7-class.

Table 2. Regression block approximation on MIMIC-III LOS prediction.

Model	Full R^2	Approx. (w=3)	Approx. (w=4)
XGBoost	.204	—	—
LightGBM	.209	—	—
TabPFN	.172	.134	.110
TabICL	.200	**.196**	**.196**
TabDPT	.144	.131	.134
Mitra	.169	.164	.165

4. Conclusion and future work

In this work, we addressed a critical bottleneck for the adoption of TFMs in clinical environments: the high inference cost of deep transformer architectures that precludes CPU-only, on-premise deployment. We demonstrated that pre-trained TFMs (such as TabPFNv2, TabICL, TabDPT, and Mitra) are vastly over-parametrized for tabular tasks. By substituting up to $\sim$94% of contiguous transformer blocks with a simple, closed-form linear translator (fitted on just 500 calibration samples) we successfully recovered downstream AUROC across diverse clinical and general tabular benchmarks. Remarkably, our approach allows state-of-the-art TFMs to be compressed to a single active transformer block, avoiding the catastrophic failure associated with naive layer dropping (Identity) and consistently outperforming early-exit strategies.

Our findings carry significant practical implications. By drastically reducing the computational footprint of TFMs, we bridge the gap between the zero-shot, tuning-free convenience of foundation models and the lightweight inference of traditional methods like GBDTs. This makes local, privacy-preserving deployment of modern tabular models more feasible for hospitals with constrained hardware. Furthermore, our analysis of layer-wise sensitivity and attention entropy reveals fundamentally different representational dynamics compared to NLP or vision transformers, highlighting that early blocks perform the vast majority of the meaningful computation in tabular contexts.

Future work. The extreme compressibility of current TFMs naturally raises questions about their architectural design. Future work should investigate whether tabular foundation models can be explicitly pre-trained to be inherently shallower, or if this massive over-parametrization is a strict prerequisite for successful optimization during the pre-training phase. Additionally, we plan to extend our evaluation to real-time inference latency on standard clinical hardware, and explore whether structurally dynamic routing could allow TFMs to adapt their depth dynamically based on the complexity of individual patient records. Ultimately, unlocking lightweight, CPU-friendly inference for TFMs will be instrumental in bringing their benefits directly to the bedside.

References

Arik, S., Pfister, T., et al. Attentive interpretable tabular learning. *arXiv preprint arXiv:1908.07442*, 20, 2019.

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., and Hoffman, J. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Campos, D. and Bernardes, J. Cardiotocography. UCI Machine Learning Repository, 2000. DOI: https://doi.org/10.24432/C51S4N.

Cannistraci, I., Antonelli, S., Palumbo, E., Sutter, T. M., Rodolà, E., Rieck, B., and Vogt, J. E. Toast: Transformer optimization using adaptive and simple transformations. *arXiv preprint arXiv:2410.04941*, 2024.

Chen, J., Yan, J., Chen, Q., Chen, D. Z., Wu, J., and Sun, J. Excelformer: A neural network surpassing gbdts on tabular data. *arXiv preprint arXiv:2301.02819*, 2023.

Dalvi, F., Sajjad, H., Durrani, N., and Belinkov, Y. Analyzing redundancy in pretrained transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4908–4926, 2020.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34: 18932–18943, 2021.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Investigators, S. P. et al. A controlled trial to improve care for seriously hospitalized patients: The study to understand prognoses and preferences for outcomes and risks of treatments. *Jama*, 274:1591–1598, 1995.

Jacobs, M., Fel, T., Hakim, R., Brondetta, A., Ba, D. E., and Keller, T. A. Block recurrent dynamics in vision transformers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gH3HhnfWLC.

Janosi, A., Steinbrunn, W., Pfisterer, M., and Detrano, R. Heart Disease. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C52P4X.

Jayawardhana, M., Dooley, S., Cherepanova, V., Wilson, A. G., Hutter, F., White, C., Goldstein, T., Goldblum, M., et al. Transformers boost the performance of decision trees on tabular data across sample sizes. *arXiv preprint arXiv:2502.02672*, 2025.

Johnson, A., Pollard, T., and Mark, R. MIMIC-III Clinical Database. *PhysioNet*, September 2016. doi: 10.13026/C2XW26. URL https://doi.org/10.13026/C2XW26. Version 1.4.

Kahn, M. Diabetes. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5T59G.

Ma, J., Thomas, V., Hosseinzadeh, R., Labach, A., Cresswell, J. C., Golestan, K., Yu, G., Caterini, A. L., and Volkovs, M. TabDPT: Scaling tabular foundation models on real data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=pIZxEOZCId.

McElfresh, D., Khandagale, S., Valverde, J., Ramakrishnan, G., Prasad, V., Goldblum, M., and White, C. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems*, 2023.

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):180178, 2018.

QU, J., Holzmüller, D., Varoquaux, G., and Morvan, M. L. TabICL: A tabular foundation model for in-context learning on large data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=0VvD1PmNzM.

Quinlan, R. Thyroid Disease. UCI Machine Learning Repository, 1986. DOI: https://doi.org/10.24432/C5D010.

Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., and Goldstein, T. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

Song, J., Oh, K., Kim, T., Kim, H., Kim, Y., and Kim, J.-J. Sleb: Streamlining llms through redundancy verification and elimination of transformer blocks. *arXiv preprint arXiv:2402.09025*, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zeng, Y., Dinh, T., Kang, W., and Mueller, A. C. Tabflex: Scaling tabular learning to millions with linear attention. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=d60cmFf89H.

Zhang, X., Maddix, D. C., Yin, J., Erickson, N., Ansari, A. F., Han, B., Zhang, S., Akoglu, L., Faloutsos, C., Mahoney, M. W., Hu, C., Rangwala, H., Karypis, G., and Wang, B. Mitra: Mixed synthetic priors for enhancing tabular foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=t8YRsWY6HM.

Zwitter, M. and Soklic, M. Breast Cancer. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C51P4M.

A. Implementation Details

A.1. Datasets

Clinical (binary). Diabetes (Pima Indians, 768 samples, 8 numerical features; target: diabetes onset). Heart Disease (UCI, 270 samples, 13 features; target: disease presence). Breast Cancer (Wisconsin, 699/9; malignancy). SUPPORT2 (9,105 / 46 features incl. 8 categorical, 12.5% missing; hospital mortality). MIMIC-III (26,538 / 74 features; 48-hour in-hospital ICU mortality (Johnson et al., 2016)). eICU-CRD (30,000 / 74 features; multi-center mortality from 200+ U.S. hospitals (Pollard et al., 2018)).

Clinical (multi-class). Cardiotocography (CTG, 2,126 / 35 numerical, 3-class fetal health). Thyroid (3,772 / 21 numerical, 3-class thyroid function; highly imbalanced).

Non-medical TabZilla controls. Jannis (4-class, 54 features), Higgs (binary, 28), Covertype (7-class, 54 mixed); each subsampled to 30K for tractable in-context inference. We use 68/12/20% train/val/test splits, mean over 3 seeds for clinical and 5 seeds for TabZilla. Numerical features are standardised on the training set; categoricals are integer-encoded; missing numericals are imputed with zero after scaling. The validation set doubles as the block-approximating calibration set.

A.2. GBDT Hyperparameters

XGBoost and LightGBM use standard defaults without per-dataset tuning: `n_estimators=500`, `max_depth=6`, `learning_rate=0.1`, `subsample=0.8`, `colsample_bytree=0.8`, with early stopping after 20 rounds on the validation set. These are intentionally not optimised: the goal is a reasonable GBDT baseline for context.

A.3. Foundation Models

We use `n_estimators=1` for TabPFNv2 to enable deterministic per-block representation extraction (the ensembling of n_{est}=8 averages over independent forward passes, which would conflate per-block residuals across passes). Compressibility analysis measures *relative* block redundancy, which is invariant to the ensembling configuration. TabICL, TabDPT, and Mitra are used at their default configurations. All four models are pre-trained exclusively on synthetic data (TabPFNv2, TabICL) or non-overlapping real datasets (TabDPT's 123 real-data training sets, Mitra's curated real-data collection), ruling out data contamination as an explanation for the observed redundancy.

A.4. Attention entropy

Where model architectures permit (TabICL), we measure attention diversity via per-head entropy: $\mathcal{H}(\mathbf{a}) = -\sum_j a_j \log a_j$, averaged across heads, queries, and samples. Entropy near $\log(S)$ (where S is the sequence length) signals uniform attention; low entropy signals specialisation. TabPFNv2 and Mitra use dual-attention; TabDPT's hooks did not yield self-attention weights in our framework.

A.5. Translator family per architecture

The *linear translator* of Cannistraci et al. (2024) is a closed-form affine map $\mathbf{H}_e \approx \mathbf{H}_s W + b$ that approximates the action of a contiguous stack of transformer blocks. The right parameterisation depends on the per-block representation structure of the underlying model. We instantiate it as follows for the four ICL foundation models in this paper:

- **TabICL, TabDPT (single-axis self-attention).** Block outputs are 2D, (seq, d) (with optional pass/batch leading dimensions whose entries share the same block parameters). A single shared $(W \in \mathbb{R}^{d\times d}, b \in \mathbb{R}^d)$ fit by closed-form least squares on $(\text{seq} \cdot \cdots, d)$ flattened calibration data. $d^2{+}d$ free parameters per skip.
- **TabPFNv2 (dual-axis attention).** Block outputs are $(n_{\text{items}}, n_{\text{features}}, d)$ from alternating between-items and between-features attention. Mixing the two axes destroys channel correspondence between source and target, so we fit a separate $(W_f \in \mathbb{R}^{d\times d}, b_f \in \mathbb{R}^d)$ per feature index f and apply it as $\mathbf{H}'[\cdot, f, \cdot] = \mathbf{H}[\cdot, f, \cdot]\, W_f + b_f$. $n_{\text{features}} \cdot (d^2{+}d)$ free parameters per skip.
- **Mitra (dual-axis 2D attention with separate streams).** Block outputs are a tuple $(\mathbf{H}^{\text{sup}}, \mathbf{H}^{\text{qry}})$, each of shape $(n_{\text{items}}, n_{\text{features}}, d)$. The two streams share block parameters but have distinct activations (a translator that only updates support and bypasses query would put them at incompatible depths in subsequent blocks, producing spurious block-0

sensitivity not present in Mitra itself). We fit per-feature affines independently per stream: $2 \cdot n_{\text{features}} \cdot (d^2{+}d)$ free parameters per skip.

With $n_{\text{features}} = 1$, the per-feature variant reduces exactly to the single-(W, b) case, so this is a generalisation of the original linear translator rather than a different method. Concrete feature counts on our datasets: TabPFNv2's internal n_{features} ranges from 6 (Diabetes, Breast Cancer) to 69 (Jannis); Mitra's ranges from 31 (Breast Cancer) similarly.

Conditioning. Each per-feature solve has n_{items} rows and $d{+}1$ columns. For TabICL/TabDPT the single (W, b) is fit on flattened $(n_{\text{items}} \cdot \text{seq}, d)$, giving thousands of rows per skip — well over-determined. For TabPFNv2 ($d{=}192$, calibration capped at 500) the per-feature fit has $n_{\text{items}} {\geq} 500 \gg d{+}1{=}193$, also comfortably over-determined. For Mitra's support stream the fit uses the in-context training set ($3{,}000$ items per feature) and is over-determined; for Mitra's query stream ($d{=}512$, $n_{\text{items}} \leq 500$) it is marginally under-determined by at most 13 unknowns per output column. We use the `gelsd` driver, which returns the stable minimum-norm solution in either regime. The reported ε is the translator's training residual on calibration activations and is therefore not a generalisation metric: generalisation is measured by test AUROC, and the translator never sees test data or any labels. All solves complete in < 50 ms per skip on a single GPU.

B. Additional Results

B.1. Why the saturation matrix is asymmetric

Each block writes an additive residual update $\Delta_{s \to e} = \mathbf{H}_e - \mathbf{H}_s$, so predicting $\mathbf{H}_s$ from $\mathbf{H}_e$ only requires subtracting a low-rank update (linearly easy), while predicting $\mathbf{H}_e$ from $\mathbf{H}_s$ requires synthesising that update from information not yet present (linearly hard whenever intermediate blocks injected non-linear content). The empirical asymmetry of Figure 1 is therefore a signature of the redundancy we exploit: later-block updates are predominantly low-rank corrections, not novel non-linear computation.

B.2. Foundation model full results

Table 3 reports the per-(model, dataset, width) skip-compression result with the best skip start position selected by calibration-set AUROC, then averaged over 3 seeds (medical) or 5 seeds (TabZilla); the corresponding aggregate "best width $\in \{3, 4\}$" numbers are in main-text Table 1. All models tolerate 3–4 approximated blocks with $|\Delta\text{AUROC}| \leq 2.0\%$ (worst: Mitra on Diabetes, -2.0% at width 3). SUPPORT2 is among the most resistant to compression across all models. At mild compression (width 3–4), position matters mostly via block 0: approximating block 0 is consistently the worst start; excluding start=0, TabICL and TabDPT have max test-AUROC spread $\sim 9\%$ across (dataset, width). TabPFNv2 is more sensitive (spread up to 66% on Thyroid at width 4), making calibration-based selection essential.

Table 3. Foundation model block approximation: best skip position per (model, dataset, width), mean ± std over 3 seeds (medical) or 5 seeds (TabZilla). Linear translator. Selection by calibration-set AUROC.

Model	Dataset	Width	Full AUROC	Skip AUROC	ΔAUROC
TabPFN	Diabetes	3	.841	.844±.028	+.003
	Diabetes	4	.841	.838±.031	−.003
	Heart	3	.897	.900±.031	+.003
	Heart	4	.897	.898±.027	+.001
	Br. Cancer	3	.994	.994±.006	+.000
	Br. Cancer	4	.994	.994±.007	+.001
	SUPPORT2	3	.982	.981±.001	−.001
	SUPPORT2	4	.982	.980±.000	−.002
	MIMIC-III	3	.758	.763±.005	+.005
	MIMIC-III	4	.758	.764±.001	+.006
	eICU	3	.758	.763±.009	+.004
	eICU	4	.758	.763±.009	+.004
	CTG†	3	.998	.999±.000	+.001
	CTG†	4	.998	.997±.001	−.000
	Thyroid†	3	1.00	.999±.000	−.000
	Thyroid†	4	1.00	.998±.001	−.001
	Jannis†	3	.846	.838±.008	−.008
	Jannis†	4	.846	.839±.011	−.007
	Higgs	3	.780	.778±.011	−.002
	Higgs	4	.780	.779±.011	−.001
	Covertype‡	3	.957	.954±.003	−.003

continued on next page

Table 3 continued

Model	Dataset	Width	Full AUROC	Skip AUROC	ΔAUROC
	Covertype‡	4	.957	.954±.004	−.002
TabICL	Diabetes	3	.845	.843±.035	−.002
	Diabetes	4	.845	.843±.037	−.002
	Heart	3	.904	.905±.042	+.000
	Heart	4	.904	.908±.046	+.004
	Br. Cancer	3	.994	.995±.005	+.000
	Br. Cancer	4	.994	.995±.005	+.001
	SUPPORT2	3	.983	.982±.001	−.001
	SUPPORT2	4	.983	.981±.002	−.001
	MIMIC-III	3	.800	.798±.006	−.003
	MIMIC-III	4	.800	.793±.010	−.007
	eICU	3	.786	.783±.013	−.004
	eICU	4	.786	.783±.011	−.003
	CTG†	3	.999	.999±.001	+.000
	CTG†	4	.999	.999±.001	+.000
	Thyroid†	3	1.00	1.00±.000	−.000
	Thyroid†	4	1.00	1.00±.000	−.000
	Jannis†	3	.866	.863±.007	−.003
	Jannis†	4	.866	.863±.007	−.003
	Higgs	3	.790	.787±.009	−.003
	Higgs	4	.790	.788±.011	−.002
	Covertype‡	3	.962	.961±.002	−.001
	Covertype‡	4	.962	.962±.003	−.001
TabDPT	Diabetes	3	.841	.841±.036	+.000
	Diabetes	4	.841	.843±.034	+.002
	Heart	3	.898	.892±.046	−.007
	Heart	4	.898	.897±.043	−.001
	Br. Cancer	3	.994	.994±.006	+.000
	Br. Cancer	4	.994	.993±.007	−.000
	SUPPORT2	3	.979	.978±.002	−.001
	SUPPORT2	4	.979	.978±.002	−.001
	MIMIC-III	3	.761	.759±.009	−.002
	MIMIC-III	4	.761	.758±.009	−.003
	eICU	3	.757	.757±.011	−.000
	eICU	4	.757	.757±.011	+.000
	CTG†	3	.996	.998±.002	+.002
	CTG†	4	.996	.998±.002	+.002
	Thyroid†	3	.998	.998±.001	−.000
	Thyroid†	4	.998	.998±.001	+.000
	Jannis†	3	.813	.811±.005	−.002
	Jannis†	4	.813	.812±.005	−.001
	Higgs	3	.734	.734±.013	+.000
	Higgs	4	.734	.733±.014	−.001
	Covertype‡	3	.942	.941±.002	−.001
	Covertype‡	4	.942	.940±.003	−.002
Mitra	Diabetes	3	.839	.819±.026	−.020
	Diabetes	4	.839	.830±.035	−.009
	Heart	3	.906	.904±.045	−.002
	Heart	4	.906	.901±.044	−.005
	Br. Cancer	3	.994	.995±.005	+.000
	Br. Cancer	4	.994	.995±.006	+.000
	SUPPORT2	3	.980	.979±.001	−.001
	SUPPORT2	4	.980	.979±.001	−.002
	MIMIC-III	3	.781	.780±.002	−.001
	MIMIC-III	4	.781	.779±.006	−.002
	eICU	3	.772	.770±.010	−.001
	eICU	4	.772	.768±.011	−.004
	CTG†	3	.998	.997±.003	−.001
	CTG†	4	.998	.998±.001	+.000
	Thyroid†	3	.999	.999±.000	−.000
	Thyroid†	4	.999	.999±.000	−.000
	Jannis†	3	.828	.823±.006	−.004
	Jannis†	4	.828	.821±.004	−.007
	Higgs	3	.773	.771±.011	−.002
	Higgs	4	.773	.769±.010	−.004
	Covertype‡	3	.927	.927±.001	−.000
	Covertype‡	4	.927	.927±.002	−.001

†Multi-class (macro-OVR AUROC; CTG=3-class, Thyroid=3-class, Jannis=4-class). ‡Covertype (7-class).

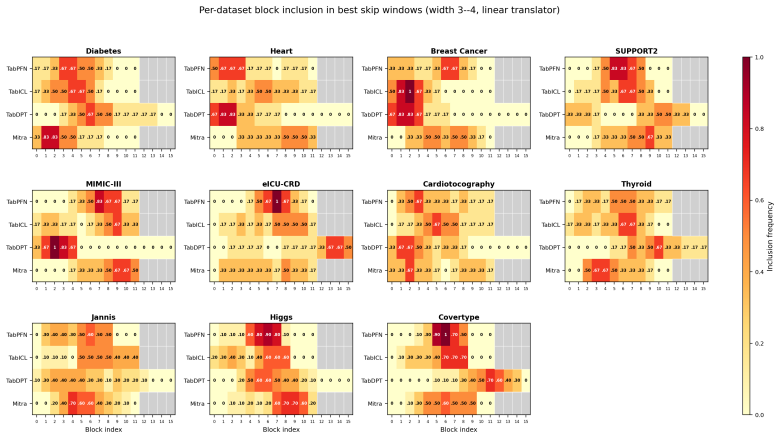

Figure 5. Per-dataset block inclusion frequency in best skip windows (width 3–4, linear translator, averaged over 3 seeds). Gray cells indicate blocks beyond the model's depth (e.g., blocks 12–15 for 12-block models). Cells annotated "0" (lightest yellow) indicate that the block was never selected in the best skip window for that dataset. TabICL and TabDPT shows near-uniform inclusion across all datasets, confirming that the aggregate position-insensitivity in Figure 3 is not an averaging artefact.

B.3. Skip-position consistency (per dataset)

Per-dataset inclusion frequencies confirm the aggregate pattern (Figure 3): TabICL and TabDPT are near-uniform on each dataset, TabPFNv2 clusters at blocks 7–9 most strongly on SUPPORT2 and MIMIC-III, and Mitra avoids block 0 across all datasets.

B.4. Extreme compression and naive layer dropping

The Linear (intermediate, "8/12" or "11/16"), Approx. max, and Identity rows of main-text Table 1 provide the per-architecture identity-vs-linear contrast at maximum compression. The retained block is selected by calibration-set AUROC. For TabDPT it lands on **block 0** on $10/11$ datasets and for TabICL on $7/11$, confirming that under single-axis self-attention block 0 carries the dominant computation. Under the per-feature translator (TabPFNv2, Mitra), the retained block varies by dataset (TabPFNv2 $1/11$, Mitra $3/11$ retain block 0): the per-feature parameterisation equalises the contribution of different blocks, so any contiguous window can be replaced.

TabICL retains 1 of 12 blocks within $|\Delta| \leq 2.3\%$ on every dataset evaluated, including multi-class Jannis (4-class) and Covertype (7-class). TabDPT keeps 1 of 16 within -1% on all clinical datasets (worst -4.3% on Higgs). TabPFNv2 and Mitra, under the per-feature translator appropriate to their dual-axis architecture (Appendix A.5), retain 1 of 12 blocks within $|\Delta| \leq 1\%$ on Breast Cancer (the smallest clinical dataset) and within -3 to -10% on the larger clinical datasets (Diabetes, Heart, SUPPORT2, MIMIC, CTG, Thyroid); they trail TabICL/TabDPT by 9–19 AUROC points on the larger TabZilla benchmarks (Jannis, Higgs, Covertype). At moderate compression (4 of 12 retained, "Linear (8/12)"), TabPFNv2 matches or beats its full model on 4 of 10 datasets and is within 1.7% on every clinical dataset; Mitra is within 1.3% on every clinical dataset. Identity dropping at the same compression collapses every model to AUROC $\in [0.16, 0.99]$ depending on architecture and dataset (median around 0.5), isolating what the translator learns: a low-rank linear correction that maps early representations into the subspace expected by later blocks.

B.5. NLP comparison: attention entropy

To contextualise tabular attention behaviour, we compare attention entropy profiles between TabICL on a medical (MIMIC-III) and a non-medical (Higgs) benchmark and Qwen3-4B (Yang et al., 2025), a 36-layer language model with grouped-query attention (32 query / 8 KV heads), evaluated on *two* corpora: WikiText-2 and IMDB movie reviews. This 2-vs-2 comparison guards against the contrast being an artefact of any single dataset on either side.

TabICL's per-block mean entropy is essentially flat on both MIMIC-III and Higgs (range $\approx$ 1.23–1.38 nats, std ≤ 0.06 on each), consistent with persistent, weakly differentiated attention at every depth and across qualitatively different inputs. Qwen3-4B is qualitatively different on both NLP corpora: per-layer entropy varies substantially across the 36 layers (std $\approx$ 0.4–0.5 nats), with high entropy in the early layers, a sharp dip around layers 7–9, recovery in the mid layers, and a

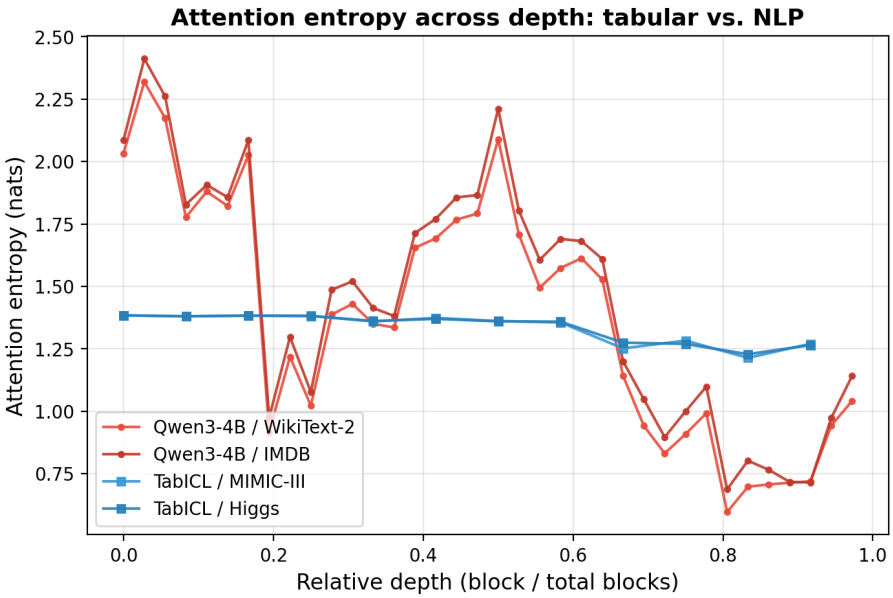

Figure 6. Per-layer mean attention entropy across two tabular settings (TabICL on MIMIC-III and Higgs) and two NLP corpora (Qwen3-4B on WikiText-2 and IMDB). Tabular curves are flat across blocks and on both datasets; NLP curves vary non-trivially with depth on both corpora—high in early layers, a dip around layers 7–9, recovery in mid layers, and a marked decline in the last quarter.

pronounced decline in the last quarter. The contrast is not in the absolute entropy value (which depends on sequence length, $\mathcal{H}_{\max} = \ln S$) but in its *flatness*: tabular blocks all sit at essentially the same entropy regardless of dataset, while NLP blocks span a wide, depth-dependent range that is consistent across corpora, indicative of depth-wise specialisation.

B.6. Attention does not compensate for approximated blocks

The per-block mean attention entropy on surviving blocks before vs. after block approximation is shown in Figure 4 (main text) for the canonical skip 4–8; the per-dataset breakdown is reported here.

Table 4. TabICL attention entropy on surviving blocks before vs. after block approximation. ε is the linear-translator residual; $|\Delta\mathcal{H}|$ is the absolute change in per-block mean entropy (nats), aggregated across surviving blocks. Pre/post overall = mean across surviving blocks.

Dataset	skip	ε	mean $\lvert\Delta\mathcal{H}\rvert$	max $\lvert\Delta\mathcal{H}\rvert$	pre $\overline{\mathcal{H}}$	post $\overline{\mathcal{H}}$
Diabetes	4–8	.192	.013	.038	1.343	1.331
Heart	4–8	.110	.007	.029	1.338	1.331
Br. Cancer	4–8	.090	.017	.054	1.331	1.314
SUPPORT2	4–8	.280	.008	.037	1.300	1.296
Diabetes (early window)	2–6	.168	.048	.108	1.340	1.293
Heart (late window)	6–10	.062	.001	.009	1.352	1.350

A direct test of the redundancy interpretation is whether surviving blocks reorganise their attention to compensate for the approximated sub-stack. Per-block entropy changes by less than 0.02 nats on average and at most 0.06 nats on any single surviving block—about 1% of the entropy value—indicating the approximated sub-stack contributed negligibly to begin with. The change is larger for an early skip window (blocks 2–6 on Diabetes: max $\Delta = 0.11$ nats) because approximating closer to block 0 perturbs the input distribution to all later blocks more, and smaller for a late skip window (blocks 6–10 on Heart: max $\Delta = 0.009$ nats). This rules out attention-collapse explanations and contrasts with NLP transformers where later blocks specialise (Dalvi et al., 2020).

C. Ablation Studies

C.1. Per-patient redundancy and fairness

A natural clinical concern is whether block approximation disproportionately harms certain patient subpopulations. We compute per-sample $\varepsilon(0,1)$ by fitting a global linear translator $\mathbf{W}^*$ on all test samples and evaluating per-patient residuals.

Table 5. Per-patient $\varepsilon(0,1)$ stratified by outcome (mean $\pm$ std over 3 seeds). $\Delta = \varepsilon_{\text{survived}} - \varepsilon_{\text{died}}$. Negligible differences confirm that compression does not disproportionately affect any clinical subgroup.

Model	Dataset	Survived ε	Died ε	Δ
TabPFNv2	SUPPORT2	.236±.055	.234±.051	+.002
	MIMIC-III	.241±.073	.282±.095	−.041
	eICU	.263±.072	.300±.098	−.038
TabICL	SUPPORT2	.034±.011	.026±.010	+.007
	MIMIC-III	.030±.006	.034±.007	−.004
	eICU	.032±.008	.034±.008	−.003
TabDPT	SUPPORT2	.057±.037	.057±.038	+.000
	MIMIC-III	.079±.055	.088±.061	−.009
	eICU	.096±.067	.092±.063	+.004
Mitra	SUPPORT2	.173±.022	.171±.022	+.002
	MIMIC-III	.170±.023	.182±.025	−.012
	eICU	.179±.023	.181±.024	−.002

Across all four models, the difference in per-patient $\varepsilon(0,1)$ between outcome classes is negligible ($|\Delta| \leq 0.007$ on SUPPORT2, ≤ 0.041 on MIMIC-III, ≤ 0.038 on eICU-CRD). TabPFNv2 shows the largest class difference on both MIMIC-III and eICU ($\Delta \approx -0.04$), meaning deceased patients require slightly more first-block transformation, but the *subsequent* blocks remain redundant for both groups ($\varepsilon(1,2) < 0.05$). Redundancy is universal, not patient-dependent: no subpopulation is disadvantaged by compression.

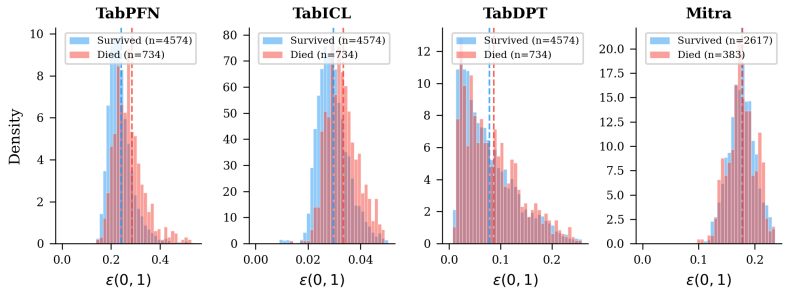

Figure 7. Distribution of per-patient $\varepsilon(0,1)$ on MIMIC-III, split by outcome (survived vs. died). Dashed lines show class means. Distributions overlap almost completely, confirming block redundancy is not severity-dependent.

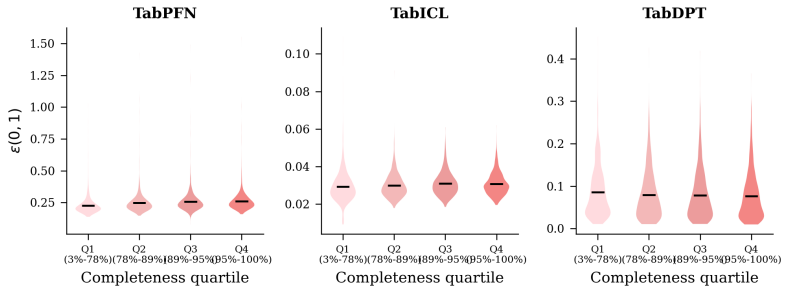

Figure 8. Per-patient $\varepsilon(0,1)$ stratified by feature completeness quartile (Q1 = sparsest, Q4 = densest) on MIMIC-III. No quartile gradient is observed—patients with sparse records do not require more blocks than those with complete records.

C.2. Robustness under feature corruption

We test compressed models under random feature corruption (10–50% of features replaced with column-wise random values). Table 6 reports the full sweep on SUPPORT2 and eICU-CRD.

On SUPPORT2, all four models' compressed variants closely track or outperform the full model under increasing corruption. TabICL is the most compression-tolerant ($|\Delta| \leq 0.007$ at all levels); Mitra's skip model is *more robust* than the full model at 10–50% corruption ($\Delta = +0.023$ at 50%). On eICU-CRD, TabDPT's skip model outperforms the full model at all corruption levels ($\Delta = +0.008$ at 50%).

C.3. Structured clinical missingness

Real clinical missing patterns are structured (whole device fails, lab batch lost) rather than random. We define clinical scenarios that zero out entire feature groups (all labs, all vitals, all variability statistics, all severity scores) and re-evaluate.

On SUPPORT2 all four models' skip variants closely match the full model ($|\Delta| \leq 0.005$); even at 30% missing (labs + vitals), skip is within 0.5% of full across all models. On MIMIC-III, the "drop all variability" scenario (81% of features zeroed) is most challenging—TabPFNv2 and TabICL degrade by ~ 0.05, while TabDPT and Mitra remain within 0.02. Under realistic single-group scenarios (drop labs or vitals), TabICL/TabDPT/Mitra skip models are within 1.5% of full across all three clinical datasets.

C.4. Early-exit baseline

A natural alternative to block approximation is direct classification from intermediate representations: *early exit* discards the original pre-trained head and trains a new classifier on $\mathbf{H}_L$. We sweep *every* block per model.

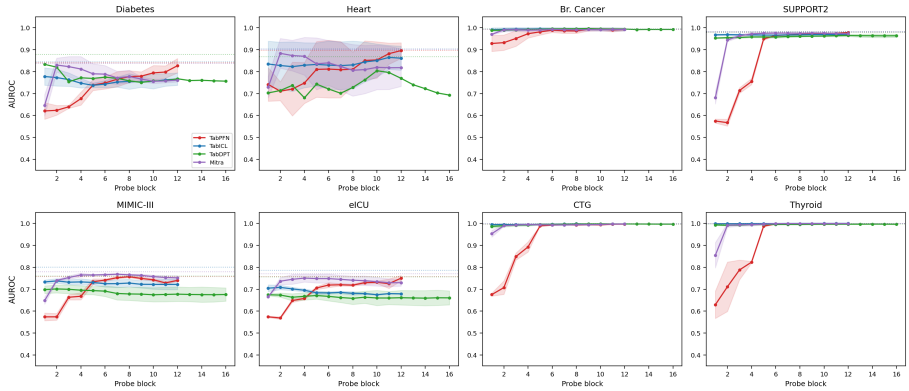

Figure 9. Early-exit logistic-regression probe AUROC vs. probe block index, mean ± std over 3 seeds. Dotted horizontal lines show the corresponding full-model AUROC. TabPFNv2 (red) rises monotonically with depth; TabICL (blue) and TabDPT (green) are non-monotonic. Mitra (purple) shows an intermediate pattern: probe AUROC rises quickly in the first few blocks then plateaus, with best probes at early depths (L2–L4) on most datasets.

We then run a head-to-head comparison: at each depth L, early exit fits logistic regression on $\mathbf{H}_L$, while block approximation bridges $\mathbf{H}_L \rightarrow \mathbf{H}_N$ with a linear translator and feeds the result through the original pre-trained head—same calibration set, same intermediate representation; only the classifier differs (refitted vs. retained).

Aggregating across all 416 (model, dataset, depth) configurations: block approximation matches or exceeds early exit in **83**% of cells; per-model wins are 94% (TabICL), 91% (TabDPT), 84% (Mitra), 61% (TabPFNv2). The advantage holds even at very early depths ($L = 1, 2$) for TabICL/TabDPT, where skip already approaches full-model AUROC but a fresh classifier on the same $\mathbf{H}_L$ loses 5–10 points (e.g. TabICL/MIMIC at $L = 1$: skip .787 vs. probe .733; TabDPT/Heart at $L = 1$: skip .909 vs. probe .703). Retaining the pre-trained head and bridging with a linear translator is uniformly preferable to discarding the head and refitting, except at extreme compression for TabPFNv2/Mitra (the two models with higher residual transformation), where the translator residual eventually dominates and refitting becomes the better option.

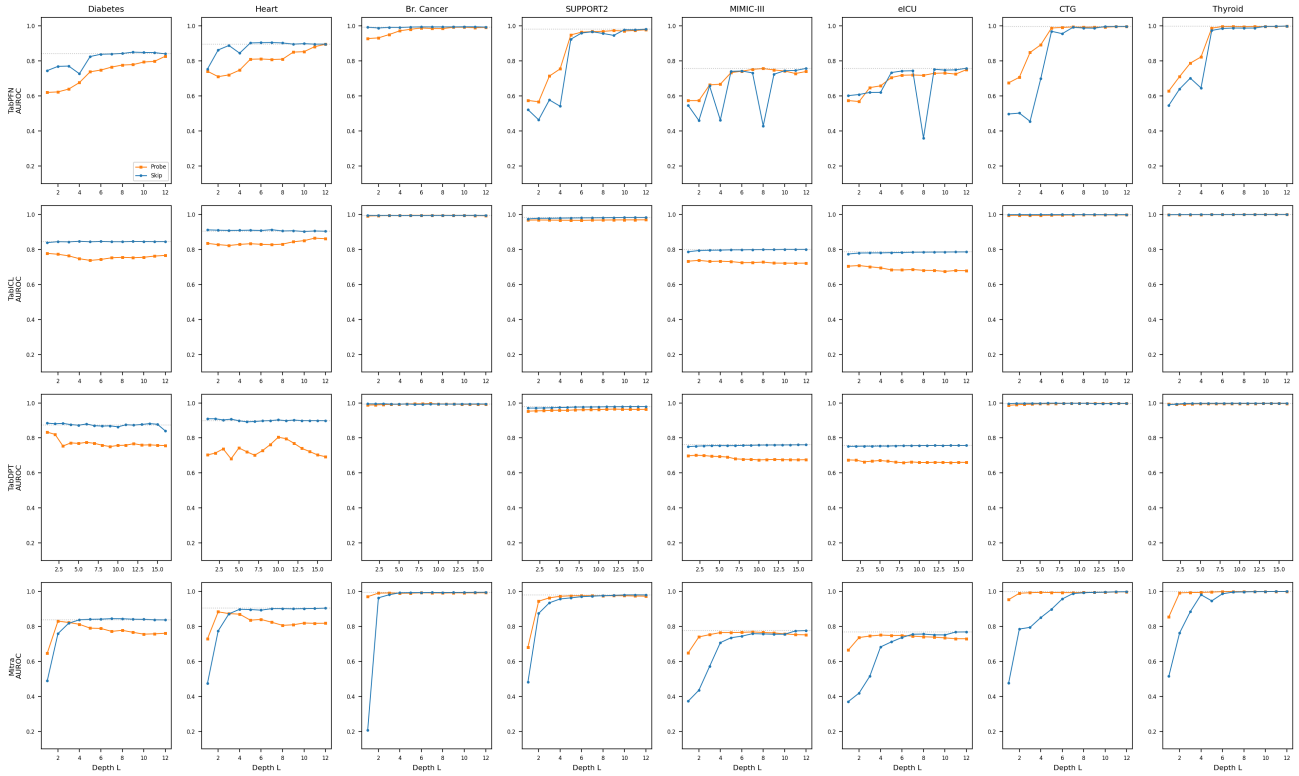

Figure 10. Block approximation vs. early-exit at each depth L, on the same intermediate representation. block approximation keeps the pre-trained head and bridges $\mathbf{H}_L \to \mathbf{H}_N$ with a linear translator; early exit replaces the head with a logistic regression fit on $\mathbf{H}_L$. Mean over 3 seeds; gray dotted line is the full-model AUROC.

C.5. Related Work

GBDTs vs. tabular transformers. Grinsztajn et al. (2022) shows that tree-based models outperform deep learning on typical tabular data, attributing this to lack of rotation invariance and uninformative features. McElfresh et al. (2023) introduced TabZilla and showed neural nets beat GBDTs only on specific subsets. Jayawardhana et al. (2025) recently demonstrated that combining foundation models with decision trees closes this performance gap across sample sizes. Our analysis offers a complementary explanation: tabular transformers waste over 90% of their depth, so their effective capacity is closer to a 1–2 block model than the architecture suggests.

Tabular transformers and foundation models. Tabular transformers include FT-Transformer (Gorishniy et al., 2021), SAINT (Somepalli et al., 2021), ExcelFormer (Chen et al., 2023), and TabFlex (Zeng et al., 2025). TabPFNv2 (Hollmann et al., 2025), TabICL (QU et al., 2025), TabDPT (Ma et al., 2026), and Mitra (Zhang et al., 2026) are recent ICL foundation models with diverse pre-training and attention designs. TabNet (Arik et al., 2019) uses sequential sparse attention to select different features at each step; such explicit per-step routing may mitigate block-level redundancy and is an interesting target for future analysis.

Transformer redundancy. Dalvi et al. (2020) found 85–92% of BERT/XLNet neurons redundant. (Jacobs et al., 2026) rewrites vision transformers with $k \ll L$ recurrent blocks; Token Merging (Bolya et al., 2023) and SLEB (Song et al., 2024) exploit token and block redundancy. Block-level translators (Cannistraci et al., 2024) fit training-free linear maps between block representations. LoRA (Hu et al., 2022) demonstrated that pre-trained weight updates reside in a low-rank subspace; our finding that a linear translator suffices to bridge skipped blocks is conceptually related, excess depth is captured by a linear approximation.

Table 6. Robustness under test-time feature corruption (SUPPORT2 and eICU, mean over 3 seeds $\times$ 5 corruption seeds). ΔAUROC = Skip $-$ Full.

Model	Corruption	Full AUROC	Skip AUROC	ΔAUROC
		SUPPORT2		
XGBoost	0%	.982±.001	—	—
	10%	.971±.002	—	—
	25%	.940±.004	—	—
	50%	.858±.004	—	—
LightGBM	0%	.982±.001	—	—
	10%	.970±.001	—	—
	25%	.935±.004	—	—
	50%	.847±.006	—	—
TabPFNv2	0%	.982	.974	−.009
	10%	.966	.958	−.008
	25%	.928	.919	−.009
	50%	.840	.830	−.011
TabICL	0%	.983	.981	−.002
	10%	.971	.970	−.001
	25%	.941	.940	−.001
	50%	.876	.869	−.007
TabDPT	0%	.979	.978	−.001
	10%	.964	.962	−.002
	25%	.929	.924	−.005
	50%	.855	.843	−.012
Mitra	0%	.980	.979	−.001
	10%	.964	.967	+.003
	25%	.926	.937	+.011
	50%	.845	.868	+.023
		eICU-CRD		
TabPFNv2	0%	.758	.732	−.026
	10%	.748	.722	−.025
	25%	.730	.707	−.023
	50%	.691	.672	−.019
TabICL	0%	.786	.782	−.004
	10%	.772	.763	−.009
	25%	.752	.739	−.013
	50%	.717	.699	−.018
TabDPT	0%	.749	.751	+.002
	10%	.736	.739	+.004
	25%	.721	.728	+.007
	50%	.684	.693	+.008
Mitra	0%	.769	.769	−.000
	10%	.763	.765	+.002
	25%	.749	.755	+.006
	50%	.717	.731	+.013

Table 7. Structured clinical missingness: Full vs. Skip AUROC under realistic missing-data scenarios (mean over 3 seeds). Δ = Skip − Full.

Model	Scenario	% dropped	Full	Skip	Δ
		SUPPORT2			
TabPFNv2	Drop all labs	17%	.946	.936	−.010
	Drop all vitals	13%	.982	.974	−.009
	Drop labs+vitals	30%	.946	.936	−.011
	Drop severity scores	20%	.980	.973	−.006
TabICL	Drop all labs	17%	.962	.962	+.000
	Drop all vitals	13%	.983	.981	−.002
	Drop labs+vitals	30%	.962	.962	−.000
	Drop severity scores	20%	.979	.978	−.001
TabDPT	Drop all labs	17%	.948	.949	+.000
	Drop all vitals	13%	.979	.979	−.001
	Drop labs+vitals	30%	.950	.952	+.001
	Drop severity scores	20%	.977	.976	−.001
Mitra	Drop all labs	17%	.941	.945	+.003
	Drop all vitals	13%	.980	.979	−.001
	Drop labs+vitals	30%	.941	.945	+.003
	Drop severity scores	20%	.978	.977	−.001
		MIMIC-III			
TabPFNv2	Drop all labs	54%	.644	.616	−.028
	Drop all vitals	43%	.748	.734	−.015
	Drop all variability	81%	.722	.675	−.047
TabICL	Drop all labs	54%	.679	.679	−.001
	Drop all vitals	43%	.779	.777	−.002
	Drop all variability	81%	.756	.707	−.048
TabDPT	Drop all labs	54%	.657	.657	−.000
	Drop all vitals	43%	.736	.733	−.003
	Drop all variability	81%	.706	.687	−.019
Mitra	Drop all labs	54%	.665	.657	−.007
	Drop all vitals	43%	.760	.760	−.001
	Drop all variability	73%	.691	.728	+.038
		eICU-CRD			
TabPFNv2	Drop all labs	60%	.667	.658	−.009
	Drop all vitals	38%	.743	.745	+.002
	Drop lab variability	45%	.720	.724	+.004
	Drop vital variability	28%	.748	.750	+.001
	Drop all variability	73%	.701	.700	−.001
TabICL	Drop all labs	60%	.680	.691	+.011
	Drop all vitals	38%	.764	.764	+.000
	Drop lab variability	45%	.746	.724	−.021
	Drop vital variability	28%	.773	.773	+.000
	Drop all variability	73%	.728	.697	−.032
TabDPT	Drop all labs	60%	.632	.634	+.002
	Drop all vitals	38%	.735	.742	+.008
	Drop lab variability	45%	.724	.731	+.007
	Drop vital variability	28%	.732	.741	+.009
	Drop all variability	73%	.682	.677	−.005
Mitra	Drop all labs	60%	.673	.681	+.008
	Drop all vitals	38%	.749	.750	+.001
	Drop lab variability	45%	.725	.738	+.013
	Drop vital variability	28%	.756	.756	+.000
	Drop all variability	73%	.704	.725	+.020

Table 8. Head-to-head AUROC at three reference depths: probe (early-exit logistic regression on $\mathbf{H}_L$) vs. skip (linear translator $\mathbf{H}_L \rightarrow \mathbf{H}_N$, original head retained). Mean over 3 seeds. **Bold** = winner at that cell.

			$L=1$		$L=\lfloor N/2 \rfloor$		$L=N$	
Model	Dataset	Full	probe	skip	probe	skip	probe	skip
TabPFNv2	Diabetes	.841	.621	**.745**	.748	**.839**	.827	**.841**
	Heart	.897	.742	**.754**	.812	**.905**	.896	**.897**
	Br. Cancer	.994	.928	**.993**	.989	**.995**	.992	**.994**
	SUPPORT2	.982	**.575**	.523	**.965**	.960	**.979**	**.982**
	MIMIC-III	.758	**.574**	.547	.742	**.743**	.740	**.758**
	eICU	.758	.574	**.602**	.719	**.743**	.750	**.758**
	CTG†	.998	**.675**	.498	**.992**	.956	.997	**.998**
	Thyroid†	1.00	**.629**	.547	**.997**	.986	.999	**1.00**
TabICL	Diabetes	.845	.778	**.839**	.743	**.846**	.766	**.845**
	Heart	.904	.835	**.912**	.830	**.908**	.861	**.904**
	Br. Cancer	.994	.991	**.995**	**.995**	.994	.993	**.994**
	SUPPORT2	.983	.968	**.975**	.966	**.980**	.969	**.983**
	MIMIC-III	.800	.733	**.787**	.725	**.798**	.722	**.800**
	eICU	.786	.705	**.774**	.683	**.783**	.679	**.786**
	CTG†	.999	.994	**.999**	.996	**.999**	.998	**.999**
	Thyroid†	1.00	.998	**.999**	.999	**.999**	1.00	**1.00**
TabDPT	Diabetes	.879	.833	**.884**	.758	**.868**	.756	**.841**
	Heart	.898	.703	**.909**	.728	**.897**	.693	**.898**
	Br. Cancer	.994	.988	**.995**	**.995**	.992	.992	**.994**
	SUPPORT2	.979	.953	**.971**	.961	**.977**	.963	**.979**
	MIMIC-III	.761	.698	**.751**	.678	**.758**	.676	**.761**
	eICU	.757	.675	**.752**	.658	**.756**	.661	**.757**
	CTG†	.997	.986	**.996**	.997	**.997**	.997	**.997**
	Thyroid†	.998	.993	.990	.995	**.998**	.996	**.998**
Mitra	Diabetes	.837	.663	**.688**	.788	**.841**	.724	**.837**
	Heart	.900	.714	**.867**	.801	**.916**	.808	**.900**
	Br. Cancer	.994	.983	**.993**	.995	**.995**	.993	**.994**
	SUPPORT2	.980	.664	**.704**	.975	**.977**	.973	**.980**
	MIMIC-III	.778	.645	**.645**	.763	**.770**	.753	**.778**
	eICU	.770	.664	**.670**	.746	**.763**	.726	**.770**
	CTG†	.998	**.967**	.801	.992	**.998**	.996	**.998**
	Thyroid†	.999	**.853**	.804	**.999**	.998	.999	**.999**

† Three-class datasets (macro-averaged OVR AUROC).