CHARACTERIZING TABULAR FOUNDATION MODELS IN EARLY LITERACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular foundation models show strong performance across diverse domains and settings, yet few studies examine their use for early literacy prediction—a consequential domain with scarce longitudinal data and heterogeneous, incomplete real-world records. We ask whether this promise carries over to K–4 literacy prediction in realistic settings. To answer this question, we introduce *RealLit*, a large-scale, longitudinal multi-cohort benchmark comprising 4,582 students across 53 schools and six U.S. states. A large portion of the benchmark comes from real-world school-system data, preserving naturally occurring missingness and heterogeneity in assessment practice, while providing fine-grained student-level demographic attributes across cohorts. Our benchmarking surfaces several key findings while offering guidance for evaluating performance beyond aggregate metrics using student-level metadata. For instance, we find that model ensembling can have uneven effects across student groups, reducing performance for some. Newer model variants can also underperform earlier ones. Surprisingly, in our missing-data analysis, learned imputation also provides little consistent downstream benefit over simpler missing-data handling strategies. By making such behaviors measurable, RealLit provides a standardized testbed for understanding and advancing the reliable use of tabular foundation models in real-world early literacy prediction. Our code and data are available at https://github.com/anonymous-submission-site/reallit.

1 INTRODUCTION

Reading failure is often identified only after children have accumulated years of difficulty with print, during which gaps in reading and language skills may have already widened, contributing to broader academic, motivational, and socioemotional consequences (Juel, 1988; Torgesen, 2000; Vellutino et al., 2006; Morgan et al., 2008). The implications extend beyond literacy itself, as early reading performance is closely associated with later achievement in areas such as mathematics and science (Duncan et al., 2007; Grimm, 2008; Barnard-Brak et al., 2017). The scale of the challenge is substantial: In 2024, the National Assessment of Educational Progress reported that 40% of U.S. fourth-grade students performed below the Basic level in reading (National Center for Education Statistics, 2025).

At the core of a preventive response is the precise and timely identification of student needs, so that those needs can be addressed before gaps form and widen, a goal that scalable machine learning (ML) models Qu et al. (2025); Hollmann et al. (2025) could potentially support.

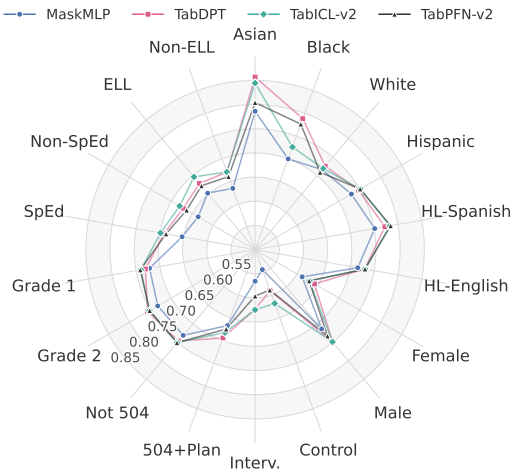

Figure 1: **Subgroup Variation in Early Literacy Modeling.** Classification accuracy averaged *across RealLit cohorts* (when available). HL: home language, ELL: English-language learner status, SpEd: special education status, Interv.: intervention, and 504 disability plan (U.S. Department of Education, Office for Civil Rights, 2026).

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.

Using longitudinal data, models could learn to help identify students at risk early and, in intervention settings, estimate which supports may benefit them (Varol et al., 2009; Shangguan et al., 2024). Realizing this potential, however, requires models to remain reliable under the conditions that characterize real educational data, including variation in data characteristics (*e.g.*, missingness) and performance across student subgroups (Juel, 1988; Torgesen, 2000; Connor et al., 2009).

Recently, tabular foundation models have expanded the possibility of scalable prediction across heterogeneous datasets and tasks Hollmann et al. (2025); Qu et al. (2025). While these models may offer useful tools for early literacy outcome prediction, their performance in early childhood educational settings remains largely untested. A key barrier is the lack of representative longitudinal benchmarks for evaluation. For instance, TabArena Erickson et al. (2025) provides a standard evaluation suite for general tabular prediction, but does not include longitudinal early literacy data. Critically, prior early literacy datasets may report only aggregate demographic statistics Shangguan et al. (2024), which limits finer-grained analysis of model reliability across students. Beyond aggregate performance, we aim to understand and advance reliable early-literacy models across diverse student populations and educational contexts.

Key Contribution—Understanding Tabular Foundation Models for Early Literacy: RealLit brings together longitudinal early-literacy data collected across realistically diverse educational settings, including intervention-linked outcomes, naturally occurring missingness, and fine-grained student-level demographic variation. Despite substantial heterogeneity across settings, we standardize these datasets within a ***common evaluation framework***, *i.e.*, for tabular foundation models, facilitating consistent analysis across cohorts, data conditions, and student populations. Beyond large-scale evaluation of aggregate model performance, RealLit also provides the first ***demographic-aware testbed for early literacy outcome prediction***, targeting the development of educational foundation models that remain reliable across student populations and data conditions. While we take an initial step, the longer-term goal is stable predictive performance across datasets and collection settings, missingness patterns, imputation strategies, ensemble configurations, and demographic subgroups. This distinction is central to more responsible early-literacy modeling, as models may achieve strong aggregate AUC (*e.g.*, in the case of model ensembling) while failing under structured missingness or performing unevenly across student populations.

Key Insights: Our analysis shows how aggregate performance can conceal substantial variation across student populations, where models with similar overall AUC can behave quite differently across demographic subgroups. While this phenomenon has been documented in other tabular settings (Cruz & Hardt, 2024), we confirm its extent in early-literacy contexts. The analysis further reveals more nuanced insights; for instance, newer tabular foundation model releases are not necessarily better. TabPFN-v3.5 (Jäger et al., 2026), despite extending the model family toward relevant temporal settings, does not improve over TabPFN-v2 Grinsztajn et al. (2025) on our benchmark. TabPFN-v2 achieves higher mean performance on seven of eight dataset and metric comparisons. Another insight concerns *inference-time ensembling*, with varying effects across student subgroups, further highlighting the need for subgroup-aware validation before models are used to inform screening or intervention decisions. More broadly, RealLit surfaces failure modes that aggregate evaluation can obscure, while supporting the development of models that are more robust and better aligned with responsible educational decision support. To facilitate reproducibility and future research, the benchmark, evaluation code, and documentation will be released (currently hosted on an anonymous repository).

2 Related Work

Educational Datasets in ML: Educational datasets remain limited in widely used ML benchmarks, and existing examples focus on higher education rather than the longitudinal, intervention-linked structure of early literacy prediction in K-12 settings. TabArena (Erickson et al., 2025), an aggregate benchmark for tabular models, includes *only one* education-related dataset, on dropout and academic success in higher education (Martins et al., 2021; Asuncion et al., 2007). TALENT (Ye et al., 2024) includes six, its smallest domain category, but these are also largely higher-education, institutional, or education-adjacent tasks such as dropout prediction, law-school admission, and course evaluations. Yet earlier (*i.e.*, K-4) identification is critical, as reading difficulties compound over time while timely support can change a student's learning trajectory. The closest prior benchmark, ECRI (Shangguan et al., 2024), studies early childhood reading performance prediction but does not evaluate tabular

foundation models or generative imputation pipelines, and provides only per-school demographic statistics, which limits subgroup reliability analysis. RealLit, in contrast, provides *additional student-level demographic attributes*, with coverage varying by cohort according to the data collected in each study and school setting.

Tabular Foundation Models: Tabular foundation models aim to provide strong predictors for structured data without extensive task-specific tuning. Recent methods, such as TabPFN (Hollmann et al., 2025; Grinsztajn et al., 2025) and TabICL (Qu et al., 2025), leverage large-scale pretraining and in-context learning to transfer across heterogeneous tabular tasks. TabPFN has also been identified as a strong model in the few higher-education contexts studied in prior work (Ye et al., 2024). These models are especially appealing for education, where datasets are often modest in size, heterogeneous across sites, and difficult to collect. However, it remains unclear how effectively such models transfer to early reading contexts, particularly under varying data incompleteness regimes and across student-level demographic groups.

Dealing with Incomplete Data: Incomplete data is common in educational datasets because assessments are collected over time, under school constraints, and often with different measures administered to different students (Peugh & Enders, 2004; Enders et al., 2006; Peng et al., 2006; Cheema, 2014). In the related ECRI benchmark (Shangguan et al., 2024), for example, only 14% of the 6,916 reported students have valid end-of-year assessments. Standard approaches address incomplete data by removing incomplete samples, adding missingness indicators, or imputing missing values with statistical methods such as MICE (Van Buuren & Groothuis-Oudshoorn, 2011), MissForest (Stekhoven & Bühlmann, 2012), and recent LLM-based imputers (You et al., 2020; Lipton et al., 2016; Hayat & Hasan, 2024; He et al., 2025; Xu et al., 2025). More recent approaches use masked reconstruction or generative modeling to infer missing entries (Yoon et al., 2018; Shahbazian & Trubitsyna, 2022; Camino et al., 2019), including masked autoencoding (Du et al., 2024; Shangguan et al., 2024), diffusion models (Zhang et al., 2025; Zheng & Charoenphakdee, 2022), and flow matching-based imputers (Simkus & Gutmann, 2025). However, imputation can obscure informative missingness patterns when missing values reflect attendance, mobility, assessment eligibility, or intervention context. We therefore use RealLit to evaluate when incomplete-data strategies improve early literacy prediction under realistic missingness.

3 DATA

We evaluate methods over *four longitudinal early-literacy datasets*: three collected by researchers in controlled studies and one ecologically valid school-based dataset collected by teachers, as described below. Tab. 1 summarizes their main characteristics. To our knowledge, no prior ML benchmark enables systematic early-literacy evaluation at this scale across multiple longitudinal cohorts, collection settings, and student populations. Because these data were collected across multiple sites, years, and study protocols, the datasets naturally differ in grade coverage, assessment batteries, and the granularity of student-level information. In our analysis (Sec. 5), we preserve these differences by evaluating each dataset separately. At the same time, there are shared structures that create an opportunity for future work on educational foundation models that can learn across cohorts while generalizing to distinct student populations and settings. Additional information on the available outcome measures and targets is provided in Sec. A.5.

D1-ECRI: D1 is the publicly available ECRI benchmark (Shangguan et al., 2024). In the publicly released data, we identify 946 examples with valid outcome labels, with substantial missingness across features. D1 also *lacks student-level demographic information*, limiting subgroup reliability analyses. D1 is thus primarily included as a reference benchmark (the original study does not evaluate tabular foundation models).

Introduced Datasets: Our D2 comes from a brief, child-centered set of iPad-administered assessments measuring reading and language skills. Of the 4,099 students in an original calibration study (public, private, and virtual schools across five U.S. states), our subset includes 1,039 kindergarteners with student-level demographics, 515 of whom have at least one valid outcome. Features include fall screening measures of phonological awareness, rapid naming, oral language, verbal memory, and cognitive skills (Parrila et al., 2004; Melby-Lervåg et al., 2012; McWeeny et al., 2022; Parkin, 2021),

Table 1: Dataset Overview. We introduce a benchmark comprising three diverse cohorts (D2-D4) collected across distinct educational settings. The benchmark enables analysis of cross-cohort generalization and variation across student populations. We further employ D1 (Shangguan et al., 2024), the only prior publicly available early-literacy benchmark, despite its substantial missingness and limited student-level information. n_{valid} denotes number of students with at least one valid measured outcome label. *Meas.* gives the number of outcome measures available in the released data with the number benchmarked in our experiments in parentheses (label definitions in App. A.6). *Miss. (%)* is missing data percentage. *Demo. level* indicates whether demographic attributes are available at the school or student level. D3 appears as two rows because student identifiers were anonymized separately for each grade band and cannot be linked. *G* denotes grades.

Dataset	n_{total}	n_{valid}	Features	Grades	Meas.	Miss. (%)	Demo. level
D1 (Shangguan et al., 2024)	6,916	946	16	1	2	40.8	School
D2	1,039	515	14	K	4	47.3	Student
D3 (G1–G2)	2,088	2,088	22	1–2	3	10.3	Student
D3 (G3)	1,137	1,137	12	3	1	9.1	Student
D4	318	285	349	1–4	4	20.4	Student

with realistic blockwise missingness. Targets are end-of-year standardized scores in decoding, word recognition, comprehension, and sentence repetition.

D3 is an anonymized student-level operational dataset from a Florida school district with two complementary segments: a longitudinal Grade 1 and 2 panel (2,088 students across 14 schools) and a Grade 3 English language arts (ELA) segment (1,137 students). Unlike a controlled research assessment, D3 reflects *ecologically valid school-system data*, with student-level attributes including race/ethnicity, English learner status, exceptional student education status, 504 status, and gifted status, free or reduced-price lunch eligibility, homelessness, and home language. As student identifiers were anonymized separately for each segment, the two cannot be linked and are treated as separate tasks. Features are district records, including early literacy screeners, state and district progress-monitoring and diagnostic assessments (e.g., DIBELS, FAST/STAR, and i-Ready (Hagan-Burke et al., 2006; McBride et al., 2010; Bulut & Cormier, 2018)), online instructional usage, attendance, discipline, and student services. As targets, we leverage spring reading proficiency measures, three for the Grade 1 and 2 panel and one for Grade 3.

Finally, D4 is a longitudinal reading and writing dataset following one cohort of 318 students from Grade 1 through Grade 4 across six Florida schools, with gender and race demographics. Features are repeated measures of reading, spelling, writing, and language skills from the early grades, and targets are four later literacy outcomes, including word reading and spelling. As a compact but deeply longitudinal cohort, D4 complements D2 and D3 by testing whether models remain reliable as outcomes, assessments, and missingness patterns evolve over time.

As RealLit spans distinct geographic regions, schools, grade levels, and assessment settings, substantial cross-cohort heterogeneity is a natural property of the benchmark. Our analysis therefore generally reports dataset-specific results, while also discussing trends that generalize across datasets.

4 Experimental Setup

Formulation Overview: Each benchmark instance is a supervised tabular prediction task with observed features $\mathbf{x} \in \mathbb{R}^d$, an observation mask $\mathbf{m} \in \{0,1\}^d$, and a binary early-literacy target $y \in \{0,1\}$. From the original set of reading measures, we follow prior work Shangguan et al. (2024) and benchmark Woodcock Reading Mastery Word Identification and Word Attack non-response in D1. For the introduced datasets, we leverage KTEA-3 Nonsense Word Decoding and Letter and Word Recognition risk (D2), FAST ELA (D3), and WRAT-3 Spelling (D4) recovery. We benchmark both classification, presented in the main paper, and regression, reported in the appendix, to assess whether the observed trends persist across prediction settings. App. A.5 and App. A.6 provide full target and label definitions, as well as a discussion on shared measures among the highly diverse datasets.

4.1 Tabular Foundation Models

One key question lies in how well recent tabular foundation models transfer to early-literacy prediction when used off the shelf. To answer this question, we compare three foundation models

against the MaskMLP (Shangguan et al., 2024) baseline, which is a self-supervised MLP baseline developed for early-literacy prediction. The foundation models are TabPFN-v2 (Hollmann et al., 2025), an in-context prior-data fitted network, TabICL-v2 (Qu et al., 2025), an in-context model designed to scale to larger tables, and TabDPT (Ma et al., 2025), a retrieval-augmented tabular foundation model. We additionally evaluate six conventional non-foundation baselines as references: RandomForest (Breiman, 2001), XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017), HistGB and LogReg (Pedregosa et al., 2011), and DeepConvLSTM (Ordóñez & Roggen, 2016).

Within each dataset and evaluation fold, all classifiers use identical training and test partitions, while missingness, imputation, and ensembling are examined systematically in subsequent experiments. In our primary evaluation, TabPFN-v2, TabICL-v2, and TabDPT are used directly from their released public checkpoints with no gradient updates on RealLit. Only MaskMLP is trained on each cohort's training fold. This setup lets us study how well tabular foundation models transfer to early-literacy data without conflating transfer with task-specific adaptation, and separately examine the role of fine-tuning in the appendix. Model-specific preprocessing and implementation details are provided in App. C.1.

Following (Shangguan et al., 2024), MaskMLP pre-trains an MLP encoder with a cosine-embedding loss over randomly masked inputs (per-row mask rate 25%, sentinel -1 for natively missing entries) and then fine-tunes it end-to-end with binary cross-entropy on the binary target. As the code of Shangguan et al. (2024) is not publicly available, we reimplement the method (further described in Tab. 19), yet we are unable to reproduce the results reported in the original study (see App. I.1). TabPFN (Hollmann et al., 2025) is a transformer-based prior-data fitted network (PFN) trained once on a large corpus of synthetic tabular tasks. At inference, the training fold and the test query are processed in a single in-context forward pass with no gradient updates. The released checkpoint supports up to $50{,}000$ rows and $2{,}000$ features, comfortably above every RealLit cohort's row and feature count. TabICL-v2 (Qu et al., 2025) is an in-context tabular foundation model that encodes each row as a token sequence and attends jointly across training and test rows. It is designed for larger row counts, which suits the larger D3 cohorts. Finally, TabDPT (Ma et al., 2025) is a retrieval-augmented foundation model that builds a per-query in-context set by retrieving the most relevant training rows, giving it an adaptive effective context size.

Imputation: Missingness is a defining feature of real educational data, arising from factors such as student absence, mobility, and assessment schedules. Importantly, these patterns need not be random and may themselves carry information about student outcomes, making the choice of missing-data strategy potentially consequential for downstream prediction. Under each missingness mechanism, classifiers can be paired with an *imputer* g_ϕ, trained on the training split, that maps an incomplete record $(\mathbf{x}, \mathbf{m})$ to $\tilde{\mathbf{x}} = g_\phi(\mathbf{x}, \mathbf{m})$ before classification by f_θ. We evaluate two learned generative imputers for tabular data: DiffPuter, based on EM-driven latent-space diffusion (Zhang et al., 2025), and CFMI, based on continuous flow matching (Simkus & Gutmann, 2025; Lipman et al., 2023). We also evaluate standard imputation baselines that replace missing entries with zero in the original feature units, denoted as Zeros, or with column means computed from the observed training data, denoted as Mean (Lipton et al., 2016). We compare these strategies against classifiers applied directly to incomplete inputs to quantify the benefit of explicit imputation. Implementation details and additional imputer variants are provided in App. C.2.

Permutation Ensembling: We study whether averaging predictions across equivalent presentations of the same tabular task improves robustness. For a classifier ensemble of size K, member k samples a feature permutation π_k uniformly from S_d and a class-label permutation σ_k uniformly from S_C, where S_n denotes the set of permutations of $\{1, \ldots, n\}$, d is the number of features, and C is the number of classes. After applying the inverse label permutation to return predictions to the canonical label space, we average the resulting probability vectors:

$$\hat{\mathbf{p}}(y \mid \mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} \sigma_k^{-1} \left(f_\theta(\pi_k(\mathbf{x})) \right).$$

We sweep $K \in \{1, 2, 4, 8, 16, 32\}$, with $K{=}1$ corresponding to the non-ensembled baseline.

Table 2: Classifier performance with and without imputation. The lower blocks show imputation results for RandomForest and TabPFN-v2, selected from the non-foundation and foundation classifiers, respectively. Values are mean ± standard deviation across three evaluation folds. **Bold** marks the highest mean in each dataset and metric column.

Classifier	AUC				ACC			
	D1	D2	D3	D4	D1	D2	D3	D4
DeepConvLSTM	0.639±0.030	0.768±0.007	0.843±0.022	0.903±0.054	0.574±0.029	0.739±0.009	0.784±0.025	0.778±0.031
XGBoost	0.657±0.003	0.753±0.008	0.819±0.027	0.896±0.088	0.607±0.019	0.693±0.025	0.749±0.030	0.778±0.168
HistGB	0.665±0.012	0.779±0.025	0.825±0.029	0.890±0.068	0.613±0.014	0.712±0.037	0.748±0.038	0.822±0.102
LightGBM	0.670±0.009	0.771±0.031	0.826±0.030	0.896±0.057	0.616±0.022	0.725±0.052	0.739±0.025	0.800±0.067
LogReg	0.681±0.024	0.791±0.010	0.839±0.025	0.792±0.025	0.634±0.021	0.722±0.030	0.765±0.021	0.756±0.102
RandomForest	0.682±0.011	0.809±0.016	0.841±0.029	0.887±0.055	0.634±0.033	0.745±0.010	0.768±0.026	0.756±0.077
MaskMLP	0.685±0.007	0.804±0.021	0.846±0.025	0.835±0.082	0.614±0.008	0.735±0.043	0.775±0.024	0.733±0.133
TabDPT	0.697±0.031	0.816±0.004	0.854±0.027	0.903±0.044	0.632±0.032	0.732±0.011	0.780±0.027	**0.867±0.067**
TabICL-v2	0.681±0.034	0.829±0.021	0.856±0.030	0.902±0.012	0.625±0.034	0.752±0.006	0.776±0.037	0.822±0.038
TabPFN-v2	0.696±0.026	**0.832±0.013**	**0.860±0.024**	0.914±0.022	0.643±0.038	0.748±0.015	**0.789±0.027**	**0.867±0.000**
RandomForest (Zeros)	0.677±0.017	0.806±0.000	0.843±0.026	0.907±0.103	0.632±0.032	0.739±0.020	0.773±0.025	0.756±0.077
RandomForest (Mean)	0.680±0.032	0.806±0.015	0.843±0.030	0.917±0.088	0.637±0.012	0.742±0.034	0.765±0.032	0.778±0.102
RandomForest (DiffPuter)	0.669±0.002	0.816±0.034	0.843±0.037	0.917±0.094	0.632±0.013	**0.755±0.010**	0.777±0.038	0.756±0.077
RandomForest (CFMI)	0.687±0.007	0.809±0.017	0.848±0.029	0.923±0.093	**0.657±0.017**	0.732±0.023	0.777±0.019	0.756±0.077
TabPFN-v2 (Zeros)	0.696±0.022	0.830±0.029	0.856±0.026	0.914±0.029	0.648±0.024	0.751±0.010	0.783±0.027	**0.867±0.000**
TabPFN-v2 (Mean)	0.697±0.030	0.820±0.001	0.859±0.023	0.920±0.029	0.643±0.034	0.735±0.026	0.787±0.034	0.844±0.038
TabPFN-v2 (DiffPuter)	0.691±0.024	0.814±0.029	0.859±0.022	0.921±0.012	0.639±0.036	**0.755±0.017**	0.780±0.034	0.844±0.038
TabPFN-v2 (CFMI)	**0.702±0.024**	0.823±0.006	0.857±0.025	**0.933±0.029**	0.639±0.034	0.729±0.037	0.784±0.030	**0.867±0.038**

When stochastic imputation is used, we likewise generate K completed versions of the input by re-seeding the imputer and average them before classification:

$$\tilde{\mathbf{x}} = \frac{1}{K}\sum_{j=1}^{K} \tilde{\mathbf{x}}^{(j)}, \qquad \tilde{\mathbf{x}}^{(j)} = g_\phi^{(j)}(\mathbf{x}, \mathbf{m}).$$

More precisely, the classifier and imputer ensemble sizes are K_{clf} and K_{imp}. Throughout our experiments, however, we use a diagonal sweep with $K_{\text{imp}} = K_{\text{clf}}$ (note, we thus slightly abuse notation for brevity at times, denoting both as K). Increasing K jointly increases the number of imputation and classifier ensembles.

Metrics: We use ROC-AUC as our primary metric and classification accuracy (ACC) as a secondary metric. ROC-AUC measures how well a model ranks positive examples above negative ones across decision thresholds, making it well suited to our setting where positive-class prevalence varies across cohorts and outcomes. ACC complements this with thresholded predictive performance and is particularly useful for subgroup analysis. We report results over 3 repetitions of 5-fold cross-validation, jointly stratified by the at-risk target and, where available, school identifier (App. C.5). When student-level demographic attributes are available, we additionally report subgroup-specific performance, retaining missing or unknown demographic values as explicit groups so that these students remain represented in the audit (App. F). Full implementation details, hyperparameters, and compute are provided in App. C.

5 Results and Discussion

We report observed performance differences and provide statistical tests and confidence intervals for selected comparisons in App. H.

Baseline Performance: Tab. 2 shows that performance varied substantially across cohorts. D4 was the most predictable cohort, followed by D3, D2, and D1. This ordering likely reflects differences in developmental timing, feature richness, label quality, and missingness structure. D4 includes the oldest students and a deeply longitudinal feature space, making later literacy status more predictable from prior reading and writing profiles. D3 also includes older students and operational records closely tied to spring reading proficiency, including screening, progress-monitoring, usage, attendance, and service data. D2 is more challenging because it predicts end-of-kindergarten outcomes from fall screening profiles, when reading skills are still emerging and trajectories are less stable. D1 is the least predictable, consistent with its smaller number of valid labeled examples, high missingness, and lack of student-level demographic attributes.

Learned Imputation Gains Vary Across Datasets and Classifiers: Relative to Mean and Zero filling, learned imputation provides more consistent gains for RandomForest than for TabPFN-v2 (Tab. 2). In Tab. 2, for RandomForest, CFMI achieves higher AUC than both simple baselines across all four datasets, and learned imputers improve accuracy on three of the four datasets. For TabPFN-v2, the improvements are primarily in AUC: CFMI achieves the highest AUC on D1 and D4, but its accuracy is lower than Zero on D1 and tied on D4. Thus, learned imputation can improve classification performance in some settings, although the extent of improvement varies across datasets, classifiers, and evaluation metrics.

Permutation Ensembling Is The Most Effective Lever for Both Accuracy and Subgroup Reliability: We next ask whether inference-time ensembling intervenes on this unevenness. The right panel of Fig. 2 shows the same demographic radar at $K{=}8$: every classifier polygon expands outward across the spokes. Per-subgroup accuracy improves uniformly at $K{=}8$ relative to $K{=}1$, and ensembling does not regress on subgroups where absolute accuracy is already lower (Apps. F.1 and F.2). The aggregate version of this effect is quantified in Tab. 10: at injected rate 30%, MaskMLP improves from 0.741 AUC at $K{=}1$ to 0.779 at $K{=}16$ (0.038), and TabPFN-v2 from 0.766 to 0.785 (0.019). The lift is largest under MNAR and at higher injected rates, which is the regime where presentation-order sensitivity of in-context learners would be expected to matter most. $K{=}16$ is the headline configuration. $K{=}8$ recovers most of the lift and is a defensible practical default when compute is constrained. The full $K \times \mathrm{rate} \times \mathrm{classifier}$ grid is in App. D.4, with the per-mechanism breakdown for TabPFN-v2 in Tab. 17.

Ensembling Effects Vary Across Subgroups: Tabular foundation models provide competitive baselines for early-literacy prediction (Tab. 2). Among the baseline configurations in the upper blocks, TabPFN-v2 achieves the highest AUC on D2–D4, while the four primary classifiers perform similarly on D1. However, differences among the strongest methods are generally modest, motivating an examination of performance beyond aggregate rankings. As shown in Fig. 2, most subgroups recorded across multiple cohorts change only modestly from $K{=}1$ to $K{=}8$, whereas race and ethnicity exhibit the largest and most model-dependent shifts. For students identified as Black, the change ranges from -0.049 for TabDPT to $+0.095$ for MaskMLP, while TabICL-v2 improves by 0.051. A similar pattern appears for Asian students, who are represented in only one cohort: MaskMLP and TabPFN-v2 improve by 0.086 and 0.074, respectively, whereas TabICL-v2 changes by only -0.006. These results indicate that the effect of increasing K depends on both the classifier and the subgroup, with the largest variation appearing in groups with limited cohort support. RealLit effectively showcases how ensembling, often treated as a reliable way to improve performance, can have uneven effects across student groups and even reduce performance for some. Because these shifts are largely hidden in aggregate metrics, evaluating with student-level metadata is necessary to determine whether a configuration choice benefits all students or only some of them.

TabPFN-v2 Fine-Tuning: Modest Gains, Limited Transfer: K-ensembling and imputation are inference-time levers. We now ask whether training-time adaptation provides what they cannot. We compare four TabPFN-v2 variants across all four cohorts at $K \in \{1, 8\}$ and under all three missingness mechanisms (full setup in App. G): Zero-shot (released checkpoint, no further training), FT-on-target (fine-tuned on the current cohort's training fold), FT-from-D1 (fine-tuned on D1 and used for inference only), and FT-init-D1 (initialised from FT-from-D1 and then fine-tuned a second time on the current cohort). No fine-tuning variant beats Zero-shot consistently across cohorts: scattered cell-level wins exist (mostly for FT-on-target on D1 under MAR/MCAR), but no variant wins across both K values and all three mechanisms on more than a single cohort (Tabs. 16 and 17).

Longitudinal Imputation: We examine whether explicitly using repeated measurements improves downstream classification. D4 is the only cohort in which the same instruments are measured at consecutive grades, allowing us to evaluate two simple, training-free longitudinal references, LOCF and LinGrowth, on its G1+G2 panel. Both achieve higher mean ROC-AUC than mean-fill across all four classifiers at native missingness and match or exceed it under 30% injected missingness, with MCAR, MAR, and MNAR pooled. Accuracy gains are less consistent and depend on the classifier. Results are detailed in Sec. D.5.

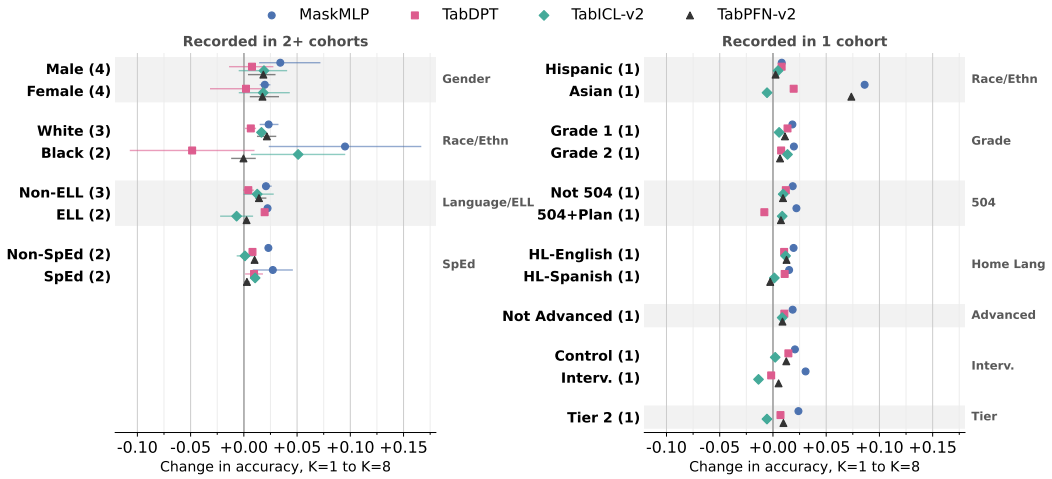

Figure 2: **Subgroup Accuracy Change with K.** Change in per-subgroup accuracy from K=1 to K=8 for each classifier, pooled across the RealLit cohorts that record each subgroup (count in parentheses). Markers show the average change. Lines span the smallest to largest change within any single cohort, so a line crossing zero means cohorts disagree on direction. Left: subgroups recorded in two or more cohorts. Right: subgroups recorded in one cohort.

RealLit as a Tool Toward Equitable Early Literacy Prediction: Early literacy outcome prediction can inform which students are flagged for screening, intervention, or additional support, such that precision is crucial. Here, a missed prediction can delay help for a struggling reader during the years when learning assistance could be most effective, and if errors concentrate in particular student groups, predictive tools risk reinforcing gaps that they are meant to close. Within this context, our findings suggest four implications. First, subgroup behavior should be audited before deployment. Aggregate AUC differences of ± 0.005 between classifiers routinely accompany per-subgroup ACC gaps an order of magnitude larger, and ensembling can reduce accuracy for some student groups even as it helps others (Fig. 2). The largest and most model-dependent shifts appear across race and ethnicity. Second, newer model releases should be benchmarked against their predecessors rather than assumed to improve on them, as TabPFN-v3 underperforms TabPFN-v2 on RealLit under the baseline setting (Sec. D.3). Third, when compute is constrained, K-ensembling is a better investment than learned imputation, subject to the subgroup audit above. At K=8, the trivial Mean baseline matches the diffusion- and flow-based imputers on the three foundation classifiers. Finally, the ECRI release (D1) is best treated as a sample-size-limited reference rather than a figure of merit for modeling progress (Tab. 2). These claims are scoped to the four English-language K–Grade 4 cohorts of RealLit. Because each record carries student-level metadata, performance can be analyzed across models, ensemble sizes, and missing-data strategies. RealLit can be viewed as a prerequisite for equitable use of machine learning for early literacy prediction. We emphasize that despite our best efforts to maintain diverse and representative settings, bilingual and dual-language cases, non-U.S. educational systems, and reading constructs below the kindergarten grain are appropriate targets for future benchmark expansion, as discussed next.

6 Limitations

While RealLit expands beyond prior coarse early literacy datasets through fine-grained, demographic-aware student-level labels, several dimensions remain uncovered. It spans multiple U.S. states and school contexts but is not nationally representative, and its outcomes are limited to English-language reading. Bilingual and dual-language settings are absent, even though they are central to many classrooms and students' literacy trajectories. We hope future work expands multilingual data collection and intersectional analyses. Moreover, while the student-level demographic attributes enable new forms of auditing, they should not be mistaken for complete or neutral descriptions of students. Some groups are also too small for stable subgroup estimates, so subgroup results should be interpreted as signals for *further investigation* rather than final claims about fairness. In deployment,

demographic-aware evaluation should not become demographic targeting. Our experiments use a binary label for clear evaluation, but this compresses a child's reading development into a thresholded outcome. A model score should never replace teacher judgment, family knowledge, or evaluation by trained professionals. Responsible use would require local validation, transparent uncertainty communication, monitoring across student groups, and policies that keep educators and families in the decision process. RealLit is therefore best understood as a step toward more responsible evaluation, not a complete solution for deploying foundation models in schools.

7 Conclusion

Using RealLit, we show that current tabular foundation models can perform competitively on longitudinal early-literacy prediction, but their behavior varies meaningfully across cohorts, missingness conditions, and student populations. Aggregate metrics can conceal subgroup-level differences, newer model variants do not consistently improve over earlier ones, learned imputation provides inconsistent gains over simple missing-data handling, and task-specific fine-tuning offers limited benefit. In contrast, inference-time ensembling emerges as a useful practical intervention, although its effects must still be examined across subgroups.

These findings point to several concrete next steps. Models should be evaluated across schools and cohorts rather than on a single dataset, missingness should be treated as part of the prediction problem rather than only as a preprocessing nuisance, and subgroup behavior should be audited alongside aggregate performance. Methodologically, there is also room to develop models that better exploit longitudinal structure, jointly reason about missingness and prediction, and adapt across cohorts without sacrificing reliability. Broadly, RealLit creates a testbed for moving beyond whether tabular foundation models work on early-literacy data to understanding when they work, for whom, and under which data conditions. We hope this benchmark supports the development of educational models that remain reliable under the heterogeneity and incompleteness that characterize real school settings.

AI use statement

Large language models were used solely to assist with spelling, grammar, and language editing. All suggested edits were reviewed by the authors, who retain full responsibility for the final manuscript.

Reproducibility statement

To support reproducibility of RealLit, we document cohort composition, outcome definitions, preprocessing, and data provenance in Secs. A and K. Sec. C describes classifier and imputer configurations, training-fold-only fitting of learned imputers, cross-validation, random-seed control, missingness stress tests, and permutation ensembling. Our complete software environment, including the used computing resources, computational cost, and evaluation-code and data-access information are provided in Sec. C.8. D1 is obtained through the original public ECRI release. While Shangguan et al. (2024) report 6,916 samples for D1, we were able to identify only 6,803 students in the publicly available data. We are unsure of the source of this discrepancy, as we employed the released CSV file. The TabPFN fine-tuning setup and ablations are described in Sec. G. We also document the implementation choices and known reproducibility limitations of our MaskMLP baseline in Sec. I.1. Resources are available at our anonymous repository: https://github.com/anonymous-submission-site/real-lit.

References

Arthur Asuncion, David Newman, et al. UCI machine learning repository, 2007.

Lucy Barnard-Brak, Tara Stevens, and William Ritter. Reading and mathematics equally important to science achievement: Results from nationally-representative data. *Learning and Individual Differences*, 2017.

Leo Breiman. Random forests. *Machine Learning*, 2001.

Okan Bulut and Damien C. Cormier. Validity evidence for progress monitoring with star reading: Slope estimates, administration frequency, and number of data points. *Frontiers in Education*, 2018.

Ramiro D Camino, Christian A Hammerschmidt, and Radu State. Improving missing data imputation with deep generative models. *arXiv preprint arXiv:1902.10666*, 2019.

Jehanzeb R Cheema. A review of missing data handling methods in education research. *Review of Educational Research*, 2014.

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

Carol McDonald Connor, Shayne B Piasta, Barry Fishman, Stephanie Glasney, Christopher Schatschneider, Elizabeth Crowe, Phyllis Underwood, and Frederick J Morrison. Individualizing student instruction precisely: Effects of child $\times$ instruction interactions on first graders' literacy development. *Child Development*, 2009.

André F Cruz and Moritz Hardt. Unprocessing seven years of algorithmic fairness. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2024.

Tianyu Du, Luca Melis, and Ting Wang. Remasker: Imputing tabular data with masked autoencoding. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2024.

Greg J. Duncan, Chantelle J. Dowsett, Amy Claessens, Katherine Magnuson, Aletha C. Huston, Pamela Klebanov, Linda S. Pagani, Leon Feinstein, Mimi Engel, Jeanne Brooks-Gunn, Holly Sexton, Kathryn Duckworth, and Crista Japel. School readiness and later achievement. *Developmental Psychology*, 2007.

Craig Enders, Samantha Dietz, Marjorie Montague, and Jennifer Dixon. Modern alternatives for dealing with missing data in special education research. *Advances in Learning and Behavioral Disabilities*, 2006.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A living benchmark for machine learning on tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021.

Kevin J. Grimm. Longitudinal associations between reading and mathematics achievement. *Developmental Neuropsychology*, 2008.

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 arXiv:2511.08667*, 2025.

Shanna Hagan-Burke, Mack D. Burke, and Clay Crowder. The convergent validity of the dynamic indicators of basic early literacy skills and the test of word reading efficiency for the beginning of first grade. *Assessment for Effective Intervention*, 2006.

Ahatsham Hayat and Mohammad Rashedul Hasan. A context-aware approach for enhancing data imputation with pre-trained language models. *arXiv preprint arXiv:2405.17712*, 2024.

Xinrui He, Yikun Ban, Jiaru Zou, Tianxin Wei, Curtiss Cook, and Jingrui He. LLM-Forest: Ensemble learning of LLMs with graph-augmented prompts for data imputation. In *Findings of the Association for Computational Linguistics (ACL)*, 2025.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, Frank Hutter, et al. Accurate predictions on small data with a tabular foundation model. *Nature*, 2025.

Connie Juel. Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, 1988.

Benjamin Jäger, Nick Erickson, Léo Grinsztajn, Felix Birkel, Klemens Flöge, Oscar Key, Kürşat Kaya, Jonas Kübler, Adèle Frankel, Tobias Schröder, Anurag Garg, Jan Hendrik Metzen, David Salinas, Simon Bing, Kristina Collins, Tuana Çelik, Vahid Balazadeh, Lydia Sidhoum, Tomás Pereda, Brendan Roof, Andrej Tschalzev, Siyuan Guo, Philipp Singer, Lennart Purucker, Jake Robertson, Marie Salmon, Philipp Jund, Jerry Chen, Diana Kriuchkova, Arthur Cahu, Eliott Kalfon, Adrian Hayler, Georg Grab, Vitor Monteiro, Lilly Wehrhahn, Dominik Safaric, Clara Cornu, Alan Arazi, Rylee Grace, Simone Alessi, Mihir Manium, Bernhard Schölkopf, Yann LeCun, Madelon Hulsebos, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-3.5: Technical Report, September 2026.

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 (NeurIPS)*, 2017.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2023.

Zachary C Lipton, David C Kale, Randall Wetzel, et al. Modeling missing data in clinical time series with RNNs. In *Proc. of Machine Learning for Healthcare (MLHC)*, 2016.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Hamidreza Kamkari, Alex Labach, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony L. Caterini. TabDPT: Scaling tabular foundation models on real data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Mónica V Martins, Daniel Tolledo, Jorge Machado, Luís MT Baptista, and Valentim Realinho. Early prediction of student's performance in higher education: A case study. In *World Conference on Information Systems and Technologies*, 2021.

James R. McBride, Jim Ysseldyke, Michael Milone, and Eric Stickney. Technical adequacy and cost benefit of four measures of early literacy. *Canadian Journal of School Psychology*, 2010.

Sean McWeeny, Soujin Choi, June Choe, Alexander LaTourrette, Megan Y. Roberts, and Elizabeth S. Norton. Rapid automatized naming (RAN) as a kindergarten predictor of future reading in english: A systematic review and meta-analysis. *Reading Research Quarterly*, 2022.

Monica Melby-Lervåg, Solveig-Alma Halaas Lyster, and Charles Hulme. Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin*, 2012.

Paul L Morgan, George Farkas, Paula A Tufis, and Rayne A Sperling. Are reading and behavior problems risk factors for each other? *Journal of learning disabilities*, 2008.

National Center for Education Statistics. 2024 NAEP reading assessment: Results at grades 4 and 8 for the nation, states, and districts. Technical Report NCES 2024218, U.S. Department of Education, Institute of Education Sciences, 2025.

Francisco Javier Ordóñez and Daniel Roggen. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors*, 2016.

Jason R. Parkin. The simple views of reading and writing in the kaufman test of educational achievement, third edition. *Journal of Psychoeducational Assessment*, 2021.

Rauno Parrila, John R. Kirby, and Lynn McQuarrie. Articulation rate, naming speed, verbal short-term memory, and phonological awareness: Longitudinal predictors of early reading development? *Scientific Studies of Reading*, 2004.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research (JMLR)*, 2011.

Chao-Ying Joanne Peng, Michael Harwell, Show-Mann Liou, and Lee H. Ehman. Advances in missing data methods and implications for educational research. In *Real Data Analysis*. Information Age Publishing, 2006.

James L Peugh and Craig K Enders. Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research (RER)*, 2004.

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

Ivan Rubachev, Akim Kotelnikov, Nikolay Kartashev, and Artem Babenko. On finetuning tabular foundation models. *arXiv preprint arXiv:2506.08982*, 2025.

Donald B. Rubin. Inference and missing data. *Biometrika*, 1976.

Reza Shahbazian and Irina Trubitsyna. DEGAIN: Generative-adversarial-network-based missing data imputation. *Information*, 2022.

Zhongkai Shangguan, Zanming Huang, Eshed Ohn-Bar, Ola Ozernov-Palchik, Derek Kosty, Michael Stoolmiller, and Hank Fien. Scalable early childhood reading performance prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Vaidotas Simkus and Michael U Gutmann. CFMI: Flow matching for missing data imputation. *arXiv preprint arXiv:2506.09258*, 2025.

Jean Louise M Smith, Nancy J Nelson, Hank Fien, Keith Smolkowski, Derek Kosty, and Scott K Baker. Examining the efficacy of a multitiered intervention for at-risk readers in grade 1. *The Elementary School Journal (ESJ)*, 2016.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for Mixed-Type data. *Bioinformatics*, 2012.

Aditya Tanna, Pratinav Seth, Mohamed Bouadi, and Vinay Kumar Sankarapu. Exploring fine-tuning for tabular foundation models. In *Proceedings of the ACM Web Conference (W4A)*, 2026.

Joseph K Torgesen. Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice*, 2000.

U.S. Department of Education, Office for Civil Rights. Section 504. https://www.ed.gov/laws-and-policy/individuals-disabilities/section-504, 2026. Accessed 2026.

Stef Van Buuren and Karin Groothuis-Oudshoorn. MICE: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 2011.

H. Atakan Varol, Subramani Mani, Donald L. Compton, Lynn S. Fuchs, and Douglas Fuchs. Early prediction of reading disability using machine learning. In *AMIA Annual Symposium Proceedings*, 2009.

Frank R Vellutino, Donna M Scanlon, Sheila Small, and Diane P Fanuele. Response to intervention as a vehicle for distinguishing between children with and without reading disabilities: Evidence for the role of kindergarten and first-grade interventions. *Journal of Learning Disabilities*, 2006.

Maxwell A Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A Tailor, Ahmed Metwally, A Ali Heydari, Yuwei Zhang, Jake Garrison, et al. LSM-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025.

Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. A closer look at deep learning methods on tabular datasets. *arXiv preprint arXiv:2407.00956*, 2024.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. *International Conference on Machine Learning (ICML)*, 2018.

Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J Kochenderfer, and Jure Leskovec. Handling missing data with graph representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jiahao Yu, Qizhen Ying, Leyang Wang, Ziyue Jiang, and Song Liu. Missing data imputation by reducing mutual information with rectified flows. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Hengrui Zhang, Liancheng Fang, Qitian Wu, and Philip S. Yu. DiffPuter: An EM-driven diffusion model for missing data imputation. In *International Conference on Learning Representations (ICLR)*, 2025.

Shuhan Zheng and Nontawat Charoenphakdee. Diffusion models for missing value imputation in tabular data. In *Workshop on Table Representation Learning, Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

APPENDIX CONTENTS

CONTENTS

A Dataset Details **16**

A.1 D1 – ECRI (Enhanced Core Reading Instruction) 16

A.2 D2 – Kindergarten Screening Study 16

A.3 D3 – District Records 17

A.4 D4 – Longitudinal Cohort 17

A.5 Reading Measures and Assessment Instruments 17

A.6 Target Label Construction 18

A.7 Cohort Target Construction 19

A.8 Demographic Composition 20

B Cohort Embedding Visualizations **20**

B.1 Pooled t-SNE Projection 21

B.2 Per-Cohort Zoom 21

B.3 Demographic Embedding Grid 23

C Implementation Details **23**

C.1 Classifiers 23

C.2 Imputers 24

C.3 Missingness Indicator Augmentation 25

C.4 Permutation Ensembling 25

C.5 Cross-Validation Protocol 26

C.6 Missingness Stress-Test Mechanisms 26

C.7 Hyperparameters 26

C.8 Compute and Reproducibility 27

D Additional Results **28**

D.1 Dataset Differences 28

D.2 Regression 30

D.3 Comparison of TabPFN Versions 30

D.4 K-sweep $\times$ Rate $\times$ Classifier 30

D.5 Longitudinal Imputation 32

E Imputation Reconstruction Quality **33**

E.1 Reconstruction Mean Absolute Error (MAE) 33

E.2 Reconstruction Root Mean Squared Error (RMSE) 33

F Demographic Subgroup Analysis **33**

F.1 Subgroup K-Sweep – Accuracy 34
F.2 Subgroup K-Sweep – AUC 34
F.3 Imputation Quality across Subgroups – MAE 35
F.4 Imputation Quality across Subgroups – RMSE 35

G TabPFN-v2 Fine-Tuning **36**
G.1 Setup 36
G.2 Per-Dataset Accuracy 37
G.3 Per-Mechanism AUC across Variants 39
G.4 Discussion: Modest Gains, Poor Transfer 40

H Statistical Significance **40**

I Limitations and Failure Modes **41**
I.1 MaskMLP Reproducibility 41
I.2 Subgroup Sample-Size Caveats 41
I.3 At-Risk Threshold Cross-Dataset Comparability 42
I.4 Imputation Coverage 42
I.5 Geographic and Linguistic Scope 42

J Ethics and Responsible Use **42**
J.1 Responsible Use of Student Records 42
J.2 Demographic Auditing Principles 42
J.3 Deployment Caveats 42

K Datasheet for RealLit **43**

A Dataset Details

This appendix provides the full dataset documentation for the four longitudinal early-literacy datasets that comprise RealLit: D1 (ECRI, public), D2 (kindergarten screening study), D3 (district records), and D4 (longitudinal cohort). We describe data collection, reading measures, label construction, demographic coverage, and missingness structure.

A.1 D1 – ECRI (Enhanced Core Reading Instruction)

D1 is the publicly available ECRI benchmark from Shangguan *et al.* (Shangguan et al., 2024). The parent study spans 44 elementary schools, 6,916 students, and 172 teachers, with multi-day assessor training, scoring fidelity checks, and intervention-linked classroom context. The released benchmark configuration that we evaluate contains 6,803 rows, 16 predictors, and two released prediction targets: Woodcock Reading Mastery Word Identification and Word Attack non-response, with 40.8% cell-wise missingness. Only 946 students carry valid end-of-year labels for the released tasks. We therefore retain D1 as a prior-work reference point. The released version reports demographics only in aggregate at the school level, which precludes the per-student subgroup analysis that motivates RealLit. Including D1 in our evaluation grid nonetheless allows direct comparison to prior tabular early-literacy work at the level of average AUC and serves as a control for whether trends observed on D2-D4 also hold on the only previously released benchmark in this domain. Shangguan *et al.* (Shangguan et al., 2024) report 6,916 samples for D1. However, we were able to identify only 6,803 students in the publicly available data. We are unsure of the source of this discrepancy, as we employed their released CSV file.

A.2 D2 – Kindergarten Screening Study

D2 is an IRB-approved longitudinal kindergarten study. The parent study originally recruited 4,099 students across kindergarten through Grade 3 from 33 public, private, and virtual schools in five U.S. states (Florida n=2,185, Georgia n=243, South Carolina n=1,030, Massachusetts n=352, Oregon n=289); by grade, the parent sample includes 1,927 kindergarten, 1,034 first-grade, 688 second-grade, and 450 third-grade students. The benchmarked subset that we evaluate contains 1,039 kindergarten students with fall screening measures, partially observed per-student demographic attributes, and end-of-year KTEA-3 and CELF-5 outcomes. Assessments were administered through iPads in one-to-one sessions by trained assessors who completed written and in-person training, guided practice, and a fidelity check requiring 100% accuracy before any student-facing session. Students provided verbal assent and could withdraw from assessment at any point.

The fall screening battery covers early phonological awareness, vocabulary, verbal short-term memory, rapid naming, and a partially administered WPPSI-4 cognitive battery (Parrila et al., 2004; Melby-Lervåg et al., 2012; McWeeny et al., 2022; Parkin, 2021). The four end-of-year outcomes are KTEA-3 Nonword Decoding (NWD), KTEA-3 Letter and Word Recognition (LWR), CELF-5 Sentence Comprehension (SC), and CELF-5 Recalling Sentences (RS). Outcome coverage varies by target: 515 students have valid KTEA-3 NWD labels, 514 have KTEA-3 LWR, 205 have CELF-5 SC, and 488 have CELF-5 RS. Mean cell-wise missingness is 47.3%, dominated by blockwise co-missingness from study design and assessment battery selection.

D2 provides student-level demographic fields that are themselves partially observed, reflecting realistic study participation and survey coverage. Gender is recorded for 543 students (273 male, 270 female), ethnicity for 523 students (472 not Hispanic, 51 Hispanic), race for 540 students (297 White, 160 Black or African American, 60 multi-racial, 15 Asian, 6 Other, 1 Native Hawaiian or Other Pacific Islander, 1 American Indian or Alaska Native), and home language for 553 students (507 English-only, 46 with another language spoken at home). We retain partially observed demographic records rather than filtering to complete cases. Complete-case filtering would produce a cleaner table at the cost of removing the students whose data patterns are most informative for responsible evaluation, since selection-bias risk increases precisely when demographic observability correlates with assessment access.

A.3 D3 – District Records

D3 contains anonymized student-level records from a single Florida school district, contributed across two complementary cohorts. The Grade 1-2 panel spans 2,088 students across 14 schools and links demographics, DIBELS fall benchmark scores, FAST STAR progress monitoring, i-Ready diagnostic results, and weekly i-Ready usage telemetry (Hagan-Burke et al., 2006; McBride et al., 2010; Bulut & Cormier, 2018). The Grade 3 cohort spans 1,137 students and combines attendance, discipline, reading assessments, and a fully observed spring FAST ELA outcome. D3 captures contemporary district operational data, and missingness reflects record linkage, assessment windows, and operational data availability (mean cell-wise missingness 10.3% in the Grade 1-2 panel and 9.1% in the Grade 3 cohort). This makes D3 the most realistic stress test in RealLit for whether models and imputers behave reliably on the kind of multi-source linked records that schools and districts actually maintain.

D3 provides the richest student-level demographic and service indicators in RealLit. The Grade 1-2 cohort includes gender, grade, ethnicity, Limited English Proficient (LEP) services, Exceptional Student Education (ESE) status, Section 504 status, gifted status, free or reduced lunch (FRL), homelessness, and primary home language, with all displayed demographic fields fully populated. The Grade 1-2 cohort comprises 1,121 male and 967 female students. There are 863 White, 655 Hispanic, 383 Black or African American, 147 Other, 37 Asian, 2 American Indian or Alaska Native, and 1 Native Hawaiian or Other Pacific Islander students, 265 students currently receiving LEP services, 415 students receiving ESE services, 1,154 students directly certified for lunch assistance, and 376 students with Spanish as the primary home language. The Grade 3 cohort has fully observed indicators for economic disadvantage, ESE, Section 504, and English Language Learner (ELL) status: 773 economically disadvantaged students, 218 receiving ESE services, 112 with Section 504 status, and 93 ELL students. These attributes attach to individual student records, which enables the subgroup reliability analysis reported in App. F.

A.4 D4 – Longitudinal Cohort

D4 is a longitudinal reading and writing dataset that follows the same cohort across Grades 1-4 in six Florida schools. The benchmarked release contains 318 students, with yearly measures from 16 assessment instruments covering oral reading fluency, decoding, vocabulary, comprehension, working memory, spelling, and related constructs (349 features after harmonization). D4 is the smallest cohort in RealLit but it is also the most deeply longitudinal one: it asks whether models trained on early-grade literacy profiles remain reliable as outcomes, assessment batteries, and missingness change across years.

Demographic coverage is high for the available fields, with race and sex observed for 316 of 318 students. The dominant missingness mechanism in D4 is longitudinal attrition: grade-level observation decreases from 318 students in Grade 1 to 220 in Grade 4, and mean field missingness rises from 3.8% in Grade 1 to 36.4% in Grade 4. This pattern matters because it stresses an assumption that learned imputers and tabular foundation models often share: observation patterns are roughly homogeneous across rows. In D4, observation patterns are determined by the year a student left the study. School transitions, mobility, and family-level factors that drive attrition also predict literacy trajectories.

A.5 Reading Measures and Assessment Instruments

RealLit aggregates assessments from four cohorts that were not designed to share a single battery. We list each instrument together with the construct it measures, the cohorts in which it appears, and the role it plays in the prediction task. Where an instrument contributes a target, we say so explicitly.

KTEA-3 – Kaufman Test of Educational Achievement, 3rd Ed. (D2): A standardized individually administered achievement battery. We use two subtests as outcomes in D2: Nonword Decoding (NWD) measures phonological decoding of pseudowords (e.g., word-attack ability), and Letter and Word Recognition (LWR) measures sight-word and real-word reading. Both are reported as age- or grade-based standard scores (M=100, SD=15). KTEA-3 standard scores are also used to define the static at-risk indicator in D2 (App. A.6).

CELF-5 – Clinical Evaluation of Language Fundamentals, 5th Ed. (D2): A clinician-administered language battery. We use two subtests as outcomes in D2: Sentence Comprehension (SC) probes receptive processing of sentence-level structure, and Recalling Sentences (RS) probes verbal working memory by asking students to repeat sentences of increasing length and complexity. Scores are reported on the standard CELF-5 scaled-score metric.

WPPSI-IV – Wechsler Preschool and Primary Scale of Intelligence, 4th Ed. (D2): A partially administered cognitive battery from the WPPSI-IV is included in the D2 fall screening. We do not use these subtests to construct labels—they enter as predictors that capture verbal reasoning, visual-spatial, and working-memory constructs.

Phonological Awareness, Vocabulary, Verbal STM, and Rapid Naming (D2): The D2 fall battery additionally includes brief measures of early phonological awareness, vocabulary, verbal short-term memory, and rapid automatized naming (Parrila et al., 2004; Melby-Lervåg et al., 2012; McWeeny et al., 2022; Parkin, 2021). These measures are well-attested early predictors of later decoding and word-recognition outcomes and serve as the primary predictors entering the four KTEA-3/CELF-5 outcome tasks.

DIBELS – Dynamic Indicators of Basic Early Literacy Skills (D3): DIBELS is an individually administered curriculum-based measurement battery used as a fall benchmark in the D3 Grade 1-2 panel (Hagan-Burke et al., 2006). DIBELS subtests assess phoneme segmentation, nonword reading fluency, and oral reading fluency. Scores enter as predictors.

FAST STAR – Florida Assessments for Instruction in Reading, STAR (D3): A computer-adaptive progress-monitoring assessment used in the Grade 1-2 panel as a periodic predictor of reading growth (Bulut & Cormier, 2018).

i-Ready Diagnostic and Usage Telemetry (D3): i-Ready provides a computer-adaptive diagnostic with sub-domain scaled scores, plus operational usage telemetry (minutes of practice per week). We treat the diagnostic scaled scores as standard predictors. The usage telemetry is itself a feature: it is informative about engagement and instructional dosage, and its missingness encodes whether a student used the platform at all in a given week.

FAST ELA – Florida Assessments for Instruction in Reading, ELA (D3): A standardized end-of-year ELA assessment used as the outcome in both D3 cohorts. The Grade 3 cohort uses spring FAST ELA proficiency as the supervised target (McBride et al., 2010), and the Grade 1-2 panel uses the prior FAST ELA achievement level as the source-window predictor for the longitudinal recovery target (App. A.6).

WRAT-3 Spelling – Wide Range Achievement Test, 3rd Ed. (D4): A standardized spelling subtest reported as a standard score (M=100, SD=15). The Grade 3 WRAT-3 Spelling standard score defines the longitudinal recovery target in D4 (App. A.6).

D4 Longitudinal Battery (16 instruments): The D4 yearly battery covers oral reading fluency, decoding, vocabulary, comprehension, working memory, spelling, and writing constructs across 16 named instruments harmonized into a 349-feature panel.

A.6 Target Label Construction

The four cohorts use binary targets that share a common purpose, flagging students who may benefit from additional instructional support, but differ in instrument, threshold, and follow-up window. We retain the per-cohort definitions used by each parent study rather than imposing a unified threshold, because doing so would distort the screening conventions under which each dataset was originally collected and would erase the construct-level differences across reading outcomes.

D2 (static threshold, KTEA-3). Following the screening convention of the parent study (Smith et al., 2016), kindergarten students are classified as at-risk when their KTEA-3 standard score falls below 90 (normal curve equivalent below the 25th percentile, M=100, SD=15). The benchmarked D2 cohort exposes two independent at-risk targets at this threshold: KTEA-3 Nonsense Word Decoding

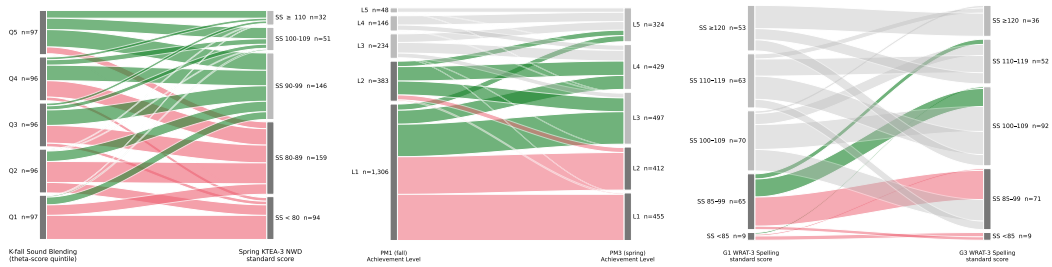

Figure 3: Per-cohort target construction. Each panel shows the flow from the source-window predictor (left axis) to the target-window outcome (right axis). Color key shared across panels: green flows = $Y=1$ (recovered / proficient at follow-up), pink flows = $Y=0$ (still below threshold at follow-up), grey flows = rows excluded from the supervised cohort because they were already at-or-above the threshold at the source window. D2 (left) uses a static at-risk indicator on the KTEA-3 Nonword Decoding (NWD) standard score (threshold < 90), and flows are quintiles of the K-fall Sound Blending predictor. D3 (center) and D4 (right) use longitudinal recovery indicators (proficient vs. not-proficient on spring FAST ELA achievement level and G3 WRAT-3 Spelling SS, respectively), and excluded rows are those already proficient (D3) or already at/above average (D4) at the source window.

(NWD, the static at-risk indicator visualized in Fig. 3) and KTEA-3 Letter and Word Recognition (LWR).

D3 (longitudinal recovery, FAST ELA). A student in the Grade 1-2 panel or the Grade 3 cohort is labeled $Y=1$ ("recovered / proficient") if they reach the proficient FAST ELA achievement level at the target spring window, and $Y=0$ otherwise. Students who are already proficient at the source window are excluded from the supervised cohort because the recovery target is undefined for them.

D4 (longitudinal recovery, WRAT-3 Spelling). A student is labeled $Y=1$ if the Grade 3 WRAT-3 Spelling standard score is at or above average at the target window, and $Y=0$ otherwise. Students already at or above average at the source window are excluded for the same reason.

Two consequences follow. First, raw AUC values are not directly comparable across cohorts: the labels carry different positive-class prevalences (Tab. 1), and reflect different constructs (decoding in D2, ELA proficiency in D3, spelling in D4) measured against different reference points. Cross-cohort summary numbers therefore aggregate over heterogeneous tasks. Second, these labels are screening signals, not diagnostic decisions. They should not be interpreted as proxies for eligibility, placement, or denial of services, and any deployment context would require local validation against decisions that schools and families actually make.

A.7 Cohort Target Construction

Fig. 3 visualizes the target-construction logic for D2, D3, and D4 using a shared color key: green flows are $Y=1$ (recovered or proficient at the target window), pink flows are $Y=0$ (still below the threshold at the target window), and grey flows are rows excluded from the supervised cohort because they were already at or above the threshold at the source window.

The D2 panel (left) bins kindergarten-fall Sound Blending into quintiles on the left axis (the source-window predictor) and maps each quintile to the static at-risk indicator on the KTEA-3 Nonword Decoding standard score on the right axis (threshold < 90). Because D2 uses a static threshold, no rows are excluded.

The D3 panel (center) maps prior FAST ELA achievement level on the left axis to spring FAST ELA proficiency on the right axis. Green flows correspond to students who reach proficient at the target window, pink flows correspond to students still below proficient, and grey flows correspond to students who were already proficient at the source window and are therefore excluded from the supervised cohort because the recovery target is undefined for them.

The D4 panel (right) maps a Grade 1-2 WRAT-3 Spelling standard-score quintile on the left axis to the Grade 3 WRAT-3 Spelling standard score on the right axis (at or above average vs. below average). Grey flows are students already at or above average at the source window, excluded for the same recovery-construct reason.

The shared color key makes recovery (green) versus persistence (pink) directly comparable across panels even though the underlying constructs differ.

A.8 Demographic Composition

Fig. 4 summarizes the pooled demographic composition of RealLit across the 14 canonical demographic axes that we audit in App. F: gender, ethnicity, race, home language, English-learner status, special-education status, free-or-reduced-lunch status, gifted status, Section 504 status, homelessness, grade, and three further axes that vary by cohort availability. Each donut shows the count and percentage per subgroup for one axis, pooled over the cohorts that contribute that field, and subgroup labels are kept consistent with the per-cohort documentation in Apps. A.2, A.3 and A.4. Subgroups whose pooled support is below the reliability threshold used in our subgroup tables ($n < 8$ on the mean test fold) are flagged separately in App. I.2.

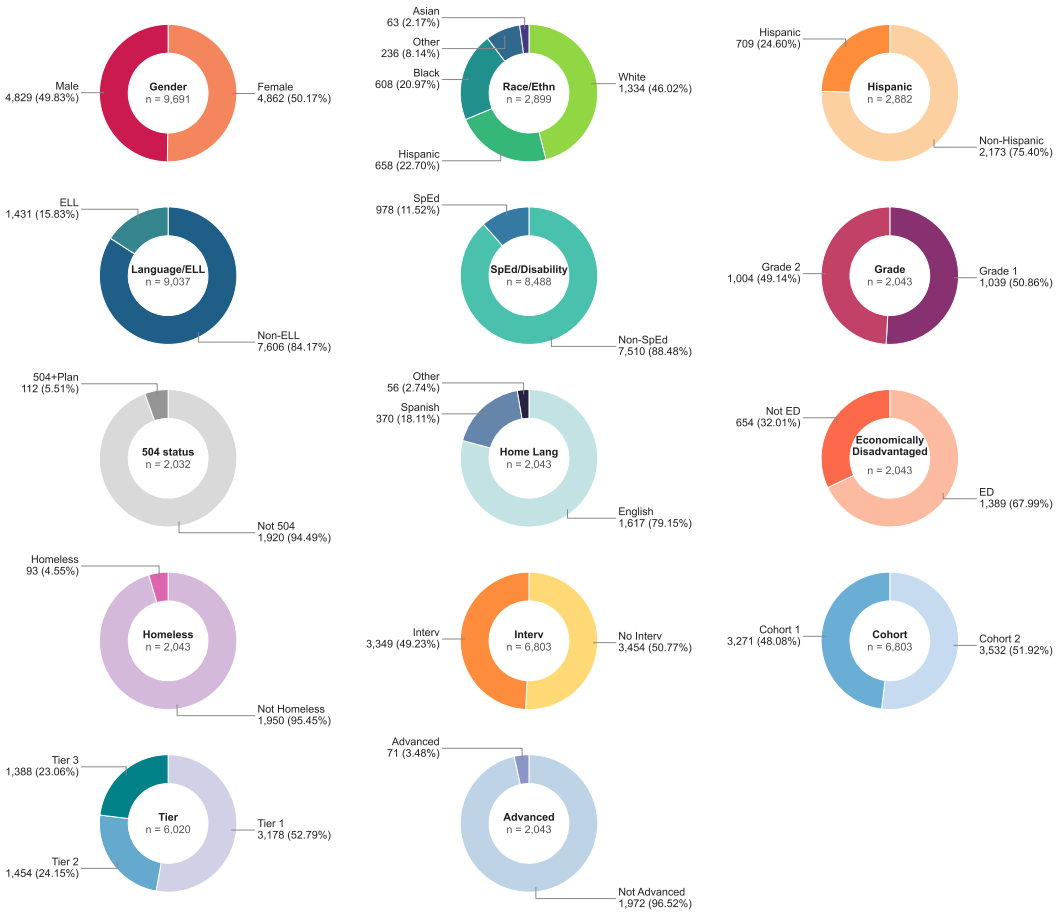

Figure 4: Demographic composition of RealLit across all 14 canonical demographic axes. Each donut shows count and percentage per subgroup for one axis.

B Cohort Embedding Visualizations

We visualize the joint feature space across cohorts using TabPFN-v2 embeddings projected to two dimensions via t-SNE.

B.1 Pooled t-SNE Projection

Fig. 5 shows the cohort-faceted t-SNE projection colored by at-risk status. Two structural observations are useful for the rest of the appendix. First, the four cohorts occupy substantially distinct regions of the shared TabPFN-v2 embedding. Cohort identity is not provided as a feature; the separation is a consequence of cohort-specific reading constructs, assessment batteries, grade ranges, and missingness patterns documented in App. A. Second, within each cohort, the at-risk overlay (filled markers) is visibly non-uniform: at-risk students concentrate in identifiable sub-regions of their cohort cloud, which is consistent with the moderate-to-strong per-cohort AUC values reported in Tab. 2 and the rate-stratified results in Tab. 10.

Taken together, these two facts have a direct implication for cross-cohort transfer. A test row drawn from one cohort sits in a region of the embedding that is only sparsely populated by training rows from the other cohorts, and in-context learners that score a test row by reference to nearby support points therefore behave very differently when the in-context support is in-cohort versus out-of-cohort. This is consistent with the brittle D1→{D2,D3,D4} transfer reported in App. G, where TabPFN-v2 fine-tuned on D1 fails to improve on the zero-shot prior for the other cohorts and in some configurations regresses below it. The cohort separation visible in Fig. 5 predicts the transfer behavior we measure.

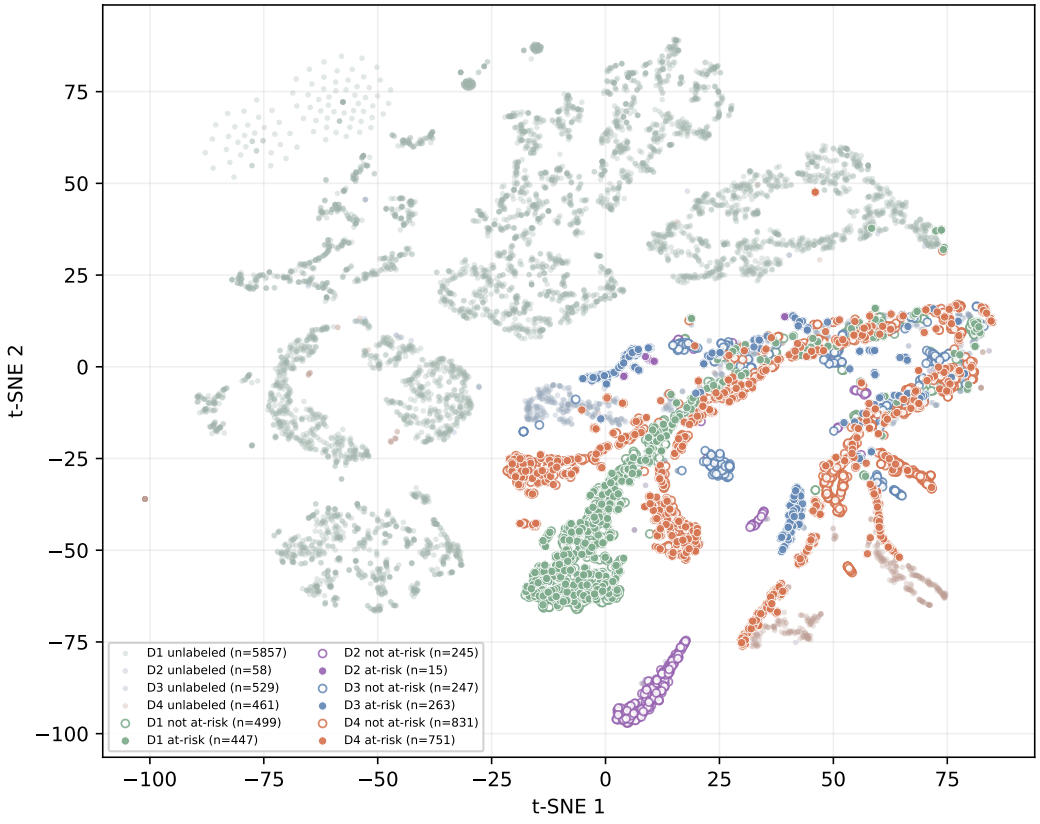

Figure 5: TabPFN-v2 t-SNE projection colored by at-risk status, cohort-faceted (D1-D4).

B.2 Per-Cohort Zoom

Figs. 6 and 7 zoom into the D2 and D3 sub-clouds with labelled feature-space neighbourhoods. The neighbourhoods correspond to centroids in the underlying predictor space: each labelled region groups students whose screening or operational measures co-vary in interpretable ways (Sound Blending and KTEA-3 sub-scores in D2; DIBELS, FAST STAR, and i-Ready sub-scores in D3). The two zooms therefore show that the embedding is internally structured along the predictor axis, and that this structure is visible at the resolution of dozens of students.

This sub-cohort structure is what makes in-context tabular foundation models effective on RealLit: a small in-context support (e.g., the $K=8$ ensemble configurations of App. D.4) is enough to localize a test row within the right neighbourhood when the support comes from the same cohort. The fact that D2 and D3 both exhibit interpretable sub-cohort geometry, despite using different assessment batteries, helps explain why the relative ranking of classifiers is broadly stable across cohorts even when absolute AUC differs. We restrict the zoom to D2 and D3 because they are the two cohorts with enough valid-labeled rows to make sub-cohort centroids meaningful, D1's 946 valid rows produce noisier local geometry, and D4's 318 students are too few to support a stable sub-cohort decomposition.

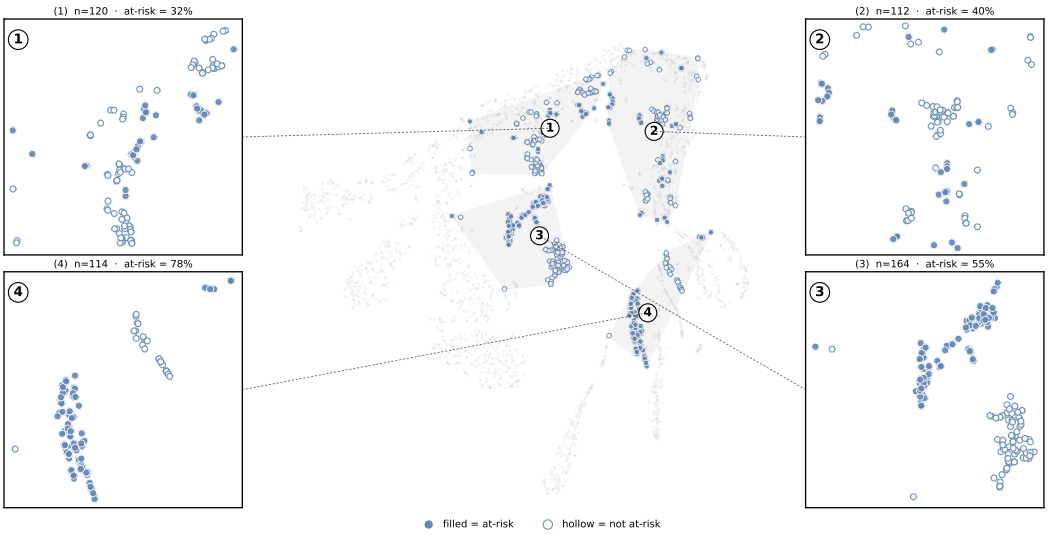

Figure 6: Cohort zoom restricted to the D2 sub-cloud, with feature-space neighbourhoods labelled by predictor centroid id.

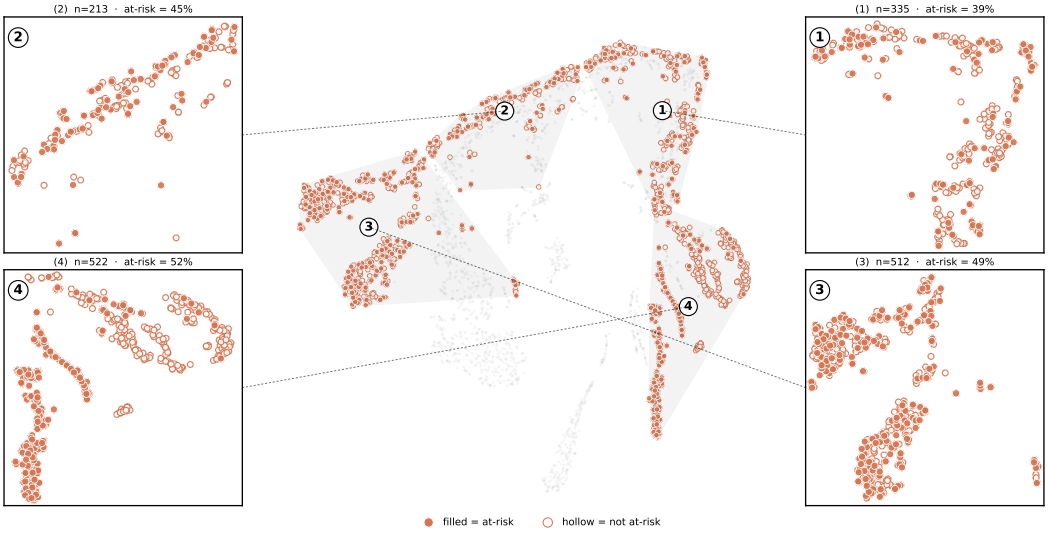

Figure 7: Cohort zoom restricted to the D3 sub-cloud, with feature-space neighbourhoods labelled by predictor centroid id.

B.3 Demographic Embedding Grid

Fig. 8 overlays each of the 14 canonical demographic axes on the shared embedding to expose where demographic structure aligns with representational structure. The 14 panels separate into two qualitatively different regimes. Some axes are visibly aligned with the principal directions of the embedding, these are typically axes that co-vary with cohort identity, grade range, or service-status indicators that themselves shape which assessments were administered. Other axes are roughly uniform across the embedding, partitioning each cohort cloud into mixtures. This contrast is informative regardless of the absolute number of students per subgroup.

The distinction matters because it changes how a subgroup AUC gap should be interpreted. When a demographic axis is representationally aligned with the embedding, downstream subgroup differences can be mediated by feature-availability and feature-distribution differences. The model is then operating in different parts of feature space for different subgroups. When a demographic axis is representationally uniform but a subgroup AUC gap is still observed, the gap is a stronger candidate for downstream-only intervention because feature context is held roughly constant across subgroups.

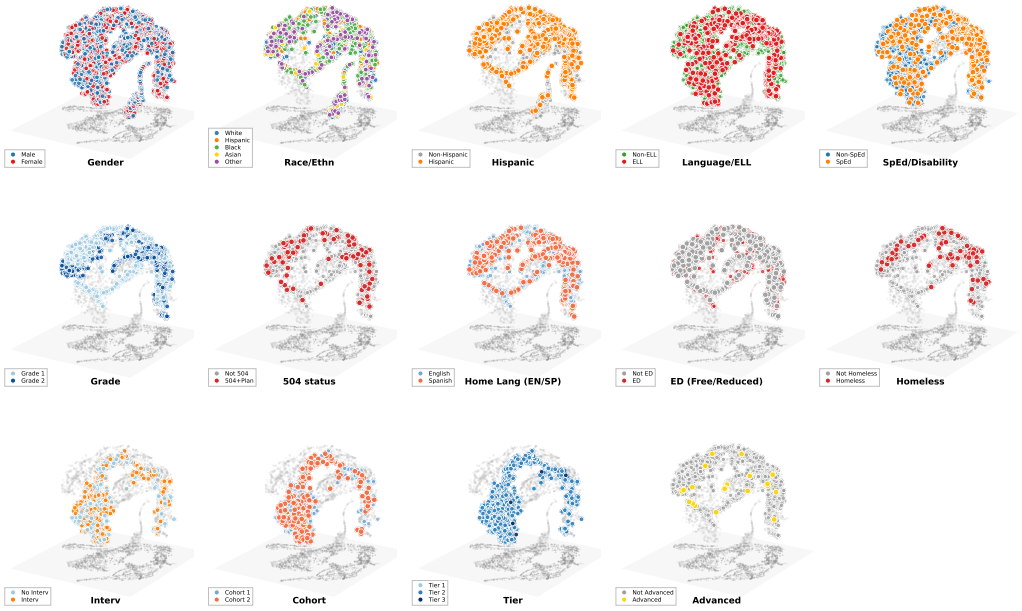

Figure 8: t-SNE colored by demographic axis (3D, labelled rows). Fourteen panels share one TabPFN-v2 embedding. Each overlays a single demographic field.

C Implementation Details

This section provides full implementation details for every component of the RealLit evaluation pipeline: classifiers, imputers, permutation ensembling, cross-validation, missingness stress tests, hyperparameters, and compute.

C.1 Classifiers

C.1.1 MaskMLP

We follow the masked-pretraining recipe introduced in ECRI (Shangguan et al., 2024) as a baseline, with full reimplementation details tabulated in Tab. 18. The encoder is a 3-layer MLP with hidden size 64 (Linear $\rightarrow$ ReLU $\rightarrow$ BatchNorm) and a linear projection to logits, trained in two stages. In the self-supervised pre-training stage, missing entries in the input table are flagged with the sentinel

value -1 and observed entries are randomly masked at a rate of 25% per row, per minibatch. The encoder is trained with a cosine-embedding loss that pulls the representation of a masked view toward the representation of the unmasked view of the same row. This lets the encoder learn to recover information from observed entries without relying on an explicit imputation step. In the supervised stage, the pre-trained encoder is fine-tuned end-to-end with binary cross-entropy with logits on the at-risk target. Both stages use Adam ($\mathrm{lr}{=}10^{-3}$), batch size 16, and 50 epochs. Fine-tuning uses a stratified 15% validation split with validation-AUC early stopping (patience 10) and the best-val weights are restored before evaluation. The same encoder architecture and hyperparameters are used across the four cohorts, and the encoder is re-fit per cross-validation fold so each MaskMLP result is fitted to its training fold rather than to a fixed checkpoint.

C.1.2 TabPFN-v2

We use the public TabPFN-v2 checkpoint (version v2.6) (Hollmann et al., 2025) without further training on RealLit. At inference time, both the training fold and the test query are passed in a single in-context forward pass; no gradient updates occur. The released checkpoint supports up to $50{,}000$ samples and $2{,}000$ features, which is well above the size of every RealLit cohort, so no truncation or down-sampling is needed. Categorical fields are encoded with the published TabPFN preprocessing pipeline (label encoding for low-cardinality columns and the model's built-in handling for missing values), and continuous fields are passed through TabPFN's standard scaler. Permutation ensembling (App. C.4) is the only inference-time aggregation used. We do not enable any TabPFN-specific test-time augmentation beyond that.

C.1.3 TabICL-v2

We use the public TabICL-v2 checkpoint (Qu et al., 2025) without further training on RealLit. TabICL encodes each row as a sequence and uses a transformer to attend jointly over training and test rows; relative to TabPFN, it scales more gracefully to larger row counts, which is useful for the D3 cohorts. As with TabPFN, both the training fold and the test query are processed in-context and no gradient updates occur. We use the published preprocessing pipeline to convert mixed continuous and categorical columns into the TabICL token vocabulary, with the column-then-row attention pattern unchanged from the released configuration.

C.1.4 TabDPT

We use the public TabDPT checkpoint (Ma et al., 2025) without further training on RealLit. TabDPT is a retrieval-augmented tabular foundation model: for each test query, an in-context support set is constructed dynamically by retrieving the most relevant training rows rather than passing the entire training fold. This makes TabDPT's effective context size adaptive, which is particularly useful in our setting because relevant neighbours for a test row often share the same observation pattern. We use the retrieval, distance, and top-k defaults from the released configuration, and we keep these defaults fixed across the four cohorts so that performance differences cannot be attributed to cohort-specific retrieval tuning.

C.2 Imputers

C.2.1 MICE (Statistical Reference)

We run MICE (Van Buuren & Groothuis-Oudshoorn, 2011) as a statistical reference imputer with 10 chained-equation iterations and a LightGBM base learner (via the `miceforest` library, version ≥ 5.7, see App. C.8) for mixed continuous-ordinal columns. The same LightGBM-based learner is also used for categorical columns. Despite the library's name, `miceforest` uses gradient-boosted trees rather than random forests as its default base learner since v3+. Iterations terminate either at the iteration cap or when the per-column change between iterations falls below the default tolerance of the implementation, whichever comes first. MICE is fit on the training fold only and applied to the test fold so that imputation cannot leak label information across the cross-validation split. All other imputers are compared against MICE via Cohen's d effect sizes computed over targets and classifiers.

C.2.2 MIRI

MIRI (Yu et al., 2025) is a rectified-flow imputer that reduces the mutual information between the imputed values and the missingness mask. We use the public reference implementation with the default training objective, network architecture, and integration scheme from the published configuration. The model is trained per training fold and applied to the test fold under the same train/test split used for every other component of the pipeline. As with all learned imputers in RealLit, MIRI is fit only on observed entries of the training fold to avoid label leakage.

C.2.3 TABCSDI

TabCSDI (Zheng & Charoenphakdee, 2022) adapts the CSDI score-based diffusion framework to tabular data. The conditional score network takes observed entries as conditioning context and iteratively denoises the missing entries through a Langevin-dynamics reverse process. We use the public reference implementation with the released score-network architecture, noise schedule, and number of reverse-process steps. The model is trained per training fold on observed entries only, and the trained imputer is applied to the test fold under the same train/test split as the rest of the pipeline.

C.2.4 DIFFPUTER

DiffPuter (Zhang et al., 2025) is an EM-driven diffusion imputer that alternates between an expectation step (sampling imputed values from the current diffusion model) and a maximization step (refitting the diffusion model on the completed data). We use the public reference implementation with the released autoencoder architecture, latent-space diffusion schedule, and default number of EM iterations. As with the other learned imputers, DiffPuter is fit on the training fold's observed entries and applied to the test fold under the same split.

C.2.5 CFMI

CFMI (Simkus & Gutmann, 2025) replaces the diffusion reverse process with a straight-line flow-matching trajectory (Lipman et al., 2023) from a noise distribution to the data manifold. We use the public reference implementation with the released flow-matching network, training objective, prior distribution, and integration scheme (number of integration steps). The imputer is fit on the training fold's observed entries and applied to the test fold under the same split as the rest of the pipeline.

C.3 MISSINGNESS INDICATOR AUGMENTATION

For all imputation-based pipelines we evaluate two variants of each classifier: a "base" variant that consumes only the imputed feature matrix $\tilde{\mathbf{x}}$, and a "+M" variant that additionally consumes the binary missingness mask $\mathbf{m} \in \{0,1\}^d$ as a separate set of features. The mask is appended after the imputed feature columns in a fixed column order so that any classifier that is sensitive to feature position sees the mask block in the same place across folds. Mask columns are not scaled or rescaled. They are passed as $\{0,1\}$ indicators. Categorical encoders see only the imputed feature columns, they never receive a mask column to encode, so the "+M" augmentation is invariant to the choice of categorical encoding. The "+M" variant is the default for results reported in Apps. D and F when block-MNAR patterns are expected to carry signal. The base variant is used as the foil that asks whether the mask is providing information the classifier cannot recover from $\tilde{\mathbf{x}}$ alone.

C.4 PERMUTATION ENSEMBLING

We use the shared ensemble size K defined in the main text, with $K_{\text{imp}} = K_{\text{clf}} = K$. Feature and class-label permutations are sampled with fixed random seeds for reproducibility, and the same set of K classifier presentations is shared across classifiers within each evaluation fold. For stochastic imputation, the imputer is independently re-seeded K times and the resulting completed matrices are aggregated by a uniform per-cell mean before being passed to the K-member classifier ensemble. Because imputers operate on standardized numerical representations, the same averaging rule is used across features. Full results for the diagonal K-sweep are reported in Tabs. 9 and 10.

C.5 Cross-Validation Protocol

All results are reported as 3 repetitions of 5-fold stratified cross-validation (3×5 CV). Stratification is applied jointly on the binary at-risk target and, where available, on school identifier. Joint stratification serves two purposes: (i) it preserves class balance across folds, which matters because positive-class prevalence differs sharply across cohorts (App. A.6), and (ii) it ensures that students from the same school are not split across folds, which prevents school-level data leakage in cohorts where school identifier is a strong source of feature-distribution structure (notably D2 and D3). Folds are constructed with a fixed RNG seed so that all classifiers, imputers, and ensemble configurations see the same fold partition and any performance differences cannot be attributed to fold variance alone. For each (cohort, target, classifier, imputer, K, mechanism, rate) cell we report mean $\pm$ standard error across the CV folds.

C.6 Missingness Stress-Test Mechanisms

We follow the classical Rubin taxonomy of missing-data mechanisms (Rubin, 1976). For each cohort and each target rate $r \in \{0.10, 0.20, 0.30\}$, we inject additional missingness into the test fold's observed feature entries under each of three mechanisms. The training fold is never modified by the stress-test injector, and demographic columns are never masked. Let X_{ij} denote the test-fold value at row i and feature column j, and let $M_{ij} \in \{0, 1\}$ be the stress-test mask we sample, with $M_{ij}{=}1$ indicating that the already-observed value X_{ij} is hidden from the model.

MCAR (Missing Completely At Random): Each eligible (i, j) entry is masked independently with probability r: $P(M_{ij}{=}1) = r$. This is the simplest baseline and serves as a control: any imputer or model whose performance degrades sharply under MCAR is sensitive to overall missingness mass rather than to mechanism per se.

MAR (Missing At Random): The mask probability for entry (i, j) depends on the observed portion of row i (i.e., on the values of other features that were not themselves masked): $P(M_{ij}{=}1 \mid X^{\text{obs}}_{i,\cdot})$. We parameterize this by a logistic function of a small number of pivot features chosen per cohort, calibrated so that the marginal mask rate matches the target r.

MNAR (Missing Not At Random): The mask probability for entry (i, j) depends on the value of X_{ij} itself: $P(M_{ij}{=}1 \mid X_{ij})$. Concretely, for continuous features we mask with probability that is monotone in the standardized value of X_{ij}, biased toward masking values away from the column median. For categorical features, we mask with a category-specific probability. As with MAR, the per-feature mask rates are calibrated so that the marginal rate matches the target r. In all three mechanisms, the injected mask is composed with the cohort's native missingness pattern: an entry that was already missing in the test row remains missing, and only entries that were originally observed are eligible to be additionally masked. This composition is what makes the stress-test results in Tabs. 9 and 10 additive on top of the cohort's native missingness rate rather than replacing it.

C.7 Hyperparameters

RealLit is designed so that classifier and imputer choices can be compared without per-cohort hyperparameter tuning confounding the results. Every classifier and every imputer in our evaluation grid uses the published default configuration of its released implementation, fixed across cohorts. The only trainable classifier we run is MaskMLP, which is re-fit per fold, its defaults match the original ECRI paper recipe (Tab. 18). The other three classifiers (TabPFN-v2, TabICL-v2, TabDPT) are inference-only via public checkpoints, with no gradient updates on RealLit. This is a deliberate evaluation choice: cohort-specific tuning would inflate the apparent transfer ability of a method by silently adapting it to each cohort, which is exactly the failure mode that the cross-cohort comparison in Tab. 2 is designed to expose. The only knob we sweep across our grid is the ensemble size K, a presentation-level aggregator.

C.8 Compute and Reproducibility

The full RealLit evaluation grid spans four cohorts $\times$ four classifiers $\times$ five imputers (including the native incomplete-input baseline and the MICE statistical reference) $\times$ six ensemble sizes $\times$ three missingness mechanisms $\times$ three injected rates $\times$ 5 CV folds, 3 seeds. Foundation-model inference is performed with the published checkpoints. Learned imputers are trained per fold from scratch. Random seeds are fixed at the (fold, K, classifier, imputer, mechanism, rate) level so that every cell of the reported tables is reproducible end-to-end. Experiments were run on RunPod serverless GPU workers, totaling approximately $1{,}421$ GPU-hours across the full evaluation grid. The bulk of compute used NVIDIA GeForce RTX 4090 workers ($\approx 1{,}201$ GPU-hours, $\approx 84.5\%$ of total). The remaining ≈ 220 GPU-hours were distributed across NVIDIA L4 (≈ 69h), A100 80GB PCIe (≈ 65h), RTX 2000 Ada (≈ 37h), RTX A4500 (≈ 33h), A100-SXM4-80GB (≈ 8h), H100 80GB HBM3 (≈ 6h), and RTX 5090 (≈ 3h). The heterogeneous SKU mix reflects RunPod's serverless scheduler matching workers to availability. The main RTX 4090 worker is sufficient for both foundation-model in-context inference and per-fold imputer training at the cohort sizes of RealLit.

Software Stack: All experiments ran in a single Docker image based on pytorch/pytorch:2.6.0-cuda12.6-cudnn9-devel (PyTorch 2.6.0, CUDA 12.6, cuDNN 9, Python 3.11, `linux/amd64`). The Python environment pins: `numpy` ≥ 1.26 (< 2.0), `pandas` ≥ 2.2, `scikit-learn` ≥ 1.4, `scipy` ≥ 1.12, `lightgbm` ≥ 4.0, `pytorch_lightning` ≥ 2.0, plus the released foundation-model packages `tabpfn` $\geq 0.1.10$, `tabicl` $\geq 0.2.0$, and `tabdpt` $\geq 0.0.4$, and the chained-equation imputer `miceforest` ≥ 5.7 used as our MICE reference (App. C.2.1).

Code and Dataset Release: We release the introduced cohorts of RealLit (D2, D3, D4) under the Creative Commons Attribution 4.0 (CC BY 4.0) license, together with code for running the evaluation harness, at https://github.com/anonymous-submission-site/real-lit. D1 (ECRI) is not redistributed by RealLit. The harness loads it from the public ECRI release of (Shangguan et al., 2024). The release covers the introduced cohorts and the evaluation harness. Reproducing the trained imputers is a matter of running the public reference implementations cited in App. C.2 on the released training folds.

D Additional Results

D.1 Dataset Differences

Table 3: Cross-dataset overlap in assessment categories. Overlap is computed over pairs of datasets, where multiple metrics may map to the same assessment domain. PA denotes phonological awareness and Vocab. denotes vocabulary. Percentages are reported in the order of the dataset pair (e.g., for D3 / D4, 98% of decoding metrics are covered in D3). Shared groups have near-complete coverage in both datasets, whereas partially shared groups appear in both datasets but have incomplete or uneven coverage. Non-overlapping groups appear in only one dataset.

Dataset Pair	Shared Groups	Partially Shared Groups	Non-overlapping Groups
D3 / D4	Decoding (98% / 100%), Fluency (98% / 100%), Vocab. (98% / 99%)	—	Alphabetics, PA
D1 / D4	Decoding (100% / 100%), Fluency (99% / 100%)	—	Alphabetics, PA, Vocab.
D1 / D3	Decoding (100% / 98%), Fluency (99% / 98%)	Alphabetics (99% / 48%)	PA, Vocab.
D2 / D3	—	PA (99% / 63%), Vocab. (72% / 98%), Alphabetics (55% / 48%), Fluency (18% / 98%)	Decoding
D2 / D4	—	Vocab. (72% / 99%)	Alphabetics, Decoding, Fluency, PA
D1 / D2	—	Alphabetics (99% / 55%), Fluency (99% / 18%)	Decoding, PA, Vocab.

Table 4: Gender-based performance across cohorts. Values are AUC mean ± standard deviation across three evaluation folds, reported for all students (Overall) and for the at-risk subgroup (At-Risk). Female performance is lower than male performance on D1, D2, and D4, while the direction reverses on D3.

Category	D1 AUC	D2 AUC	D3 AUC	D4 AUC
Male				
Overall	$0.712_{\pm 0.038}$	$0.849_{\pm 0.040}$	$0.853_{\pm 0.037}$	$0.907_{\pm 0.013}$
At-Risk	$0.534_{\pm 0.020}$	$0.659_{\pm 0.140}$	$0.672_{\pm 0.016}$	$0.778_{\pm 0.069}$
Female				
Overall	$0.673_{\pm 0.007}$	$0.569_{\pm 0.207}$	$0.873_{\pm 0.020}$	$0.778_{\pm 0.069}$
At-Risk	$0.521_{\pm 0.030}$	$0.471_{\pm 0.143}$	$0.693_{\pm 0.017}$	$0.437_{\pm 0.073}$

Table 5: Gender distribution, missingness, and outcome severity for each dataset. Dataset share is the percentage of students of each gender. Missingness is the percentage of missing values for students of that gender. At-risk is the percentage of students of that gender labeled at risk, and the at-risk reading score is the mean ± standard deviation reading score among those at-risk students.

Dataset	Dataset Share (%)	Missingness (%)	At-Risk (%)	At-Risk Reading Score
Male				
D1	52.3	2.2	45	$37.4_{\pm 9.6}$
D2	52.5	41.8	36	$80.3_{\pm 4.7}$
D3	53.5	16.4	44	$140.4_{\pm 30.6}$
D4	51.2	3.2	51	$84.0_{\pm 4.72}$
Female				
D1	47.7	1.9	47	$37.5_{\pm 8.0}$
D2	47.5	45.7	48	$82.2_{\pm 4.0}$
D3	46.5	15.2	48	$142.8_{\pm 28.7}$
D4	48.5	3.3	38	$85.71_{\pm 2.75}$

Table 6: Dataset composition by cohort. Categories are reported as percentages of students with a recorded value for that field, and % recorded is the share of students for whom the field is available. D3 is reported separately for grades 1 and 2 (G1–2) and grade 3 (G3), since the two grade bands record different fields. LEP denotes limited English proficiency, EL and ELL denote English learner and English language learner, ESE denotes exceptional student education, and ED denotes economically disadvantaged.

Cohort	Field	Categories (% of recorded)	% Recorded
D2	Gender	Male 50.3% · Female 49.7%	52.3%
D2	Ethnicity	Not Hispanic 90.2% · Hispanic 9.8%	50.3%
D2	Race	White 55.0% · Black or African American 29.6% · Multi-racial 11.1% · Asian 2.8% · Other 1.1% · Native Hawaiian or Other Pacific Islander 0.2% · American Indian or Alaska Native 0.2%	52.0%
D2	Language	Non-EL 91.7% · EL 8.3%	53.2%
D3 (G1–2)	Gender	Male 53.7% · Female 46.3%	100%
D3 (G1–2)	Race / Ethnicity	White 41.3% · Hispanic 31.4% · Black or African American 18.3% · Other 7.0% · Asian 1.8% · American Indian or Alaska Native 0.1% · Native Hawaiian or Other Pacific Islander 0.05%	100%
D3 (G1–2)	Home language	English 79.1% · Spanish 18.0% · 13 other languages 2.9%	100%
D3 (G1–2)	LEP status	Not applicable 86.5% · LEP in LEP classes 12.7% · Two-year follow-up 0.6% · Awaiting testing 0.1%	100%
D3 (G1–2)	ESE (Special Ed.)	No 80.1% · Yes 19.9%	100%
D3 (G1–2)	Advanced	No 96.6% · Yes 3.4%	100%
D3 (G1–2)	Section 504	Not eligible 94.0% · Eligible + plan 5.5% · 3 other statuses 0.5%	100%
D3 (G1–2)	Lunch status (FNR)	Direct cert/free 55.3% · Paid 31.9% · Free-eligible 8.0% · Reduced-eligible 4.8%	100%
D3 (G3)	ED (Econ. disadv.)	Yes 68.0% · No 32.0%	100%
D3 (G3)	ESE (Special Ed.)	No 80.8% · Yes 19.2%	100%
D3 (G3)	Section 504	No 90.1% · Yes 9.9%	100%
D3 (G3)	ELL	No 91.8% · Yes 8.2%	100%
D4	Gender	Male 51.6% · Female 48.4%	99.4%
D4	Race / Ethnicity	White 59.8% · Black 24.7% · Multi-racial 6.0% · Hispanic 4.4% · Asian 3.8% · Other 1.2%	99.4%

This section reports classifier performance under additional configurations not shown in the main paper: native missingness (rate $= 0\%$) across imputer choices, and the full $K \times$ rate $\times$ classifier grid.

D.2 REGRESSION

We also evaluate regression performance on D1–D4 using RMSE, MAE and R^2 (Table 7). The results show similar trends to those observed in classification. We also include TabPFN-v3 with a temporal prior (TabPFN-v3-tp), designed to better capture the longitudinal structure of the data.

Table 7: Regression performance on D1–D4. Each cell shows the mean above and ± standard deviation below. Bold green cells indicate the best reported mean per dataset and metric.

Model	D1			D2			D3			D4		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
XGBoost	8.78 ±0.19	6.97 ±0.20	0.206 ±0.023	10.65 ±0.45	7.94 ±0.30	0.184 ±0.056	21.37 ±0.57	15.43 ±0.39	0.580 ±0.022	7.24 ±0.22	5.81 ±0.22	0.638 ±0.029
DeepConvLSTM	8.41 ±0.20	6.69 ±0.21	0.271 ±0.022	10.19 ±0.40	7.51 ±0.23	0.249 ±0.061	20.27 ±0.15	14.25 ±0.10	0.623 ±0.005	8.02 ±0.30	6.37 ±0.20	0.551 ±0.050
HistGB	8.26 ±0.22	6.57 ±0.20	0.298 ±0.014	10.32 ±0.48	7.68 ±0.34	0.235 ±0.044	20.40 ±0.58	14.55 ±0.41	0.617 ±0.020	6.95 ±0.25	5.41 ±0.24	0.665 ±0.033
LightGBM	8.25 ±0.22	6.56 ±0.19	0.301 ±0.009	10.31 ±0.42	7.67 ±0.32	0.234 ±0.047	20.34 ±0.56	14.64 ±0.37	0.619 ±0.020	6.74 ±0.22	5.36 ±0.18	0.685 ±0.030
Ridge	8.20 ±0.24	6.47 ±0.20	0.307 ±0.029	9.61 ±0.44	7.39 ±0.27	0.338 ±0.034	20.73 ±0.24	14.67 ±0.16	0.606 ±0.003	10.24 ±0.73	8.02 ±0.27	0.281 ±0.073
RandomForest	8.12 ±0.28	6.50 ±0.25	0.323 ±0.019	10.03 ±0.51	7.65 ±0.41	0.279 ±0.043	20.00 ±0.43	14.30 ±0.35	0.632 ±0.014	6.88 ±0.22	5.46 ±0.19	0.672 ±0.027
MaskMLP	8.74 ±1.53	6.93 ±0.50	0.142 ±0.329	11.76 ±1.14	8.08 ±0.27	0.263 ±0.279	20.70 ±0.52	14.60 ±0.39	0.607 ±0.013	7.69 ±0.67	5.88 ±0.33	0.581 ±0.076
TabPFN-v2	8.01 ±0.25	6.37 ±0.25	0.339 ±0.025	9.59 ±0.52	7.13 ±0.37	0.341 ±0.036	19.56 ±0.22	14.08 ±0.18	0.649 ±0.008	6.55 ±0.26	5.28 ±0.25	0.700 ±0.037
TabPFN-v3	**7.95** ±0.24	**6.30** ±0.24	**0.349** ±0.030	9.45 ±0.52	**7.10** ±0.37	0.360 ±0.039	**19.32** ±0.22	13.74 ±0.20	**0.657** ±0.006	6.36 ±0.27	5.12 ±0.29	0.717 ±0.037
TabPFN-v3-tp	7.97 ±0.24	6.34 ±0.22	0.345 ±0.029	**9.44** ±0.54	7.14 ±0.38	**0.362** ±0.038	19.35 ±0.20	**13.68** ±0.18	0.656 ±0.005	**6.34** ±0.25	**5.06** ±0.27	**0.720** ±0.033

RMSE/MAE: ↓; R^2: ↑. `tp`: temporal prior.

D.3 COMPARISON OF TABPFN VERSIONS

We compare TabPFN-v2 with the newer TabPFN-v3 and TabPFN-v3.5 variants under the same baseline setting. As shown in Tab. 8, the newer versions do not consistently improve performance over TabPFN-v2. For example, TabPFN-v2 achieves the highest AUC on D1, D2, and D4, and the highest accuracy on D2–D4. The newer variants provide only small gains in isolated cases, such as D1 accuracy and D3 AUC, while showing lower performance on several other dataset–metric combinations.

Table 8: Comparison of TabPFN versions under the baseline setting at K = 1. Values are mean ± standard deviation across three evaluation folds. **Bold** marks the highest mean in each dataset–metric column.

Classifier	AUC				ACC			
	D1	D2	D3	D4	D1	D2	D3	D4
TabPFN-v2	**0.696**±0.026	**0.832**±0.013	0.860±0.024	**0.914**±0.022	0.643±0.038	**0.748**±0.015	**0.789**±0.027	**0.867**±0.000
TabPFN-v3	0.686±0.031	0.819±0.029	**0.861**±0.027	0.884±0.029	**0.648**±0.030	0.735±0.000	0.786±0.030	0.822±0.038
TabPFN-v3.5	0.688±0.039	0.823±0.012	**0.861**±0.026	0.878±0.046	0.637±0.039	0.745±0.026	0.778±0.039	0.822±0.038

D.4 K-SWEEP × RATE × CLASSIFIER

Tabs. 9 and 10 report the diagonal K-sweep across injected missingness rates $10/20/30\%$, pooled across cohorts, imputers, mechanisms, and seeds. Three patterns are consistent across both metrics and all rates.

First, K=16 is the best K for almost every (rate, classifier) cell. The exceptions are minor (e.g., K=8 ties or narrowly leads for TabICL-v2 at the 20% rate) and the absolute differences between K=8 and K=16 rarely exceed 0.005 AUC. This makes K=8 a defensible practical choice when

compute is constrained, while $K{=}16$ is the headline configuration when one is willing to pay the additional inference cost.

Second, the gain from increasing K is largest where the conditions are hardest. At 10% rate, $K{=}16$ improves over $K{=}1$ by approximately $0.006{-}0.030$ AUC depending on classifier; at 30% rate, the same lift is approximately $0.019{-}0.038$ AUC. Permutation ensembling acts as a robustness lever: it pays off most under elevated missingness, which is exactly the regime where presentation-order sensitivity of in-context learners would be expected to matter.

Third, MaskMLP gets the largest absolute lift from ensembling, the foundation models the smallest. At 30% rate AUC, MaskMLP improves from 0.741 at $K{=}1$ to 0.779 at $K{=}16$ (a lift of 0.038), while TabPFN-v2 improves from 0.766 to 0.785 (a lift of 0.019). This is consistent with the foundation models already aggregating multiple presentations implicitly through their in-context attention; MaskMLP, which has no such built-in aggregation, captures it externally through the permutation ensemble. Diminishing returns kick in around $K{=}8$ for all classifiers, with $K{=}8 \rightarrow K{=}16$ typically yielding ≤ 0.005 AUC.

The breakdown of these aggregate trends across the three missingness mechanisms (MCAR/MAR/MNAR) is reported separately in Tab. 17 for TabPFN-v2 and is broadly consistent with the picture above: the relative ranking of K values is stable across mechanisms, with the largest absolute gaps showing up under MNAR.

Table 9: Rate $\times$ classifier $\times$ K table, Accuracy. Diagonal K-sweep cells (imputer-K matched to classifier-K) pooled across the 4 cohorts, 4 imputers (MIRI/TabCSDI/DiffPuter/CFMI), 3 missingness mechanisms (MCAR/MAR/MNAR). Green marks the best K per row; blue is second-best.

Rate	Classifier	$K{=}1$	$K{=}2$	$K{=}4$	$K{=}8$	$K{=}16$
10%	MaskMLP	0.693 ± 0.088	0.705 ± 0.091	0.715 ± 0.092	0.718 ± 0.092	0.718 ± 0.089
	TabDPT	0.719 ± 0.079	0.727 ± 0.079	0.727 ± 0.075	0.729 ± 0.075	0.731 ± 0.077
	TabICL-v2	0.730 ± 0.092	0.731 ± 0.091	0.733 ± 0.091	0.737 ± 0.090	0.738 ± 0.090
	TabPFN-v2	0.729 ± 0.086	0.734 ± 0.086	0.733 ± 0.084	0.733 ± 0.087	0.738 ± 0.090
20%	MaskMLP	0.684 ± 0.088	0.696 ± 0.089	0.705 ± 0.086	0.709 ± 0.086	0.715 ± 0.084
	TabDPT	0.712 ± 0.081	0.715 ± 0.069	0.721 ± 0.074	0.724 ± 0.071	0.727 ± 0.073
	TabICL-v2	0.716 ± 0.079	0.726 ± 0.085	0.725 ± 0.077	0.725 ± 0.078	0.724 ± 0.076
	TabPFN-v2	0.717 ± 0.081	0.718 ± 0.080	0.719 ± 0.082	0.725 ± 0.087	0.728 ± 0.091
30%	MaskMLP	0.675 ± 0.097	0.687 ± 0.100	0.701 ± 0.104	0.701 ± 0.093	0.711 ± 0.101
	TabDPT	0.705 ± 0.077	0.707 ± 0.085	0.711 ± 0.078	0.711 ± 0.076	0.718 ± 0.077
	TabICL-v2	0.703 ± 0.069	0.705 ± 0.078	0.718 ± 0.086	0.716 ± 0.082	0.721 ± 0.078
	TabPFN-v2	0.695 ± 0.082	0.706 ± 0.083	0.715 ± 0.089	0.709 ± 0.086	0.715 ± 0.089

Table 10: Rate $\times$ classifier $\times$ K table, ROC-AUC. Diagonal K-sweep cells (imputer-K matched to classifier-K) pooled across the 4 cohorts, 4 imputers (MIRI/TabCSDI/DiffPuter/CFMI), 3 missingness mechanisms (MCAR/MAR/MNAR). Green marks the best K per row; blue is second-best.

Rate	Classifier	$K{=}1$	$K{=}2$	$K{=}4$	$K{=}8$	$K{=}16$
10%	MaskMLP	0.766 ± 0.112	0.782 ± 0.111	0.787 ± 0.111	0.793 ± 0.108	0.796 ± 0.106
	TabDPT	0.792 ± 0.091	0.796 ± 0.089	0.801 ± 0.088	0.799 ± 0.085	0.803 ± 0.085
	TabICL-v2	0.792 ± 0.099	0.798 ± 0.102	0.797 ± 0.097	0.799 ± 0.098	0.800 ± 0.097
	TabPFN-v2	0.796 ± 0.099	0.796 ± 0.100	0.799 ± 0.102	0.801 ± 0.103	0.802 ± 0.102
20%	MaskMLP	0.751 ± 0.107	0.775 ± 0.107	0.780 ± 0.102	0.788 ± 0.100	0.791 ± 0.099
	TabDPT	0.784 ± 0.090	0.790 ± 0.084	0.795 ± 0.085	0.797 ± 0.084	0.798 ± 0.083
	TabICL-v2	0.788 ± 0.092	0.793 ± 0.090	0.795 ± 0.087	0.802 ± 0.094	0.801 ± 0.090
	TabPFN-v2	0.787 ± 0.099	0.792 ± 0.098	0.795 ± 0.097	0.797 ± 0.098	0.798 ± 0.099
30%	MaskMLP	0.741 ± 0.116	0.754 ± 0.120	0.766 ± 0.122	0.775 ± 0.116	0.779 ± 0.117
	TabDPT	0.771 ± 0.100	0.781 ± 0.101	0.784 ± 0.099	0.784 ± 0.097	0.790 ± 0.096
	TabICL-v2	0.767 ± 0.090	0.776 ± 0.101	0.785 ± 0.100	0.787 ± 0.101	0.791 ± 0.099
	TabPFN-v2	0.766 ± 0.108	0.776 ± 0.111	0.781 ± 0.108	0.779 ± 0.108	0.785 ± 0.110

D.5 LONGITUDINAL IMPUTATION

Every imputer in the main grid treats a student's record as an unordered feature vector. D4 is the one cohort with a time axis that retains the Grade 1 and Grade 2 administrations of the same instruments (d=170; 85 columns per grade, 48 instruments present in both), and its missingness is longitudinal attrition rather than item-level nonresponse (App. A.4). We add two training-free references that impute along that axis, each operating only within a student's own record. **LOCF** carries the last observed value for an instrument forward to the missing grade; **LinGrowth** fits a per-(student, instrument) linear trend over the observed grades and evaluates it at the missing one.

Simple Longitudinal References Are Competitive: At native missingness, LOCF achieves the highest mean ROC-AUC among the evaluated imputers for all four classifiers, with LinGrowth matching it for TabICL-v2 and TabPFN-v2 (Tab. 11). Both longitudinal references also achieve the highest mean accuracy for these two classifiers. Under an additional 30% injected missingness, with MCAR, MAR, and MNAR pooled, LOCF retains the highest mean ROC-AUC for MaskMLP, TabDPT, and TabICL-v2, whereas TabPFN-v2 performs best with NATIVE handling (Tab. 12). Across both settings, LinGrowth matches or falls below LOCF in mean ROC-AUC. These results suggest that explicitly using repeated measurements provides a strong, simple reference against the learned imputers evaluated here.

Benefits Depend on the Classifier and Evaluation Metric: The improvements in ROC-AUC do not consistently translate into higher accuracy. For MaskMLP, Native yields the highest mean accuracy at both missingness levels despite LOCF's higher ROC-AUC. Under 30% injected missingness, LOCF ties Native for the highest TabICL-v2 accuracy, while Native leads on both metrics for TabPFN-v2 and Mean achieves the highest accuracy for TabDPT. Thus, D4 highlights the value of including longitudinal references when repeated assessments are available, while showing that their benefits depend on the downstream classifier, missingness setting, and evaluation metric. These comparisons are descriptive and specific to D4, so the relative ordering may differ across individual missingness mechanisms.

Table 11: Longitudinal imputation on D4 at native missingness only (rate 0%). Imputer × classifier cross-tab on Grade 1 and Grade 2 administrations of the same instruments (d=170). Green marks the best imputer per column; ties at the displayed precision are both marked.

(a) ROC-AUC.

Imputer	MaskMLP	TabDPT	TabICL-v2	TabPFN-v2
LOCF	0.943 ± 0.056	0.874 ± 0.095	0.970 ± 0.026	0.968 ± 0.046
LinGrowth	0.940 ± 0.056	0.867 ± 0.110	0.970 ± 0.026	0.968 ± 0.046
DiffPuter	0.892 ± 0.123	0.873 ± 0.072	0.954 ± 0.040	0.947 ± 0.047
CFMI	0.913 ± 0.065	0.860 ± 0.083	0.958 ± 0.035	0.952 ± 0.046
Native	0.916 ± 0.080	0.864 ± 0.089	0.963 ± 0.035	0.952 ± 0.053
Mean	0.911 ± 0.096	0.861 ± 0.103	0.963 ± 0.030	0.953 ± 0.048

(b) Accuracy.

Imputer	MaskMLP	TabDPT	TabICL-v2	TabPFN-v2
LOCF	0.877 ± 0.077	0.928 ± 0.047	0.946 ± 0.033	0.953 ± 0.032
LinGrowth	0.862 ± 0.091	0.932 ± 0.045	0.946 ± 0.033	0.953 ± 0.028
DiffPuter	0.882 ± 0.076	0.928 ± 0.039	0.936 ± 0.037	0.944 ± 0.032
CFMI	0.860 ± 0.087	0.932 ± 0.038	0.938 ± 0.033	0.940 ± 0.031
Native	0.904 ± 0.057	0.929 ± 0.049	0.941 ± 0.033	0.941 ± 0.039
Mean	0.856 ± 0.070	0.932 ± 0.044	0.944 ± 0.034	0.940 ± 0.036

Table 12: Longitudinal imputation on D4 at injected rate 30%, MCAR/MAR/MNAR pooled. Imputer × classifier cross-tab on Grade 1 and Grade 2 administrations of the same instruments (d=170). Green marks the best imputer per column. Ties at the displayed precision are both marked.

(a) ROC-AUC.

Imputer	MaskMLP	TabDPT	TabICL-v2	TabPFN-v2
LOCF	0.906 ± 0.119	0.899 ± 0.059	0.964 ± 0.032	0.942 ± 0.055
LinGrowth	0.903 ± 0.106	0.898 ± 0.068	0.963 ± 0.032	0.935 ± 0.066
DiffPuter	0.865 ± 0.166	0.885 ± 0.061	0.948 ± 0.039	0.938 ± 0.050
CFMI	0.856 ± 0.175	0.886 ± 0.079	0.955 ± 0.036	0.933 ± 0.062
Native	0.826 ± 0.170	0.889 ± 0.063	0.948 ± 0.050	0.956 ± 0.046
Mean	0.885 ± 0.107	0.889 ± 0.076	0.959 ± 0.038	0.935 ± 0.065

(b) Accuracy.

Imputer	MaskMLP	TabDPT	TabICL-v2	TabPFN-v2
LOCF	0.866 ± 0.082	0.934 ± 0.041	0.947 ± 0.034	0.943 ± 0.033
LinGrowth	0.867 ± 0.087	0.934 ± 0.039	0.942 ± 0.033	0.941 ± 0.035
DiffPuter	0.879 ± 0.084	0.936 ± 0.041	0.939 ± 0.037	0.943 ± 0.038
CFMI	0.859 ± 0.079	0.937 ± 0.040	0.940 ± 0.038	0.944 ± 0.031
Native	0.900 ± 0.049	0.935 ± 0.040	0.947 ± 0.035	0.945 ± 0.032
Mean	0.866 ± 0.086	0.938 ± 0.039	0.945 ± 0.034	0.943 ± 0.036

E Imputation Reconstruction Quality

This section reports per-cohort imputer reconstruction error on the artificial-mask cells and discusses why reconstruction quality is decoupled from downstream classification.

E.1 Reconstruction Mean Absolute Error (MAE)

Tab. 13 reports per-cohort, per-imputer reconstruction MAE on the artificial-mask test cells at injected rate 30%, swept over imputer ensemble size $K \in \{1, 2, 4, 8, 16\}$ (lower is better). The diffusion- and flow-based imputers (TabCSDI, DiffPuter, CFMI) achieve substantially lower error than MIRI at K=1 across all four cohorts, but MIRI catches up as K_{imp} grows: by K=16 the four learned imputers are within approximately 0.05 MAE on D1, D2, and D3, and within approximately 0.10 MAE on D4. D4 produces the largest absolute errors across imputers, consistent with its longitudinal-attrition missingness pattern documented in App. A.4.

E.2 Reconstruction Root Mean Squared Error (RMSE)

Tab. 14 reports the same comparison under RMSE, which weights large reconstruction errors more heavily than MAE. The relative ordering of imputers is preserved: TabCSDI, DiffPuter, and CFMI are the strongest reconstructors at K=1, MIRI is the weakest at K=1 but converges as K_{imp} grows, and D4 again produces the largest absolute errors. The fact that the MAE and RMSE rankings agree to this degree suggests that the imputers do not differ markedly in tail behavior on these cohorts, only in the mean accuracy of their imputed values.

F Demographic Subgroup Analysis

This section reports per-subgroup classifier and imputer performance across all 14 canonical demographic axes documented in App. A.8, both pooled and stratified by injected missingness rate. The pooled figures (Secs. F.1 to F.4) show how the aggregate findings of Secs. D and E distribute across subgroups.

Table 13: Imputer reconstruction error on the diagonal K-sweep @ rate= 30%, metric MAE (test set, lower is better). Rows = dataset × imputer-base, columns = $K \in \{1, 2, 4, 8, 16\}$ where imputer = <base>+AVGK (or bare base for K=1). Green marks the best K per row (lowest error), blue marks the second-best.

Dataset	Imputer	K=1	K=2	K=4	K=8	K=16
D1	MIRI	1.314 ± 0.071	1.061 ± 0.072	0.917 ± 0.070	0.830 ± 0.078	0.790 ± 0.076
	TabCSDI	0.836 ± 0.079	0.826 ± 0.079	0.820 ± 0.080	0.819 ± 0.083	0.817 ± 0.085
	DiffPuter	0.843 ± 0.088	0.834 ± 0.078	0.825 ± 0.080	0.821 ± 0.082	0.818 ± 0.084
	CFMI	0.840 ± 0.082	0.829 ± 0.083	0.824 ± 0.083	0.820 ± 0.083	0.818 ± 0.083
D2	MIRI	1.321 ± 0.075	1.117 ± 0.070	0.957 ± 0.060	0.868 ± 0.063	0.823 ± 0.066
	TabCSDI	0.856 ± 0.066	0.853 ± 0.067	0.850 ± 0.068	0.847 ± 0.067	0.844 ± 0.066
	DiffPuter	0.865 ± 0.070	0.855 ± 0.067	0.851 ± 0.065	0.848 ± 0.065	0.846 ± 0.066
	CFMI	0.869 ± 0.061	0.854 ± 0.067	0.851 ± 0.065	0.848 ± 0.065	0.846 ± 0.065
D3	MIRI	1.174 ± 0.079	0.929 ± 0.083	0.787 ± 0.076	0.691 ± 0.082	0.649 ± 0.089
	TabCSDI	0.875 ± 0.090	0.873 ± 0.087	0.868 ± 0.089	0.867 ± 0.089	0.866 ± 0.090
	DiffPuter	0.871 ± 0.090	0.873 ± 0.085	0.869 ± 0.088	0.866 ± 0.088	0.865 ± 0.090
	CFMI	0.883 ± 0.088	0.877 ± 0.085	0.874 ± 0.087	0.869 ± 0.088	0.867 ± 0.089
D4	MIRI	1.199 ± 0.100	0.987 ± 0.068	0.889 ± 0.096	0.826 ± 0.089	0.796 ± 0.093
	TabCSDI	1.362 ± 0.064	1.160 ± 0.092	1.036 ± 0.107	0.952 ± 0.109	0.897 ± 0.114
	DiffPuter	1.246 ± 0.096	1.079 ± 0.087	0.973 ± 0.088	0.901 ± 0.092	0.877 ± 0.094
	CFMI	1.174 ± 0.098	1.050 ± 0.090	0.976 ± 0.107	0.930 ± 0.109	0.907 ± 0.110

Table 14: Imputer reconstruction error on the diagonal K-sweep @ rate= 30%, metric RMSE (test set, lower is better). Rows = dataset × imputer-base, columns = $K \in \{1, 2, 4, 8, 16\}$ where imputer = <base>AVGK (or bare base for K=1). Green marks the best K per row (lowest error), blue marks the second-best.

Dataset	Imputer	K=1	K=2	K=4	K=8	K=16
D1	MIRI	1.704 ± 0.090	1.383 ± 0.104	1.214 ± 0.102	1.113 ± 0.107	1.060 ± 0.106
	TabCSDI	1.118 ± 0.106	1.105 ± 0.108	1.099 ± 0.109	1.098 ± 0.109	1.097 ± 0.110
	DiffPuter	1.122 ± 0.112	1.117 ± 0.107	1.104 ± 0.108	1.100 ± 0.107	1.098 ± 0.108
	CFMI	1.122 ± 0.110	1.109 ± 0.111	1.100 ± 0.109	1.098 ± 0.109	1.098 ± 0.109
D2	MIRI	1.735 ± 0.201	1.468 ± 0.209	1.266 ± 0.229	1.163 ± 0.239	1.103 ± 0.248
	TabCSDI	1.141 ± 0.233	1.136 ± 0.237	1.133 ± 0.237	1.131 ± 0.236	1.128 ± 0.236
	DiffPuter	1.154 ± 0.231	1.144 ± 0.233	1.137 ± 0.234	1.133 ± 0.236	1.130 ± 0.236
	CFMI	1.166 ± 0.227	1.141 ± 0.235	1.136 ± 0.235	1.131 ± 0.236	1.130 ± 0.237
D3	MIRI	1.514 ± 0.119	1.202 ± 0.132	1.022 ± 0.133	0.915 ± 0.146	0.861 ± 0.158
	TabCSDI	1.110 ± 0.140	1.107 ± 0.137	1.103 ± 0.141	1.102 ± 0.141	1.100 ± 0.142
	DiffPuter	1.109 ± 0.142	1.109 ± 0.139	1.104 ± 0.140	1.100 ± 0.140	1.099 ± 0.142
	CFMI	1.122 ± 0.141	1.116 ± 0.135	1.111 ± 0.138	1.106 ± 0.139	1.102 ± 0.141
D4	MIRI	1.546 ± 0.167	1.298 ± 0.147	1.185 ± 0.193	1.116 ± 0.196	1.081 ± 0.198
	TabCSDI	1.731 ± 0.141	1.495 ± 0.143	1.335 ± 0.181	1.232 ± 0.186	1.179 ± 0.198
	DiffPuter	1.600 ± 0.171	1.389 ± 0.179	1.262 ± 0.172	1.185 ± 0.188	1.152 ± 0.195
	CFMI	1.508 ± 0.159	1.363 ± 0.178	1.276 ± 0.196	1.226 ± 0.198	1.195 ± 0.203

F.1 Subgroup K-Sweep – Accuracy

Fig. 9 extends the aggregate classifier comparison from Tab. 9 to a per-subgroup view at injected rate 30%. Each panel is one canonical demographic subgroup. Each line within a panel is one classifier, and the horizontal axis is the diagonal K-sweep with $K_{\text{imp}} = K_{\text{clf}}$. Two patterns are visible across panels: classifier ranking is broadly stable across subgroups, but absolute accuracy varies between the highest- and lowest-performing subgroups for the same classifier and K.

F.2 Subgroup K-Sweep – AUC

Fig. 10 reports the same per-subgroup comparison under ROC-AUC at injected rate 30%. The within-classifier K trend matches the aggregate finding in Tab. 10, monotone improvement up to

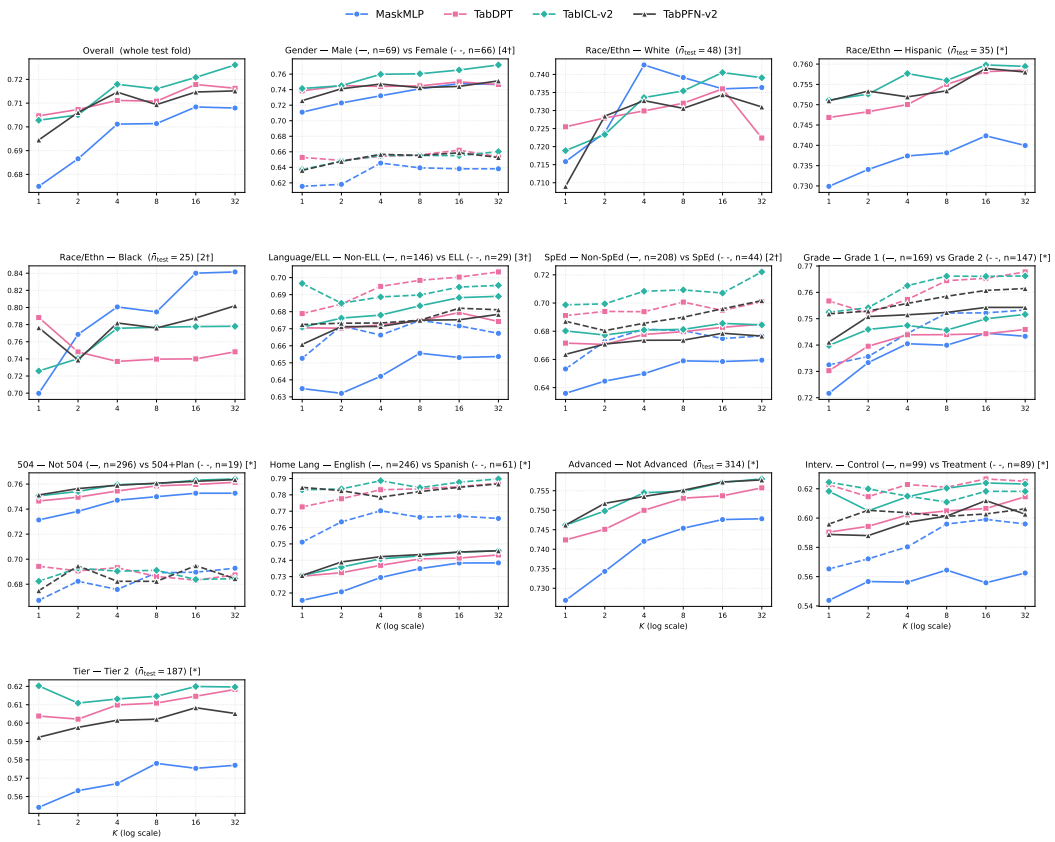

Figure 9: Each panel is one canonical demographic subgroup. Eeach line is one classifier (MaskMLP, TabDPT, TabICL-v2, TabPFN-v2). The K-axis (log) is the diagonal averaging order, where the imputer ensemble size matches the classifier ensemble size ($K_{\text{imp}} = K_{\text{clf}}$). Cells are pooled across the four cohorts (per-cohort mean, then averaged across contributing cohorts) with accuracy at rate $= 30\%$.

K=16 with diminishing returns past K=8, and this trend holds at the subgroup level: ensembling does not produce per-subgroup regressions, even in subgroups where absolute AUC is lower.

F.3 Imputation Quality across Subgroups – MAE

Fig. 11 extends the aggregate imputer comparison from Tab. 13 to a per-subgroup view. Reading down a panel shows whether reconstruction error varies systematically across subgroups for a given imputer. Reading across panels at fixed K shows whether one imputer is more or less robust to subgroup composition than another. The pattern matches the aggregate one: TabCSDI, DiffPuter, and CFMI achieve uniformly lower MAE than MIRI at K=1, with the four imputers converging by K=16.

F.4 Imputation Quality across Subgroups – RMSE

Fig. 12 reports the same per-subgroup comparison under RMSE. The relative ordering of imputers is preserved with respect to MAE in Fig. 11, consistent with the aggregate observation in Sec. E.2 that the imputers do not differ markedly in tail behavior on these cohorts.

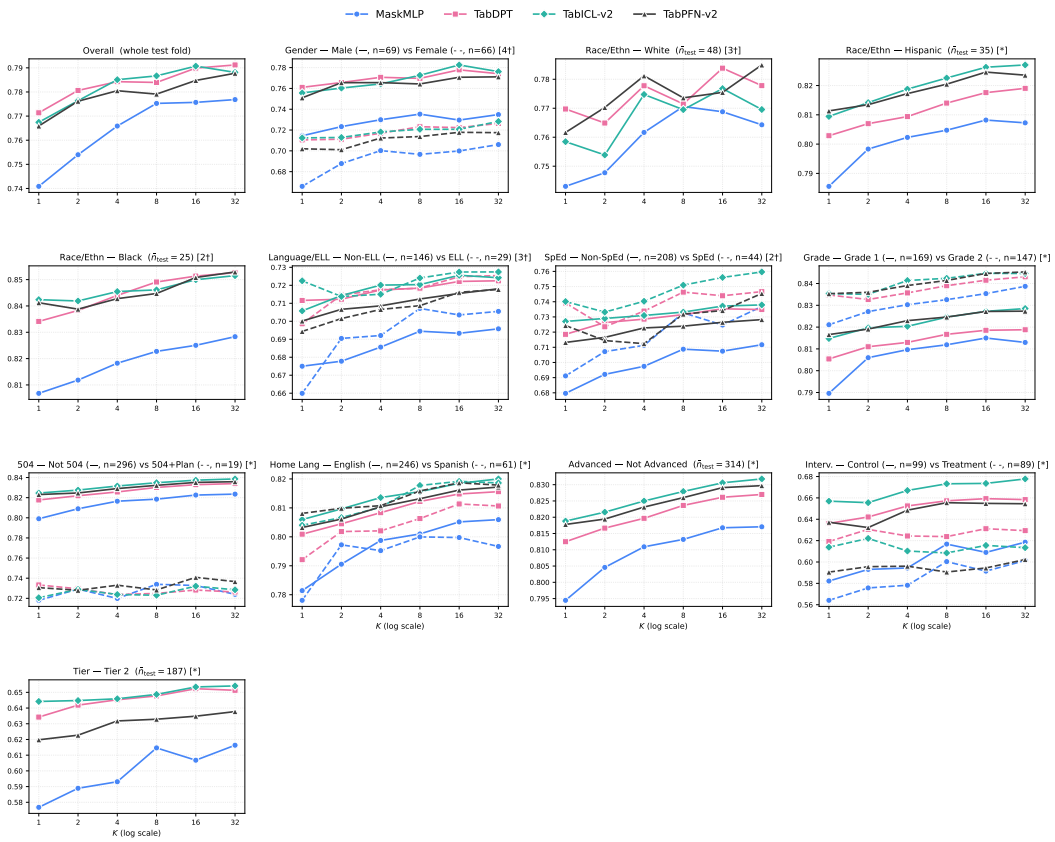

Figure 10: Same axes and pooling as Fig. 9, AUC instead of accuracy.

G TabPFN-v2 Fine-Tuning

We assess whether fine-tuning the TabPFN-v2 prior on RealLit cohorts improves over the zero-shot in-context predictor.

G.1 Setup

We compare four TabPFN-v2 inference modes to test whether per-cohort or cross-cohort fine-tuning improves over the bare in-context prior used everywhere else in the appendix.

Zero-shot is bare TabPFN-v2 checkpoint (version v2.6) with no further training, exactly as described in Sec. C.1.2: the training fold is passed as in-context support and no gradient updates occur. FT-on-target fine-tunes TabPFN-v2 end-to-end on the current cohort's training fold starting from the off-the-shelf released prior—this tests whether the published prior leaves headroom that per-cohort supervised training can recover. FT-from-D1 fine-tunes TabPFN-v2 once on D1's training fold and is then used for inference on the current cohort with no further fine-tuning—this isolates pure cross-cohort transfer from the public benchmark. FT-init-D1 initializes from the FT-from-D1 checkpoint and then fine-tunes a second time on the current cohort, testing whether a public-prior-plus-cohort-polish regime beats either ingredient alone.

We use D1 as the transfer source for the FT-from-D1 and FT-init-D1 variants for two reasons: (i) D1 is the public ECRI benchmark and so represents the most realistic transfer scenario for downstream researchers, who can replicate the source training without an access agreement, and (ii) D1 lacks student-level demographics by design, which means D1→{D2,D3,D4} transfer also tests whether a demographic-blind source prior remains useful in demographic-aware target settings. All fine-tuned variants are trained per fold on the same 10×5 cross-validation split used for every other component of the pipeline (Sec. C.5).

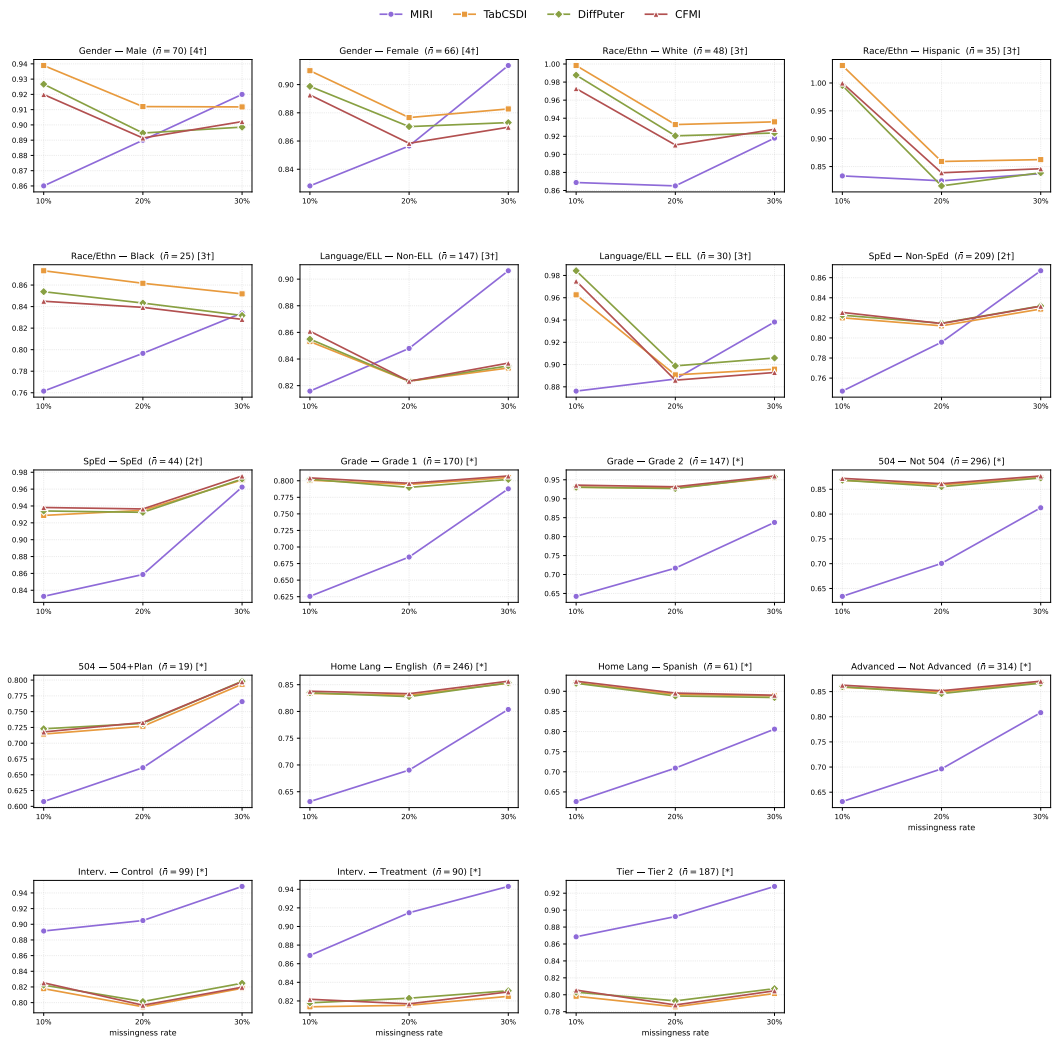

Figure 11: Each panel is one canonical demographic subgroup. Each line is one imputer (MIRI, TabCSDI, DiffPuter, CFMI). The K-axis (log) is the diagonal averaging order, where the imputer ensemble size matches the classifier ensemble size ($K_{\text{imp}} = K_{\text{clf}}$). Cells are pooled across the four cohorts (per-cohort mean, then averaged across contributing cohorts) with reconstruction MAE at rate $= 30\%$.

G.2 Per-Dataset Accuracy

Fine-tuning ablations. Fine-tuning provides only modest and inconsistent gains over the zero-shot TabPFN-v2 baseline. In the learning-rate ablation (Table 15), reducing the learning rate from 2×10^{-5} to 2×10^{-6} recovers or slightly improves performance on some cohorts, most notably D3, but does not yield a consistent advantage across datasets. The cross-dataset ablations show a similar pattern: the best fine-tuning source varies by target cohort and ensemble size, while zero-shot inference remains competitive and is best for several settings, including D2 at $K = 1$ and D4 at both $K = 1$ and $K = 8$. These results suggest that the limited benefit of fine-tuning is not explained solely by the default learning rate or by the choice of target-only adaptation. Instead, adaptation appears sensitive to the source and target cohorts, consistent with the substantial heterogeneity across datasets.This observation is also consistent with recent work showing that fine-tuning tabular foundation models can provide limited or unstable improvements under heterogeneous or distribution-shifted settings (Rubachev et al., 2025; Tanna et al., 2026).

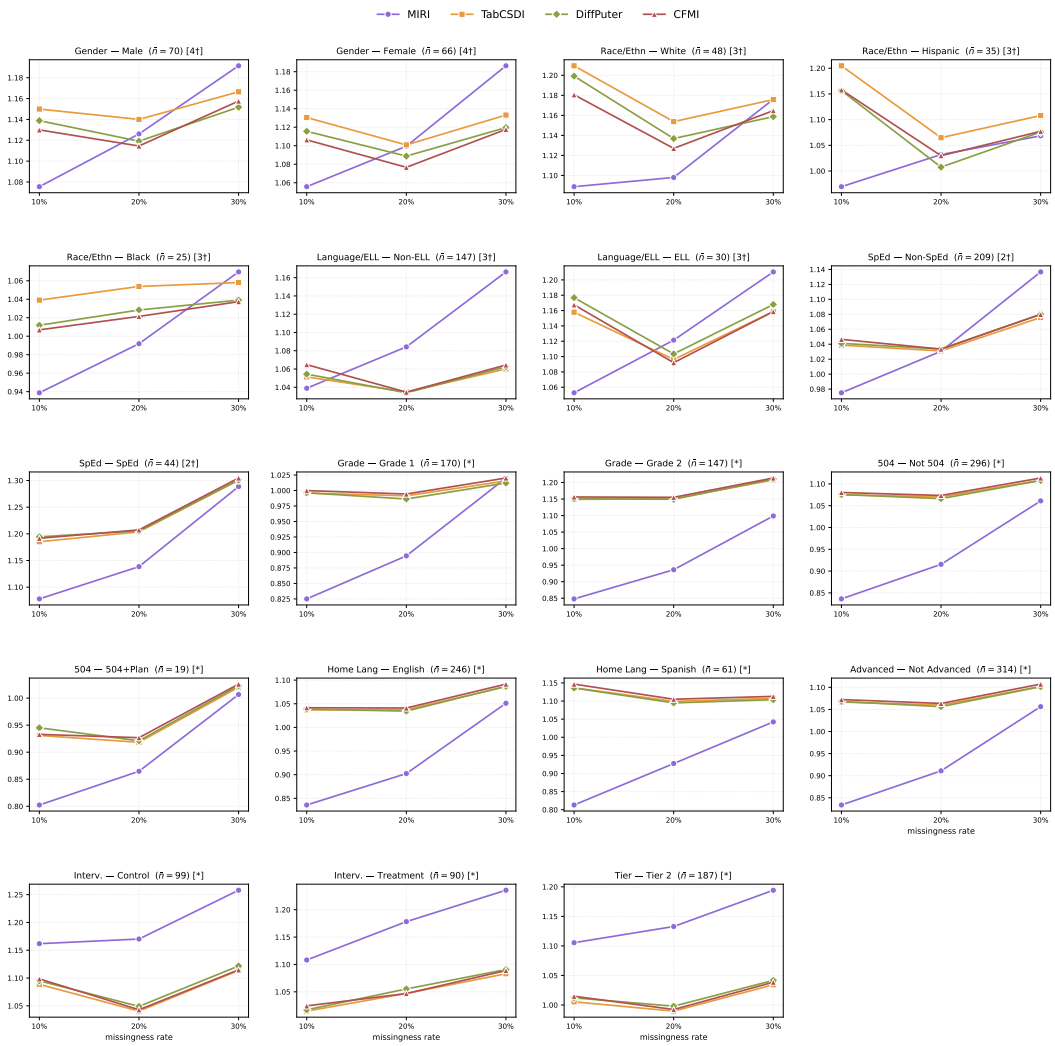

Figure 12: Same axes and pooling as Fig. 11, RMSE instead of MAE.

Table 15: Learning-rate ablation for TabPFN-v2. Results are reported as mean ± standard deviation. Bold indicates the best reported mean for each dataset.

Dataset	TabPFN-v2	TabPFN-v2-FT ($\mathrm{lr} = 2 \times 10^{-5}$)	TabPFN-v2-FT ($\mathrm{lr} = 2 \times 10^{-6}$)
D1	$\mathbf{0.696 \pm 0.026}$	0.692 ± 0.023	0.694 ± 0.020
D2	$\mathbf{0.832 \pm 0.013}$	0.828 ± 0.017	0.832 ± 0.016
D3	0.860 ± 0.024	0.857 ± 0.029	$\mathbf{0.861 \pm 0.024}$
D4	$\mathbf{0.914 \pm 0.022}$	0.914 ± 0.057	$\mathbf{0.914 \pm 0.022}$

Note. The default fine-tuning learning rate is 2×10^{-5}. The alternative learning rate was selected based on internal validation using the training set.

Tab. 16 reports test accuracy at injected rate 30% for seven TabPFN-v2 variants across D1–D4 at $K \in \{1, 8\}$. A variant is omitted (-) when the target cohort is one of its source cohorts: D1 therefore has no FT-from-D1, FT-init-D1 or FT-D1→D2 entry, D2 has no FT-from-D2 or FT-D1→D2 entry, and D3 has no FT-from-D3 entry.

Table 16: TabPFN-v2 fine-tuning comparison @ rate= 30%, metric ACC. Rows = dataset × K (ensemble size), columns = TabPFN variant. Green marks the best variant per row. Dashes mark variants undefined for that target: FT-from-Di and FT-init-Di do not apply when the target *is* Di, and FT-D1→D2 applies only to targets outside its two source cohorts.

Dataset	K	Zero-shot	FT-on-target	FT-from-D1	FT-init-D1	FT-from-D2	FT-from-D3	FT-D1→D2
D1	K=1	0.607±0.032	0.612±0.039	—	—	0.615±0.028	0.614±0.037	—
	K=8	0.609±0.033	0.617±0.036	—	—	0.620±0.044	0.613±0.049	—
D2	K=1	0.716±0.039	0.704±0.057	0.704±0.044	0.706±0.048	—	0.710±0.057	—
	K=8	0.709±0.030	0.715±0.044	0.716±0.051	0.714±0.041	—	0.721±0.040	—
D3	K=1	0.768±0.036	0.766±0.022	0.760±0.034	0.770±0.026	0.762±0.033	—	0.762±0.036
	K=8	0.766±0.030	0.767±0.028	0.765±0.031	0.766±0.029	0.768±0.031	—	0.766±0.029
D4	K=1	0.830±0.068	0.807±0.070	0.815±0.073	0.815±0.065	0.815±0.073	0.815±0.073	0.807±0.070
	K=8	0.822±0.082	0.800±0.075	0.807±0.078	0.807±0.078	0.800±0.075	0.807±0.070	0.807±0.078

G.3 Per-Mechanism AUC across Variants

Tab. 17 reports the same comparison under ROC-AUC, broken out by missingness mechanism (MAR/MCAR/MNAR) at injected rate 30%. The per-mechanism breakdown shows which fine-tuning variant wins under each mechanism and supplies the per-mechanism numbers cited in Sec. G.4 below.

Table 17: TabPFN-v2 fine-tuning vs zero-shot, per mechanism, across all four datasets (NATIVE imputer, rate=30%). *Zero-shot* = bare TabPFN-v2. *FT-on-target* = fine-tuned on this dataset's training fold from the off-the-shelf TabPFN-v2 prior. *FT-from-D1* = TabPFN-v2 fine-tuned on D1, used for inference only on this dataset. *FT-init-D1* = D1-fine-tuned weights as init, then a further round of fine-tuning on this dataset. Green marks the best variant per (dataset, K, mechanism) cell. *FT-from-D1* and *FT-init-D1* are N/A on D1 itself.

Dataset	Variant	K	MAR	MCAR	MNAR
D1	Zero-shot	K=1	0.646 ± 0.035	0.637 ± 0.058	0.660 ± 0.036
	FT-on-target	K=1	0.658 ± 0.048	0.654 ± 0.060	0.651 ± 0.011
	FT-from-D1	K=1	—	—	—
	FT-init-D1	K=1	—	—	—
	Zero-shot	K=8	0.646 ± 0.042	0.648 ± 0.056	0.669 ± 0.029
	FT-on-target	K=8	0.648 ± 0.046	0.668 ± 0.079	0.662 ± 0.020
	FT-from-D1	K=8	—	—	—
	FT-init-D1	K=8	—	—	—
D2	Zero-shot	K=1	0.790 ± 0.010	0.805 ± 0.034	0.769 ± 0.045
	FT-on-target	K=1	0.794 ± 0.011	0.788 ± 0.023	0.770 ± 0.062
	FT-from-D1	K=1	0.790 ± 0.015	0.800 ± 0.020	0.777 ± 0.058
	FT-init-D1	K=1	0.790 ± 0.011	0.792 ± 0.016	0.771 ± 0.057
	Zero-shot	K=8	0.797 ± 0.009	0.798 ± 0.023	0.782 ± 0.035
	FT-on-target	K=8	0.798 ± 0.005	0.795 ± 0.025	0.772 ± 0.061
	FT-from-D1	K=8	0.798 ± 0.006	0.803 ± 0.022	0.768 ± 0.068
	FT-init-D1	K=8	0.796 ± 0.003	0.796 ± 0.028	0.772 ± 0.059
D3	Zero-shot	K=1	0.839 ± 0.025	0.847 ± 0.032	0.848 ± 0.036
	FT-on-target	K=1	0.834 ± 0.029	0.846 ± 0.032	0.847 ± 0.034
	FT-from-D1	K=1	0.838 ± 0.021	0.839 ± 0.034	0.844 ± 0.033
	FT-init-D1	K=1	0.838 ± 0.021	0.848 ± 0.030	0.848 ± 0.034
	Zero-shot	K=8	0.841 ± 0.026	0.846 ± 0.030	0.851 ± 0.032
	FT-on-target	K=8	0.841 ± 0.024	0.845 ± 0.027	0.855 ± 0.034
	FT-from-D1	K=8	0.839 ± 0.024	0.842 ± 0.031	0.852 ± 0.033
	FT-init-D1	K=8	0.843 ± 0.025	0.846 ± 0.027	0.855 ± 0.034
D4	Zero-shot	K=1	0.933 ± 0.029	0.920 ± 0.092	0.926 ± 0.074
	FT-on-target	K=1	0.915 ± 0.059	0.914 ± 0.088	0.878 ± 0.054
	FT-from-D1	K=1	0.924 ± 0.046	0.938 ± 0.091	0.890 ± 0.050
	FT-init-D1	K=1	0.927 ± 0.048	0.926 ± 0.065	0.878 ± 0.057
	Zero-shot	K=8	0.932 ± 0.047	0.933 ± 0.088	0.920 ± 0.095
	FT-on-target	K=8	0.908 ± 0.081	0.920 ± 0.092	0.877 ± 0.076
	FT-from-D1	K=8	0.932 ± 0.057	0.926 ± 0.098	0.890 ± 0.057
	FT-init-D1	K=8	0.914 ± 0.071	0.920 ± 0.092	0.883 ± 0.066

G.4 Discussion: Modest Gains, Poor Transfer

Reading Tabs. 16 and 17 together, three patterns are consistent across both metrics, both ensemble sizes, and all three missingness mechanisms.

First, fine-tuning produces at best modest gains on the source dataset. On D1, the source for FT-on-target, accuracy improves over zero-shot by 0.005 at $K{=}1$ and 0.008 at $K{=}8$. AUC gains under MAR and MCAR are similarly small ($0.012-0.017$), and FT-on-target is actually worse than zero-shot under MNAR (-0.009 at $K{=}1$). These are within or near the run-to-run standard deviation of the zero-shot baseline, so the practical improvement from per-cohort fine-tuning on this cohort is small even before considering compute cost.

Second, transfer from D1 to D2, D3, and D4 fails to add value over the zero-shot prior. Across the three target cohorts, FT-on-target, FT-from-D1, and FT-init-D1 are all within ± 0.005 accuracy of zero-shot on average, and each of them regresses below zero-shot on at least one (cohort, K, mechanism) cell. The most pronounced failure is on D4, where every fine-tuning variant underperforms zero-shot at both $K{=}1$ (-0.015 to -0.023 ACC) and $K{=}8$ (-0.015 to -0.022 ACC), and where AUC drops by $\approx 0.04-0.05$ under MNAR for FT-on-target.

Third, the variant that comes closest to working consistently is FT-init-D1, but only on D3, where it ties zero-shot at $K{=}1$ and $K{=}8$ on most cells. Even there, the absolute gain over zero-shot is bounded by 0.002 AUC. No fine-tuning variant strictly dominates zero-shot on more than one (cohort, K, mechanism) configuration.

We attribute these patterns to two structural facts about the setting. The TabPFN-v2 prior has been pretrained on millions of synthetic tabular tasks and operates by in-context inference, so a few hundred to a few thousand training rows from a single cohort do not add information that the prior did not already capture, *i.e.*, the pretrained prior is already close to a Bayes predictor on tabular tasks of this size. At the same time, the four cohorts in RealLit differ along feature schema, grade range, assessment battery, and missingness pattern, the cross-cohort separation visible in the t-SNE projection of Fig. 5 is a direct consequence of these differences. A fine-tune step that adapts the prior to D1's geometry therefore moves away from the geometry of D2, D3, and D4 rather than toward a shared regularity. The takeaway for practitioners is that for TabPFN-v2 in this domain, in-context inference is the right default, and fine-tuning should be reserved for the source cohort only, not relied on for transfer.

H Statistical Significance

The improvements for increasing ensemble size ($K = 1$ vs. $K = 16$) for tabPFN are statistically supported, e.g.,

for D2: $t = 3.6$, $p = 6 \times 10^{-4}$ and 95% CI: $[+0.009, +0.029]$

for D3 (G1–G2): $t = 4.56$, $p = 3 \times 10^{-5}$ and 95% CI: $[+0.005, +0.012]$

for D3 (G3): $t = 4.67$, $p = 2 \times 10^{-5}$ and 95% CI: $[+0.005, +0.013]$.

The estimates are also positive for D1 and D4, but their confidence intervals include zero, so the improvements are not statistically supported in those cohorts.

We also identify statistically supported variability for several subgroups, including at-risk students via bootstrap analysis, showing that the models were generally less accurate for the students most in need of identification. Specifically, $p = 6.96 \times 10^{-4}$ for D1, $p = 0.00327$ for D2, $p = 0.00874$ for D3, and $p = 0.00294$ for D4. Other supported subgroups include gender, Black students, Section 504 eligibility, advanced or gifted status, and socioeconomic status.

I Limitations and Failure Modes

I.1 MaskMLP Reproducibility

As the code of (Shangguan et al., 2024) is not publicly available, we reimplement the method as summarized in Tab. 18. Our reimplementation matches the reported AUC (0.023 over the original) but underperforms on accuracy (-0.057). The full comparison is in Tab. 19.

Table 18: MaskMLP reimplementation recipe. We follow the masked-pretraining description in (Shangguan et al., 2024). Specific architecture and optimization values that the original paper does not document explicitly are filled in for our reimplementation.

Component	Setting
Backbone architecture	3-layer MLP encoder, hidden size 64, Linear $\rightarrow$ ReLU $\rightarrow$ BatchNorm
Output head	Linear projection to logits (no sigmoid, BCE-with-logits applied externally)
Missing-entry sentinel	-1
Pre-training objective	Cosine-embedding loss between row x and a randomly masked view x_{mask}
Pre-training mask rate	25% of observed entries per row, sampled per minibatch
Pre-training epochs	50
Fine-tuning objective	Binary cross-entropy with logits on the at-risk target
Fine-tuning epochs	50 (with validation-AUC early stopping, see below)
Optimizer (both stages)	Adam, $\text{lr}=10^{-3}$, $\text{betas}=(0.9, 0.999)$, $\text{eps}=10^{-8}$, no weight decay
LR schedule	None (constant LR, cosine warm-restart available but disabled by default)
Batch size (both stages)	16
Validation split	Stratified 15% of the labeled training fold
Early-stopping rule	Best validation AUC with patience 10 epochs, best-val weights restored
Per-fold refit	Yes (each CV fold trains a new encoder, no shared checkpoint)

Table 19: MaskMLP reimplementation on D1 (ECRI) compared to the numbers reported in (Shangguan et al., 2024), Table 1. Original numbers are averaged across the two D1 reading targets (Word Identification, Word Attack) at the School Split, All Students setting – the configuration that most closely matches our school-stratified cross-validation protocol (App. C.5). Cells from our reimplementation are at the original-data, $K=1$ baseline configuration of Tab. 2.

Metric	(Shangguan et al., 2024) (reported)	Our reimplementation
AUC	0.662	0.685 ± 0.007
ACC	0.671	0.614 ± 0.008

Beyond the D1 reimplementation gap, during our cross-cohort evaluation we were unable to fully reproduce MaskMLP under every (dataset, target, mechanism) configuration. We observed cells in which the encoder failed to converge on a usable representation, producing AUC consistent with random guessing rather than the encoder's typical $0.7-0.8$ range visible in Tab. 10. The failure mode appears related to the interaction between block-MNAR cohorts and the cosine-embedding pre-training objective rather than to a single hyperparameter, but we have not been able to characterize it precisely. Reported MaskMLP numbers are pooled over runs that did converge. Readers should treat MaskMLP results as informative rather than definitive, and the foundation-model classifiers (TabPFN-v2, TabICL-v2, TabDPT) carry the central comparison weight in our evaluation. The evaluation code release (App. C.8) includes the per-cell logs needed for community debugging of the affected configurations.

I.2 Subgroup Sample-Size Caveats

Subgroups with mean test-fold support $n < 8$ are dropped from the subgroup-stratified results in App. F, because per-subgroup AUC and MAE estimates become unstable below that threshold. Two consequences follow. First, dropped subgroups can be small in the cross-validation test fold even when they are large in the underlying cohort, since joint stratification on target and school identifier (App. C.5) can place a subgroup's students into only one fold. Second, intersectional subgroups (e.g.,

race × disability status, gender × home language) fall below the threshold for every combination in our cohorts, so RealLit does not support intersectional reliability claims.

I.3 At-Risk Threshold Cross-Dataset Comparability

As discussed in App. A.6, the binary at-risk targets across cohorts use different instruments and follow-up windows: D2 uses a static KTEA-3 standard-score threshold ($SS < 90$), D3 uses spring FAST ELA proficiency recovery, and D4 uses Grade 3 WRAT-3 Spelling recovery. Positive-class prevalence varies from approximately 15% in the kindergarten cohort to $44-52\%$ in the first- and third-grade cohorts.

Raw AUC values are therefore not directly comparable across cohorts. The cross-cohort summary numbers in Tab. 2 aggregate over heterogeneous label constructions, and the apparent superiority of one cohort's AUC over another may reflect prevalence and label-construct differences rather than model behavior. Cross-cohort claims in this paper rely on the relative ranking of methods within each cohort, not on absolute cross-cohort AUC values.

I.4 Imputation Coverage

Our imputation evaluation focuses on five methods: MICE (statistical reference, App. C.2.1) and four learned methods: MIRI (rectified-flow regularization), TabCSDI (score-based diffusion), DiffPuter (latent-space diffusion), and CFMI (flow matching). Notable methods we do not benchmark as primary baselines include MissForest at non-zero rates, HyperImpute, and the VAE-based family (MIWAE, not-MIWAE). We also do not evaluate longitudinal joint imputation models that explicitly use a cohort's time-indexed structure, which is a missed opportunity for D4's Grades 1-4 panel.

I.5 Geographic and Linguistic Scope

RealLit spans six U.S. states and is limited to English-language reading outcomes. Bilingual and dual-language program settings are absent, non-U.S. contexts are not represented, and the four cohorts draw predominantly from metropolitan and suburban districts, leaving rural, remote, and resource-constrained schools under-represented. Subgroup reliability findings should not be presumed to generalize to these contexts.

J Ethics and Responsible Use

J.1 Responsible Use of Student Records

RealLit is intended to support the methodological evaluations described in the main paper, benchmarking tabular foundation models, imputation pipelines, and subgroup reliability under realistic missingness. We trust researchers to use the benchmark in good faith and consistent with the spirit of the consent under which the underlying records were collected.

J.2 Demographic Auditing Principles

Student-level demographic attributes are included to make subgroup behavior measurable, not to justify demographic targeting. The subgroup analyses in Sec. F are audits, not performance metrics for selecting a deployed model: a subgroup gap is a signal for further investigation rather than a final claim about fairness, and we do not recommend using demographic attributes as model inputs without independent validation in the deployment setting.

J.3 Deployment Caveats

Models trained on RealLit-style data should not be used as substitutes for teacher judgment, family knowledge, or professional evaluation. Deployment in real schools requires local validation against the decisions schools and families actually make, transparent uncertainty surfaced to the educator using a score, and human-in-the-loop workflows where a score informs but does not determine

eligibility, placement, or denial of services. High-stakes contexts will typically require additional safeguards developed with the district, the local IRB, and the affected families.

K DATASHEET FOR REALLIT

This section provides a single suite-level datasheet for RealLit following the template of Gebru *et al.* (Gebru et al., 2021). RealLit aggregates D1 (the publicly released ECRI benchmark of (Shangguan et al., 2024)) and three further cohorts (D2-D4) under a common evaluation harness. We therefore document the suite as one artifact, with per-cohort distribution and maintenance breakouts where governance differs. Full per-cohort descriptions are in App. A, and the provided entries below provide an outline.

Purpose: RealLit is a benchmark suite for evaluating tabular foundation models, imputation pipelines, and subgroup reliability under realistic missingness in early-literacy outcome prediction. The suite combines a previously-released cohort (D1, ECRI (Shangguan et al., 2024)) with three additional cohorts that expand grade range, assessment-battery diversity, and student-level demographic coverage.

Composition: Four cohorts.

- D1 (ECRI): $6{,}803$ students, 16 features, 2 binary outcomes (Word Identification, Word Attack), with 946 students having valid end-of-year labels (at 40.8% cell-wise missingness). D1 is used as a reference cohort via the public ECRI release, not redistributed by RealLit.
- D2 (Kindergarten Screening): $1{,}039$ students, 14 features, 4 outcomes (KTEA-3 NWD/LWR, CELF-5 SC/RS), valid-label counts $515/514/205/488$, 47.3% missingness, and partial student-level demographics.
- D3 (District Records): two cohorts, a $2{,}088$-student Grade 1-2 panel ($22+$ features, 3 outcomes, 10.3% missingness) and a $1{,}137$-student Grade 3 cohort (12 features, 1 outcome, 9.1% missingness), fully observed student-level demographics.
- D4 (Longitudinal Cohort): 318 students, 349 harmonized features, 4 outcomes, missingness rises from 3.8% in Grade 1 to 36.4% in Grade 4 due to longitudinal attrition, and demographics observed for 316 of 318 students.

See App. A for full per-cohort details.

Collection: D1 was collected by (Shangguan et al., 2024) as part of an IRB-approved multi-school study spanning 44 elementary schools, $6{,}916$ students, and 172 teachers. D2 is a longitudinal kindergarten study across 33 public, private, and virtual schools in five U.S. states, with iPad-administered one-to-one sessions and parental consent. D3 is anonymized student-level operational records from a single Florida school district, de-identified at the source under a district-level data sharing agreement. D4 is a longitudinal reading and writing study across six schools in a metropolitan Florida school district, with trained-tester administration in $30-60$-minute sessions and parental consent, and cohort followed from Grade 1 (fall 2007) through Grade 4 (school years 2007/08–2010/11).

Preprocessing: Each cohort is used as released. Year-by-year measures in D4 are harmonized into a single 349-feature panel. Missingness, label conventions, and feature schema are preserved per cohort. Cross-cohort heterogeneity is part of what RealLit is designed to evaluate.

Recommended uses: Benchmarking tabular foundation models on early-literacy outcome prediction, imputation evaluation under realistic missingness mechanisms (MCAR/MAR/MNAR stress tests), subgroup reliability auditing on student-level demographic attributes (D2-D4), and longitudinal robustness studies under attrition-driven missingness (D4).

Distribution: D1 (ECRI) is not redistributed by RealLit, as users access it directly from the public ECRI release of (Shangguan et al., 2024). The introduced cohorts are partially released as part of RealLit under the Creative Commons Attribution 4.0 (CC BY 4.0) license. Access procedures and the public release URL are documented in App. C.8.

Maintenance: D1 is maintained by the original ECRI authors. D2 and D4 by the parent-study collaborators. D3 operationally by the contributing district. RealLit freezes the per-cohort snapshots used in our evaluation and bundles them with the harness (App. C.8).