Schema-Adaptive Tabular Representation Learning with LLMs for Generalizable Multimodal Clinical Reasoning

Anonymous ACL submission

Abstract

Machine learning for tabular data remains constrained by poor schema generalization, a challenge rooted in the lack of semantic understanding of structured variables. This challenge is particularly acute in domains like clinical medicine, where electronic health record (EHR) schemas vary significantly. To solve this problem, we propose Schema-Adaptive Tabular Representation Learning, a novel method that leverages large language models (LLMs) to create transferable tabular embeddings. By transforming structured variables into semantic natural language statements and encoding them with a pretrained LLM, our approach enables zero-shot alignment across unseen schemas without manual feature engineering or retraining. We integrate our encoder into a multimodal framework for dementia diagnosis, combining tabular and MRI data. Experiments on NACC and ADNI datasets demonstrate state-of-the-art performance and successful zero-shot transfer from NACC to ADNI, two large-scale dementia cohorts with distinctly different schemas, significantly outperforming clinical baselines, including board-certified neurologists, in retrospective diagnostic tasks. These results validate our LLM-driven approach as a scalable, robust solution for heterogeneous real-world data, offering a pathway to extend LLM-based reasoning to structured domains. Code will be released upon publication.

1 Introduction

Machine learning excels in structured data modeling but struggles with cross-domain generalization, particularly in tabular domains where varying schemas cause models to fail across datasets (Topol, 2019; Rajkomar et al., 2019; Miotto et al., 2018; Zhi et al., 2026). This limits AI scalability and reliability in fields like healthcare with diverse electronic health records. At the root of this generalization crisis lies *schema heterogeneity*. Real-world data rarely share consistent column names, coding systems, or data formats. For instance, a key biomarker may appear as a continuous value in one database but as a categorical code in another. Conventional machine learning models are trained on fixed, syntactic representations and lack the ability to reconcile these schema variations, resulting in brittle, non-transferable embeddings (Lee et al., 2020; Yang et al., 2023; Ellis et al., 2022). Manual feature harmonization offers a partial remedy but is inherently non-scalable, error-prone, and unsustainable in complex real-world pipelines.

To overcome these limitations, we advocate a new approach from *schema-dependent learning* to *semantic schema understanding*. This work introduces **Schema-Adaptive Tabular Representation Learning**, a framework that leverages the powerful semantic reasoning of Large Language Models (LLMs) to align heterogeneous structured data through natural language. Instead of treating column-value pairs as numeric tokens, our approach converts them into semantic text, capturing variable metadata and context. Encoded by a pretrained LLM, these texts produce schema-agnostic embeddings, enabling zero-shot transfer across new datasets without manual alignment or retraining (Liu et al., 2022; Hegselmann et al., 2023; Koloski et al., 2025; Lee et al., 2025).

We situate this methodological contribution within a challenging, multimodal setting: differential dementia diagnosis. This task serves as a rigorous *stress test* for our framework, requiring the integration of semantically encoded tabular features with neuroimaging data and the prediction of co-occurring etiologies (e.g., Alzheimer's and vascular pathologies). By embedding our schema-adaptive encoder into a cross-modal transformer with a multi-objective contrastive learning scheme, we test its robustness under real-world heterogeneity, label imbalance, and limited data availability (Johnson et al., 2019; Vale-Silva and Rohr, 2021;

Yang et al., 2022). Importantly, the medical domain here is not our end goal but a high-stakes validation ground for schema-level generalization. This is a core capability essential to scalable, language-driven machine learning.

Our contributions are summarized as follows:

1. We propose a schema-adaptive representation framework that reformulates tabular data as semantically compositional text, allowing pre-trained LLMs to achieve zero-shot schema alignment without explicit feature harmonization or fine-tuning.

2. We integrate this encoder into a multimodal architecture to evaluate its robustness under extreme heterogeneity and limited supervision, achieving state-of-the-art performance and outperforming human experts in retrospective diagnostic settings.

3. Through comprehensive analysis, we demonstrate how LLM-based semantic encoding enhances multimodal learning, improves sample efficiency, and yields interpretable decision patterns grounded in domain knowledge.

2 Related Work

2.1 Tabular Representation Learning

Classical tabular models like Gradient Boosted Decision Trees are purely syntactic and fail to generalize across schemas (Beekly et al.; Mueller et al., 2005). Recent deep learning approaches, including TabPFN (Hollmann et al., 2025),TIP (Du et al., 2025) and TransTab (Wang and Sun, 2022), have begun to incorporate semantic information but remain largely schema-dependent, limiting their robustness under significant schema shifts. The advent of LLMs has catalyzed a new direction. Seminal works in NLP like TaPas (Herzig et al., 2020) and other table pretraining methods (Wang et al., 2021) have focused on deep, intra-table reasoning for tasks like question answering. Models such as TableGPT (Ren et al., 2025; Su et al., 2024) and TableDreamer (Zheng et al., 2025) extend this to generative tasks, primarily on clean, single-source benchmarks. However, the critical challenge of robust generalization across noisy, heterogeneous real-world schemas remains largely unaddressed. Our work is positioned to fill this specific gap. We focus on cross-schema generalization, proposing a simple yet robust method that leverages a pre-trained LLM as a semantic encoder. By transforming column-value pairs into descriptive statements, our framework achieves semantic alignment across diverse, unseen table structures, a key capability underexplored by prior models focused on single-schema understanding.

2.2 Multimodal Learning with Structured Inputs

Multimodal learning seeks to unify representations across diverse data types, such as images, text, and structured records, to enable holistic reasoning. Early approaches relied on late-fusion strategies, combining predictions from modality-specific encoders (Warner et al., 2024), but these often failed to capture cross-modal dependencies during joint representation learning. Recent methods, inspired by CLIP (Radford et al., 2021), use contrastive objectives for modality alignment (Hager et al., 2023; Huang, 2023; Du et al., 2025), but image-tabular and text-tabular integration lags due to structured data's lack of semantic continuity. Our framework reinterprets tabular data as linguistic tokens, enabling integration into language-aligned multimodal architectures via cross-modal attention, enhancing performance through language-driven schema representation while ensuring comparability with prior work.

2.3 Multi-Label Learning and Optimization under Imbalance

Multi-label classification, prevalent in clinical and biomedical tasks, faces challenges from class imbalance and correlated outputs. Standard supervised contrastive learning (SCL) (Khosla et al., 2020) performs well for single-label tasks but struggles with co-occurring or hierarchical labels. Extensions such as MulSupCon (Zhang and Wu, 2024), lREG (Audibert et al., 2024), and JSCL (Lin et al., 2023) incorporate label-aware reweighting to address these issues, yet their performance falters on highly imbalanced datasets. Prototype-based methods like C-GMVAE (Bai et al., 2022) and Proto (Gupta et al., 2023) model label distributions explicitly but are often confounded by high inter-label correlations.

Our framework uses multi-label classification to evaluate schema generalization, employing a composite objective with focal loss, contrastive regularization, and multi-objective optimization, ensuring performance gains via an LLM-driven,

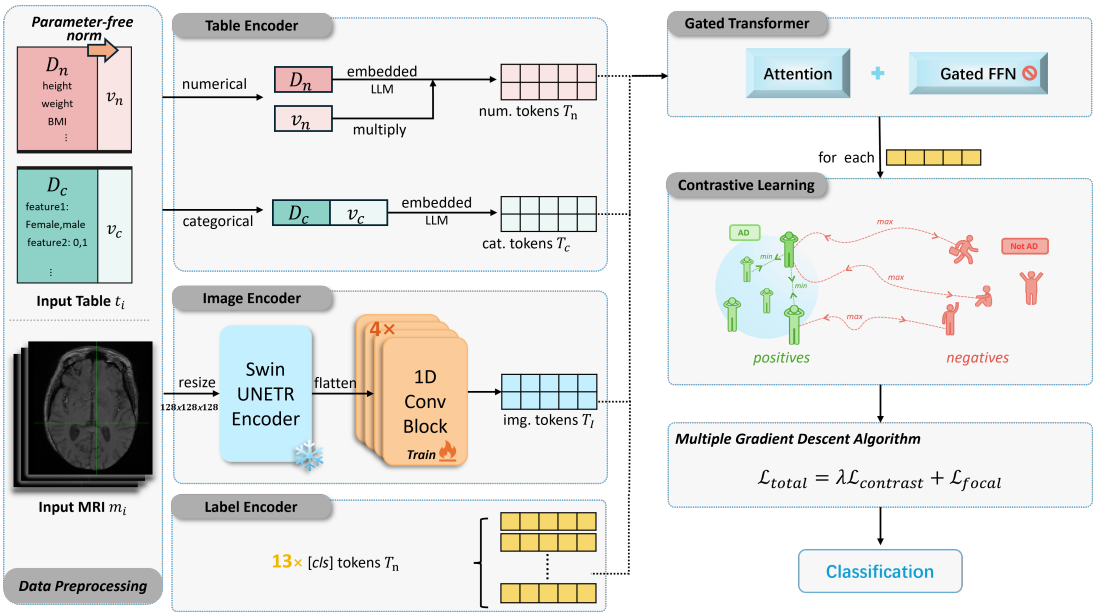

Figure 1: Overview of our proposed architecture. Patient records comprising tabular and imaging data are processed via modality-specific encoders. The LLM-based table encoder handles schema heterogeneity, while a frozen Swin UNETR encodes images. These representations are fused in a gated transformer with label-specific [CLS] tokens. A composite loss, optimized via MGDA, enables robust multi-label learning under class imbalance.

schema-adaptive encoder for robust structured-data learning.

3 Methodology

We present a generalizable framework for learning schema-invariant representations from heterogeneous structured data by leveraging the semantic priors of large language models (LLMs). The framework rests on three key components: (1) a schema-adaptive tabular encoder that performs language-based semantic tokenization of structured inputs; (2) a modality-specific encoder for auxiliary data (e.g., neuroimaging) to assess cross-modal robustness; and (3) a unified transformer-based fusion backbone with a multi-objective optimization scheme for complex multi-label prediction. Figure 1 illustrates the overall architecture. While our evaluation focuses on multimodal dementia diagnosis, the proposed architecture is designed to test a broader hypothesis: *can natural language embeddings serve as a universal representational interface for heterogeneous tabular schemas?*

3.1 Problem Formulation

We formalize the task as a multi-label classification problem over multimodal inputs. Given a dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$, each instance X_i comprises structured tabular data x_i^t and a complementary modality x_i^m (e.g., an MRI volume). The ground-truth label vector $Y_i = [y_{i1}, y_{i2}, \ldots, y_{iL}] \in \{0,1\}^L$ denotes the presence or absence of each condition or class. The learning objective is to find a mapping function $f_\theta(X_i)$ that minimizes an aggregate loss over all labels:

$$\min_{\theta} \; \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\ell(f_\theta(X), Y)\right]. \tag{1}$$

A central requirement is that f_θ remains robust to arbitrary schema variations in x_i^t, enabling zero-shot transfer across unseen datasets without explicit feature harmonization.

3.2 LLM-Powered Schema-Adaptive Tabular Encoder

The schema-adaptive encoder serves as the linguistic core of our framework. It transforms each column–value pair of a table into a structured textual statement, thereby grounding numerical and categorical features in the semantic space of an LLM. Formally, for a tabular input x_i^t, we process categorical (x_c) and numerical (x_n) attributes as follows.

$x_c \in$ **Categorical Features:** For a column c with description D_c and value v_c^i:

1. *Description Refinement*: A lightweight rewrite function L augments metadata to form human-readable context, e.g., L("SEX") $\rightarrow$ "Gender of the subject:".

2. *Statement Construction*: Construct a natural language statement $S_c^i = L(D_c) \parallel v_c^i$, where $\parallel$ denotes string concatenation, e.g., "Gender of the subject: Female".

3. *Semantic Embedding*: Encode S_c^i using a pretrained LLM embedding model to obtain $\boldsymbol{E}_c^i = \mathrm{Emb}(S_c^i) \in \mathbb{R}^d$.

This process, a form of *semantic tokenization*, maps schema-specific categorical attributes into a shared language-driven latent space.

$x_n \in$ **Numerical Features:** For a column n with description D_n and value v_n^i:

1. *Normalization*: Normalize the raw value as $\tilde{v}_n^i = 1 + \frac{v_n^i - \mu_n}{R_n}$, where μ_n and R_n denote the mean and range of the feature.

2. *Description Embedding*: Embed the refined column description using the same LLM encoder: $\boldsymbol{E}_D = \mathrm{Emb}(L(D_n)) \in \mathbb{R}^d$.

3. *Value-Weighted Embedding*: Combine magnitude and semantics through element-wise scaling, yielding $\boldsymbol{E}_n^i = \tilde{v}_n^i \cdot \boldsymbol{E}_D \in \mathbb{R}^d$.

All embeddings $\{\boldsymbol{E}_c^i, \boldsymbol{E}_n^i\}$ are projected through a shared linear layer to dimension $d_{\text{model}} = 256$, forming a unified token sequence T_{tab}^i. This encoder effectively decouples model performance from schema syntax, enabling generalization through semantic abstraction.

3.3 Auxiliary Modality Encoder

To assess the generality of our schema-adaptive design in multimodal settings, we introduce a secondary encoder for image data. Given an MRI volume $x_i^m \in \mathbb{R}^{128\times128\times128}$ (after standard preprocessing such as skull-stripping and intensity normalization), we employ a frozen Swin UNETR backbone (Hatamizadeh et al., 2021) to extract latent representations $\mathbf{Z} \in \mathbb{R}^{768\times4\times4\times4}$. Swin UNETR integrates hierarchical attention and U-Net structures to capture both global and local anatomical context. We freeze its pretrained weights to preserve general visual semantics while avoiding overfitting. A lightweight projection block comprising four 1D convolutional layers maps $\mathbf{Z}$ into an image token sequence $T_{\text{img}}^i \in \mathbb{R}^{d_{\text{model}}}$. Although the imaging modality is domain-specific, it provides a stringent testbed for evaluating the adaptability of language-grounded tabular embeddings in cross-modal reasoning.

3.4 Multimodal Fusion Backbone

The token sequences from tabular (T_{tab}) and image (T_{img}) modalities are concatenated along with L learnable `[CLS]` tokens—one per label—and passed into a gated transformer. This *mid-level fusion* design enables deep inter-modality interaction while maintaining representational independence. Each `[CLS]` token $T_k \in \mathbb{R}^{d_{\text{model}}}$, $k \in [1, L]$, acts as a dedicated classifier query. After transformer propagation, the contextualized state of each token is used to produce the corresponding label prediction. This architecture naturally extends standard text–image transformers (e.g., CLIP) into a text–table–image tri-modal context, where tabular information is expressed in linguistic form.

3.5 Training Objective for Robust Evaluation

To manage interdependent labels and severe class imbalance, we formulate training as a multi-objective optimization problem following (Sener and Koltun, 2018). The total objective comprises $2L$ components—L Focal losses and L contrastive losses—balanced by the Multiple Gradient Descent Algorithm (MGDA). MGDA dynamically reweights gradients across objectives, ensuring that high-magnitude tasks do not dominate optimization. Formally, we solve:

$$\min_{\alpha^1,\ldots,\alpha^{2L}} \left\| \sum_{t=1}^{2L} \alpha^t \nabla_{\theta^{sh}} \mathcal{L}_t \right\|_2^2 \quad \text{s.t.} \quad \sum_{t=1}^{2L} \alpha^t = 1,$$

where θ^{sh} are shared parameters of the fusion backbone, $\mathcal{L}_t$ denotes individual loss terms (L focal losses and L contrastive losses), and α^t are dynamically adjusted weights assigned by MGDA to balance gradient magnitudes across objectives.

Focal Loss. To counteract label imbalance, we adopt the focal loss (Lin et al., 2017):

$$\begin{aligned}\ell_{\text{focal}}(p_{i,k}, y_{i,k}) = & -\alpha_k [y_{i,k}(1-p_{i,k})^\gamma \log(p_{i,k}) \\ & + (1-y_{i,k}) p_{i,k}^\gamma \log(1-p_{i,k})],\end{aligned} \tag{2}$$

where $p_{i,k}$ denotes the predicted probability, γ the focusing parameter, and α_k a class-balancing weight.

Multi-Label Contrastive Loss. To further regularize the representation space, we employ a multi-label extension of supervised contrastive learning (Khosla et al., 2020), augmented

with the hardness-aware dual-temperature mechanism (Zhang et al., 2022). For an anchor representation r_i^k corresponding to label k, the contrastive objective is:

$$\mathcal{L}_{r_i^k} = -\text{sg}\left(\frac{W_\beta^i}{W_\alpha^i}\right) \log\left(\frac{\sum_{j=1}^{N} \mathbb{I}_{[y_{jk}=y_{ik}]} \exp\left(\frac{r_i^k \cdot r_j^k}{\tau_\alpha}\right)}{\sum_{j=1}^{N} \exp\left(\frac{r_i^k \cdot r_j^k}{\tau_\alpha}\right)} \right), \quad (3)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator, τ_α the temperature, and W_α, W_β are hardness coefficients. This contrastive formulation encourages semantic clustering of samples sharing similar label semantics, reinforcing alignment between schema-derived language embeddings and target-level supervision.

4 Experiments

We design our experiments to examine how language-grounded representations derived from LLMs enable schema-level generalization and multimodal consistency in structured data modeling. Specifically, we evaluate our framework along three research questions (RQs) that probe complementary aspects of its representational capabilities:

- **RQ1: Schema-Level Generalization.** Can the proposed schema-adaptive encoder transfer across entirely unseen tabular schemas without fine-tuning, demonstrating genuine zero-shot representational alignment?
- **RQ2: Multimodal Consistency and In-Domain Robustness.** Within a single, complex schema, does linguistic grounding improve in-domain discriminability, and how does the integration of auxiliary modalities (e.g., imaging) contribute to semantic consistency?
- **RQ3: Representation Efficiency and Interpretability.** How efficiently can the learned schema-agnostic representations adapt to new domains with limited data, and do the resulting embeddings exhibit interpretable, clinically coherent semantics?

4.1 Experimental Setup

Datasets for Schema Heterogeneity Evaluation. To evaluate schema-level generalization, we use two large-scale dementia datasets with distinct feature schemas. The **National Alzheimer's Coordinating Center (NACC)** dataset (Beekly et al.) serves as our training and in-domain benchmark, containing over 200,000 visit records from 44,656 subjects with 390 heterogeneous features and annotations for 13 dementia etiologies (e.g., AD, LBD, VD); a subset provides MRI scans for 308 subjects. The **Alzheimer's Disease Neuroimaging Initiative (ADNI)** dataset (Mueller et al., 2005) is reserved solely for zero-shot evaluation, comprising approximately 20,000 records from 3,392 subjects with a distinct schema of 110 features, and is strictly excluded from all training or validation phases to ensure unbiased unseen-schema testing. (Details are provided in Appendix A)

Task Definition. We formulate a 13-label dementia etiology prediction task as a high-dimensional, imbalanced, and multimodal classification challenge. Rather than focusing on clinical outcome, this serves as a stress test for schema-level and modality-level generalization under realistic heterogeneity.

Evaluation Metrics. We report macro-averaged AUROC as the principal metric, alongside Balanced Accuracy, AUC-PR, and macro F1 Score. These jointly measure discriminative ability, robustness to imbalance, and overall predictive consistency across multiple labels.

Implementation Details. Models are trained for 256 epochs using AdamW (lr=0.003) with cosine annealing. The schema-adaptive encoder employs `text-embedding-3-large` for semantic representations. While we use OpenAI's model for primary experiments, our modular framework remains effective with open-source alternatives, ensuring applicability in privacy-sensitive environments (see Appendix D). All experiments run on 4×NVIDIA A40 GPUs under identical computational budgets across baselines to ensure fair comparison. Data was split using official partitions or an 80/10/10 random split stratified by patient ID.

Baselines. We benchmark against four representative baselines that capture different perspectives of generalization: (1) **Human Experts (Neurologists)**—a panel of 12 board-certified neurologists independently assessed 100 multimodal cases. This serves as a real-world clinical benchmark for in-domain diagnostic accuracy. (2) **TabPFN** (Hollmann et al., 2025)—a schema-dependent Trans-

Model	AD	MCI	Avg-All
No-LLM (Rand Emb.)	0.512	0.508	0.513
No-LLM (Pret Emb.)	0.625	0.611	0.611
Ours (Schema-Adap.)	**0.789**	**0.765**	**0.727**

Table 1: Zero-shot AUROC on unseen ADNI. "Avg-All" = macro-averaged AUROC over 13 diagnostic labels. Our schema-adaptive model outperforms non-semantic baselines.

former for tabular data, providing a strong syntactic baseline; (3) **Gemini-2.5** (Comanici et al., 2025)—a large-scale general-purpose multimodal LLM used to evaluate cross-domain reasoning capacity; and (4) **LLaVA-Med** (Li et al., 2023)—a domain-adapted medical vision-language model that tests whether specialized finetuning can compensate for schema variance.

4.2 Results and Analysis

We now present the empirical results that rigorously evaluate our central hypothesis: language-grounded representations derived from LLMs enable robust schema-level generalization, in-domain multimodal consistency, and interpretable semantic alignment. Results are organized by the three research questions introduced earlier.

4.2.1 RQ1: Validating Zero-Shot Schema Generalization

The first question investigates whether our LLM-powered encoder can bridge the semantic gap between heterogeneous tabular schemas by grounding structured variables in a shared linguistic space. We perform a strict zero-shot cross-dataset evaluation: models are trained exclusively on the NACC dataset and directly tested on the unseen ADNI schema without any fine-tuning or exposure to its feature names or distributions. This setting isolates the model's intrinsic ability to perform schema-level transfer through semantic abstraction.

We compare our full model against two ablations that explicitly remove linguistic grounding: (1) **No-LLM (Random Emb.)**, which replaces our encoder with randomly initialized embeddings—representing a purely syntactic baseline dependent on schema-specific structure; and (2) **No-LLM (Pretrained Emb.)**, which uses a pretrained sentence transformer to embed only raw column names but lacks contextual understanding of feature values.

As summarized in Table 1, both non-LLM baselines collapse when evaluated on the unseen ADNI schema, with performance barely above random, confirming their brittleness to schema shifts. In contrast, our schema-adaptive model achieves an average AUROC of 0.727, a substantial margin demonstrating its capacity to align heterogeneous feature semantics. For instance, the model successfully maps *"MMSE Total Score"* from NACC and *"MMSCORE"* from ADNI into a coherent semantic embedding. This finding provides strong empirical evidence that linguistic tokenization allows the model to perform cross-schema translation, enabling zero-shot reasoning across disparate tabular structures—a core capability long considered unattainable for standard tabular models. Furthermore, our framework demonstrates consistent bidirectional zero-shot generalization (ADNI $\rightarrow$ NACC), confirming it learns intrinsic transferable representations rather than dataset-specific artifacts (see Appendix C).

4.2.2 RQ2: In-Domain Performance and Multimodal Consistency

Having established zero-shot transfer, we next evaluate whether the proposed schema-adaptive encoder maintains strong discriminative performance within a single domain and how linguistic grounding influences multimodal representation quality. The evaluation uses the NACC dataset as an in-domain benchmark, comparing our model against human experts and state-of-the-art AI baselines.

Comparison with Human Experts. To anchor performance in a real-world context, we benchmark our model against the diagnostic decisions of 12 board-certified neurologists. As shown in Table 2, our model achieves a macro-averaged AUROC of 0.904, representing a strong 32.9% relative improvement over the human experts' average of 0.680. The advantage is particularly pronounced in complex etiologies with ambiguous symptoms—e.g., Systemic and Endocrine Factors (SEF), where the model (AUROC 0.771) significantly exceeds the expert baseline (0.517). This demonstrates that language-grounded embeddings enable integrative reasoning over subtle, high-dimensional patterns that are difficult for human clinicians to aggregate consistently.

Comparison with AI Baselines. Table 3 compares our framework to state-of-the-art tabular and multimodal models. Our schema-adaptive architecture consistently outperforms all baselines, achiev-

		NC	MCI	DE	AD	LBD	VD	PRD
Neurologist	AUROC	0.930	0.699	0.914	0.761	0.833	0.613	0.517
Ours	AUROC	0.972	0.908	0.984	0.921	0.965	0.910	0.980
		FTD	NPH	SEF	PSY	TBI	ODE	**Avg**
Neurologist	AUROC	0.708	0.719	0.517	0.613	0.497	0.516	0.680
Ours	AUROC	0.957	0.876	0.771	0.846	0.889	0.779	**0.904**

Table 2: In-domain diagnostic performance on the NACC dataset, measured by AUROC across 13 neurological conditions. Our schema-adaptive model consistently outperforms expert neurologists, especially in less prevalent or complex conditions (e.g., SEF, ODE). "Avg" denotes the macro-average AUROC across all 13 labels, reflecting the overall diagnostic robustness of the model. (Detailed metrics with standard deviations are provided in Appendix B)

Model	AUROC	Bal Acc	AUC(PR)	F1
TabPFN	0.868	0.683	0.505	0.439
Gemini-2.5	-	0.663	-	-
LLaVA-Med	-	0.589	-	-
Ours	**0.904**	**0.785**	**0.533**	**0.476**

Table 3: In-domain performance (NACC dataset). AUROC, Bal Acc, AUC(PR), and F1 are evaluated. TabPFN/ours output probabilities (full metrics); Gemini-2.5/LLaVA-Med only output binaries (- = not applicable).

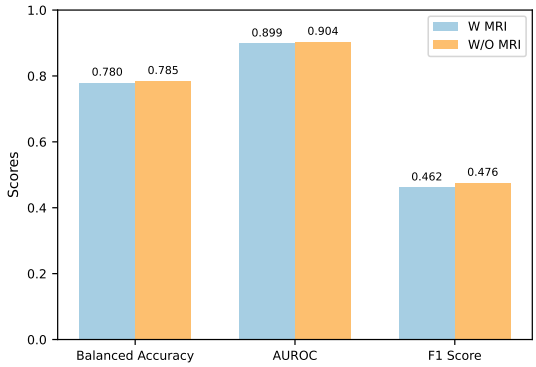

Figure 2: Comparison between table-only and multimodal variants.

ing 0.904 AUROC versus 0.868 for TabPFN, confirming that linguistic grounding not only supports schema transfer but also improves in-domain representation quality. The model's Balanced Accuracy (0.785) also exceeds large-scale general-purpose multimodal LLMs such as Gemini-2.5 (0.663) and LLaVA-Med (0.589), illustrating that scale alone cannot compensate for semantic misalignment in structured data. Together, these results show that our approach achieves both precision and interpretive stability within complex multimodal settings, establishing a new standard for schema-adaptive tabular reasoning. Critically, for clinical decision support, our model exhibits significantly superior calibration (ECE=0.042) compared to generalist LLMs like Gemini-2.5, avoiding catastrophic overconfidence (see Appendix E).

4.2.3 RQ3: Representation Efficiency and Interpretability

Low-Resource Adaptation and Sample Efficiency. An essential theoretical advantage of schema-grounded representations is efficient adaptation under data scarcity. We fine-tune the NACC-pretrained encoder on progressively smaller subsets of the ADNI dataset to assess how well linguistic priors accelerate convergence and transfer. As shown in Table 4, even with only 300 samples, our fine-tuned model achieves an AUROC of 0.9362, surpassing both a model trained from scratch on ADNI (0.7206) and even one trained on the full dataset (0.8943). This indicates that schema-level linguistic alignment transfers semantic knowledge compactly, allowing the model to achieve near-optimal performance with minimal labeled data—a highly desirable property for medical and scientific domains with limited annotations.

Train Samples	NACC-FewShot	ADNI-Train
30	**0.7389**	0.6982
100	**0.7561**	0.7176
300	**0.9362**	0.7206
1000	**0.9520**	0.8943
2713	**0.9532**	0.9320

Table 4: Few-shot generalization (AUROC) on ADNI dataset. NACC-FewShot: fine-tuned schema-grounded model; ADNI-Train: model trained from scratch.

In-Depth Analysis: The Role of Multimodality. A nuanced trend emerges in multimodal integration. While adding MRI data yields only marginal

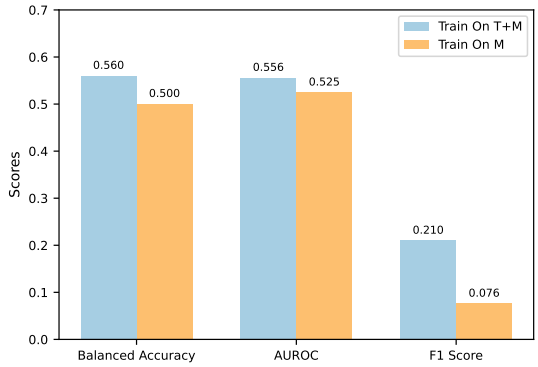

Figure 3: Image-only vs. multimodal training performance.

Metric	Transformer Depth			Projection Architecture	
	1 Layer	2 Layers (Ours)	3 Layers	Single Linear (Ours)	2-Layer MLP
F1 Score	0.4170	**0.4757**	0.3959	**0.4757**	0.4133
Bal Acc	0.7423	**0.7853**	0.7072	**0.7853**	0.7435
AUROC	0.8672	**0.9044**	0.8537	**0.9044**	0.8691
AUC(PR)	0.4793	**0.5334**	0.4553	**0.5334**	0.4799

Table 5: Ablation studies on core architectural components: (left) transformer encoder depths, and (right) projection layer topologies.

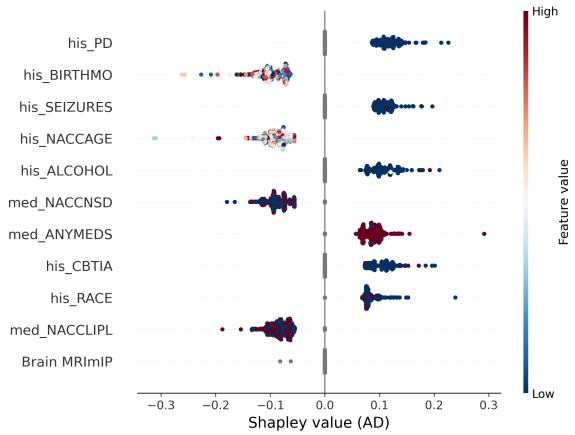

Figure 4: SHAP summary plot for AD prediction showing the top 10 features and an MRI-derived feature. Each point represents a patient; the x-axis indicates Shapley value impact and color denotes feature value (blue=low, red=high).

overall improvement (Figure 2), it plays a critical regularization role. As shown in Figure 3, models trained on MRI alone overfit severely, whereas co-training with semantically grounded tabular representations constrains the image encoder to align with language-informed feature semantics. This demonstrates that linguistic priors act as a stabilizing anchor for weak or data-scarce modalities, reinforcing the hypothesis that language-based abstraction can unify heterogeneous signals within a coherent representational manifold.

Architectural Ablations. We further examine the sensitivity of architectural design choices. Table 5 shows that a 2-layer Transformer achieves the best trade-off between model capacity and generalization, outperforming both shallower and deeper variants. Similarly, Table 5 reveals that a single linear projection outperforms a deeper MLP, suggesting that our LLM-based encoder already provides a structured and semantically meaningful latent space. These findings confirm that the observed performance gains originate from the language-driven representation itself rather than architectural complexity.

4.2.4 Interpretability: What Does the Model Learn?

Finally, to verify that the learned embeddings reflect meaningful, human-interpretable semantics, we employ SHAP (Lundberg and Lee, 2017) analysis for Alzheimer's Disease (AD) prediction. Figure 4 shows that the model highlights clinically coherent predictors such as seizure history (`his_SEIZURES`) as positive indicators and Parkinson's history (`his_PD`) as a negative predictor. Notably, MRI-derived features are absent from the top ten contributors, whereas aggregated medical and medication histories dominate. For this specific task and dataset, the model's predictions are primarily driven by clinical history variables—a finding that aligns with domain knowledge, though the role of imaging may become more prominent with larger multimodal cohorts or fine-tuned encoders. Overall, our schema-adaptive LLM encoder yields interpretable representations grounded in semantic clinical reasoning.

5 Conclusion

We presented Schema-Adaptive Tabular Representation Learning, a framework that addresses schema heterogeneity by transforming structured variables into semantically grounded natural language representations. By leveraging pretrained language embeddings, the proposed method enables cross-schema alignment without manual feature harmonization or retraining. Experiments on multimodal dementia diagnosis demonstrate consistently strong and reliable zero-shot transfer across heterogeneous real-world clinical cohorts, suggesting that language-driven semantic representations therefore offer a practical direction for further improving broader generalization in structured clinical learning settings.

Limitations

Our framework relies on the availability of semantically meaningful column metadata for cross-schema alignment. Although the model remains relatively robust under moderate header corruption (Appendix F), performance degrades substantially when semantic identifiers are entirely removed.

In addition, while our experiments include both proprietary and open-source embedding models (Appendix D), we do not perform a comprehensive analysis of embedding-related bias, privacy, or domain shift effects across different pretrained encoders.

Finally, our empirical evaluation is limited to dementia-related clinical cohorts. Although the proposed framework shows strong cross-schema transfer in this setting, its effectiveness on structured data from other domains remains to be validated.

References

Alexandre Audibert, Aurélien Gauffre, and Massih-Reza Amini. 2024. Multi-label contrastive learning: A comprehensive study. *arXiv preprint arXiv:2412.00101*.

Junwen Bai, Shufeng Kong, and Carla P Gomes. 2022. Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In *international conference on machine learning*, pages 1383–1398. PMLR.

Duane L Beekly, Erin M Ramos, Gerald van Belle, Woodrow Deitrich, Amber D Clark, Mary E Jacka, and Walter A Kukull. The national alzheimer's coordinating center (nacc) database: An alzheimer disease database. *Alzheimer Disease & Associated Disorders*, 18(4):270–277.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Siyi Du, Shaoming Zheng, Yinsong Wang, Wenjia Bai, Declan P. O'Regan, and Chen Qin. 2025. TIP: Tabular-Image Pre-training for Multimodal Classification with Incomplete Data. In *Computer Vision – ECCV 2024*, pages 478–496, Cham. Springer Nature Switzerland.

Randall J. Ellis, Ryan M. Sander, and Alfonso Limon. 2022. Twelve key challenges in medical machine learning and solutions. *Intelligence-Based Medicine*, 6:100068.

Rohit Gupta, Anirban Roy, Claire Christensen, Sujeong Kim, Sarah Gerard, Madeline Cincebeaux, Ajay Divakaran, Todd Grindal, and Mubarak Shah. 2023. Class prototypes based contrastive learning for classifying multi-label and fine-grained educational videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19923–19933.

Paul Hager, Martin J Menten, and Daniel Rueckert. 2023. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23924–23935.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics*, pages 5549–5581. PMLR.

Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly Supervised Table Parsing via Pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

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

Weichen Huang. 2023. Multimodal contrastive learning and tabular attention for automated alzheimer's disease prediction. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2465–2474.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Boshko Koloski, Andrei Margeloiu, Xiangjian Jiang, Blaž Škrlj, Nikola Simidjievski, and Mateja Jamnik. 2025. Llm embeddings for deep learning on tabular data. *arXiv preprint arXiv:2502.11596*.

Dongha Lee, Xiaoqian Jiang, and Hwanjo Yu. 2020. Harmonized representation learning on dynamic ehr graphs. *Journal of biomedical informatics*, 106:103426.

Simon A Lee, Sujay Jain, Alex Chen, Kyoka Ono, Arabdha Biswas, Ákos Rudas, Jennifer Fang, and Jeffrey N Chiang. 2025. Clinical decision support using pseudo-notes from multiple streams of ehr data. *npj Digital Medicine*, 8(1):394.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Nankai Lin, Guanqiu Qin, Gang Wang, Dong Zhou, and Aimin Yang. 2023. An effective deployment of contrastive learning in multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8730–8744.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Guang Liu, Jie Yang, and Ledell Wu. 2022. Ptab: Using the pre-trained language model for modeling tabular data. *arXiv preprint arXiv:2209.08060*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246.

Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. 2005. Ways toward an early diagnosis in alzheimer's disease: The alzheimer's disease neuroimaging initiative (adni). *Alzheimer's Dementia*, 1(1):55–66.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. 2019. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358.

Yi Ren, Chenglong Yu, Weibin Li, Wei Li, Zixuan Zhu, TianYi Zhang, ChenHao Qin, WenBo Ji, and Jianjun Zhang. 2025. TableGPT: a novel table understanding method based on table recognition and large language model collaborative enhancement. *Applied Intelligence*, 55(5):311.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, Haoxuan Lan, Jiaming Tian, Jing Yuan, Junbo Zhao, Junlin Zhou, Kaizhe Shou, Liangyu Zha, Lin Long, and 14 others. 2024. TableGPT2: A Large Multimodal Model with Tabular Data Integration. *arXiv preprint*. ArXiv:2411.02059 [cs].

Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56.

Luís A Vale-Silva and Karl Rohr. 2021. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1):13505.

Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Bing Yin, and Yan Zhang. 2021. TUTA: Tree-based Transformers for Generally Structured Table Pretraining. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 1780–1790, New York, NY, USA. Association for Computing Machinery.

Zifeng Wang and Jimeng Sun. 2022. TransTab: Learning Transferable Tabular Transformers Across Tables. *Advances in Neural Information Processing Systems*, 35:2902–2915.

Elisa Warner, Joonsang Lee, William Hsu, Tanveer Syeda-Mahmood, Charles E Kahn Jr, Olivier Gevaert, and Arvind Rao. 2024. Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects. *International Journal of Computer Vision*, 132(9):3753–3769.

Siyue Yang, Paul Varghese, Ellen Stephenson, Karen Tu, and Jessica Gronsbell. 2023. Machine learning approaches for electronic health records phenotyping: a methodical review. *Journal of the American Medical Informatics Association*, 30(2):367–381.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. 2022.

Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *Preprint*, arXiv:2203.03540.

Chaoning Zhang, Kang Zhang, Trung X Pham, Axi Niu, Zhinan Qiao, Chang D Yoo, and In So Kweon. 2022. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14441–14450.

Pingyue Zhang and Mengyue Wu. 2024. Multi-label supervised contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16786–16793.

Mingyu Zheng, Zhifan Feng, Jia Wang, Lanrui Wang, Zheng Lin, Yang Hao, and Weiping Wang. 2025. Tabledreamer: Progressive and weakness-guided data synthesis from scratch for table instruction tuning. *arXiv preprint arXiv:2506.08646*.

Weihai Zhi, Jiayan Guo, and Shangyang Li. 2026. Medgr2: Breaking the data barrier for medical reasoning via generative reward learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(34):28901–28909.

A Schema Heterogeneity Quantification

To rigorously demonstrate the severity of the cross-schema generalization challenge, we provide a quantitative comparison of the NACC and ADNI schemas in Table 6. Following initial quality control and the removal of irrelevant administrative variables, the curated feature spaces of NACC and ADNI consist of 390 and 110 clinical features, respectively. Despite both repositories targeting the same dementia research domain, their remaining feature spaces exhibit substantial divergence.

Schema Heterogeneity Metric	Value
Approximate usable variables in NACC	∼390
Approximate usable variables in ADNI	∼110
Manually aligned shared clinical concepts	49
Identical feature names across datasets	11
Semantically matched but differently named variables	38
Shared variables requiring unit/code conversion	27
Variables unique to NACC	∼340
Variables unique to ADNI	∼60

Table 6: Quantitative overview of schema heterogeneity between the curated NACC and ADNI cohorts after preprocessing. Despite targeting the same dementia diagnosis domain, the two datasets exhibit substantial divergence in feature naming conventions, coding systems, and variable coverage.

B Human Expert Evaluation Protocol

To establish a robust clinical baseline, we recruited 12 board-certified neurologists (average clinical experience: 8.2 years, range: 5–15 years) to independently evaluate 100 multimodal cases.

B.1 Case Selection

The 100 cases were selected via stratified random sampling from the held-out NACC test set, ensuring balanced representation across all 13 diagnostic categories. The exact sample sizes per evaluation label are: NC=12, MCI=12, DE=10, AD=10, LBD=8, VD=6, PRD=6, FTD=6, NPH=6, SEF=6, PSY=6, TBI=6, ODE=6.

B.2 Evaluation Procedure

- Neurologists were strictly blinded to model predictions and patient identifiers.
- Each expert reviewed the identical tabular and MRI modalities provided to the model.
- No consensus discussion was permitted; all diagnostic assessments were performed independently.
- The study exclusively used publicly available de-identified datasets under their respective data use agreements.

B.3 Statistical Analysis

- **Inter-rater agreement:** The inter-rater reliability among the 12 experts yielded a Fleiss' kappa of 0.52, indicating moderate agreement. This is highly consistent with previously reported clinical variability in multi-class dementia differential diagnosis.
- **Statistical significance:** Based on Delong's test for AUROC comparisons, our model achieved higher AUROC scores than the average neurologist performance in 11 out of 13 diagnostic categories ($p < 0.05$, Delong's test).
- **Per-label performance:** The full per-label AUROC with standard deviations (denoted as $\pm$ SD) for both the model and the human baseline is detailed in Table 7.

		NC	MCI	DE	AD	LBD	VD	PRD
Neurologist	AUROC	0.930 ± 0.021	0.699 ± 0.035	0.914 ± 0.018	0.761 ± 0.029	0.833 ± 0.025	0.613 ± 0.042	0.517 ± 0.051
Ours	AUROC	$\mathbf{0.972 \pm 0.007}$	$\mathbf{0.908 \pm 0.012}$	$\mathbf{0.984 \pm 0.005}$	$\mathbf{0.921 \pm 0.009}$	$\mathbf{0.965 \pm 0.006}$	$\mathbf{0.910 \pm 0.015}$	$\mathbf{0.980 \pm 0.004}$
		FTD	NPH	SEF	PSY	TBI	ODE	**Avg**
Neurologist	AUROC	0.708 ± 0.038	0.719 ± 0.036	0.517 ± 0.041	0.613 ± 0.043	0.497 ± 0.045	0.516 ± 0.042	0.680 ± 0.023
Ours	AUROC	$\mathbf{0.957 \pm 0.008}$	$\mathbf{0.876 \pm 0.018}$	$\mathbf{0.771 \pm 0.032}$	$\mathbf{0.846 \pm 0.021}$	$\mathbf{0.889 \pm 0.017}$	$\mathbf{0.779 \pm 0.028}$	$\mathbf{0.904 \pm 0.008}$

Table 7: In-domain diagnostic performance on the NACC dataset, measured by AUROC across 13 neurological conditions with standard deviations ($\pm$ SD). Our schema-adaptive model consistently outperforms expert neurologists, especially in less prevalent or complex conditions (e.g., SEF, ODE). "Avg" denotes the macro-average AUROC across all 13 labels, reflecting the overall diagnostic robustness of the model.

Model	NACC $\rightarrow$ ADNI	ADNI $\rightarrow$ NACC	Average
No-LLM (Random Emb.)	0.513 ± 0.021	0.489 ± 0.025	0.502 ± 0.023
No-LLM (Pretrained Emb.)	0.611 ± 0.018	0.572 ± 0.024	0.592 ± 0.021
Ours (Schema-Adaptive)	$\mathbf{0.727 \pm 0.015}$	$\mathbf{0.661 \pm 0.023}$	$\mathbf{0.694 \pm 0.019}$

Table 8: Bidirectional zero-shot transfer performance under cross-schema evaluation. Results are reported as macro-AUROC averaged across five runs.

C Bidirectional Zero-Shot Transfer

To rule out dataset-specific overfitting and rigorously test cross-schema robustness, we conducted bidirectional zero-shot transfer experiments: training on NACC and testing directly on ADNI, and vice versa. As demonstrated in Table 8, our schema-adaptive model achieves good performance in both transfer directions, whereas baselines lacking explicit semantic alignment show substantial performance degradation.

These results suggest that the proposed framework learns transferable clinical representations that generalize across heterogeneous schemas, rather than relying solely on dataset-specific structural regularities.

D Embedding Model Choice Ablation

Although our primary experiments leverage OpenAI's `text-embedding-3-large`, the proposed framework is fully modular and inherently compatible with any pretrained text embedding architecture. To evaluate this structural flexibility, we benchmarked three state-of-the-art open-source alternatives across the `NACC` $\rightarrow$ `ADNI` zero-shot transfer task, as demonstrated in Table 9.

All open-source alternatives retain highly competitive cross-schema generalization capability. The relatively small performance gap between proprietary and open-source embedding models suggests that the primary benefit of the proposed framework arises from semantically grounded schema transformation rather than dependence on a specific embedding API. Crucially, this modularity facilitates seamless on-premise deployment in privacy-sensitive clinical environments where cloud-bound data transmission is strictly prohibited under institutional regulations.

E Statistical Significance and Calibration

To suppress optimization variance, all reported macro-averaged metrics represent the mean performance across 5 independent training runs initialized with distinct random seeds. The corresponding uncertainty estimates (denoted as standard deviations, $\pm$ SD) presented in our tables were computed via non-parametric bootstrap resampling on the evaluation sets.

E.1 Model Calibration

In clinical decision support, raw predictive accuracy is insufficient; a model's predicted probability must reliably reflect the true empirical likelihood of the diagnosis. We evaluated this property using the Expected Calibration Error (ECE) across 15 bins on the `NACC` test set, benchmarking our framework against both tabular experts and generalist foundation models:

- **Ours (Schema-Adaptive):** ECE $= 0.042$
- **TabPFN:** ECE $= 0.078$
- **Gemini-2.5:** ECE $= 0.112$

Our framework achieves a significantly lower ECE compared to all baselines. General-purpose

Embedding Model	Zero-shot ADNI AUROC	Relative Performance
OpenAI `text-embedding-3-large` (Proprietary)	0.727 ± 0.014	100.0%
Qwen-Text-Embedding-V2 (Open-source)	0.709 ± 0.022	97.5%
E5-large-v2 (Open-source)	0.694 ± 0.018	95.5%
BGE-M3 (Open-source)	0.701 ± 0.026	96.4%

Table 9: Ablation study on different embedding models, measured by macro-averaged AUROC with standard deviations ($\pm$ SD).

Header Corruption Condition	Zero-shot ADNI AUROC
Original headers	0.727 ± 0.015
Clinical abbreviations (e.g., MMSE, CDR)	0.714 ± 0.017
25% headers replaced with random strings	0.683 ± 0.021
50% headers replaced with random strings	0.641 ± 0.028
All headers removed (numerical indices only)	0.531 ± 0.031

Table 10: Robustness analysis under systematic schema metadata degradation. Results are reported as macro-AUROC averaged across five runs.

foundation models exhibit comparatively larger calibration errors in this setting, whereas the proposed framework maintains more reliable probabilistic confidence estimates.

F Header Quality Robustness Test

Our framework assumes the availability of semantically meaningful column identifiers for cross-schema alignment. To evaluate the sensitivity of the proposed method to schema metadata quality, we conduct a controlled header degradation experiment under the NACC $\rightarrow$ ADNI zero-shot transfer setting.

Specifically, we progressively corrupt column headers by replacing subsets of feature names with random strings or removing semantic identifiers entirely. The results are summarized in Table 10.

The proposed framework remains relatively stable under moderate schema corruption. In particular, replacing original feature names with common clinical abbreviations (e.g., MMSE, CDR) leads to only minor performance degradation, suggesting that the embedding encoder can effectively interpret widely used clinical terminology.

As schema corruption increases, performance gradually declines, indicating that semantic feature descriptors contribute substantially to cross-schema transfer. When all column names are removed and replaced with numerical indices, performance drops significantly, approaching the behavior of non-semantic baselines. This observation supports the hypothesis that the proposed schema-adaptive representation primarily benefits from semantically grounded feature alignment rather than memorizing dataset-specific structural patterns.