Rethink Representation Learning for Questionnaire Data

Guanhua Ye[1], Jifeng He[1], Yan Li[2], Junping Du[1], Xue Zhe[1], Yingxia Shao[1], Meiyu Liang[1], Yawen Li[2*]

[1]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China.
[2]School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing, China.
{g.ye, hejifeng, liyanly, junpingd, xuezhe, shaoyx, meiyu1210, warmly0716}@bupt.edu.cn

Abstract

Questionnaire data serve as a valuable resource across numerous scientific domains, offering insights into human behavior, health, and social trends. Traditional downsampling-based representation learning methods—such as standardization and one-hot encoding—reformat these data into tabular structures that inherently discard semantic richness and obscure inter-sample and inter-feature relationships. Consequently, advanced deep learning models often underperform compared to simpler approaches like gradient-boosted decision trees (GBDT), due to their limited capacity to extract meaningful representations from semantically sparse inputs. To address this limitation, we introduce SemantiQ, a novel upsampling-based representation learning framework that embeds questionnaire responses into a unified semantic space. Leveraging Retrieval-Augmented Generation (RAG) in conjunction with large language models (LLMs), SemantiQ transforms question text, option text, and external knowledge into semantically enriched natural language statements. These statements are then encoded into semantic embeddings, which are further refined through a three-stage training mechanism and test-time training (TTT), enabling the model to capture complex sample- and feature-wise dependencies. Extensive experiments on multiple real-world datasets demonstrate that SemantiQ consistently outperforms state-of-the-art baselines.

1 Introduction

Questionnaire data play a central role in empirical resear across domains such as mental health (McCabe et al. 20 Linden, Boyes, and Stuart 2021; Schick et al. 2023), edu tion (Bach et al. 2023; Khan and Ghosh 2021), and mar analysis (Zhao et al. 2023; Dolnicar 2022; Davis and Farrell 2016; Li et al. 2025), offering large-scale insights into subjective experience and behavior (Farber et al. 2021; Yuan et al. 2025; Ang, Yawen, and Xue 2026).

While deep learning offers potential for modeling such data (Jin et al. 2021; Chen et al. 2022; Wang et al. 2021; Chen et al. 2025), conventional methods typically treat questionnaire responses as generic tabular inputs, relying on preprocessing techniques like one-hot encoding and normalization (Fei et al. 2021) that discard semantic structure. As a result, deep models often underperform compared to simpler approaches like gradient-boosted decision trees (GBDTs) (Friedman 2001; Chen and Guestrin 2016), which remain competitive due to their robustness under this limited representation.

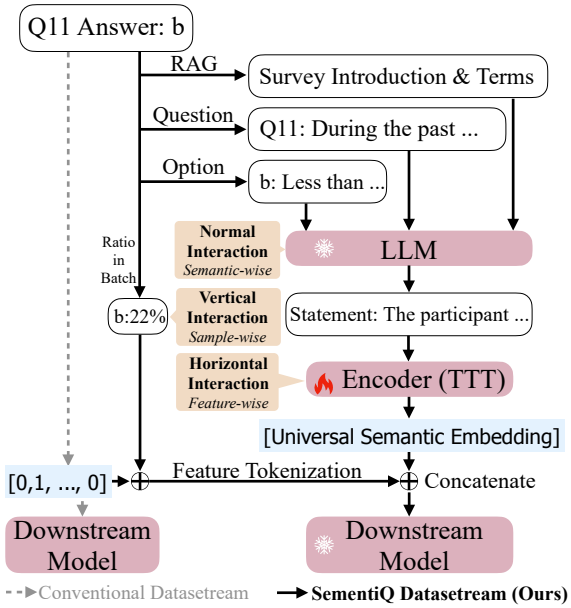

Figure 1: Overview of the SemantiQ. SemantiQ enhances representation through three interaction levels: *Normal* (semantic-wise, e.g., RAG and LLM enrichment), *Horizontal* (feature-wise, cross-feature modeling), and *Vertical* (sample-wise, batch statistics).

Recent work has realized the limitations of downsampling-based representation learning and attempted to mitigate these issues (Zhang et al. 2025)., yet the core idea remains constrained within the paradigm of "compressing redundancy" (Nguyen and Zeadally 2021). For instance, Zhang et al. (Zhang et al. 2020) unified categorical and numerical data by increasing the representation dimensions of numerical features. However, each feature's representation space remains isolated, relying on fixed feature order encoding and lacking a deep understanding of feature content. FT-Transformer (Gorishniy

*Corresponding author.

et al. 2021) addresses this by incorporating connections between different feature encodings, attempting to map one-hot encoded inputs to high-dimensional spaces through supervised training and fine-tuning with self-supervised methods. Nevertheless, this approach overlooks the fact that one-hot encoding inherently loses all semantic information, resulting in suboptimal performance when learning associations between multiple one-hot representations (Vaswani 2017). Another approach, TabNet (Arik and Pfister 2019), considers feature interactions but fails to achieve comprehensive semantic integration, as it does not fully exploit the semantic relationships between features.

To bridge this gap, we propose a novel *upsampling* strategy that introduces semantic redundancy into tabular data, enabling deep models to better capture complex feature associations. Inspired by the role of redundancy in language and vision tasks (Radford et al. 2019), we apply this idea to questionnaire data with a "question–option–response" format. Using question text, option distributions, and Retrieval-Augmented Generation (RAG) (Lewis et al. 2020), we generate declarative sentences that are encoded by large language models (e.g., GPT) into unified semantic embeddings (Long et al. 2025; Ren et al. 2024; Zhou et al. 2025). This framework, **SemantiQ**, enhances the representational depth of questionnaire inputs.

Unlike most LLM-based tabular approaches that operate at the row level, SemantiQ adopts a *cell-level* modeling paradigm, where each cell independently integrates question text, option semantics, and external knowledge before aggregation. This design preserves fine-grained structural semantics and allows heterogeneous features to be enriched individually. SemantiQ integrates three complementary information flows (Fig. 1): *semantic-wise* interactions for individual response semantics, *vertical* interactions for cross-sample feature patterns, and *horizontal* interactions for intra-sample dependencies via autoregressive TTT. A unified semantic embedding acts as a soft positional encoding, aligning semantically related options across questions. Benefiting from its cell-level design, SemantiQ invokes the LLM only once for the questionnaire schema, with embeddings reused for all responses, eliminating repeated high-cost calls.

In summary, our work makes the following contributions:

- We identify a key limitation of existing row-level and downsampling-based methods—the lack of semantic redundancy and explicit inter-feature interactions—and formalize how this limits deep model performance on questionnaire data.
- We introduce SemantiQ, an upsampling-based LLM-computation-friendly framework that incorporates the unified semantic embedding, facilitating the extraction of deeper, more complex semantic associations.
- We validate SemantiQ on diverse real-world datasets, demonstrating its effectiveness, flexibility, and robustness to feature order variations.

2 Related Work

Representation Learning for Tabular Data. Traditional methods for tabular data often rely on manually designed features or basic transformations such as one-hot or label encoding (Džeroski and Ženko 2004). Classical gradient boosting techniques, such as XGBoost (Chen and Guestrin 2016), LightGBM (Ke et al. 2017), and CatBoost (Dorogush, Ershov, and Gulin 2018), remain strong baselines for many industrial and academic applications. Recently, deep learning-based architectures have emerged for tabular data, including TabNet (Arik and Pfister 2021), SAINT (Somepalli et al. 2021), FT-Transformer (Gorishniy et al. 2021), and TabTransformer (Huang et al. 2020). These models, leveraging attention and gating mechanisms, capture complex feature interactions. However, they typically treat features in isolation and lack semantic understanding of feature content. To address this, TabPFN-v2 (Hollmann et al. 2025) uses synthetic pretraining and Bayesian inference to outperform GBDTs in low-data settings. In parallel, large language models (LLMs) have shown strong capabilities in modeling tabular data through natural language prompts. TabLLM (Hegselmann et al. 2023) reformulates rows as textual sequences and enables few-shot classification without task-specific architecture.

Semantic Augmentation and Data Expansion. The concept of enriching data with semantic or textual descriptors has been widely explored in NLP and information retrieval (IR) (Xu, Warin, and Robb 2020; Gururangan et al. 2020; Wang et al. 2025). Pretrained LLMs can generate synthetic text to enhance limited training data (Feng et al. 2021; Yu et al. 2018). In IR, for example, artificial query-document pairs are used to strengthen retrieval robustness (Nogueira, Lin, and Epistemic 2019; Ma et al. 2021). These methods primarily focus on unstructured textual data. Our work extends this idea to structured, form-based data, which lacks verbose textual context. By mapping each feature value to an LLM-generated semantic descriptor, we introduce a form of data expansion that moves beyond traditional one-hot or label encoding schemes, thereby enriching the feature space with meaningful textual augmentations.

Comparison with Existing LLM-based Tabular Models. Recent LLM-based approaches to tabular learning, such as TabLLM (Hegselmann et al. 2023), TaBERT (Padhi et al. 2021), TabPFN-v2 (Hollmann et al. 2025), and TabR (Gorishniy et al. 2023), typically serialize entire rows into natural language for prompt-based inference. While sometimes effective, these row-level methods treat heterogeneous features as a single sequence, limiting fine-grained structural modeling and applicability to datasets with minimal text. In contrast, SemantiQ models at the cell level, enriching each cell with question text, option text, and domain-specific terminology. This design captures subtle feature interactions, remains architecture-agnostic, and—by encoding the questionnaire structure only once—avoids repeated LLM calls for every sample, greatly reducing inference latency and computation cost.

3 Preliminary

In this section, the variables and mathematical formulations involved in the model are introduced, forming the foundation for understanding SemantiQ's approach to proceed

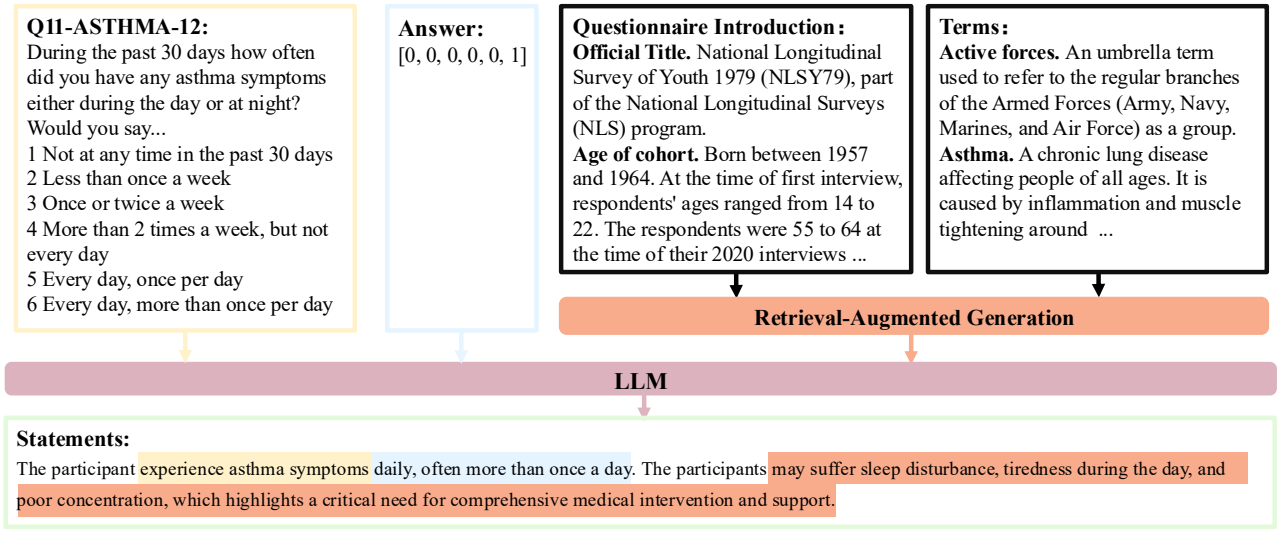

Figure 2: Visualization of the semantic statement generation workflow. Key components—derived from survey questions (yellow), participant responses (blue), and RAG-enhanced contextual knowledge (orange)—are color-coded for clarity.

Questionnaire analysis. Finally, we formulate the problem studied in this work.

3.1 Questionnaire Data

In this work, we consider a questionnaire consisting of N questions, denoted as $Q = \{q_1, q_2, \ldots, q_N\}$, with each question q_i having a set of options $O_i = \{o_{i1}, o_{i2}, \ldots, o_{im_i}\}$, where m_i is the number of options for question q_i. The set of respondents is denoted by $U = \{u_1, u_2, \ldots, u_K\}$, and each respondent u_k answers the questions, providing a sequence of responses $A_k = \{a_{k1}, a_{k2}, \ldots, a_{kN}\}$, where $a_{ki} \in O_i$ represents the answer to question q_i.

To capture the distribution of responses within a given batch, we define the batch statistics for each option o_{ij} (the j-th option of question q_i) as the proportion of respondents in the batch selecting option o_{ij}. This distribution is represented by D_i, which captures the statistical context for each question's response options within the current batch.

3.2 Semantic Enhancement

To enhance the semantic representation of questionnaire data, we adopt an *upsampling* strategy that introduces controlled semantic redundancy at the cell level. In this context, upsampling does not refer to increasing the number of numerical samples, but rather to expanding each original *Question-Response Pair* into semantically enriched statements. This allows the model to perceive heterogeneous feature attributes through multiple complementary semantic views.

Concretely, each *Question-Response Pair* $P_i = (q_i, o_{ij})$ is transformed into semantically rich statements $\{s_{i1}, s_{i2}, \ldots, s_{im_i}\}$ (Shown in Fig. 2). Each statement s_{ij} integrates the question q_i, the option o_{ij}, and relevant external knowledge $\mathcal{K}_i$ to create a more comprehensive representation of the data:

$$s_{ij} = \text{LLM}(q_i, o_{ij}, \mathcal{K}_i), \tag{1}$$

where LLM represents the large language model used to generate the statement. The resulting statement is then encoded into a latent representation d_{ij} using a semantic encoder E_d:

$$d_{ij} = \text{E}_d(s_{ij}). \tag{2}$$

This embedding d_{ij} captures the semantic content of the statement, enabling deeper analysis. To manage the high-dimensional nature of these embeddings, we apply an autoencoder-based technique to reduce their dimensionality while preserving essential semantic information.

3.3 Analysis Task

Each respondent u_k answers a series of questionnaire items. The feature representation of u_k is denoted as $X = \{x_1, x_2, \ldots, x_N\}$, where each $x_i = \{p_i, a_i, d_i\}$ represents a fused feature that integrates the positional encoding p_i, the answer a_i, and the corresponding statement embedding d_i. For simplicity, indices k and j are omitted.

We propose a model architecture consisting of an encoder E_f and two decoders: D_g and D_h, parameterized by θ_f, θ_g, and θ_h, respectively. The encoder E_f extracts feature embeddings, while the first decoder D_g performs self-supervised learning by reconstructing input features from these embeddings. The second decoder D_h produces the final prediction $\hat{y}$, optimized for downstream tasks such as classification or regression, depending on the application.

The objective function for the analysis task is:

$$\mathcal{L} = \frac{1}{K} \sum_{k=1}^{K} \ell\left(\text{D}_h(\text{E}_f(x_k; \theta_f); \theta_h), y_k\right), \tag{3}$$

where $\ell(\cdot)$ represents the loss function for the task. The parameters θ_f and θ_h are learned during training to mini-

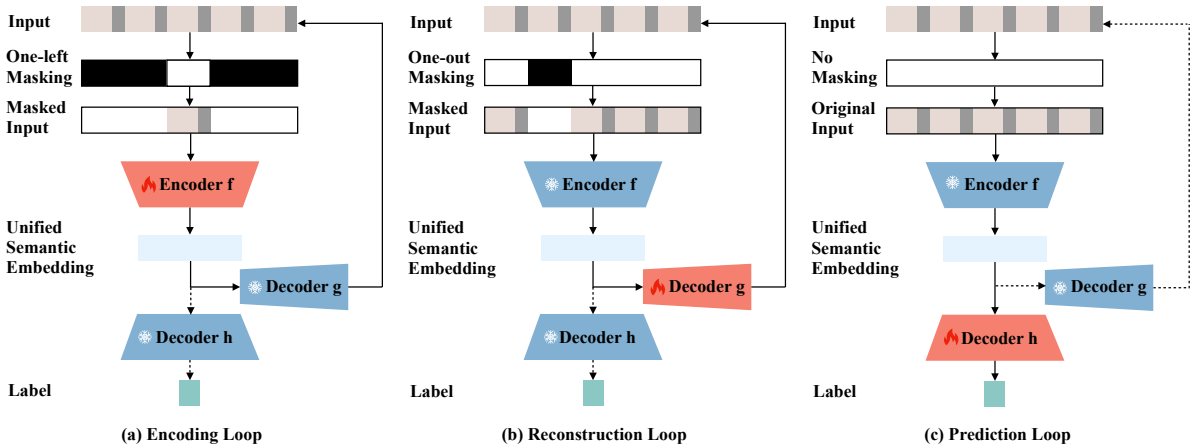

Figure 3: **Three training loops in SemantiQ.** (a) the Encoding Loop, where the encoder (represented by the flame symbol) learns feature representations from masked input, (b) the Reconstruction Loop, where the decoder (represented by the snowflake symbol, indicating frozen parameters) reconstructs the missing parts of the input, and (c) the Prediction Loop, where the model makes predictions based on the unified semantic embeddings.

mize the loss and improve prediction accuracy. The training process for θ_g is discussed in the subsection 3.4.

3.4 Model Architecture and Training Loops

SemantiQ consists of three core modules: an encoder $\mathbf{E}_f$ that learns semantically enriched embeddings at the *cell level*, a reconstruction decoder $\mathbf{D}_g$ for recovering masked questionnaire content, and a prediction decoder $\mathbf{D}_h$ for downstream tasks. Training proceeds through three complementary loops (Fig. 3), each targeting a distinct learning objective: semantic representation, information recovery, and task-specific optimization. Together, these loops enable the model to jointly capture fine-grained local semantics and global inter-question dependencies.

Encoding Loop. This loop trains the encoder to generate context-aware embeddings $d_k = \mathbf{E}_f(X_k^{\phi}; \theta_f)$ using a *one-left masking* strategy ϕ_{left}, where only one answer x_i is visible:

$$\phi_{\text{left}}(X) = ([\text{MASK}], \ldots, x_i, \ldots, [\text{MASK}]), \tag{4}$$

$$\min_{\theta_f} \frac{1}{2K} \sum_{k=1}^{K} \|\mathbf{D}_g(\mathbf{E}_f(\phi_{\text{left}}(X_k))) - X_k\|^2 \tag{5}$$

By forcing the encoder to infer global context from minimal visible input, this loop enhances the model's ability to capture latent dependencies between questions and align semantically related options across columns.

Reconstruction Loop. The second loop trains $\mathbf{D}_g$ with a *one-out masking* strategy ϕ_{out}, where all answers except one are visible:

$$\phi_{\text{out}}(X) = (x_1, \ldots, [\text{MASK}], \ldots, x_N), \tag{6}$$

$$\min_{\theta_g} \frac{1}{2K} \sum_{k=1}^{K} \|\mathbf{D}_g(\mathbf{E}_f(\phi_{\text{out}}(X_k))) - X_k\|^2 \tag{7}$$

This loop reinforces the model's ability to impute missing responses using contextual cues, improving robustness when questionnaire data is incomplete or partially corrupted.

	Training Phase			Testing Phase		
	E_f	D_g	D_h	E_f	D_g	D_h
Encoding Loop	✓	×	×	✓	×	×
Reconstruction Loop	×	✓	×	–	–	–
Prediction Loop	×	×	✓	×	×	×

Table 1: Parameter behavior comparison between training and testing phases. ✓ denotes learnable parameters; × indicates frozen parameters.

Prediction Loop. The third loop optimizes $\mathbf{D}_h$ for downstream classification or regression tasks using the full, unmasked input:

$$\hat{y}_k = \mathbf{D}_h(\mathbf{E}_f(X_k; \theta_f); \theta_h), \quad \mathcal{L}_{\text{pred}} = \frac{1}{K} \sum_{k=1}^{K} \ell(\hat{y}_k, y_k) \tag{8}$$

Here, θ_f is frozen to preserve previously learned semantic representations, ensuring that task-specific training does not overwrite general semantic knowledge.

Test-Time Training (TTT). At inference, SemantiQ can adapt to distribution shifts without full retraining by updating only $\mathbf{E}_f$ via the Encoding Loop, while $\mathbf{D}_g$ and $\mathbf{D}_h$ remain fixed:

$$e_k = \mathbf{E}_f(X_k; \theta_f'), \quad \hat{y}_k = \mathbf{D}_h(d_k; \theta_h) \tag{9}$$

This lightweight adaptation enables the model to maintain accuracy under domain changes with minimal computational overhead.

4 Experiments

In this secton, we try to answer the following Research Quesitons(RQs) by a series of experiments:

- **RQ1**: Does SemantiQ outperform existing methods on questionnaire datasets and generalize across different model architectures?

Model Type	Model	BRFSS		NLSY		MHP	
		ACC	F1	ACC	F1	ACC	F1
General DL	Mamba	66.59±0.48	66.54±0.42	56.00±0.30	56.72±0.27	74.53±0.55	75.60±0.32
	xLSTM	66.01±0.40	66.93±0.31	57.65±0.38	58.42±0.29	75.51±0.33	75.51±0.28
	Transformer	66.54±0.53	65.92±0.44	57.12±0.44	56.59±0.32	78.05±0.48	78.39±0.36
Tabular DL	TabNet	66.12±0.65	64.39±0.51	56.10±0.20	55.88±0.18	73.08±0.45	72.70±0.38
	DeepFM	67.05±0.32	66.91±0.23	56.09±0.14	55.78±0.23	75.02±0.25	75.28±0.21
	FT-T	67.54±0.46	67.28±0.32	63.50±0.31	63.31±0.26	94.01±0.39	94.33±0.34
	TabPFN-v2	69.45±0.18	68.67±0.44	62.81±0.08	62.07±0.50	95.89±0.11	96.24±0.13
GBDT	LightGBM	65.32±0.58	65.13±0.57	59.08±0.45	59.53±0.31	81.74±0.49	81.71±0.45
	XGBoost	65.02±0.50	65.08±0.42	59.04±0.64	59.84±0.65	82.05±0.55	81.89±0.48
	XGBoost*	66.20±0.38	65.90±0.36	59.40±0.52	59.42±0.53	82.04±0.37	81.78±0.30
LLM-based	TabLLM	67.64±0.37	67.01±0.33	61.44±0.60	60.72±0.62	92.83±0.45	92.81±0.40
	TaBERT	68.68±1.12	67.50±0.91	63.83±1.04	61.85±1.19	94.66±0.55	94.82±0.76
	SemantiQ-DT (Ours)	**69.42±0.53**	**68.29±0.32**	**63.77±0.29**	**62.57±0.21**	**90.23±0.57**	**88.93±0.55**
	SemantiQ (Ours)	**69.62±0.36**	**68.71±0.57**	**65.07±0.25**	**64.43±0.51**	**96.63±0.40**	**97.10±0.47**

Table 2: Model Performance on Different Questionnaire Datasets Using 5-fold Cross-validation (%).

- **RQ2**: How does SemantiQ's cell-level modeling compare with row-level LLM-based tabular models in terms of order sensitivity on questionnaire data?
- **RQ3**: Does SemantiQ exhibit advantages in few-shot scenarios?
- **RQ4**: To what extent does SemantiQ reduce computational overhead compared to existing LLM-based tabular learning methods?
- **RQ5**: To what extent do the redundancies introduced in statement length and embedding dimensionality impact final performance?
- **RQ6**: How do the semantic-wise, feature-wise, and sample-wise interactions, along with the proposed improved workflow for TTT, collectively contribute to overall model performance?
- **RQ7**: Is SemantiQ extensible to non-questionnaire datasets?

4.1 Datasets

We evaluate SemantiQ on three real-world questionnaire datasets (shown in Tab. 3), with semantic augmentation performed using Retrieval-Augmented Generation (RAG). When available, we incorporate introductory sections and glossary terms directly from the questionnaires.

BRFSS (for Disease Control and Prevention 2020): 401K health survey records; three-class classification of coronary heart disease status. **NLSY** (Martin et al. 2003): 8.9K longitudinal records; three-class prediction of self-assessed general health. **MHP** (Syeeda et al. 2024): 2K responses from university students; depression level classification based on psychological assessments.

4.2 Baseline

In this work, we compare the performance of our proposed method with a set of widely used baseline models spanning deep learning and gradient boosting approaches. All baselines are trained on structured tabular data without additional context or semantic augmentation, we reimplement each baseline model according to the original papers' architectures.

General DL Models: Mamba (Gu and Dao 2024), xLSTM (Pöppel et al. 2024), Transformer (Vaswani 2017). **Tabular-Specific DL Models:** TabNet (Arik and Pfister 2021), DeepFM (Guo et al. 2017), FT-Transformer (FT-T) (Gorishniy et al. 2021), TabPFN-v2 (Hollmann et al. 2025). **GBDT:** LightGBM (Ke et al. 2017), XGBoost (Chen and Guestrin 2016), XGBoost* (denotes a variant of XGBoost with FT-T-style feature transformations applied as preprocessing). **LLM-based Tabular Models:** TabLLM (Hegselmann et al. 2023), TaBERT (Padhi et al. 2021).

Dataset	#Samples	#Features	#Statements	Metric
BRFSS	401,958	20	107	Acc
NLSY	1,071	34	288	Acc
MHP	2,028	34	166	Acc

Table 3: Summary of Evaluation Datasets for SemantiQ. #Statements refers to the number of natural language statements generated by SemantiQ for each dataset based on the questionnaire structure and options.

4.3 Effectiveness (RQ1)

We systematically evaluate SemantiQ across three datasets to assess both its effectiveness in enhancing deep learning models for questionnaire-based tasks and its generalizability to broader structured learning scenarios, as detailed in Table 2.

On the three questionnaire datasets, SemantiQ consistently delivers state-of-the-art results in both accuracy and macro-F1. Notably, on MHP, it achieves 96.63% accuracy and 97.10% F1, surpassing the strongest deep learning baseline (FT-Transformer) by over 2 percentage points. Similar

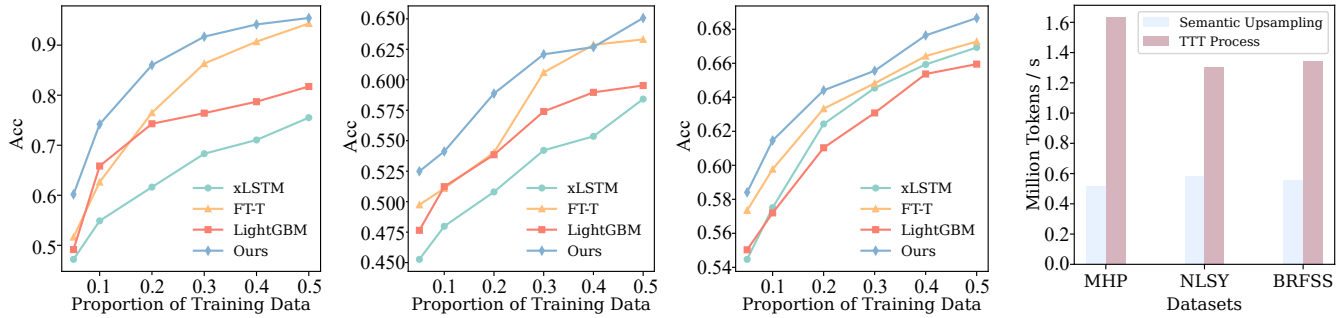

Figure 4: Few-shot learning results on BRFSS (left), NLSY (center), and MHP (right) datasets, with training data proportions from 5% to 50% while the test set remains fixed.

Figure 5: Model Efficiency.

performance margins are observed on BRFSS and NLSY, highlighting the effectiveness of SemantiQ's semantic augmentation pipeline, which leverages question prompts, answer options, and domain-specific terminology. These results underscore that semantic context is a crucial inductive bias in modeling questionnaire data.

To evaluate architectural generality, we further instantiate SemantiQ-DT, integrating the same augmentation pipeline with a tree-based learner (XGBoost). It consistently outperforms standard GBDT models across all questionnaire datasets, showing that SemantiQ's gains are not architecture-specific.

4.4 Order Sensitivity (RQ2)

To further investigate the differences between SemantiQ's cell-level modeling and existing row-level LLM-based approaches, we design an order sensitivity experiment. Without any additional fine-tuning, we directly evaluate pretrained models on three questionnaire datasets—BRFSS, NLSY, and MHP. For each dataset, we construct a shuffled-order variant by randomly permuting the column order (i.e., question order) within each sample while preserving the original cell values.

Model	BRFSS	NLSY	MHP
TaBERT	-6.73	-5.91	-7.13
TabLLM	-4.31	-7.31	-2.47
SemantiQ	**-0.98**	**-0.31**	**-0.79**

Table 4: Order Sensitivity Analysis ($\Delta\%$).

We measure the accuracy change (Δ) when switching from the original to the shuffled column order. The results show that row-level models such as TabLLM and TaBERT suffer substantial degradation after column shuffling, indicating their heavy reliance on fixed positional structures in serialized inputs. In contrast, SemantiQ exhibits a markedly smaller drop, benefiting not only from its cell-level semantic modeling but also from the adaptive capability introduced by the TTT module, which allows the encoder to adjust to altered column arrangements at inference time. This robustness is particularly valuable in scenarios where question order varies, data collection processes involve structural changes, or cross-version migration of surveys is required, as it enables SemantiQ to maintain stable predictive performance across heterogeneous input configurations.

4.5 Few-shot Enhancement (RQ3)

In Fig. 4, we compare the few-shot (low training proportion) performance of SemantiQ against three baselines on BRFSS, NLSY, and MHP. In each dataset, SemantiQ consistently achieves higher accuracy than the baselines when trained on 5%–50% of the available data. Notably, the performance gap is most pronounced in the lower-data regime (e.g., 5%–20%). This suggests that SemantiQ's semantic and structural enhancements help compensate for limited labeled examples, enabling more robust modeling under data-scarce conditions. Since other LLM-based methods require extensive pretraining, we do not include them in the few-shot comparison to ensure a fair evaluation.

4.6 Time Complexity Analysis (RQ4)

Throughput. We evaluate the computational overhead of Semantic Upsampling (workload before encoder) and Test-Time Training (TTT) across three datasets—MHP, NLSY, and BRFSS—using one Tesla A100 GPU with a batch size of 32. As illustrated in Figure 5, Semantic Upsampling processes approximately 553k tokens/s on average, whereas TTT achieves a significantly higher throughput of about 1.43M tokens/s. These results indicate that both modules are computationally efficient and suitable for large-scale for rapid processing of high-volume questionnaire data.

Pre-train & LLM-call Cost. As summarised in Table 5, SemantiQ eliminates two major expenses that burden prior LLM-for-table systems. First, unlike TaBERT—which consumes hundreds of GPU-hours for large-scale corpus pretraining—SemantiQ requires *no* extra LLM pre-training, avoiding substantial offline computation. Second, at infer-

Model	Pre-train Complexity	Inference Complexity
TaBERT	P_{pretrain}	$\mathcal{O}(n)$
TabLLM	N/A	$\mathcal{O}(n)$
SemantiQ	N/A	$\mathcal{O}(1)$

Table 5: LLM-call Time Complexity. P_{pretrain} denotes the pretraining cost. *N/A* indicates that no LLM-specific pretraining is required.

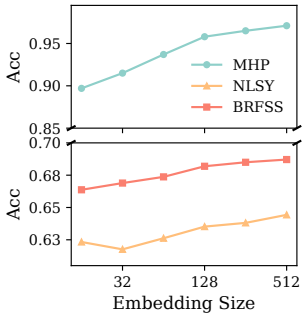

Figure 6: Impact of $|e|$

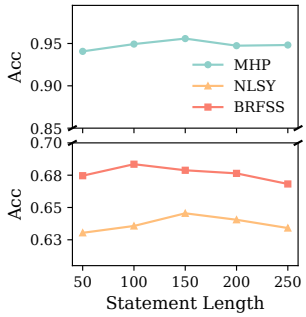

Figure 7: Impact of $|s|$

ence time TabLLM/TaBERT invoke the LLM once per *sample*, whereas SemantiQ needs only one call per *question–option* pair. For example, on the BRFSS dataset SemantiQ requires only 107 LLM calls, whereas TabLLM and TaBERT need $401,958$—yielding roughly a $3,700\times$ reduction in LLM-related latency once statements are cached. Because the number of statements is fixed by the questionnaire length and does not grow with the number of responses, the amortised LLM complexity is effectively $\mathcal{O}(1)$ with respect to #Samples, enabling efficient and flexible deployment.

4.7 Hyperparameter Analysis(RQ5)

We jointly investigate the effects of key hyperparameters and core component groups on SemantiQ's performance. For hyperparameters, we focus on the embedding dimension $|e|$ and the LLM-generated statement length $|s|$. As shown in Fig. 6, larger $|e|$ consistently boosts accuracy across MHP, NLSY, and BRFSS—e.g., BRFSS improves by about 2.3% when $|e|$ increases from 16 to 512—indicating that greater embedding capacity better captures nuanced semantics. Similarly, Fig. 7 shows that extending $|s|$ enriches semantic representation, with optimal performance at 100–150 tokens; longer statements tend to introduce redundancy or noise.

4.8 Ablation Study (RQ6)

We further analyze the contribution of SemantiQ's three core component groups: semantic-wise (Se), feature-wise (Fe), and sample-wise (Sa) interactions. Table 6 shows that removing any single group leads to a measurable performance drop, with semantic-wise components having the most pronounced effect. More importantly, removing multiple groups causes degradation exceeding the sum of individual losses, highlighting the strong *complementarity* among modules. These results indicate that SemantiQ's effectiveness benefits not only from individual components but also from their synergistic interactions.

Removed Components	BRFSS	NLSY	MHP
w/o Se	-1.97	-3.73	-3.70
w/o Fe	-1.76	-0.93	-2.54
w/o Sa	-1.60	-1.45	-1.10
w/o (Se + Fe)	-3.92	-4.89	-6.55
w/o (Se + Sa)	-3.75	-5.44	-5.04
w/o (Fe + Sa)	-3.53	-2.50	-3.82
w/o (Se + Fe + Sa)	-5.60	-6.42	-7.71

Table 6: Ablation Study ($\Delta\%$). Se (*Semantic-wise*) includes Intro & Terms, Option Text, and Question Text. Fe (*Feature-wise*) includes Test-Time Training (TTT) and Left-Out Masking. Sa (*Sample-wise*) refers to Ratio-based augmentation.

4.9 Generalization Performance (RQ7)

To further assess the versatility of SemantiQ beyond questionnaire-focused tasks, we extend our evaluation to a broader range of tabular benchmarks. Specifically, we compare against several representative LLM-based tabular models across five diverse datasets (Table 7). These datasets, all sourced from OpenML (Vanschoren et al. 2014), enabling a robust assessment of SemantiQ's adaptability to general tabular learning scenarios.

Model	Caesarian	Glass	Diabetes	Tae	Breast
TabLLM	60.14	82.02	73.67	52.16	71.71
TabPFN-v2	56.84	80.14	76.21	47.20	71.77
TaBERT	61.17	82.31	72.17	49.18	74.77
SemantiQ	**65.44**	**84.86**	**79.19**	**55.87**	**77.48**

Table 7: Accuracy (%) of LLM-based tabular models on benchmark datasets.

5 Discussion

Our semantic up-sampling strategy creates an intermediate serial structure in which each unit combines $\{$*up-sampled representation* + *original data*$\}$. At first glance, this may seem akin to the conventional $\{$*positional embedding* + *data*$\}$ formulation in many neural architectures. In reality, the injected semantic signals serve as a content-based coordinate system, not just a positional reference.

Traditional positional embeddings encode ***where*** a token resides (e.g., its index), whereas our approach centers on ***what*** a field represents. Because questionnaire data's semantic meaning does not depend on the row or column order, we attach content-driven coordinates "y" to each item, rather than positional "x." As reflected by the RQ2, these richer, context-aware embeddings reinforce ***what*** an item is, regardless of ***where*** it appears.

6 Conclusion

Contributions. We introduce SemantiQ, a framework that leverages a unified semantic space and optimized TTT to generate enriched embeddings for questionnaires. SemantiQ captures multi-level interactions and semantic redundancy, significantly outperforming SOTA methods on real-world benchmarks.

Limitations. Our method focuses on questionnaires, and its efficacy on time-series or graph data is unexplored. Furthermore, the reliance on LLMs for augmentation may inherit model-specific biases.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2023YFF0725103), the Fundamental Research Funds for the Central Universities (510224045), the National Natural Science Foundation of China (62502047, 62192784, 62422202, 62272054, U22B2038, U23A20319), and the Beijing Nova Program (20230484319).

References

Ang, L.; Yawen, L.; and Xue, Z. 2026. FedSIN: information network representation based on federated self-adaptive learning. *Frontiers of Computer Science*, 20: 2001307–.

Arik, S. O.; and Pfister, T. 2019. Tabnet: Attentive interpretable tabular learning. arxiv. *arXiv preprint arXiv:2004.13912*.

Arik, S. Ö.; and Pfister, T. 2021. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 6679–6687.

Bach, B.; Keck, M.; Rajabiyazdi, F.; Losev, T.; Meirelles, I.; Dykes, J.; Laramee, R. S.; AlKadi, M.; Stoiber, C.; Huron, S.; et al. 2023. Challenges and opportunities in data visualization education: A call to action. *IEEE Transactions on visualization and computer graphics*.

Chen, M.; Sun, Y.; Li, T.; Wang, J.; Wang, K.; Lin, X.; Zhang, Y.; and Zhang, W. 2025. Empowering Tabular Data Preparation with Language Models: Why and How? *arXiv preprint arXiv:2508.01556*.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chen, T.; Yin, H.; Long, J.; Nguyen, Q. V. H.; Wang, Y.; and Wang, M. 2022. Thinking inside the box: learning hypercube representations for group recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1664–1673.

Davis, J. A.; and Farrell, M. A. 2016. *The market oriented university: Transforming higher education*. Edward Elgar Publishing.

Dolnicar, S. 2022. Market segmentation for e-tourism. In *Handbook of e-Tourism*, 849–863. Springer.

Dorogush, A. V.; Ershov, V.; and Gulin, A. 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Džeroski, S.; and Ženko, B. 2004. Is combining classifiers with stacking better than selecting the best one? *Machine learning*, 54: 255–273.

Farber, H. S.; Herbst, D.; Kuziemko, I.; and Naidu, S. 2021. Unions and inequality over the twentieth century: New evidence from survey data. *The Quarterly Journal of Economics*, 136(3): 1325–1385.

Fei, N.; Gao, Y.; Lu, Z.; and Xiang, T. 2021. Z-score normalization, hubness, and few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 142–151.

Feng, S. Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; and Hovy, E. 2021. A survey of data augmentation approaches for NLP. *arXiv preprint arXiv:2105.03075*.

for Disease Control, C.; and Prevention. 2020. Behavioral Risk Factor Surveillance System (BRFSS) 2020.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Gorishniy, Y.; Rubachev, I.; Kartashev, N.; Shlenskii, D.; Kotelnikov, A.; and Babenko, A. 2023. Tabr: Tabular deep learning meets nearest neighbors in 2023. *arXiv preprint arXiv:2307.14338*.

Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; and Babenko, A. 2021. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34: 18932–18943.

Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *First Conference on Language Modeling*.

Guo, H.; Tang, R.; Ye, Y.; Li, Z.; and He, X. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1725–1731.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; and Sontag, D. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics*, 5549–5581. PMLR.

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

Huang, X.; Khetan, A.; Cvitkovic, M.; and Karnin, Z. 2020. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*.

Jin, D.; Yu, Z.; Jiao, P.; Pan, S.; He, D.; Wu, J.; Philip, S. Y.; and Zhang, W. 2021. A survey of community detection approaches: From statistical modeling to deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(2): 1149–1170.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Khan, A.; and Ghosh, S. K. 2021. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and information technologies*, 26(1): 205–240.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel,

T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, Y.; Zhuang, M.; Ye, G.; Li, Y.; Wang, J.; Zhou, J.; and Zhang, P. 2025. FeBT: A Feature Balancing Transformer for Corporate ESG Forecasting. *IEEE Transactions on Knowledge and Data Engineering*.

Linden, B.; Boyes, R.; and Stuart, H. 2021. Cross-sectional trend analysis of the NCHA II survey data on Canadian post-secondary student mental health and wellbeing from 2013 to 2019. *BMC Public Health*, 21: 1–13.

Long, J.; Qu, L.; Yu, J.; Chen, T.; Nguyen, Q. V. H.; and Yin, H. 2025. Harnessing Large Language Models for Group POI Recommendations. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 1968–1978.

Ma, S.; Li, X.; Tang, J.; and Guo, F. 2021. A zero-shot method for 3d medical image segmentation. In *2021 IEEE international conference on multimedia and expo (ICME)*, 1–6. IEEE.

Martin, R. A.; Puhlik-Doris, P.; Larsen, G.; Gray, J.; and Weir, K. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of research in personality*, 37(1): 48–75.

McCabe, E.; Amarbayan, M.; Rabi, S.; Mendoza, J.; Naqvi, S. F.; Thapa Bajgain, K.; Zwicker, J. D.; and Santana, M. 2023. Youth engagement in mental health research: a systematic review. *Health Expectations*, 26(1): 30–50.

Nguyen, T. N.; and Zeadally, S. 2021. Mobile crowd-sensing applications: Data redundancies, challenges, and solutions. *ACM Transactions on Internet Technology (TOIT)*, 22(2): 1–15.

Nogueira, R.; Lin, J.; and Epistemic, A. 2019. From doc2query to docTTTTTquery. *Online preprint*, 6(2).

Padhi, I.; Schiff, Y.; Melnyk, I.; Rigotti, M.; Mroueh, Y.; Dognin, P.; Ross, J.; Nair, R.; and Altman, E. 2021. Tabular transformers for modeling multivariate time series. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3565–3569. IEEE.

Pöppel, K.; Beck, M.; Spanring, M.; Auer, A.; Prudnikova, O.; Kopp, M. K.; Klambauer, G.; Brandstetter, J.; and Hochreiter, S. 2024. xlstm: Extended long short-term memory. In *First Workshop on Long-Context Foundation Models@ ICML 2024*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ren, X.; Chen, T.; Nguyen, Q. V. H.; Cui, L.; Huang, Z.; and Yin, H. 2024. Explicit knowledge graph reasoning for conversational recommendation. *ACM Transactions on Intelligent Systems and Technology*, 15(4): 1–21.

Schick, A.; Rauschenberg, C.; Ader, L.; Daemen, M.; Wieland, L. M.; Paetzold, I.; Postma, M. R.; Schulte-Strathaus, J. C.; and Reininghaus, U. 2023. Novel digital methods for gathering intensive time series data in mental health research: scoping review of a rapidly evolving field. *Psychological Medicine*, 53(1): 55–65.

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

Syeeda, M.; Rahmana, A.; Akterc, L.; Fatemaa, K.; Khana, R. H.; Karima, M. R.; Hossaina, M. S.; and Uddina, M. F. 2024. A comprehensive standardized dataset on mental health problems (mhps) of university students. *measurement*, 8: 9.

Vanschoren, J.; Van Rijn, J. N.; Bischl, B.; and Torgo, L. 2014. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2): 49–60.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, J.; Li, Y.; Shao, Y.; Xue, Z.; Guan, Z.; Li, A.; and Ye, G. 2025. Reinforcement Active Client Selection for Federated Heterogeneous Graph Learning. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 21117–21125. AAAI Press.

Wang, Z.; Luo, Y.; Qiu, R.; Huang, Z.; and Baktashmotlagh, M. 2021. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 834–843.

Xu, Y.; Warin, J.; and Robb, M. 2020. Beyond gender binaries: pedagogies and practices in early childhood education and care (ECEC).

Yu, A. W.; Dohan, D.; Luong, M.-T.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. V. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Yuan, W.; Yang, C.; Ye, G.; Chen, T.; Nguyen, Q. V. H.; and Yin, H. 2025. Fellas: Enhancing federated sequential recommendation with llm as external services. *ACM Transactions on Information Systems*, 43(6): 1–24.

Zhang, H.; Dong, L.; Gao, G.; Hu, H.; Wen, Y.; and Guan, K. 2020. DeepQoE: A Multimodal Learning Framework for Video Quality of Experience (QoE) Prediction. *IEEE Transactions on Multimedia*, 22(12): 3210–3223.

Zhang, P.-F.; Cheng, Y.; Sun, X.; Wang, S.; Zhu, L.; and Shen, H. T. 2025. A Step Toward World Models: A Survey on Robotic Manipulation. *arXiv preprint arXiv:2511.02097*.

Zhao, Y.; Peng, B.; Iqbal, K.; and Wan, A. 2023. Does market orientation promote enterprise digital innovation? Based on the survey data of China's digital core industries. *Industrial Marketing Management*, 109: 135–145.

Zhou, J.; Wang, K.; Wang, J.; Zhang, K.; and Lin, X. 2025. COMET: An interactive framework for efficient and effective community search via active learning. *INFORMS Journal on Computing*.