Semantic Embedding and Synthetic Augmentation for Longitudinal Survey Prediction (Student Abstract)

Julia Rezvani[1,6], Alina Hyk[2, 6], Thuyen Pham[3, 6], Leonardo Marciaga[4, 6], Chunyang Liao[5, 6], Raffaele Vardavas[7], Konstantinos Mitsopoulos[8]

[1]Portland State University
[2]Oregon State University
[3]University of Massachusetts Amherst
[4]Illinois Institute of Technology
[5]University of California, Los Angeles
[6]Institute for Pure and Applied Mathematics
[7]RAND Corporation
[8]Institute For Human and Machine Cognition
rezvani@pdx.edu, hyka@oregonstate.edu, thuyenpham@umass.edu, lmarciaga@hawk.illinoistech.edu, liaochunyang@math.ucla.edu, rvardava@rand.org, kmitsopoulos@ihmc.org

Abstract

Longitudinal surveys are a crucial component of behavioral research. Such surveys, however, face significant gaps in the data created by item and unit non-responses as well as semantic gaps resulting from questionnaires, assessed trends, and data collection methods evolving over time. Using 15 waves of vaccination surveys as a testbed, we demonstrate how modern AI techniques can bridge both item and unit gaps, originating from non-response, and semantic gaps, originating from instrument evolution.

We address these gaps through a two-component framework. We leverage LLM-generated semantic embeddings of survey questions to encode question meaning, enabling a Deep & Cross Network used for imputation to jointly model responses across item semantics, individual characteristics, and temporal dynamics. This structure directly addresses survey evolution by operating in learned semantic space. To overcome data scarcity, we use cluster-informed synthetic data generation via hierarchical prompting that produces synthetic responses preserving distributional properties and empirical cluster structure. Our approach achieves a strong improvement in semantic gap tasks and 80-90% synthetic data fidelity, providing practical solutions for evolving longitudinal studies.

Code — https://github.com/jlsrls/aaai-26-semantic-embedding-synthetic-augmentation

Datasets — https://alpdata.rand.org/

Problem Statement

We identify three types of gaps in longitudinal survey data: *item gaps*, created by item non-responses, where respondents fail to answer individual questions; *unit gaps*, created by unit non-responses, where respondents fail to respond to entire survey waves; and *semantic gaps*, created by the evolution of the questionnaires themselves. While established imputation methods mitigate item and unit non-response, semantic gaps present a greater challenge for conventional frameworks lacking semantic analysis capabilities.

We address these challenges through LLM-enhanced prediction models using semantic question embeddings and cluster-informed synthetic data generation. We demonstrate our approach's performance on the well-researched FluPaths and COVIDPaths datasets collected by the RAND Corporation through the American Life Panel. The datasets provide behavioral surveillance across 15 waves from Fall 2016 to Summer 2024 across 2,200 respondents. The surveys' transition from seasonal flu tracking to pandemic response introduced semantic gaps, unit, and item non-response, creating an optimal testbed for our methodology.

Synthetic Data Generation

To address data sparsity constraints originating from item and unit gaps, we employ a cluster-guided LLM approach to generate synthetic survey data.

We performed cluster analysis and identified two behavioral groups: one corresponding to generally positive attitudes towards vaccination, and the other to generally negative attitudes. We then performed a three-stage prompting method, inspired by (Wang, Wang, and Sun 2024), using a few-shot approach to generate synthetic data for each wave.

To evaluate the importance of dataset size to data generation quality, we generated two datasets. The first dataset contains LLM-generated synthetic data for three questions highly associated with cluster separation, along with age, gender, and cluster labels. The second dataset contains LLM-generated full responses for every wave, reconstructing all major variables with sufficient response density in the original data. Cluster labels are included as well. These datasets can be used to augment machine learning models by providing synthetic data during training.

Model selection varied by task complexity. The limited-variable dataset employed smaller models (GPT-4o Mini, GPT-3.5 Turbo, Gemini 1.5 Pro), while the full-response

task utilized larger models (GPT-5, GPT-4o). Claude Sonnet 4 was tested in both scenarios for comparison.

LLM-Based Embeddings for Enhanced Imputation

Our imputation model for longitudinal prediction is a conventional Deep & Cross Network (Wang et al. 2017) trained to predict individual responses within the survey data as positive or negative. Adapting a recently pioneered approach (Kim and Lee 2024) to the longitudinal paradigm, we augment our network with three sets of embeddings (Figure 1). Respondent and wave embeddings are learned by the network to characterize individual responses and survey waves, allowing for longitudinal reasoning, while question embeddings are derived from OpenAI's `text-embedding-3-large`, allowing the network to interpret the semantic content of questions while ensuring training costs remain modest; the model learns a single-layer network to reduce the large 1536-dimensional embeddings derived from the LLM to 150 dimensions. Imputation is implemented via classification: predicting whether responses in the survey data were positive or negative, given embeddings for the respondent who responded, the question that was asked of the respondent, and the survey wave in which that question appeared.

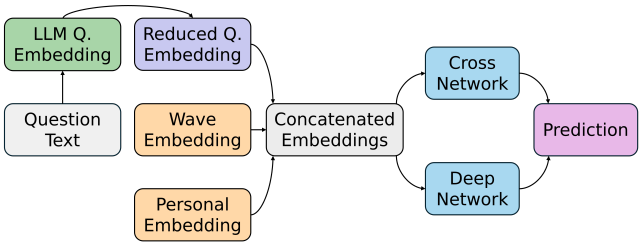

Figure 1: DCN imputation model architecture.

We evaluate our network's performance on two different missingness mechanisms: MCAR missingness and retrodiction missingness (Kim and Lee 2024), simulating the item and semantic gaps that appear in real-world datasets and allowing for a unified approach to network evaluation. We perform ablations on the use of LLM-derived question embeddings, isolating their effect on model performance.

We test synthetic data robustness on the retrodiction task using Claude Sonnet 4-generated responses. Real training data was restricted to the synthetic data's variable subset. Synthetic items underwent positive/negative classification and MCAR masking, matching the original training data.

Results

Synthetic Data Generation Quality

Synthetic responses demonstrated strong distributional fidelity, evaluated using the metrics introduced in (Shi et al. 2025). For the smaller set of questions dataset, the top performing models achieved column shape similarity scores consistently above 80% and overall similarity scores above 72%. However, column trend similarity results (C2ST) proved harder to capture due to high score variation.

For the second dataset with more variables, performance improved across all metrics. Shape similarity scores ranged predominantly between 80% and 90%, with overall similarity reaching up to 90%. C2ST values increased to approximately 40%-75%, indicating that generating full synthetic replications of the dataset captures residual artifacts substantially more effectively than focusing on a limited subset of variables. The Claude model, used for universal comparisons, further confirmed the stability of the observed improvements, achieving an increase of 7.9% in overall similarity and 20.7% in C2ST.

DCN Performance with LLM Embeddings

Our Deep & Cross Network architecture demonstrates robust prediction capabilities across varying data sparsity conditions. Under conditions of uniform (MCAR) missingness, performance degrades gracefully regardless of embedding format as the missingness level increases. Under retrodiction missingness (Kim and Lee 2024), LLM embeddings substantially enhance performance until the missingness level reaches 90%. We conclude that semantic embeddings effectively bridge semantic gaps from questionnaire evolution given a large enough dataset.

		MCAR miss.		Retrodiction miss.	
Miss. Level	Metric	LLM	no LLM	LLM	no LLM
10%	BCE Loss	**0.349**	0.355	**0.499**	0.717
	ROC AUC	**91.8%**	91.3%	**85.1%**	65.0%
25%	BCE Loss	**0.351**	0.363	**0.546**	0.672
	ROC AUC	**91.6%**	91.1%	**84.6%**	70.3%
50%	BCE Loss	0.390	**0.387**	**0.560**	0.686
	ROC AUC	**90.7%**	90.3%	**81.9%**	69.3%
75%	BCE Loss	0.467	**0.416**	**0.641**	0.638
	ROC AUC	**88.8%**	88.6%	**75.9%**	67.8%
90%	BCE Loss	0.674	**0.648**	**0.720**	0.742
	ROC AUC	85.8%	**86.2%**	**64.9%**	63.6%

Table 1: Peak test-set imputation performance with/without LLM embeddings by missingness type.

Prediction Performance with Synthetic Data

Incorporating Sonnet 4-generated synthetic data into DCN training yielded mixed results. Under 50% MCAR missingness, ROC AUC decreased marginally from 89.1% (real data only) to 88.3% (with masked synthetic data). The performance drop suggests synthetic data may oversimplify behavioral patterns, lacking the nuanced inconsistencies and response variability inherent in real longitudinal survey data.

Conclusion and Future Work

We demonstrate that LLM-enhanced survey analysis addresses item, unit, and semantic gaps in longitudinal research. Our Deep & Cross Network with semantic embeddings bridges questionnaire evolution, while synthetic data generation improves training and preserves behavioral signatures. This framework maintains analytical continuity in evolving survey instruments for behavioral surveillance. Future work includes expanding domain evaluation, advancing synthetic data authenticity, and validating real-world deployment across active survey programs.

Acknowledgments

This study was conducted as part of the Research in Industrial Projects for Students (RIPS) 2025 program at the Institute for Pure and Applied Mathematics, an NSF Mathematics Institute at UCLA. The project was supported by the RAND Corporation. The authors gratefully acknowledge the support of both the RAND Corporation and IPAM. We would also like to acknowledge the support of Susana Serna, RIPS Program Director at IPAM.

References

Kim, J.; and Lee, B. 2024. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. arXiv:2305.09620.

Shi, J.; Xu, M.; Hua, H.; Zhang, H.; Ermon, S.; and Leskovec, J. 2025. TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation. arXiv:2410.20626.

Wang, R.; Fu, B.; Fu, G.; and Wang, M. 2017. Deep & Cross Network for Ad Click Predictions. *CoRR*.

Wang, R.; Wang, Z.; and Sun, J. 2024. UniPredict: Large Language Models are Universal Tabular Classifiers. arXiv:2310.03266.