Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Representation Effects in Child and Youth Mental Health Emergency Readmission Predictions

Cassandra Czobit[†,*], Hirad Daneshvar[†], Reza Samavi[†,‡], Thomas Doyle[◇], Laura Duncan[◇], Paulo Pires[◇], Roberto Sassi[§]
[†] Toronto Metropolitan University [‡] Vector Institute [◇] McMaster University [§] University of British Columbia

Abstract

Predicting mental health–related emergency department readmission in youth remains challenging, and the role of data representation is underexplored. Using the National Survey on Drug Use and Health (ages 12–18), we compare three representations: (1) structured tabular features, (2) template-generated clinical text, and (3) LLM-derived sentence embeddings. Classical models are trained on tabular data and embeddings, while LLMs are applied to text. Results show that tabular features consistently yield the best and most stable performance. Templated text introduces a representational bottleneck and is less robust under distribution shift, while embeddings preserve some semantics but do not outperform tabular inputs. Representation choice is thus critical for predictive performance.

Keywords: Representation learning, predictive modeling, LLM, template-based representation, mental health, child and youth

1. Introduction

Child and youth mental health disorders commonly arise during adolescence and can lead to substantial functional impairments, particularly when access to and quality of care is limited [1]. Globally, studies have shown that 48.4% of neurodevelopmental disorders occur before age 18, with a peak onset of 14.5 years [2], highlighting the importance of early and accurate intervention [1, 2]. Clinicians must routinely synthesize large volumes of clinically- and patient-generated data, such as clinical notes, imaging, blood tests, non-medical sensor recordings and questionnaire responses [3] under significant time and resource constraints [4]. These challenges in achieving timely diagnosis and treatment have motivated the adoption of Artificial Intelligence (AI) systems as decision-support tools [5], which can detect diseases, optimize treatment plans, and model disease progression [5].

Classical machine learning (ML) models and Large Language Models (LLMs) have been applied to structured clinical and survey data [6]. LLMs demonstrate strong capabilities in text classification and information extraction [7], with emerging applications in mental-health contexts, including conversational agents and diagnostic support [3]. However, general-purpose LLMs often lack domain-specific knowledge required for clinical reasoning [8] and fine-tuning approaches depend heavily on the quality and structure of the input data [9, 10].

In collaboration with a clinical team, we are developing a mental-health decision support system that must operate under strict privacy constraints. This limits the development and benchmarking of models using real clinical notes, making it essential to understand how different representations of structured survey data preserve or distort clinically meaningful information. Structured tabular data treats features independently and may lose contextual relationships [11], whereas template-generated text introduces semantic structure that LLMs can leverage [12]. Sentence-embedding representations capture latent relationships but may obscure explicit details [13]. While representation learning is known to influence model behaviour, its implications for clinical prediction tasks, where outcomes reflect interacting sociodemographic, behavioural, and mental-health factors, remain underexplored [14].

[*] cczobit@torontomu.ca

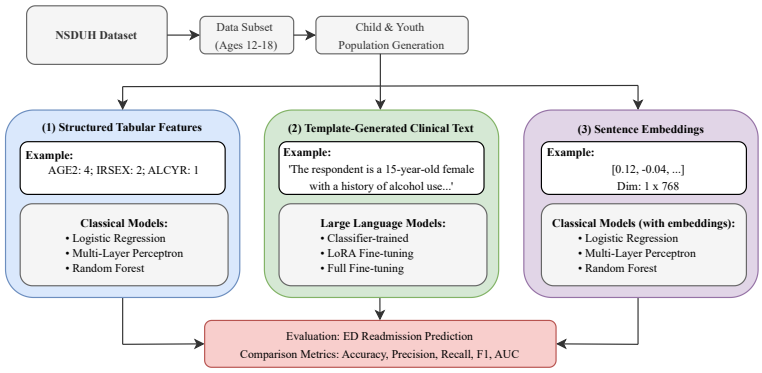

Figure 1. Overview of our proposed study workflow from the NSDUH data selection to the construction of three data representations and their evaluation.

Related Work. Prior work in mental-health prediction has applied classical machine learning models to structured survey and clinical data [15], and more recently has explored domain specific LLMs such as ClinicalBERT, MentalBERT, and MentalRoBERTa for tasks including readmission prediction, concept extraction, and diagnostic support [16, 17]. Template-based methods have been used to convert structured clinical variables into natural-language pseudo notes, primarily within electronic health record settings [12], and frameworks such as Textionnaire show that transforming fixed-response survey items into text can benefit deep learning models [18]. However, existing studies typically evaluate a single representation modality in isolation and have not explored how alternative encodings influence predictive performance or model robustness. This gap is particularly important in confidentiality-constrained mental-health contexts, where representation choice may determine whether models preserve or distort clinically meaningful co-determinants.

To address this, we evaluate three modalities derived from the same National Survey on Drug Use and Health (NSDUH) youth dataset: (1) structured tabular features, (2) template-generated clinical note style text created using a machine-readable data dictionary, and (3) LLM-derived sentence embeddings. Using these representations, we train classical machine learning models and multiple LLM configurations to predict emergency department (ED) readmission, isolating the effect of representation from the underlying model architecture. Figure 1 provides an overview of the study workflow.

Contributions. We make three contributions: (1) a machine-readable NSDUH data dictionary and modular template framework to convert categorical survey data into semantically rich text; (2) a systematic comparison of three representations—tabular features, template-generated text, and LLM-derived embeddings—evaluated with both classical models and LLMs; and (3) empirical evidence showing how representation shapes model behavior by preserving or distorting key co-determinants of youth mental health, informing model design in privacy-constrained settings.

2. Methodology

This study evaluates how three representations of the same data influence predictive performance for a clinical task. We describe the dataset, template-based text generation strategy, model configurations, and evaluation procedure.

Data Collection and Preparation. This study obtained questionnaire data from the NSDUH, an annual nationally representative survey of individuals aged 12 and older in the United States[1]. All responses to sociodemographic characteristics, substance use, mental

[1]https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health

health and medical conditions were self-reported and encoded as categorical variables. We used data collected from 2015-2019 and defined youth as respondents aged 12–18. From 2,812 available features, we selected 62 variables previously identified as influential predictors of mental-health disorders [15]. Additional derived features captured lifetime and past year major depressive episodes, and the target label (number of emergency department visits ≥ 3 in the past year). Missing, refused, and indeterminate responses were harmonized into a unified missing category, and variables without responses were removed. The resulting dataset was highly imbalanced, with approximately 10% of respondents meeting the ED readmission criterion and 90% belonging to the negative class. We used 90% of the preprocessed data for training and validation and reserved the remaining 10% for testing, with stratified 5-fold cross-validation to preserve the distribution of the minority class.

Template-based Representation. The responses of the NSDUH questionnaire are provided in a structured tabular format, with numeric codes representing categorical values. Unlike classical ML models, LLMs cannot directly infer the semantics of these values, limiting the model's reasoning and predictive capabilities. To convert these codes into semantically meaningful text, we constructed a machine-readable data dictionary with mappings from numeric values to descriptive text and from field names to human-readable labels. These mappings were used to generate modular, clinical note style sentences describing each respondent. Features were grouped by question type, and the template generation process was standardized to ensure consistency. Missing or non-substantive responses were omitted entirely. The template structure and phrasing were reviewed in consultation with collaborating clinicians to ensure that the generated text is clinically meaningful rather than synthetic or model-generated narratives. This representation enables LLMs to leverage semantic priors learned from natural-language corpora (see Appendix A for examples).

Classical Machine Learning Models. Logistic regression, a multi-layer perceptron (MLP), and random forest classifiers served as baselines for the tabular and embedding representations. Hyperparameters were tuned via grid search using the training folds, and the best configuration for each model was evaluated on the hold-out test set. Full hyperparameter grids are provided in Appendix B.

Embedding-Based Representation. Template-generated text was encoded using the pretrained MentalRoBERTa model [17]. For each respondent, hidden states were aggregated using mean pooling to produce a fixed-length dense embedding vector. These embeddings were used as inputs to the classical machine learning models to enable a direct comparison with the tabular representation.

Large Language Models. To evaluate the template-based representation, we use the publicly available pre-trained mentalRoBERTa as our baseline model [17]. We evaluate two configurations of the model: a frozen-encoder classifier head and fine-tuned variants.

Baseline Performance: To benchmark the LLM performance, we trained a classifier head on top of the frozen pretrained encoder. This isolates the contribution of the pretrained semantic representations without updating internal weights.

Fine-tuning Approaches: To adapt the pretrained mentalRoBERTa model to our prediction task, we implemented full parameter fine-tuning and LoRA-based parameter-efficient fine-tuning. The former updated all transformer parameters using a learning rate of 2e-5 with early stopping. The latter introduced trainable low-rank matrices into the attention layers [19], enabling efficient adaptation while keeping the base model frozen. Hyperparameters were tuned on validation folds. Final configurations are reported in Appendix B.

Evaluation Strategy. All models were trained on a balanced training set and evaluated on both the balanced and unbalanced test sets. The unbalanced set reflects real-world ED visit distributions, and the balanced set assesses robustness under class distribution shift. We report accuracy, precision, recall, F1 score, and area under the curve (AUC) for all models.

3. Results and Discussion

Structured Tabular Features. We evaluate the prediction-based performance of the structured tabular NSDUH features using classical machine learning models. Table 1 summarizes the performance of classical models trained on the original tabular features.

Model	Eval	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	Unbalanced	0.6620 ± 0.0012	0.1610 ± 0.0005	0.6005 ± 0.0036	0.2539 ± 0.0008	0.6837 ± 0.0013
	Balanced	0.6338 ± 0.0029	0.6436 ± 0.0032	0.6000 ± 0.0046	0.6210 ± 0.0033	0.6828 ± 0.0027
Multi-Layer Perceptron	Unbalanced	0.6735 ± 0.0274	0.1583 ± 0.0070	0.5544 ± 0.0321	0.2458 ± 0.0062	0.6624 ± 0.0023
	Balanced	0.6132 ± 0.0083	0.6214 ± 0.0096	0.5810 ± 0.0274	0.6001 ± 0.0145	0.6506 ± 0.0057
Random Forest	Unbalanced	0.6294 ± 0.0052	0.1538 ± 0.0013	0.6371 ± 0.0080	0.2477 ± 0.0017	0.6836 ± 0.0022
	Balanced	0.5054 ± 0.0007	0.9552 ± 0.0283	0.0115 ± 0.0016	0.0226 ± 0.0030	**0.6888 ± 0.0029**

Table 1. Average performance of classical models trained on structured NSDUH features. Bold values indicate the best score within each evaluation setting.

Across both evaluation settings, logistic regression and random forest achieve the strongest and most stable performance, with AUC values around 0.68. Under imbalanced test conditions, all models exhibit high recall (≈60-63%) but low precision (≈15-16%), indicating a tendency to over-predict the minority class. The MLP does not outperform simpler models, suggesting that the predictive signal is largely linearly separable.

The balanced evaluation provides additional insights into the robustness of the representation. Logistic regression and MLP maintain stable AUC values, while random forest experiences a collapse in recall (0.01), indicating sensitivity to label distribution shift and majority-class bias rather than learned knowledge from the feature space. Overall, the tabular representation provides a strong and reliable baseline, preserving class separating information across conditions.

Template-Based Unstructured Text. The LLM-based models relied on the template-generated representation generated from the NSDUH data dictionary. This representation transformed and guided the structured tabular responses into clinical-note style text representations. Table 2 reports results for LLMs trained on template-generated text, with the best performance shown in bold.

LLM Model	Eval	Accuracy	Precision	Recall	F1-Score	AUC
Classifier-trained	Unbalanced	0.6462 ± 0.0457	0.1319 ± 0.0056	0.4791 ± 0.0552	0.2062 ± 0.0027	0.6034 ± 0.0022
	Balanced	0.5708 ± 0.0031	0.5889 ± 0.0129	0.4759 ± 0.0601	0.5241 ± 0.0314	0.6039 ± 0.0024
LoRA Fine-tuned	Unbalanced	0.6143 ± 0.0197	0.1459 ± 0.0035	0.6229 ± 0.0257	0.2364 ± 0.0041	0.6601 ± 0.0056
	Balanced	0.5251 ± 0.0620	0.5153 ± 0.0842	0.6611 ± 0.3729	0.5394 ± 0.1889	0.5374 ± 0.0997
Full Fine-tuned	Unbalanced	0.6021 ± 0.0096	0.1446 ± 0.0026	0.6420 ± 0.0231	0.2361 ± 0.0045	0.6618 ± 0.0082
	Balanced	0.6191 ± 0.0074	0.6140 ± 0.0052	0.6420 ± 0.0231	0.6275 ± 0.0125	**0.6627 ± 0.0084**

Table 2. Average performance of LLM-based models trained on template-generated text.

Performance varies substantially across configurations. The frozen encoder classifier head yields the weakest results (AUC ≈ 0.60), indicating that pretrained semantic priors alone are insufficient for this task. Fine-tuning improves performance, where LoRA achieves an AUC of 0.6601, and full fine-tuning provides a modest additional gain (AUC ≈ 0.66). However, the full fine-tuned models do not surpass the classical models trained on tabular features. The rigid phrasing and limited expressiveness of the templates appear to introduce a representational bottleneck, limiting the model's ability to leverage semantic relationships. Performance becomes less stable under balanced evaluation, with large variance in precision and recall for the classifier trained and LoRA models. These findings suggest that while template-based text preserves some predictive cues, it lacks the richness required for LLMs to outperform simpler models.

Model	Eval	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	Unbalanced	0.6466 ± 0.0038	0.1558 ± 0.0022	0.6085 ± 0.0055	0.2480 ± 0.0032	**0.6778 ± 0.0019**
	Balanced	0.6308 ± 0.0022	0.6321 ± 0.0046	0.6267 ± 0.0099	0.6293 ± 0.0033	0.6634 ± 0.0032
Multi-Layer Perceptron	Unbalanced	0.6271 ± 0.0329	0.1512 ± 0.0064	0.6235 ± 0.0379	0.2430 ± 0.0053	0.6714 ± 0.0030
	Balanced	0.6040 ± 0.0049	0.5932 ± 0.0091	0.6652 ± 0.0307	0.6267 ± 0.0096	0.6181 ± 0.0060
Random Forest	Unbalanced	0.7910 ± 0.0037	0.1768 ± 0.0030	0.3236 ± 0.0107	0.2287 ± 0.0043	0.6498 ± 0.0014
	Balanced	0.5822 ± 0.0055	0.6707 ± 0.0084	0.3236 ± 0.0119	0.4364 ± 0.0120	0.6496 ± 0.0049

Table 3. Average performance of classical models trained on sentence embeddings.

Predictive Power of Embedding Representations We evaluate the LLM-derived embeddings of template-based text during classical ML model training. Table 3 summarizes the results of the classical ML models, where bold values indicate the highest AUC score.

Embedding-based models achieve intermediate performance (AUC $\approx$ 0.65–0.67), indicating that embeddings retain some discriminative structure. However, they do not improve upon the original tabular features. Precision–recall patterns reveal limitations in the embedding space, where logistic regression and MLP achieve high recall but low precision under class imbalance, suggesting that embeddings may collapse distinctions between classes. Random forest performs poorly under balanced conditions, likely due to the lack of hierarchical structure in the embedding dimensions.

Comparative Analysis of Representations Across all experiments, there is a consistent pattern in how effectively each representation captures the underlying predictive signal. Models trained on the original tabular features achieve the strongest and most stable performance, indicating that the structured variables retain information that is both discriminative and robust to shifts in class distribution. In contrast, converting these features into template-generated text introduces rigidity and noise. Although the templates preserve some cues from the original variables, their limited linguistic variability and sparse structure restrict the LLM's ability to model relationships between features. With full fine-tuning, the LLMs only partially recover the predictive signal present in the tabular data.

The embedding-based representation provides an intermediate outcome. Sentence embeddings derived from the templates capture more semantic structure than the raw text, and classical models trained on these embeddings outperform the classifier-trained LLM. However, the embeddings still do not exceed the performance of the original tabular features, suggesting that they remain constrained by the representational bottleneck introduced during template generation.

Our findings challenge the assumption that transforming structured survey data into natural language will inherently improve predictive performance. Instead, the fidelity of the representation, rather than the complexity of the model, ultimately determines how effectively the predictive signal is preserved.

Ablation Study We conducted an ablation study removing the top 10 predictors (identified via permutation importance) to investigate whether LLMs can infer relationships and co-determinants of predictive features more effectively than classical models. All representations experienced a decline in performance, and the LLM did not outperform classical baselines under either evaluation setting. Full results are provided in Appendix C.

Limitations & Future Directions The primary limitation of this study is the simplicity of the template-based text. Although modular and reproducible, the templates lack linguistic variability and contextual richness. Future work could explore generating more expressive pseudo notes using retrieval augmented generation or prompting LLMs to produce richer clinical note-style text. More broadly, mental health constructs are inherently relational in nature, shaped by demographic, behavioural, and psychological co-determinants. Representations grounded in structured knowledge such as ontologies or knowledge graphs may better capture these relationships and improve both interpretability and predictive performance.

4. Conclusion

This study examined how different representations of NSDUH survey data affect youth mental-health prediction. Tabular features consistently provided the strongest and most stable signal, while template-generated text introduced abstraction that reduced fidelity, and embeddings did not surpass the tabular baseline. These results highlight that representation quality, rather than model complexity, drives performance.

Acknowledgements

This work was supported by funding from Hamilton Health Sciences, NSERC Discovery Grant, and the Canada First Research Excellence Fund. AI-assisted tools were used solely for editing and polishing of author-written text. The authors take full responsibility for the content of the paper.

References

[1] P. J. Uhlhaas et al. "Towards a youth mental health paradigm: a perspective and roadmap". In: *Molecular Psychiatry* (2023). ISSN: 14765578. DOI: 10.1038/S41380-023-02202-Z.

[2] M. Solmi, J. Radua, et al. "Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies". In: *Molecular Psychiatry* 27 (2022), p. 22.

[3] R. Yang, T. F. Tan, et al. "Large language models in health care: Development, applications, and challenges". In: *Health Care Science* 2.4 (Aug. 2023), pp. 255–263. ISSN: 2771-1757.

[4] F. Minerva and A. Giubilini. "Is AI the Future of Mental Healthcare?" In: *Topoi* (May 2023).

[5] S. Graham, C. Depp, et al. "Artificial Intelligence for Mental Health and Mental Illnesses: an Overview". In: *Current Psychiatry Reports* 21.11 (Nov. 2019). ISSN: 15351645.

[6] S. Tutun, M. E. Johnson, et al. "An AI-based Decision Support System for Predicting Mental Health Disorders". In: *Inf Syst Front* 25.3 (June 2023), pp. 1261–1276. ISSN: 1572-9419.

[7] K. Singhal et al. "Large Language Models Encode Clinical Knowledge". In: *Nature* 620.7972 (Dec. 2022), pp. 172–180. ISSN: 14764687.

[8] A. J. Thirunavukarasu, D. S. J. Ting, et al. "Large language models in medicine". In: *Nature Medicine 2023 29:8* 29.8 (July 2023), pp. 1930–1940. ISSN: 1546-170X.

[9] E. J. Hu, Y. Shen, et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. URL: https://arxiv.org/abs/2106.09685.

[10] L. Xu et al. "Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment". In: *IEEE Trans. on Pattern Anal. Mach. Intel.* (2026).

[11] A. Rajkomar et al. "Scalable and accurate deep learning for electronic health records". In: *npj Digital Medicine* 1 (Jan. 2018). DOI: 10.1038/s41746-018-0029-1.

[12] S. A. Lee et al. "Clinical decision support using pseudo-notes from multiple streams of EHR data". In: *npj Digital Medicine* 8.1 (July 2025). DOI: 10.1038/s41746-025-01777-x.

[13] T. Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Proc. NeurIPS*. NIPS'13. Curran Associates Inc., 2013, 3111–3119.

[14] Y. Bengio et al. "Representation Learning: A Review and New Perspectives". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.8 (Aug. 2013), 1798–1828. DOI: 10.1109/TPAMI.2013.50.

[15] G. Bobashev et al. "Predictive model of multiple emergency department visits among adults: analysis of the data from the National Survey of Drug Use and Health (NSDUH)". In: *BMC Health Services Research* 21.1 (Dec. 2021). ISSN: 14726963.

[16] K. Huang, J. Altosaar, and R. Ranganath. *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*. 2020. URL: https://arxiv.org/abs/1904.05342.

[17] S. Ji et al. "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare". In: *Proc. of 13th LREC*. European Language Resources Association, June 2022.

[18] S. Rashidiani et al. "Textionnaire: An NLP-Based Questionnaire Analysis Method for Complex and Ambiguous Task Decision Support". eng. In: *Proc. IEEE ICCICC*. 2022.

[19] V. Lialin, V. Deshpande, X. Yao, and A. Rumshisky. *Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning*. 2024. URL: https://arxiv.org/abs/2303.15647.

Appendix A. Template Example

1: Sample Patient Profile

Identification: The respondent is a 18 year old Hispanic male who has completed 10th - 12th grades, with a total family income between $20,000-49,999. The respondent is not covered by health insurance and resides in a core-based statistical area with 1 million or more persons.

Mental Health: The respondent has felt restless/fidgety most of the time in the past 30 days. The respondent has not had a Major Depressive Episode in their lifetime.

Appendix B. Hyperparameter Configurations

This appendix summarizes the hyperparameter search spaces and final configurations for all models.

B.1. Classical Machine-Learning Models

Model	Parameters
Logistic Regression	$C \in \{0.01, 0.1, 1.0, 10.0\}$; penalty = L2; solver $\in$ {lbfgs, newton-cg, saga}. **Final: $C = 1.0$, L2, lbfgs**
Multi-Layer Perceptron	Hidden layers $\in \{[64, 64], [128, 64], [256, 128]\}$; learning rate $\in \{0.001, 0.005, 0.01\}$; batch size $\in \{32, 64\}$; max epochs = 150. **Final: [128, 64], LR = 0.001, batch = 32**
Random Forest	$n_{\text{estimators}} \in \{200, 300, 400\}$; max depth $\in$ {None, 5, 10, 15}; min samples split $\in \{2, 5, 10\}$; min samples leaf $\in \{1, 2, 4\}$; max features $\in$ {sqrt, log2}. **Final: 300 trees, depth = 10, min split = 2, min leaf = 1, max features = sqrt**

Table 4. Hyperparameter search space for classical models

B.2. Embedding Model

We generate fixed-length sentence embeddings using the pretrained MentalRoBERTa encoder. Each template-generated clinical note was tokenized using the model's default tokenizer, and the final hidden states were aggregated using mean pooling to produce a 768-dimensional embedding vector.

B.3. LLM Fine-Tuning

Model	Parameters
Frozen Encoder	Classifier head trained; LR=1e-3; batch=16; 5 epochs.
Full Fine-Tuning	All parameters updated; LR=2e-5; weight decay=0.01; cosine schedule with warmup.
LoRA Fine-Tuning	Rank $\in \{16, 32, 64\}$; $\alpha \in \{8, 16\}$; LR $\in$ {1e-4–5e-4}. **Final: rank=64, $\alpha = 16$, LR=2e-4, dropout=0.05**

Table 5. Hyperparameter search space for LLM variants

Appendix C. Ablation Study

This appendix summarizes the ablation experiment conducted to assess whether LLMs can recover predictive signal beyond classical models when the strongest predictors are removed. The top 10 predictors were identified using permutation importance from logistic regression across cross-validation folds. These predictors were removed from all three representations (tabular, template-based text, and embeddings), and models were retrained under identical conditions.

C.1. Experimental Setup

Permutation importance scores were averaged across folds, and the ten highest-scoring features were removed from each representation. Classical models (logistic regression, MLP, random forest) and the full fine-tuned LLM were retrained using the same hyperparameters reported in Appendix A. Performance was evaluated on both the unbalanced and balanced test sets.

C.2. Ablation Results

Model	Eval	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression (tabular)	Unbalanced	0.6409 ± 0.0039	0.1423 ± 0.0015	0.5468 ± 0.0109	0.2258 ± 0.0026	0.6431 ± 0.0010
	Balanced	0.5969 ± 0.0052	0.6076 ± 0.0041	0.5476 ± 0.0128	0.5760 ± 0.0088	**0.6406 ± 0.0024**
Full Fine-tuned (template-based text)	Unbalanced	0.5779 ± 0.0128	0.1310 ± 0.0062	0.6046 ± 0.0243	0.2154 ± 0.0096	0.6239 ± 0.0209
	Balanced	0.5898 ± 0.0150	0.5873 ± 0.0140	0.6046 ± 0.0243	0.5957 ± 0.0182	0.6237 ± 0.0218
Logistic Regression (embeddings)	Unbalanced	0.6407 ± 0.0049	0.1464 ± 0.0018	0.5694 ± 0.0098	0.2329 ± 0.0027	**0.6515 ± 0.0014**
	Balanced	0.6089 ± 0.0051	0.6191 ± 0.0061	0.5670 ± 0.0146	0.5917 ± 0.0083	0.6399 ± 0.0024

Table 6. Ablation performance of classical and LLM-based models after removing the top 10 structured predictors. The bold values indicate the highest AUC score.

Removing the strongest predictors resulted in a performance decline across all models under unbalanced conditions, indicating that these features carry substantial predictive signal. Classical models trained on tabular features or embeddings showed similar reductions in AUC, while the fine-tuned LLM did not recover additional signal beyond the classical baselines. Under balanced evaluation, performance converged to a narrow range across all representations. The tabular model achieved the highest AUC, while the embedding-based model achieved the highest accuracy and precision. These results suggest that the LLM did not infer additional relationships between features once the strongest predictors were removed. The ablation study reinforces our central findings: representation constraints, rather than model capacity, ultimately determine the predictive power of the model. The template-based text introduces structural rigidity and shortcut cues that limit the LLM's ability to perform relational reasoning, consistent with prior observations in the literature [14].