Query, Don't Train: Privacy-Preserving Tabular Prediction from EHR Data via SQL Queries

Firstname1 Lastname1 [*1] **Firstname2 Lastname2** [*1 2] **Firstname3 Lastname3** [2] **Firstname4 Lastname4** [3] **Firstname5 Lastname5** [1] **Firstname6 Lastname6** [3 1 2] **Firstname7 Lastname7** [2] **Firstname8 Lastname8** [3] **Firstname8 Lastname8** [1 2]

Abstract

Electronic health records (EHRs) contain richly structured, longitudinal data essential for predictive modeling, yet stringent privacy regulations (e.g., HIPAA, GDPR) often restrict access to individual-level records. We introduce **Query, Don't Train** (QDT): a **structured-data foundation-model interface** enabling **tabular inference** via LLM-generated SQL over EHRs. Instead of training on or accessing individual-level examples, QDT uses a large language model (LLM) as a schema-aware query planner to generate privacy-compliant SQL queries from a natural language task description and a test-time input. The model then extracts summary-level population statistics through these SQL queries and the LLM performs, chain-of-thought reasoning over the results to make predictions. "This inference-time–only approach (1) eliminates the need for supervised model training or direct data access, (2) ensures interpretability through symbolic, auditable queries, (3) naturally handles missing features without imputation or preprocessing, and (4) effectively manages high-dimensional numerical data to enhance analytical capabilities. We validate QDT on the task of 30-day hospital readmission prediction for Type 2 diabetes patients using a MIMIC-style EHR cohort, achieving F1 = 0.70, which outperforms TabPFN (F1 = 0.68). To our knowledge, this is the first demonstration of LLM-driven, privacy-preserving structured prediction using only schema metadata and aggregate statistics—offering a scalable, interpretable, and regulation-compliant alternative to conventional foundation-model pipelines.

[*]Equal contribution [1]Department of XXX, University of YYY, Location, Country [2]Company Name, Location, Country [3]School of ZZZ, Institute of WWW, Location, Country. Correspondence to: Firstname1 Lastname1 <first1.last1@xxx.edu>, Firstname2 Lastname2 <first2.last2@www.uk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

EHRs store richly structured, longitudinal data spanning diagnoses, laboratory results, procedures, medications, and outcomes—resources that are critical for predictive modeling and clinical decision support (Kim et al., 2019; Tsai et al., 2025). However, regulations such as the U.S. HIPAA Privacy Rule and the EU GDPR impose strict safeguards for protected health information, including consent, minimization, and access controls, with substantial legal and institutional constraints on data use (Cohen & Mello, 2018; Voigt & Von dem Bussche, 2017). These policies often prohibit direct access to patient-level records, creating significant barriers for model development, particularly in cross-institutional settings where data-sharing agreements are difficult to establish or enforce.

Despite these constraints, public datasets such as MIMIC-III have enabled research in EHR-driven prediction under carefully controlled conditions, supporting tasks such as mortality forecasting, hospital readmission risk, and treatment efficacy modeling (Johnson et al., 2020; Meng et al., 2022). Traditional supervised models—especially tree-based methods like XGBoost—continue to dominate tabular prediction tasks due to their robustness to heterogeneous features, irregular target functions, and missing data (Grinsztajn et al., 2022; Yu et al., 2024). Transformer-based in-context learners, such as TabPFN, offer classification via training-set conditioning, though they still require access to raw examples at inference time (Hollmann et al., 2022).

LLMs have recently demonstrated strong performance in structured reasoning tasks, including text-to-SQL translation (Gao et al., 2023), with execution accuracy exceeding 86% on cross-domain benchmarks like Spider. These capabilities suggest a new opportunity: using LLMs not just for text generation, but for **schema-aware query planning** that operates under privacy constraints. SQL serves as a controlled, interpretable interface that enables LLMs to retrieve

relevant aggregate statistics—without exposing individual-level data—thereby preserving compliance with HIPAA and GDPR (Cohen & Mello, 2018; Voigt & Von dem Bussche, 2017).

In this work, we introduce **Query, Don't Train**, a two-stage, framework for clinical tabular prediction without direct access to raw EHR data. Our approach is grounded in three pillars:

- **Privacy preservation**, by ensuring only policy-compliant SQL queries are issued and no patient-level data is revealed.
- **Structured reasoning**, which derives interpretability from two key sources: (1) LLM-mediated chain-of-thought predictions over query results, and (2) the symbolic, auditable queries themselves.
- **Robustness to missing data**, as the model dynamically selects and conditions on available features at inference without imputation.

We validate our approach on 30-day readmission prediction in a MIMIC-style cohort for Type 2 diabetes patients, showing that it obtains an F1-score of 0.70 while offering interpretability and compliance out of the box.

2. Methodology

2.1. Problem Formulation

We consider a tabular classification task under strict access constraints. Let $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a training set of patient records $x_i \in \mathbb{R}^d$ and associated outcomes $y_i \in \mathcal{Y}$. Direct access to $\mathcal{D}_{\text{train}}$ is prohibited due to regulatory or institutional privacy restrictions. Given a test-time instance x^{test}, the goal is to predict its label y^{test} by interacting with $\mathcal{D}_{\text{train}}$ exclusively via a privacy-compliant SQL interface that enforces data governance policies.

2.2. Framework Overview

Our method adopts a two-stage architecture in which an LLM serves as both a query-generation agent and a predictor through structured reasoning. The process, illustrated in Figure 1, proceeds as follows:

1. **Input:** The LLM receives (i) a natural language prompt describing the prediction task (e.g., "Predict 30-day readmission for Type 2 diabetes"), and (ii) the test-time patient record x^{test}.
2. **Query Generation:** Based on the prompt and x^{test}, the agent generates SQL queries targeting the database containing $\mathcal{D}_{\text{train}}$. These queries are designed to retrieve summary-level statistics (e.g., "average length of stay for similar patients").
3. **Privacy Filtering:** Only queries that comply with predefined privacy constraints (e.g., returning aggregates over groups of at least 2 individuals) are executed.
4. **Query Loop:** The agent may iteratively generate follow-up queries to refine its understanding of relevant cohort-level statistics.
5. **Prediction:** The outputs of the executed queries are returned to the LLM, which uses chain-of-thought reasoning to produce a prediction for y^{test}.

This inference-time-only framework enables structured prediction without accessing raw patient data. The agent implicitly performs dynamic feature selection by deciding which summary statistics to request during the Query Loop.

3. Experiments

3.1. Experimental Setup

We use OpenAI's o4-mini model as the LLM agent in our setup and for the LLM-only baseline. We leverage the LangChain library[1] to implement the agent.

For the privacy policies, we restricted queries via the system prompt to summary-level statistics, which are defined as data averaged over two or more patients. To ensure the queries do not access patient-level information, we have a seperate agent to ensure that only those queries requesting summary-level statistics proceed to execution. In practice, this validation would be enforced by a firewall to prevent unauthorized data access (Kruse et al., 2017).

3.2. Datasets

We focus on predicting 30-day hospital readmissions for patients with Type 2 Diabetes in US hospitals (Clore & Strack, 2014)[2]. The dataset consists of patient records x_i, which include demographics, laboratory results, procedures, and prior admissions, with binary outcome labels $y_i \in \{0, 1\}$.

3.3. Baselines

We compare our method against three baselines: TabPFN (Hollmann et al., 2022) is a pre-trained transformer-based predictor trained to perform tabular

[1] https://python.langchain.com/api_reference/community/agent_toolkits/langchain_community.agent_toolkits.sql.toolkit.SQLDatabaseToolkit.html

[2] https://www.kaggle.com/c/1056lab-diabetes-readmission-prediction/data

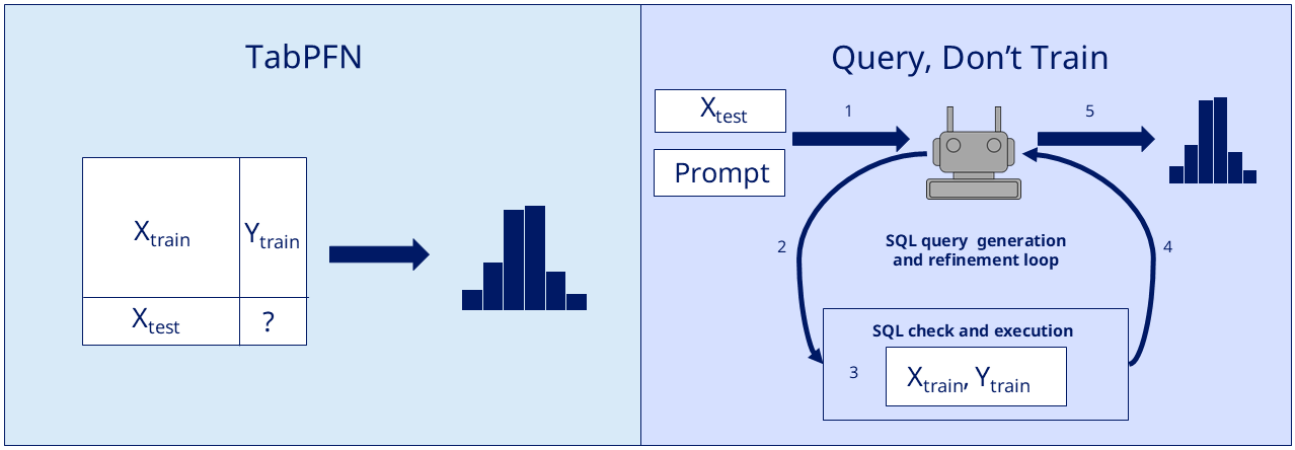

Figure 1. **Comparison of TabPFN and our "Query, Don't Train" (QDT) approach.** TabPFN uses the training set directly during inference. In contrast, QDT follows: (1) receive test record and task prompt, (2) generate SQL queries, (3) enforce compliance with privacy policies, (4) execute approved queries to retrieve summary statistics, (5) predict using chain-of-thought reasoning. QDT enables privacy-preserving, interpretable inference without raw data access.

classification by conditioning on the training set at inference time. It is particularly relevant as it accesses $\mathcal{D}_{\text{train}}$ during inference, similar in spirit to our method, albeit without privacy constraints. XGBoost (Chen & Guestrin, 2016) is a widely-used gradient boosting framework for tabular data. We train XGBoost on $\mathcal{D}_{\text{train}}$ and evaluate it on x^{test}, representing the standard supervised learning baseline with full access to training data. Additionally, we compare our method with an LLM-only baseline that receives only x^{test} and a prompt containing the problem formulation.

3.4. Classification Results

We compare our approach against TabPFN (Hollmann et al., 2022) and XGBoost (Chen & Guestrin, 2016). Despite never accessing the raw data, our method achieves competitive performance in predicting 30-day readmissions, as indicated by the metrics presented in Table 1. Specifically, our Query, Don't Train methodology demonstrates strong precision and recall, underscoring the effectiveness of structured reasoning over aggregate statistics. These results highlight the potential of our approach to provide accurate predictions while utilizing minimal training resources.

3.5. Ablation Study on Missing Features

To investigate the impact of feature availability on model performance, we conducted an ablation study by systematically removing features from x^{test}. The findings illustrate that our method maintains robust performance even with reduced feature sets. When 30% of the features were omitted, the performance metrics showed only a modest decrease in the F1-score, dropping to 0.67. This demonstrates that, despite missing features, the agent effectively utilized the remaining features in x^{test} to identify relevant similar examples, which it uses to reason for accurate predictions. However, with a substantial reduction of 70% of features, the performance was impacted more significantly, resulting in an F1-score of 0.64. These results attempt to solve the challenges posed by incomplete data in real-world EHR scenarios (Yu et al., 2024).

Table 1. Performance comparison of different models on 30-day readmission prediction for Type 2 Diabetes patients predicted for a subset of 2,000 patients. Evaluation metrics include Precision, Recall, and F1-score. Query, Don't Train (QDT) refers to using SQL queries to perform predictions without direct access to patient-level data.

Model	Precision	Recall	F1-score
TabPFN	0.63	**0.76**	0.69
XGBoost	0.65	0.68	0.66
LLM	0.54	0.51	0.52
QDT	**0.68**	0.73	**0.70**
QDT (30% less features)	0.65	0.69	0.67
QDT (70% less features)	0.62	0.65	0.64

4. Conclusion

This work introduces QDT, a new framework that reimagines structured prediction through symbolic interaction rather than model training. Our findings demonstrate that LLMs can serve as foundation models for structured data without requiring access to raw examples or parameter tun-

ing. By pairing LLM-generated SQL queries with cohort-level aggregation and chain-of-thought reasoning, QDT constructs implicit, task-conditioned table representations entirely at inference time. This paradigm offers a practical and conceptually distinct alternative to pretraining: it scales across tasks with no model updates, provides interpretability through auditable query outputs, and complies with privacy regulations by design.

The approach is particularly suited to high-stakes domains like healthcare, where individual-level data is sensitive and institutional data-sharing is often infeasible. QDT offers clear advantages in deployment flexibility, explainability, and robustness to missing data, as the system dynamically selects what to query based on feature availability. These attributes make it a compelling candidate for real-world clinical decision support under strict data governance. While demonstrated in healthcare, this abstraction readily extends to other structured domains such as finance, education, and public policy.

In sum, QDT represents a step toward a new class of foundation model interfaces for structured data—ones that emphasize reasoning over memorization, and symbolic querying over supervised optimization.

5. Limitations and Future Work

Despite these strengths, several limitations warrant discussion. First, the computational efficiency of LLM-driven query generation and execution remains an open question, especially as prediction tasks grow in complexity or require more sophisticated querying strategies. Second, while our experiments focus on structured tabular data, extending this framework to multi-modal EHRs (e.g., incorporating imaging or unstructured clinical notes) may require additional innovations in prompt engineering and query design.

The privacy constraints we implement allow access only to aggregated results for two or more patients. These constraints can be adjusted to enforce stricter censoring policies, and more fine-grained privacy-preserving mechanisms can be incorporated as needed.

Another consideration is the potential for adversarial or suboptimal queries generated by LLMs. Ensuring the reliability and safety of the query-generation process, particularly in high-stakes clinical settings, is an important direction for future work. Additionally, while our method is evaluated on cohorts in US hospitals, broader validation across diverse institutions and healthcare systems is necessary to fully establish generalizability.

References

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

Clore, John, C. K. D. J. and Strack, B. Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5230J.

Cohen, I. G. and Mello, M. M. Hipaa and protecting health information in the 21st century. *Jama*, 320(3):231–232, 2018.

Gao, D., Wang, H., Li, Y., Sun, X., Qian, Y., Ding, B., and Zhou, J. Text-to-sql empowered by large language models: A benchmark evaluation. *arXiv preprint arXiv:2308.15363*, 2023.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. Mimic-iv. *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, pp. 49–55, 2020.

Kim, E., Rubinstein, S. M., Nead, K. T., Wojcieszynski, A. P., Gabriel, P. E., and Warner, J. L. The evolving use of electronic health records (ehr) for research. In *Seminars in radiation oncology*, volume 29, pp. 354–361. Elsevier, 2019.

Kruse, C. S., Smith, B., Vanderlinden, H., and Nealand, A. Security techniques for the electronic health records. *Journal of medical systems*, 41:1–9, 2017.

Meng, C., Trinh, L., Xu, N., Enouen, J., and Liu, Y. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1): 7166, 2022.

Tsai, M.-L., Chen, K.-F., and Chen, P.-C. Harnessing electronic health records and artificial intelligence for enhanced cardiovascular risk prediction: A comprehensive review. *Journal of the American Heart Association*, 14 (6):e036946, 2025.

Voigt, P. and Von dem Bussche, A. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676): 10–5555, 2017.

Yu, Z., Chu, X., Jin, Y., Wang, Y., and Zhao, J. Smart: Towards pre-trained missing-aware model for patient health status prediction. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 63986–64009. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/751ef1e7f557a8a88f0837b61bf5070f-Paper-Conference.pdf.