An Open-Source Text-to-SQL Dataset over OMOP-Formatted MIMIC-IV

Paul Legrand[1], Kawsar Noor[1], Satyam Bhagwanani[1], and Richard J Dobson[1, 2]

[1]University College London, London, United Kingdom

[2]King's College London, London, United Kingdom

1. Introduction

Electronic Health Records (EHRs) contain valuable clinical information that are often well structured and essential for informing research and downstream applications. However, end-users of the records, such as clinicians, often lack the expertise to interact and query the record effectively. A emerging solution to this problem is the use of Text-to-SQL models which enable clinicians to query the records in natural language. To date Text-to-SQL models have focused on open-source datasets and the resulting models are not directly deployable at hospitals since they are built around schemas not in use at hospitals.

In this paper we propose an open-source Text-2-SQL dataset based on the OMOP (Observational Medical Outcomes Partnership) Common Data Model (CDM). OMOP provides a standardized way to structure and harmonize healthcare data across institutions, facilitating large-scale analytics, interoperability, and reproducible research. As such the OMOP model is a popular format for storing and sharing data between healthcare providers internationally. Our dataset contains natural language question and corresponding SQL pairs based on MIMC-IV transformed into the OMOP data standard. We extend existing work from the EHRSQL [1] templates to this schema, generate synthetic answerable examples using GPT-4 to ensure coverage, and paraphrase them for linguistic diversity. We will also incorporate recent unanswerable examples [2], enabling robust assessment of model trustworthiness. Our dataset covers unanswerable questions, which are questions that cannot be answered using the underlying database and need external knowledge, (e.g. *Does patient 64983 have any siblings?*). Including unanswerable questions is critical for evaluating whether models can recognize their own limitations and avoid generating unreliable queries.

2. Methods

An overview of the dataset generation process is shown in Figure 1.

Database Preprocessing and Sampling Before initiating data generation, we will preprocess the OMOP-formatted MIMIC-IV database by first sampling a representative subset of patient records. This subset will be selected to capture the diversity of clinical events and value distributions across the database. Inspired by EHRSQL [1], we will then apply a two-step anonymization process: (1) we will shift all time-related fields to a consistent window between 2100 and 2105 to simulate realistic time-sensitive questions and facilitate relative time expressions; and (2) we will randomly shuffle clinical values (e.g., medications, diagnoses, procedures) across patients to reinforce de-identification while preserving schema relationships.

Schema-Aligned SQL Template Construction We will first deploy the MIMIC-IV dataset into a PostgreSQL database following the OMOP CDM structure. From the original 174 answerable question templates in EHRSQL [1], we will extract those that can be adapted to OMOP and rework their corresponding SQL queries accordingly. To ensure that every OMOP table is represented, we will use GPT-4 to generate additional synthetic templates targeting underrepresented or uncovered tables. These templates will be manually reviewed and aligned with the OMOP schema.

Incorporating and Debiasing Unanswerable Questions To evaluate model reliability, we will include unanswerable questions from the original EHRSQL [1], as well as additional unanswerable

questions from the EHRSQL 2024 [2] dataset, which were originally sourced from the TrustSQL benchmark [4]. These include questions referencing unavailable data, requiring external knowledge, or containing schema-incompatible elements. Following the methodology proposed by Yang et al. (2024) [3], we will perform N-gram analysis to identify linguistic patterns disproportionately present in unanswerable questions. Based on this analysis, we will reconstruct the validation and test sets to reduce data bias—specifically, by assigning high-ratio N-gram patterns to the test set while preserving a more neutral validation set.

Temporal Normalization and De-identification To support time-sensitive queries and enhance data privacy, we will shift all timestamps in the database to fall within the range of 2100–2105. We will also shuffle clinical values across patients—such as medications, diagnoses, and procedures—to ensure full de-identification while maintaining relational consistency across tables.

Generation of Question-SQL Pairs Once the SQL templates are finalized, we will implement a data generation pipeline to populate them with realistic values drawn from the OMOP database. This will include filling patient IDs, dates, drug names, diagnosis codes, and other relevant fields. Each generated (question, SQL) pair will be validated through actual SQL execution to ensure syntactic and semantic correctness.

Paraphrasing for Linguistic Diversity To introduce linguistic variation and simulate real-world phrasing, we will generate paraphrases for all answerable questions using GPT-based models. Each template will be expanded into several paraphrased forms while preserving the underlying semantics. We will apply a similar approach to unanswerable questions to prevent overfitting to surface-level features and improve robustness in downstream model training.

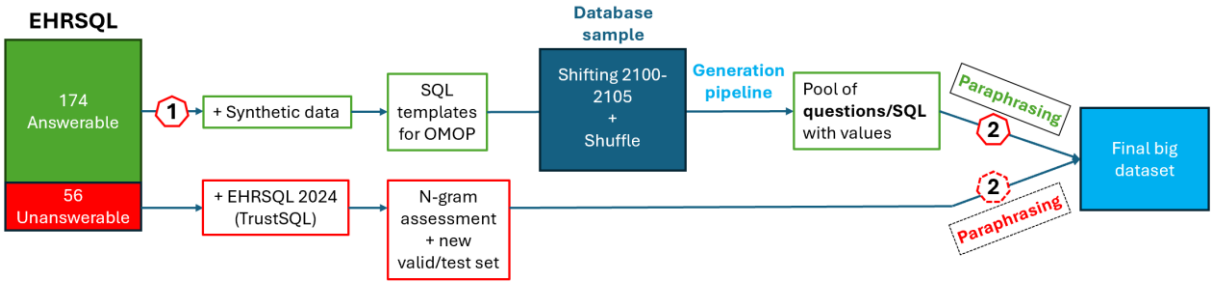

Figure 1. Dataset construction workflow combining EHRSQL [1] templates, OMOP-specific synthetic data, database preprocessing, question-SQL generation, and paraphrasing.

3. Results

To date, we have been working on generating a dataset of at least 10,000 question-SQL pairs, including both answerable and unanswerable examples, with full OMOP schema coverage and paraphrastic diversity. Planned evaluations include fine-tuning autoregressive models and assessing performance using SQL execution accuracy (F1exe), answerability classification (F1ans), and related metrics. We also aim to explore uncertainty-based refusal strategies. This work is expected to result in the first OMOP-specific text-to-SQL dataset for MIMIC-IV, supporting the development of reliable and schema-aware clinical QA systems.

4. Conclusion

This work proposes a new dataset and data generation pipeline for training and evaluating text-to-SQL models on OMOP-formatted MIMIC-IV data. By combining adapted and synthetic templates, schema coverage, paraphrased variants, and unanswerable questions, we aim to support the development of robust and trustworthy models. Informed by recent findings on data bias, we incorporate N-gram analysis to improve the evaluation of answerability. This dataset is expected to facilitate more accurate, schema-aware, and reliable clinical QA systems for real-world EHR applications.

References

[1] Lee, Gyubok, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 'EHRSQL: A Practical Text-to-SQL Benchmark for Electronic Health Records'. arXiv, 25 December 2023. https://doi.org/10.48550/arXiv.2301.07695.

[2] Lee, Gyubok, Sunjun Kweon, Seongsu Bae, and Edward Choi. 'Overview of the EHRSQL 2024 Shared Task on Reliable Text-to-SQL Modeling on Electronic Health Records'. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 644–54. Mexico City, Mexico: Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.clinicalnlp-1.62.

[3] Yang, Yongjin, Sihyeon Kim, SangMook Kim, Gyubok Lee, Se-Young Yun, and Edward Choi. 'Towards Unbiased Evaluation of Detecting Unanswerable Questions in EHRSQL'. arXiv, 29 April 2024. https://doi.org/10.48550/arXiv.2405.01588.

[4] Lee, Gyubok, Woosog Chay, Seonhee Cho, and Edward Choi. 'TrustSQL: Benchmarking Text-to-SQL Reliability with Penalty-Based Scoring'. arXiv, 2 July 2024. https://doi.org/10.48550/arXiv.2403.15879.