An Open-Source Text-to-SQL Pipeline for OMOP-Formatted Electronic Health Records

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

As more hospitals transition to recording data in electronic healthcare record (EHR) systems opportunities for fast and efficient analysis of the data are improving. However, in some cases, the end-users of this data, such as clinicians, lack the technical skills and knowledge to interact efficiently with the data and often depend on data analysts to perform this task for them. In particular querying the structured portion of the record most often always requires knowledge of SQL. This outsourcing of technical skills inevitably introduces bottlenecks to performing research and meeting deliverables. Recognising this problem, there have been ongoing efforts to investigate the possibility of a Text-to-SQL interface that enables clinicians to query their data in natural language. To date these systems have been focused on open-source datasets which use schemas and produce models that are not directly deployable to hospitals.

In this work, we propose the development and evaluation of an enhanced Text-to-SQL pipeline specifically adapted for the OMOP schema using the MIMIC-IV dataset. The OMOP schema is an internationally recognised standard, and many hospitals are already transforming their records into this format for multi-site collaborations. We will first leverage and adapt existing question templates from the EHRSQL and EHRSQL 2024 datasets, generating a novel, paraphrased dataset using GPT-4 to closely mirror realistic user queries. Additionally, recognizing the critical importance of managing unanswerable questions—questions that cannot be translated into a valid SQL query because they are either incompatible with the database schema or require external domain knowledge (e.g., *Does patient 64983 have an appointment for any test today?*) [12]—we aim to augment and refine this aspect of the dataset.

Our proposed pipeline integrates advanced entity recognition and linking tools such as MedCAT [11] to map medical entities to standardized vocabularies (e.g., SNOMED, UMLS). Using LlamaIndex, we will enable schema-aware query generation, supported by Retrieval-Augmented Generation (RAG) to enrich model understanding with relevant SQL examples. We will evaluate various fine-tuning and prompt engineering methods—including Parameter-Efficient Fine-Tuning (PEFT) and Self-Consistency techniques—to optimize query accuracy. To address unanswerable queries, we will explore rejection strategies using intent-driven prompts, entropy-based uncertainty estimation, and validation against sampled database queries.

2. Methods

Our approach focuses on developing a reliable text-to-SQL pipeline tailored to the OMOP schema (see Figure 1). For pipeline development, the process will start with user input, where users will enter their queries alongside explicit intent descriptions. Entity recognition will be implemented using MedCAT [11], a state-of-the-art clinical entity extraction tool, to detect and map entities accurately to standard medical vocabularies such as SNOMED and UMLS. To handle the database schema, we will utilize LlamaIndex to establish a schema-aware indexing mechanism, optimizing the clarity and effectiveness of prompts.

Extensive experimentation will be conducted during prompt engineering [1, 2, 6, 7, 10], comparing multiple prompting strategies to determine optimal configurations. Retrieval-Augmented Generation (RAG) [2] methods will be systematically evaluated, including approaches such as retrieving top-1 and top-2 closest SQL examples and exploring combinations with two different embedding models [2] to ascertain their impact on SQL generation accuracy.

SQL query generation itself will be carried out using a fine-tuned Large Language Model (LLM) as SQLCoder-7b-2 [6, 10]. To enhance query reliability, we will rigorously evaluate several approaches including Ensemble methods [2, 7], Self-Consistency checks [7], auto-verification techniques, and cross-validation using secondary models [4, 7]. Fine-tuning strategies will also be explored in depth, particularly Parameter-Efficient Fine-Tuning (PEFT) and Self-taught learning loops [3], aiming to identify the most efficient method to achieve high accuracy.

To address the critical task of detecting and handling unanswerable queries, a specialized rejection module will be implemented and tested [4, 9, 10]. Techniques assessed will include immediate detection using a separately fine-tuned classifier [8], uncertainty measures such as log probabilities [8, 10] or entropy-based metrics [2, 3, 5, 8] during query generation, explicit incorporation of user intent in prompts, and practical validation through executing queries [8, 9, 10] against representative samples from the database.

Finally, the impact of user query formulation on system performance will be evaluated through various preprocessing approaches. These will include direct usage of raw user queries, conversion of queries into standardized formats, and the utilization of template-based transformations prior to SQL generation, assessing each approach for effectiveness in terms of accuracy and generalizability.

3. Results

To date, we have been working on designing a robust evaluation pipeline with clearly defined metrics. For unanswerable question detection, we plan to quantify performance using F1-answerable [12] and a reliability score rated on a 0–5–10 scale [13]. For SQL query generation accuracy, we are incorporating metrics such as F1-execution [12], execution accuracy, and exact set matching accuracy. These evaluations will allow us to compare different pipeline configurations and help identify the most effective approach for accurately translating clinical questions into executable SQL queries.

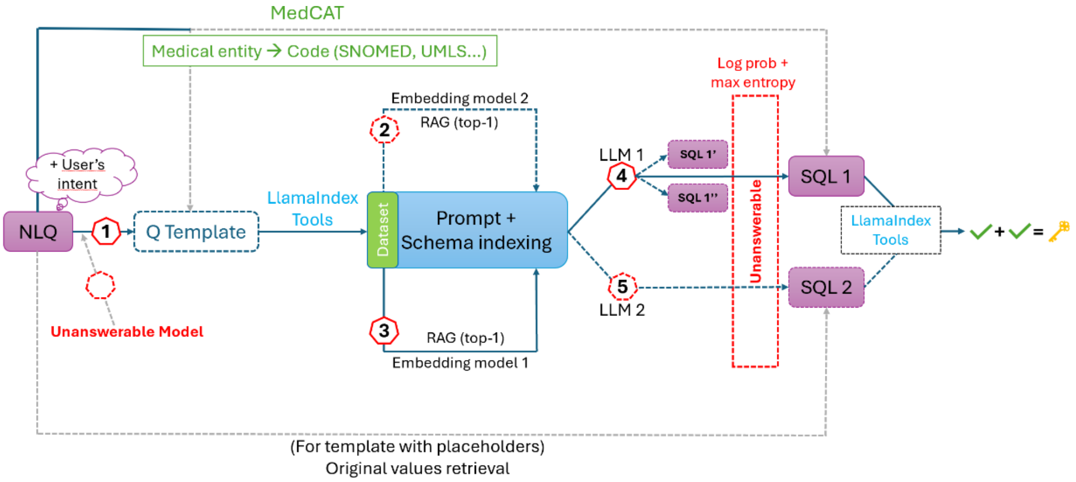

Figure 1. Overview of the proposed pipeline for robust Text-to-SQL generation. The pipeline integrates medical entity recognition (MedCAT [11]), schema-aware prompting (LlamaIndex), Retrieval-Augmented Generation (RAG) techniques, fine-tuned language models (LLMs), and uncertainty-based rejection strategies to handle unanswerable queries.

4. Conclusion

This extended abstract outlines a structured approach to building a reliable Text-to-SQL pipeline adapted to the OMOP schema. By evaluating diverse NLP techniques, prompt engineering strategies, and rejection methodologies, we aim to significantly improve query accuracy and reliability. The anticipated results will guide optimal configuration choices, enhancing clinical data accessibility and supporting more interpretable, accurate decision-making. To date, my primary focus has been generating the dataset required for these experiments.

References

[1] Gao, Dawei, Haibin Wang, Yaliang Li, Xiuyu Sun, Yichen Qian, Bolin Ding, and Jingren Zhou. ‘Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation’. arXiv, 20 November 2023. https://doi.org/10.48550/arXiv.2308.15363.

[2] Gundabathula, Satya K., and Sriram R. Kolar. ‘PromptMind Team at EHRSQL-2024: Improving Reliability of SQL Generation Using Ensemble LLMs’. arXiv, 14 May 2024. https://doi.org/10.48550/arXiv.2405.08839.

[3] He, Mingqian, Yongliang Shen, Wenqi Zhang, Qiuying Peng, Jun Wang, and Weiming Lu. ‘STaR-SQL: Self-Taught Reasoner for Text-to-SQL’. arXiv, 19 February 2025. https://doi.org/10.48550/arXiv.2502.13550.

[4] Jabir, Mohammed, Kamal Kanakarajan, and Malaikannan Sankarasubbu. ‘Saama Technologies at EHRSQL 2024: SQL Generation through Classification Answer Selector by LLM’. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 655–71. Mexico City, Mexico: Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.clinicalnlp-1.63.

[5] Jo, Yongrae, Seongyun Lee, Minju Seo, Sung Ju Hwang, and Moontae Lee. ‘LG AI Research & KAIST at EHRSQL 2024: Self-Training Large Language Models with Pseudo-Labeled Unanswerable Questions for a Reliable Text-to-SQL System on EHRs’. arXiv, 18 May 2024. https://doi.org/10.48550/arXiv.2405.11162.

[6] Joy, Sourav, Rohan Ahmed, Argha Saha, Minhaj Habil, Utsho Das, and Partha Bhowmik. ‘Project PRIMUS at EHRSQL 2024 : Text-to-SQL Generation Using Large Language Model for EHR Analysis’. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 422–27. Mexico City, Mexico: Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.clinicalnlp-1.41.

[7] Kim, Hajung, Chanhwi Kim, Hoonick Lee, Kyochul Jang, Jiwoo Lee, Kyungjae Lee, Gangwoo Kim, and Jaewoo Kang. ‘KU-DMIS at EHRSQL 2024:Generating SQL Query via Question Templatization in EHR’. arXiv, 19 June 2024. https://doi.org/10.48550/arXiv.2406.00014.

[8] Kim, Sangryul, Donghee Han, and Sehyun Kim. ‘ProbGate at EHRSQL 2024: Enhancing SQL Query Generation Accuracy through Probabilistic Threshold Filtering and Error Handling’. arXiv, 25 April 2024. https://doi.org/10.48550/arXiv.2404.16659.

[9] Somov, Oleg, Alexey Dontsov, and Elena Tutubalina. ‘AIRI NLP Team at EHRSQL 2024 Shared Task: T5 and Logistic Regression to the Rescue’. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 431–38. Mexico City, Mexico: Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.clinicalnlp-1.43.

[10] Thomas, Jerrin, Pruthwik Mishra, Dipti Sharma, and Parameswari Krishnamurthy. ‘LTRC-IIITH at EHRSQL 2024: Enhancing Reliability of Text-to-SQL Systems through Abstention and Confidence Thresholding’. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 697–702. Mexico City, Mexico: Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.clinicalnlp-1.66.

[11] Kraljevic, Zeljko, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, et al. ‘Multi-Domain Clinical Natural Language Processing with MedCAT: The

Medical Concept Annotation Toolkit’. arXiv, 25 March 2021. https://doi.org/10.48550/arXiv.2010.01165.

[12] 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.

[13] 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.