SM3-Text-to-Query: Synthetic Multi-Model Medical Text-to-Query Benchmark – Supplementary Materials –

Sithursan Sivasubramaniam*, **Cedric Osei-Akoto***, **Yi Zhang**,
Kurt Stockinger, **Jonathan Fürst**†
Zurich University of Applied Sciences
{sivassit,oseiaced}@students.zhaw.ch, {zhay,stog,fues}@zhaw.ch

In this supplementary material, we mainly provide a more detailed description of our dataset and benchmark following the "Datasheet for Dataset" guidelines [3]. The data and code of our benchmark can be accessed at `https://github.com/jf87/SM3-Text-to-Query`. The repository includes further documentation on how to use our benchmark and replicate its results. The dataset is published under the CC BY-ND 4.0 license.

A Author Statement

We bear all responsibilities for licensing, distributing, and maintaining our dataset.

B Licensing

The proposed dataset is under the CC BY-ND 4.0 license, while the code in the repository is released under the Apache 2.0 license.

C Datasheet for Datasets

We follow the documentation framework provided by "Datasheet for Datasets" [3] to answer the important questions considering this dataset.

C.1 Motivation

For what purpose was the dataset created? The dataset was created to evaluate Text-to-Query systems (e.g., Text-to-SQL, Text-to-SPARQL) across multiple query languages for the same underlying data. This is increasingly important as Text-to-Query systems have evolved from dedicated, specialized systems for a single query language (e.g., SQL) to systems that employ Large Language Models (LLMs) with in-context-learning strategies. The reliance on in-context learning and LLMs makes it much easier to solve Text-to-Query tasks in general and not just only for a single query language such as SQL. Our dataset is the first dataset that enables a fair evaluation across four important query languages (SQL, SPARQL, Cypher, MQL) in this context. We selected a medical data scenario due to an ongoing digital health project [1], but also due to the prevalence of international standards of medical ontologies (SNOMED [7]) and the general importance of this sector. Therefore, our dataset has the potential to have a positive impact on the development of better natural language interfaces in practice.

∗ Equal contribution.
† Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The dataset was created by Zurich University of Applied Sciences at the Institute of Computer Science. The idea to create the dataset was inspired by the Digital Health Zurich project, in which we strongly collaborated with the University Hospital of Zurich. This dataset addresses an important gap for enabling state-of-the-art natural language processing research on synthetic medical datasets where no individual patient personal data is exposed. Moreover, we saw the need for a new Text-to-Query dataset for multiple database models (relational, document, graph) and query languages (SQL, SPARQL, Cypher, MQL) based on our previous works [2, 4].

Who funded the creation of the dataset? The dataset construction has been partially self-funded, partly based on the work of two undergraduate students, and partially paid by a grant from the Digital Health Zurich project for senior searchers. The OpenAI experiments have been supported by an OpenAI Research grant.

Any other comments? No.

C.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? There are three parts of the dataset:

- The used data is based on output generated using Synthea [6], a data generator for synthetic patient data. We chose synthetic data due to the privacy and ethical issues involved with real medical data. Further, with Synthea, the dataset could be adapted to different patient populations to better reflect the distributions in various countries.
- We provide databases including data schemas for four systems based on this data: PostgreSQL database (relational database), Cypher database (graph database), SPARQL database (graph database), and MongoDB database (document store).
- Based on 408 template questions, we create train, dev, and test data for all four query languages (SQL, SPARQL, Cypher, MQL) of 6K train, 2K dev, and 2K test (overall 40K).

How many instances are there in total (of each type, if appropriate)? Table 1 summarizes the Synthea dataset that is the base for SM3-Text-to-Query databases. The data contains 110 patients (10 of which are deceased) and all related entities and relationships (overall 272,817). We selected this specific data size as a good compromise between reasonable query execution time and the hardware requirements of our various database systems while still being large enough to augment our query templates. We implemented Extract Transform and Load (ETL), i.e. database loading techniques, to process and transform the original data to our four target databases.

We create overall 40K Text-to-Query pairs (see Table 2).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? The database is based on synthetic data generation. It currently contains 110 patients. However, our approach is scalable and allows for generating 1K, 10K, 100K, or even more patients. The Text/Query pairs are based on 408 question templates, which could be extended in the future.

What data does each instance consist of? The actual data's attributes depend on the type of entity/relationship. Usually, each entity has a unique identifier (UUID) and several lexical and numerical values describing medical assessments, measurements, or other medical-related data. Appendix A, provided in the main paper PDF, contains visualizations of data models, including attributes. Furthermore, the data can be inspected on the shared repository.

For the annotated Text-to-Query data, the instances consist of the natural language question, the type of question, the associated entities, the expert-annotated SQL query, the expert-annotated SPARQL query, the expert-annotated Cypher query, the expert-annotated MQL query, and the retrieved results from four databases, respectively.

Is there a label or target associated with each instance? The label is the expert-annotated query for the respective query language (SQL, SPARQL, Cypher, MQL). A second person has cross-checked the expert-annotated query.

Is any information missing from individual instances? No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? We make the relationship of the question category explicit through the applied entity labels.

Are there recommended data splits (e.g., training, development/validation, testing)? We split the data into 6K train, 2K dev, and 2K test (see also Table 2). The test data will be kept hidden to allow for a fair evaluation of Text-to-Query systems as it is common practice for other Text-to-Query benchmarks [8, 5] .

Are there any errors, sources of noise, or redundancies in the dataset? Our annotation procedure involves cross-checking the queries with a second expert to ensure data quality. Our team is committed to enhancing the data even after this paper is accepted. In addition, we encourage users to provide feedback and report errors on our data website, allowing us to rectify and enhance the dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? It is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? No, the dataset is synthetic.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

Does the dataset identify any sub-populations (e.g., by age, gender)? The dataset contains age and gender, but as stated, the data is purely synthetic. The goal is to test the capabilities of Text-to-Query systems.

Table 1: Overview of created instances of various entities and relationships

Entities/Relationships	No. Values	Percentage
medications	8,346	3.06%
providers	280	0.10%
patient expenses	1,123	0.41%
payer transitions	1,123	0.41%
imaging studies	41	0.02%
supplies	1,322	0.48%
payers	10	0.00%
claims	16,122	5.91%
allergies	64	0.02%
procedures	14,688	5.38%
organizations	280	0.10%
conditions	3,886	1.42%
careplans	388	0.14%
encounters	7,776	2.85%
devices	235	0.09%
immunizations	1,571	0.58%
claims transactions	110,612	40.54%
patients	110	0.04%
observations	104,840	38.43%
Total	**272,817**	**100%**

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? No, the dataset is synthetic.

Does the dataset contain data that might be considered sensitive in any way? No, the dataset is synthetic.

Any other comments? No.

C.3 Collection Process

How was the data associated with each instance acquired? The original data is based on a synthetic data generator. Next, we created natural language/query-pairs for four different databases, as described in the main part of the paper.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? See above.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? The starting point is a synthetic data generator. Our training/dev/test dataset generator can be used to upscale the dataset. In other words, we do not apply downsampling from an existing dataset; rather, we apply upsampling using a template-based approach.

Who was involved in the data collection process (e.g., students, crowd workers, contractors), and how were they compensated (e.g., how much were crowd workers paid)? Undergraduate students and senior researchers collected and prepared the data. We did not use crowd workers or contractors. The senior researchers were partially funded by the university and partially by a grant from the Canton of Zurich, Switzerland. In short, all participants received proper compensation aligned with the academic salaries of the Canton of Zurich, Switzerland.

Over what timeframe was the data collected? The data was collected between October 2023 and June 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? No, the dataset is synthetic.

Did you collect the data from the individuals in question directly or obtain it via third parties or other sources (e.g., websites)? Not applicable since the dataset is synthetic.

Were the individuals in question notified about the data collection? Not applicable since the dataset is synthetic.

Did the individuals in question consent to the collection and use of their data? Not applicable since the dataset is synthetic.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? Not applicable since the dataset is synthetic.

Table 2: Overview of created Text-to-Query instances.

Query Language	Train	Dev	Test	Total
SQL	6,000	2,000	2,000	10,000
SPARQL	6,000	2,000	2,000	10,000
Cypher	6,000	2,000	2,000	10,000
MQL	6,000	2,000	2,000	10,000
Total	**24,000**	**8,000**	**8,000**	**40,000**

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? Not applicable since the dataset is synthetic.

Any other comments? No.

C.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? We labeled the natural language questions with an expert-annotated query in SQL, SPARQL, Cypher, and MQL. A second expert then verified the correctness of the annotated query.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Yes, all data is available since the original source is a synthetic dataset.

Is the software that was used to preprocess/clean/label the data available? Yes it is available through our software repository.

Any other comments? No.

C.5 Uses

Has the dataset been used for any tasks already? No, it is a new dataset.

Is there a repository that links to any or all papers or systems that use the dataset? No, it is a new dataset.

What (other) tasks could the dataset be used for? The dataset could be used for knowledge graph construction and completion, given a standardized medical ontology. However, the dataset could also be used for entity matching, query optimization, prediction of certain medical outcomes, etc.

In short, since the dataset can easily be extended and enriched with other medical data, there are many opportunities to use it for the development of medical studies. Algorithms could be tested on synthetic datasets without exposing any risk of leaking confidential patient data. When working with real medical data, the synthetic dataset could be used to further enrich the real dataset for testing corner cases of novel machine learning algorithms under controlled conditions.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? We are not aware of any of these since the data is synthetic.

Are there tasks for which the dataset should not be used? The data should not be used for predicting specific medical outcomes for single people in scenarios of "personalized health" since the data is synthetic. However, the data can be used to make machine learning algorithms more robust for general settings.

Any other comments? No.

C.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? The dataset will be made publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Via GitHub (`https://github.com/jf87/SM3-Text-to-Query`).

When will the dataset be distributed? With the camera-ready version of the paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? The dataset will be made publicly available (CC BY-ND 4.0 license). Our NeurIPS paper needs to be referenced when the dataset is used.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

Any other comments? No.

C.7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The dataset will be supported, hosted, and maintained as part of the multi-year project Digital Health Zurich funded by the Canton of Zurich, Switzerland [1].

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Via email: jonathan.fuerst@zhaw.ch

Is there an erratum? Currently, there is no erratum, but we will include one based on the feedback and possible bug fixes that we hope to receive when the dataset is public and used by other researchers.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Yes, the dataset will be updated.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? Not applicable since the data is synthetic.

Will older versions of the dataset continue to be supported/hosted/maintained? Yes. We will keep several versions to perform comparative studies.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes, we will provide information for contributions.

Any other comments? No.

References

[1] DIZH. Digital Health Zurich - A Practice Lab for Patient-Centred Clinical Lnnovation, 2024. URL `https://dizh.ch/en/2022/07/07/zurich-applied-digital-health-center-2/`.

[2] J. Fürst, C. Kosten, F. Nooralahzadeh, Y. Zhang, and K. Stockinger. Evaluating the data model robustness of text-to-sql systems based on real user queries. *arXiv preprint arXiv:2402.08349*, 2024.

[3] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[4] C. Kosten, P. Cudré-Mauroux, and K. Stockinger. Spider4SPARQL: A Complex Benchmark for Evaluating Knowledge Graph Question Answering Systems. In *2023 IEEE International Conference on Big Data (BigData)*, pages 5272–5281. IEEE, 2023.

[5] J. Li, B. Hui, G. Qu, J. Yang, B. Li, B. Li, B. Wang, B. Qin, R. Geng, N. Huo, et al. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] H. G. MITRE Corporation. Synthea-international, 2024. URL `https://github.com/synthetichealth/synthea-international/tree/master`. International Demographic Github Repository.

[7] SNOMED International. SNOMED, 05 2024. URL `https://www.snomed.org/`.

[8] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, 2018.