School of Engineering

# SM3-Text-to-Query: Synthetic Multi-Model Medical Text-to-Query Benchmark

*NeurIPS24 Dataset & Benchmark Track*

Zurich University of Applied Sciences, Switzerland

Sithursan Sivasubramaniam
s.sithursan24@gmail.com

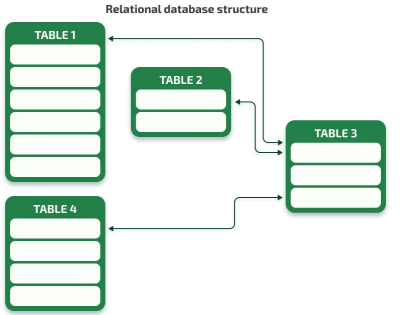
Cedric Osei-Akoto
cedricoseiakoto@gmail.com

Yi Zhang
yi.zhang@zhaw.ch

Kurt Stockinger
kurt.stockinger@zhaw.ch

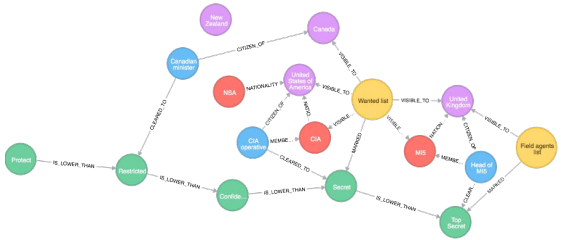
Jonathan Fürst
jonathan.fuerst@zhaw.ch

# The Modern Healthcare Challenge




Relational Model

Document Model



Graph Model

→ Text-to-Query Systems (e.g., Text-to-SQL) can provide medical experts access to health data in an intuitive way.

# Differences Between Database Models and Query Languages

**#Tokens:** 25; **#Keywords:** 8;
**#Joins/Traversals**: 2; **Nesting Depth**: 0

**SQL Query:**

```
SELECT DISTINCT p.first, p.last
FROM organizations org
LEFT JOIN encounters e ON org.id=e.organization
LEFT JOIN patients p ON e.patient=p.id
WHERE org.name='ROYAL OF FAIRHAVEN NURSING CENTER';
```

**#Tokens:** 60; **#Keywords:** 8;
**#Joins/Traversals**: 1; **Nesting Depth**: 5

**MongoDB Query:**

```
db.organizations.aggregate([
    { $match: { "NAME": "ROYAL OF FAIRHAVEN NURSING CENTER" } },
    { $lookup: { from: "patients", localField: "ORGANIZATION_ID",
      foreignField: "ENCOUNTERS.ORGANIZATION_REF", as: "op" } },
    { $unwind: "$op" },
    { $unwind: "$op.ENCOUNTERS" },
    { $match: { "NAME": "ROYAL OF FAIRHAVEN NURSING CENTER" } },
    { $group: { _id: {first: "$op.FIRST", last: "$op.LAST"}} },
    { $project: { _id: 0, first: "$_id.first", last: "$_id.last"} }])
```

"Provide me the names of patients that are linked with the organization Royal of Fairhaven Nursing Center."

**SPARQL Query:**

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX syn: <https://knacc.umbc.edu/dae-young/kim/ontologies/synthea#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX pl: <http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral>
SELECT DISTINCT ?first ?last
WHERE {
    ?organization a syn:Organization;
        syn:id ?organizationId; syn:name 'ROYAL OF FAIRHAVEN NURSING
CENTER'^^pl:.
    ?encounter a syn:Encounter;
        syn:organizationId ?organizationId;
        syn:patientId ?patientId.
    ?patient a syn:Patient;
        syn:id ?patientId; syn:first ?first;
        syn:last ?last. }
```

**Cypher Query:**

```
MATCH (o:Organization {name: 'ROYAL OF FAIRHAVEN NURSING CENTER'})-
[:IS_PERFOMED_AT]->(e:Encounter)<-[:HAS_ENCOUNTER]-(p:Patient)
RETURN DISTINCT p.firstName, p.lastName
```

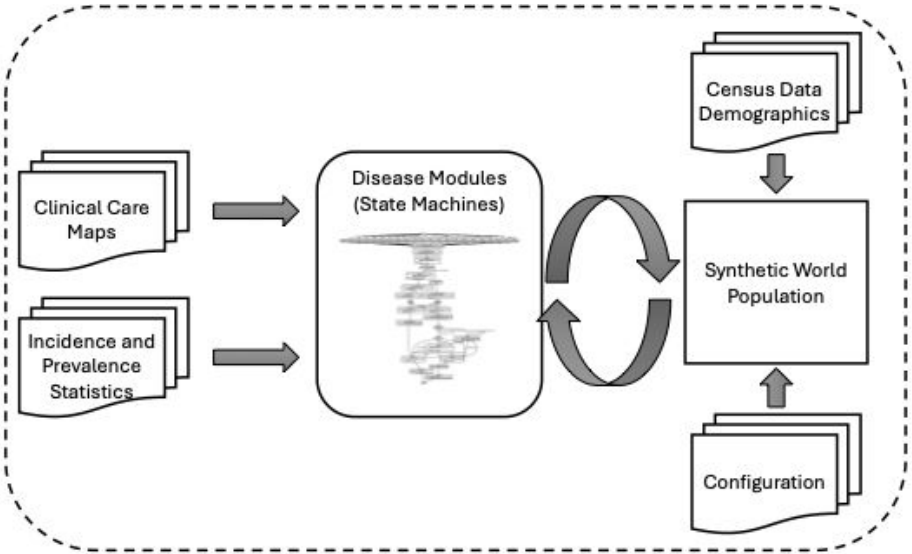
**#Tokens:** 12; **#Keywords: 3**;
**#Joins/Traversals**: 3; **Nesting Depth**: 1

**#Tokens:** 49; **#Keywords:** 8;
**#Joins/Traversals**: 4; **Nesting Depth**: 1

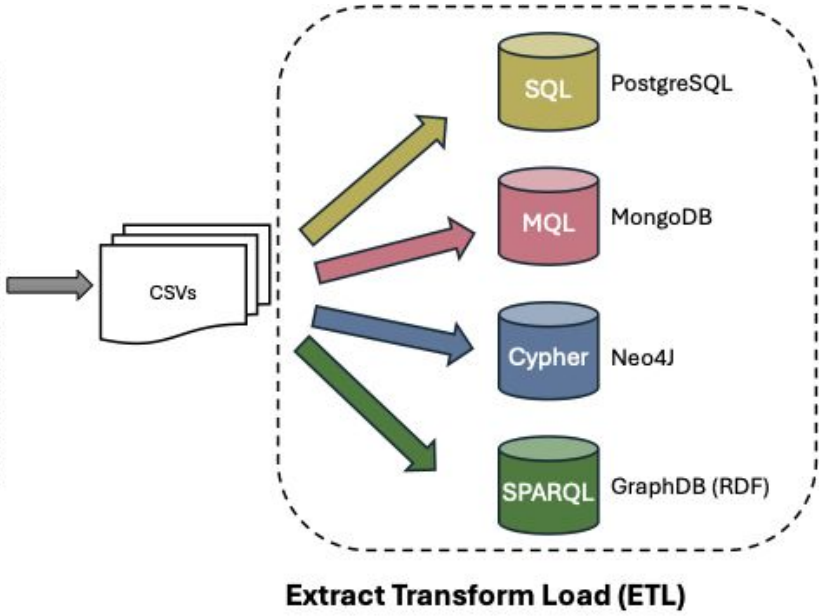
- Query languages differ greatly in their characteristics.
- No existing dataset across these different query languages.

**SM3-Text-to-Query closes that gap by providing the first benchmark across four query languages!**

# SM3 Dataset Construction




**Synthetic Patient Data with Synthea**




**Extract Transform Load (ETL)**

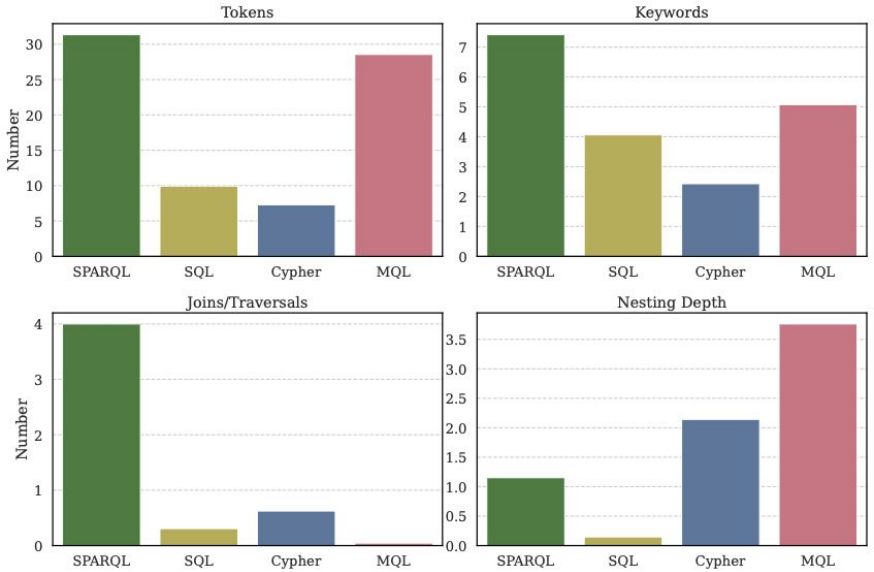
Privacy Preserving ✅

Standard Based ✅

# SM3 Dataset Analysis

**10K text/query pairs per query language**

- SPARQL: Longest and most complex queries.
- Cypher: Most compact queries.
- Language Differences: Variations in joins and nesting highlight complexity across query languages.



# SM3 Dataset Baseline Results

| Models | without schema | | with schema | | |
|---|---|---|---|---|---|
| | w/o schema 1-shot | w/o schema 5-shot | w/ schema 0-shot | w/ schema 1-shot | w/ schema 5-shot |
| ***SQL (PostgreSQL)*** | | | | | |
| Llama3-8b | 4.20 (±5.6) | 10.81 (±9.89) | 22.55 | 23.27 (±1.05) | 27.49 (±15.27) |
| Gemini 1.0 Pro | 4.47 (±4.88) | 21.65 (±11.10) | 38.60 | 38.37 (±3.31) | 49.32 (±3.63) |
| GPT 3.5 | 1.45 (±0.99) | 11.71 (±12.77) | 42.20 | 48.92 (±6.72) | 56.30 (±2.36) |
| Llama3-70b | 7.35 (±7.59) | 20.14 (±13.14) | 47.05 | 51.06 (±1.75) | 57.50 (±2.91) |
| ***SPARQL (GraphDB)*** | | | | | |
| Llama3-8b | 3.09 (±2.70) | 4.18 (±9.04) | 0.05 | 1.51 (±1.92) | 4.27 (±8.92) |
| Gemini 1.0 Pro | 3.23 (±1.95) | 11.99 (±7.87) | 2.85 | 7.76 (±4.65) | 26.32 (±5.60) |
| GPT 3.5 | 6.95 (±5.48) | 25.32 (±4.57) | 3.30 | 7.88 (±4.78) | 23.58 (±8.09) |
| Llama3-70b | 7.37 (±4.46) | 27.14 (±2.69) | 1.00 | 10.26 (±6.89) | 30.49 (±1.82) |
| ***Cypher (Neo4j)*** | | | | | |
| Llama3-8b | 9.43 (±4.12) | 19.64 (±3.35) | 2.75 | 15.31 (±11.28) | 34.89 (±5.34) |
| Gemini 1.0 Pro | 13.80 (±1.67) | 22.91 (±1.38) | 23.45 | 39.74 (±2.99) | 53.84 (±4.09) |
| GPT 3.5 | 10.37 (±4.84) | 18.08 (±1.05) | 16.35 | 29.87 (±3.44) | 41.12 (±2.85) |
| Llama3-70b | 16.04 (±2.40) | 25.25 (±5.10) | 34.45 | 43.06 (±4.53) | 57.07 (±4.41) |
| ***MQL (MongoDB)*** | | | | | |
| Llama3-8b | 2.64 (±3.35) | 4.62 (±6.56) | 9.45 | 6.71 (±6.55) | 11.33 (±15.06) |
| Gemini 1.0 Pro | 5.25 (±2.47) | 13.25 (±3.25) | 3.40 | 18.53 (±1.67) | 30.65 (±7.19) |
| GPT 3.5 | 1.49 (±3.30) | 5.36 (±5.17) | 3.50 | 26.26 (±13.64) | 35.06 (±15.74) |
| Llama3-70b | 8.86 (±2.09) | 17.91 (±4.52) | 21.55 | 33.83 (±8.54) | 40.35 (±17.03) |

Schema information helps for all query languages but not equally.

# SM3 Dataset Baseline Results

| Models | without schema | | with schema | | | |
|---|---|---|---|---|---|---|
| | w/o schema 1-shot | w/o schema 5-shot | w/ schema 0-shot | w/ schema 1-shot | w/ schema 5-shot | |
| ***SQL (PostgreSQL)*** | | | | | | |
| Llama3-8b | 4.20 (±5.6) | 10.81 (±9.89) | 22.55 | 23.27 (±1.05) | 27.49 (±15.27) | |
| Gemini 1.0 Pro | 4.47 (±4.88) | 21.65 (±11.10) | 38.60 | 38.37 (±3.31) | 49.32 (±3.63) | +26% |
| GPT 3.5 | 1.45 (±0.99) | 11.71 (±12.77) | 42.20 | 48.92 (±6.72) | 56.30 (±2.36) | |
| Llama3-70b | 7.35 (±7.59) | 20.14 (±13.14) | 47.05 | 51.06 (±1.75) | 57.50 (±2.91) | |
| ***SPARQL (GraphDB)*** | | | | | | |
| Llama3-8b | 3.09 (±2.70) | 4.18 (±9.04) | 0.05 | 1.51 (±1.92) | 4.27 (±8.92) | |
| Gemini 1.0 Pro | 3.23 (±1.95) | 11.99 (±7.87) | 2.85 | 7.76 (±4.65) | 26.32 (±5.60) | +3207% |
| GPT 3.5 | 6.95 (±5.48) | 25.32 (±4.57) | 3.30 | 7.88 (±4.78) | 23.58 (±8.09) | |
| Llama3-70b | 7.37 (±4.46) | 27.14 (±2.69) | 1.00 | 10.26 (±6.89) | 30.49 (±1.82) | |
| ***Cypher (Neo4j)*** | | | | | | |
| Llama3-8b | 9.43 (±4.12) | 19.64 (±3.35) | 2.75 | 15.31 (±11.28) | 34.89 (±5.34) | |
| Gemini 1.0 Pro | 13.80 (±1.67) | 22.91 (±1.38) | 23.45 | 39.74 (±2.99) | 53.84 (±4.09) | +379% |
| GPT 3.5 | 10.37 (±4.84) | 18.08 (±1.05) | 16.35 | 29.87 (±3.44) | 41.12 (±2.85) | |
| Llama3-70b | 16.04 (±2.40) | 25.25 (±5.10) | 34.45 | 43.06 (±4.53) | 57.07 (±4.41) | |
| ***MQL (MongoDB)*** | | | | | | |
| Llama3-8b | 2.64 (±3.35) | 4.62 (±6.56) | 9.45 | 6.71 (±6.55) | 11.33 (±15.06) | |
| Gemini 1.0 Pro | 5.25 (±2.47) | 13.25 (±3.25) | 3.40 | 18.53 (±1.67) | 30.65 (±7.19) | +453% |
| GPT 3.5 | 1.49 (±3.30) | 5.36 (±5.17) | 3.50 | 26.26 (±13.64) | 35.06 (±15.74) | |
| Llama3-70b | 8.86 (±2.09) | 17.91 (±4.52) | 21.55 | 33.83 (±8.54) | 40.35 (±17.03) | |

Adding examples improves accuracy through in-context learning for all LLMs and query languages; however, the rate of improvement varies greatly across query languages.

# SM3 Dataset Summary and Next Steps

**Next Steps**

- More complex query templates based on input from health professionals
- Multilingual extensions

**Poster presentation:**

Thu 12 Dec 11 a.m. PST — 2 p.m. PST

**Get the data and code:**

https://github.com/jf87/SM3-Text-to-Query