Towards Reasoning-Augmented Natural Language Querying for Domain-Specific NoSQL Databases

Ping Wang
Department of Computer Science
Stevens Institute of Technology
Email: ping.wang@stevens.edu
Web: https://leafnlp.org/ping

2nd NJIT Workshop on MultiMedia Intelligence (MMI 2026)

STEVENS
INSTITUTE OF TECHNOLOGY
1870

Electronic Health Records (EHR)

Relational Database

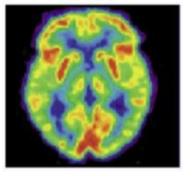
- Demographics
- Diagnosis
- Procedures
- Lab Tests
- Medications

Clinical Notes

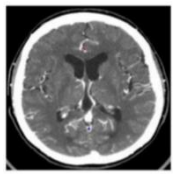
- Family History
- Social History
- Symptoms
- Discharge Instructions

Medical Images

Magnetic Resonance Imaging (MRI)



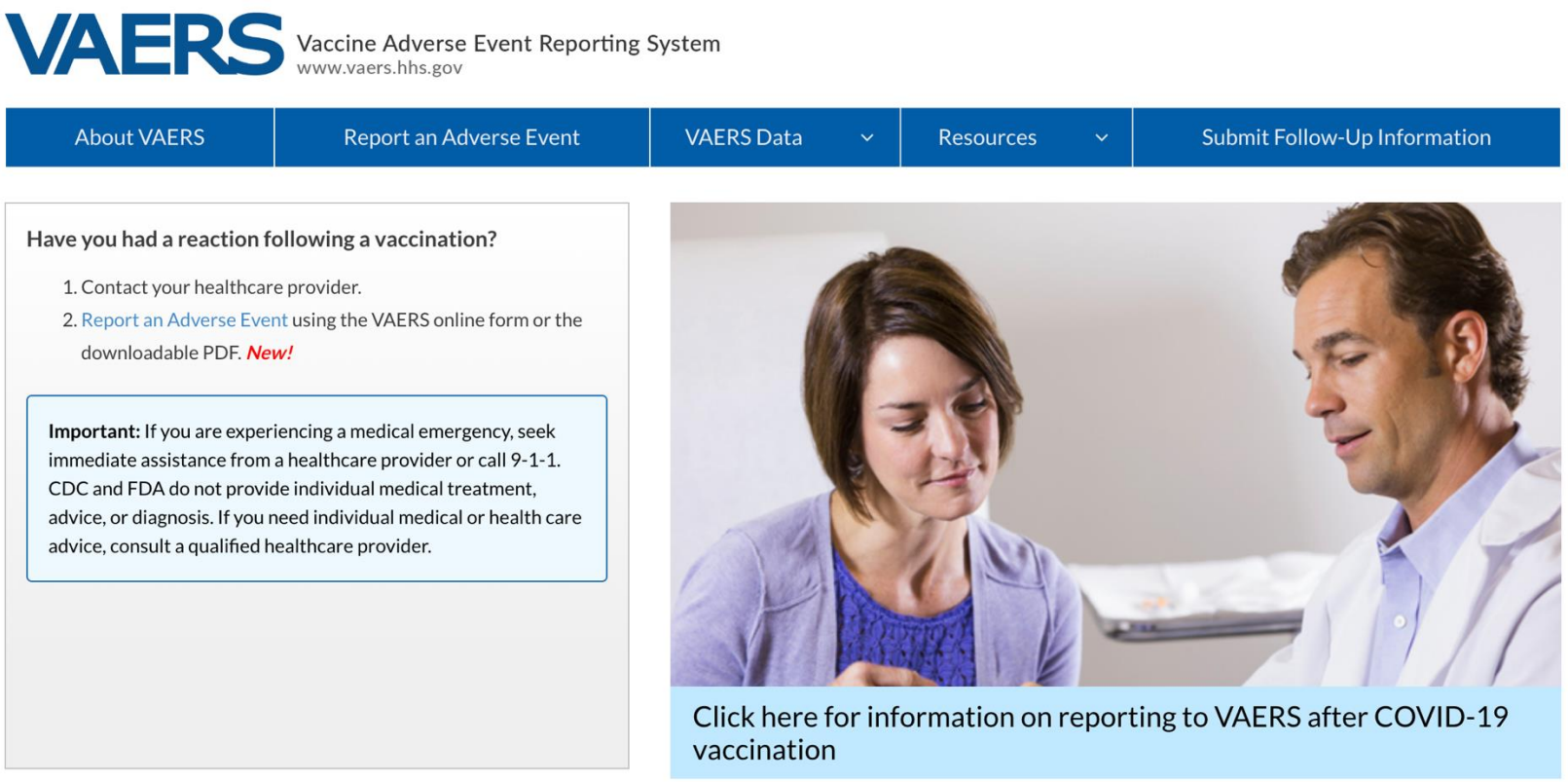
Positron Emission Tomography (PET) Scan



Computerized Tomography (CT) Scan

Medical images source: https://www.slideshare.net/VamsiIntellectual/nuclear-medicine-pet-scan

Vaccine Adverse Event Reporting System (VAERS)

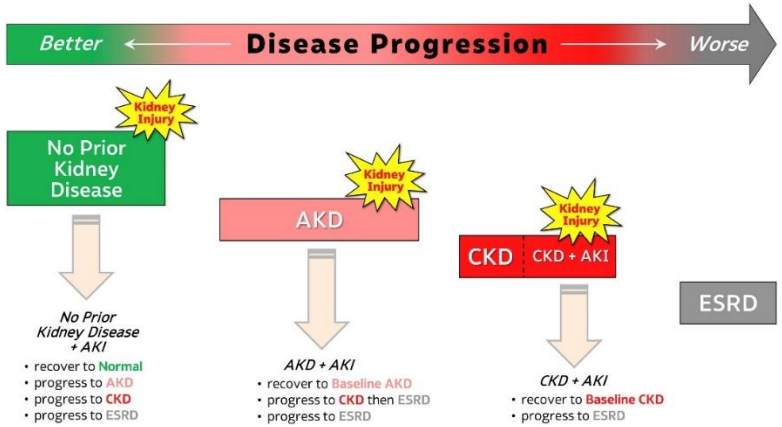


- A national early warning system established in 1990.
- Co-managed by CDC and FDA.

- **Three tables:**
  - VAERSDATA
  - VAERSVAX
  - VAERSSYMPTOMS
- **Symptom information:**
  - Free text symptom description
  - A list of adverse event coded terms

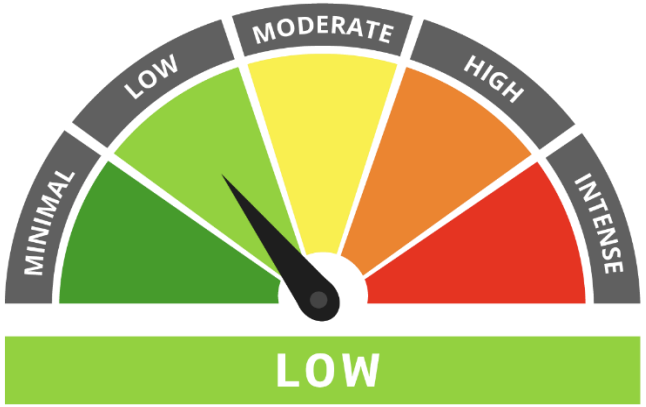
In the Big Data Era: From Records to Insights

Understanding Natural History of Disease

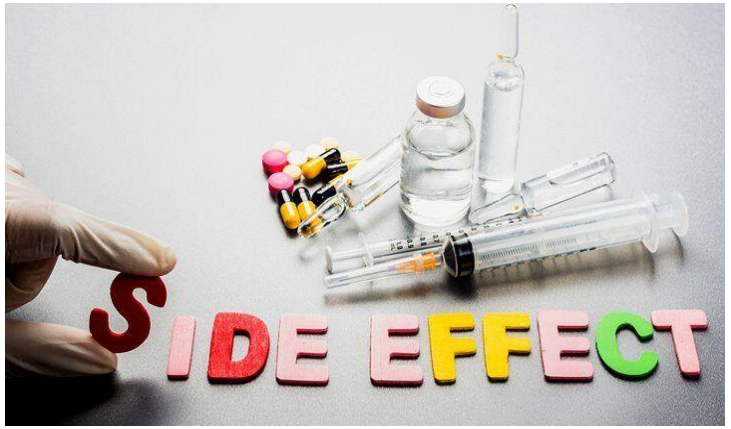


Patient Cohort Identification

Risk Prediction and Biomarker Discovery



Quantifying Effect of Intervention

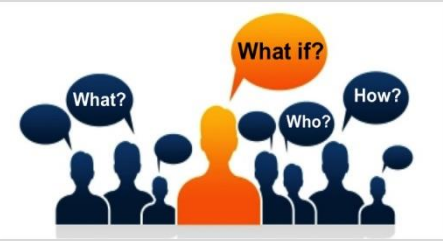


Yadav, Pranjul, et al. "Mining electronic health records (EHRs): A survey." *ACM Computing Surveys, 2018*.
Image sources: https://www.jtcvs.org/article/S0022-5223(18)30399-4/fulltext, https://www.iths.org/investigators/services/bmi/patient-cohort-identification/, https://www.subpng.com/png-9y08xn/, https://www.everydayhealth.com/rheumatoid-arthritis/treatment/ra-medication-side-effects/.

Research Topics in My Group

Natural Language Querying

- How to seek answers from various types of medical records for clinical activity related questions posed in human language without the assistance of database and natural language processing (NLP) domain experts?

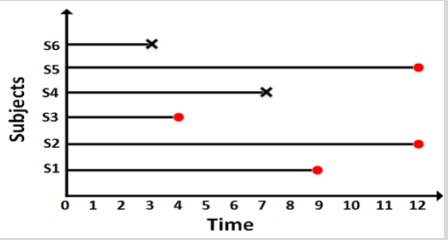


Knowledge Extraction and Discovery

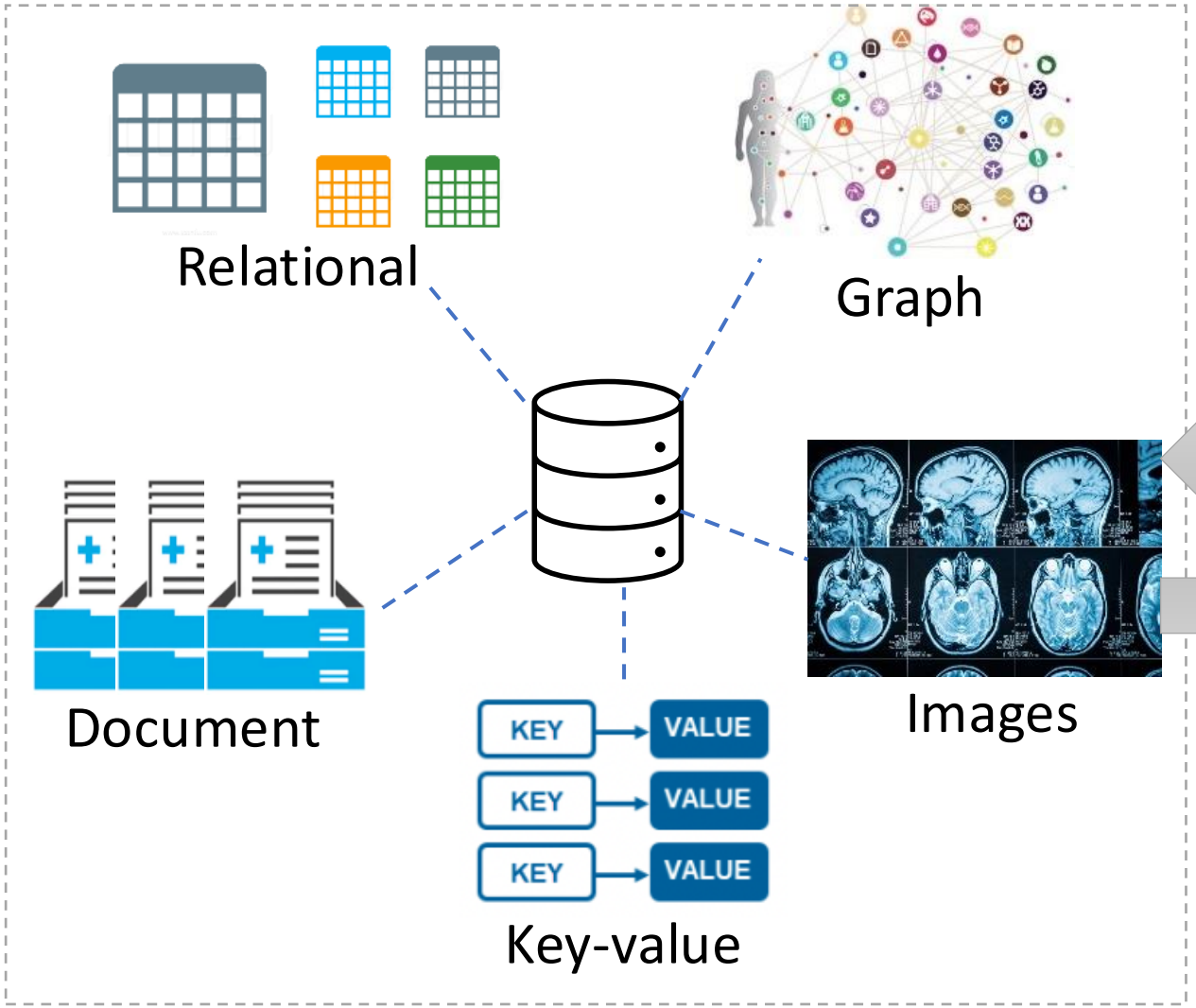
- How to discover underlying correlations of different medical events and entities in EHR?
- How to apply NLP techniques to construct structured events and knowledge bases from clinical notes?

Predictive Analysis

- How to predict when a medical event will occur and estimate its probability based on previous medical information of patients?
- How to make predictions at the early stage of the medical study?



Natural Language Querying (NLQ)



Various data types in healthcare

Question

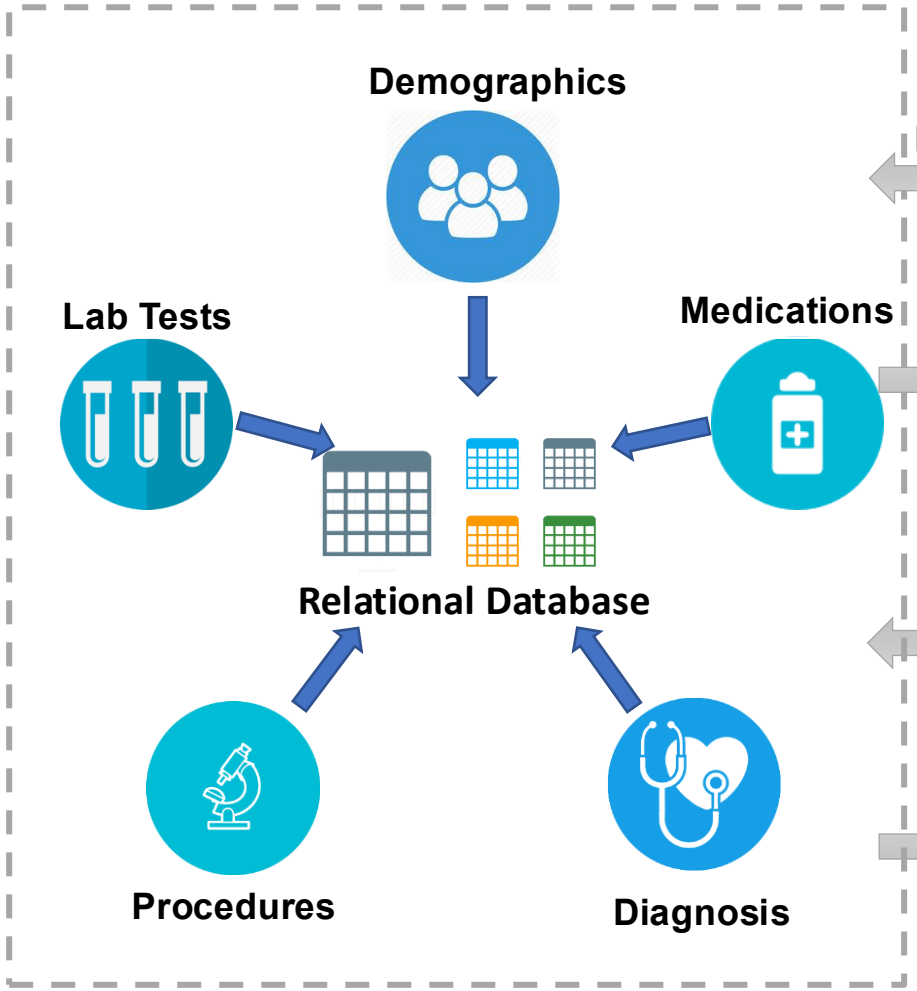
Users using human-like language to find the information in the databases.

Answer

Computers transform the human-like query into a machine-readable search query or directly return the answers.

A reasoning process

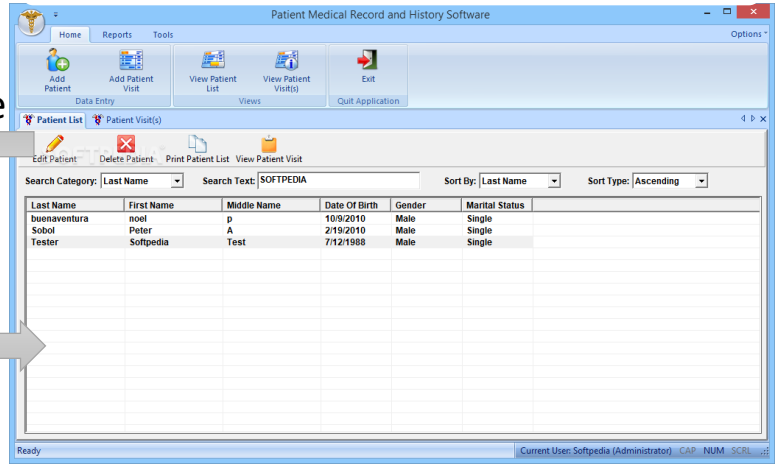
Existing Work Focuses on Text-to-SQL



Traditional Rule-based System

Query Database

Return Output

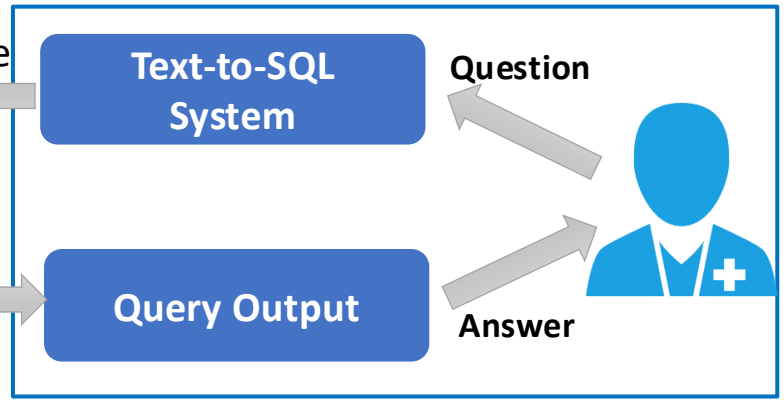


- Complicated to specify rules
- Difficult to manage
- Require training

Question Answering (QA) System

Query Database

Return Output



- No rules required
- Automatic translation
- No database or NLP knowledge required

Example of Text-to-SQL in Healthcare

- MIMICSQL dataset was designed to support the training and evaluation of Text-to-SQL in healthcare.
  - 10,000 question-SQL pairs.
  - Two types of questions: template questions and natural language (NL) questions.

Tables

DEMOGRAPHIC

SUBJECT_ID	HADM_ID	Gender	ADMISSION_TYPE	...
990	184231	F	EMERGENCY	...
17772	122127	M	NEWBORN	...
⋮	⋮	⋮	⋮	⋮
66411	178264	F	EMERGENCY	...
29961	196409	M	EMERGENCY	...

PROCEDURES

SUBJECT_ID	HADM_ID	SHORT_TITLE	...
9258	183354	Procedure-one vessel	...
28588	141664	Insert endotracheal tube	...
⋮	⋮	⋮	⋮
66411	178264	Abdomen artery incision	...
66411	178264	Venous catch NEC	...

SQL template: SELECT $AGG_OP ($AGG_COLUMN)+ FROM $TABLE WHERE ($COND_COLUMN $COND_OP $COND_VAL)+

Question: How many female patients underwent the procedure of abdomen artery incision?

SQL query:

```
SELECT COUNT ( DISTINCT DEMOGRAPHIC.SUBJECT_ID )
FROM DEMOGRAPHIC INNER JOIN PROCEDURES on DEMOGRAPHIC.HADM_ID = PROCEDURES.HADM_ID
WHERE DEMOGRAPHIC."GENDER" = "F" AND PROCEDURES."SHORT_TITLE" = "Abdomen artery incision"
```

Wang, Ping, Tian Shi, and Chandan K. Reddy. "Text-to-sql generation for question answering on electronic medical records." In *Proceedings of The Web Conference 2020*, pp. 350-361. 2020.

From SQL to NoSQL

- **Text-to-SQL:**
  - A promising research topic in both the NLP and database community.
  - Interact with relational databases without the need for NLP/database knowledge and without the help of database engineers.
  - Several large-scale Text-to-SQL datasets in a variety of domains have been created and many state-of-the-art deep learning models have also been developed.
- **Limitations:**
  - Text-to-SQL capabilities are limited by the data structures and functionality of SQL databases for full-text search.
  - Difficult to incorporate external knowledge bases (KBs) into relational tables because the complex relationships between nodes in KBs will result in too many tables or table columns in the relational database.

Our solution: Exploiting the potential of NoSQL database for natural language querying!

NLQ on NoSQL Databases

- **Overall objective:**
  - **Fill the gap**: very little work has been devoted to developing NLQ tools for NoSQL databases.
  - **Forge new research directions**: NLQ task for Elasticsearch query (Text-to-ESQ) generation.
  - **Re-designing a variety of aspects** of NLQ: including the target application, the benefit of NoSQL database, and the JSON data stored in the database, etc.

- **Advantages of Text-to-ESQ:**
  - Document-like format of Elasticsearch database: allows us to easily convert complex data formats into **nested JSON objects (like reasoning path on knowledge graphs), and integrate data from disparate sources**, such as tables, texts, graphs, and other external knowledge bases.
  - Incorporate **reasoning processes** into searching via both query clauses (e.g., Boolean relationships) and path to the target fields (e.g., DATA.SYMPTOMS.SEVERITY).

Research Challenges for Text-to-ESQ

- **Task formulation and benchmarks:**
  - Research gap in task formulation of Text-to-NoSQL.
  - Limited public datasets and benchmarks compared to SQL (e.g., Spider, WikiSQL, MIMICSQL).
  - No standard evaluation metrics for NoSQL query generation.

- **LLM-based reasoning for Text-to-ESQ:**
  - LLM pipelines are underexplored for Text-to-ESQ.
  - NoSQL-specific reasoning and schema grounding are required.

- **Complexity taxonomy and efficiency improvement:**
  - Lack of a complexity taxonomy for ESQ query generation.
  - Need complexity-aware and optimization-driven query generation.

Text-to-ESQ: A Two-Stage Controllable Approach for Efficient Retrieval of Vaccine Adverse Events from NoSQL Database

Zhang, Wenlong, Kangping Zeng, Xinming Yang, Tian Shi, and Ping Wang. "Text-to-esq: A two-stage controllable approach for efficient retrieval of vaccine adverse events from nosql database." In Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 1-10. 2023.

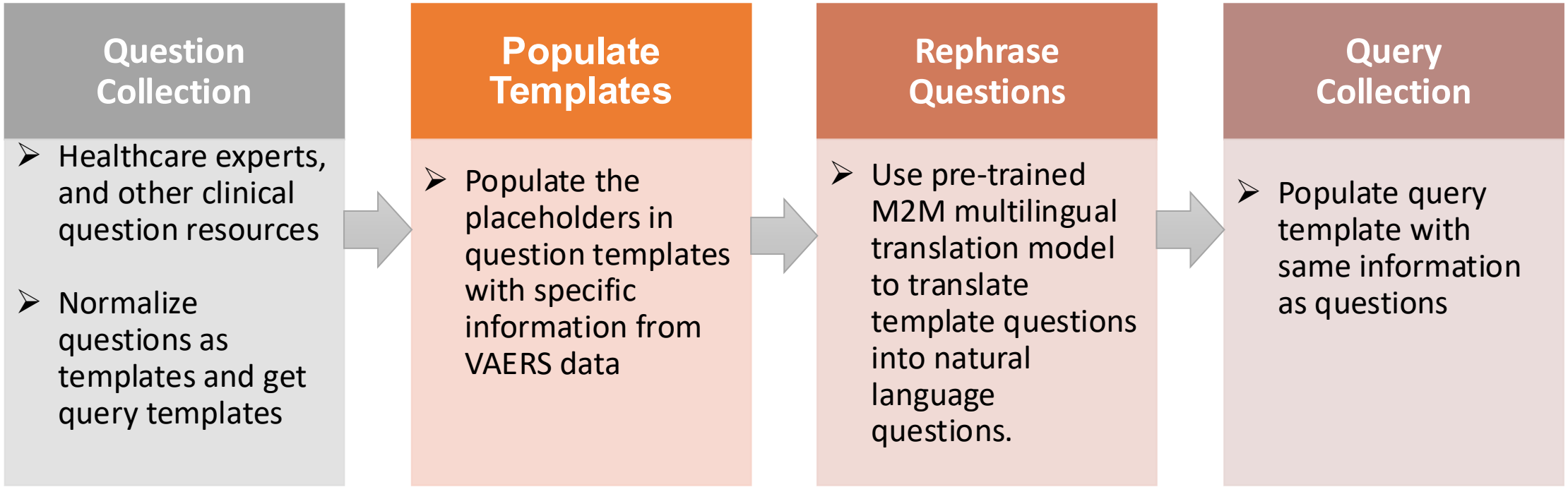
Our Contributions

- **Create a large-scale Text-to-Elasticsearch (Text-to-ESQ) dataset:**
  - Based on the **VAERS data**, we considered three categories of information:
    - Patients' vaccine relative history with side effects
    - Free text about symptom description
    - Vaccine information (e.g., series number, manufacturer)
  - VAERSESQ is the **FIRST large-scale** dataset for Text-to-ESQ generation task in healthcare.

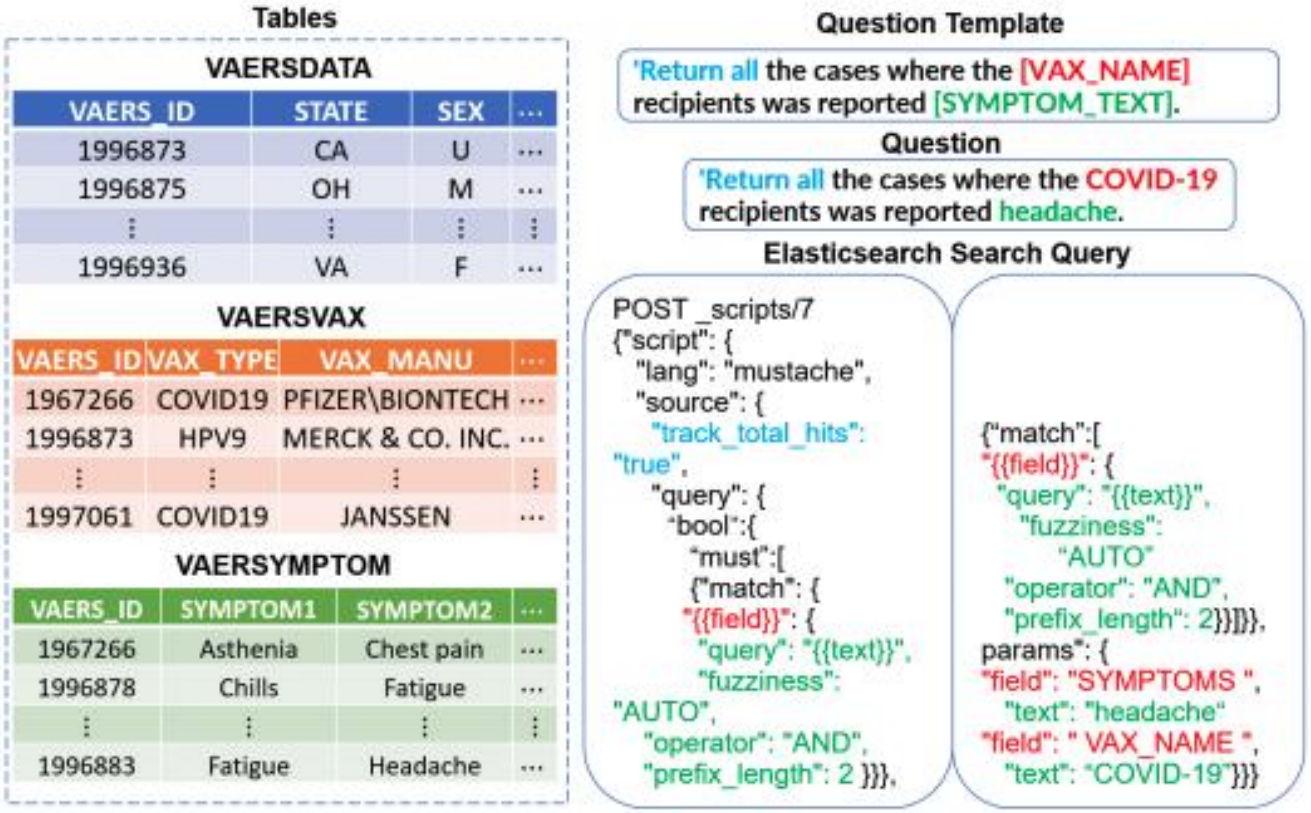
- **Explore initial method for Elasticsearch query generation:**
  - Propose a two-stage controllable (TSC) framework.
  - Conduct an extensive experimental analysis of Text-to-ESQ on the VAERSESQ dataset.

VAERSESQ Data Generation

- **Question-query pairs generation:**
  - **Human Annotation**: questions are natural, but the process is time-consuming and expensive.
  - **Machine Generation**: efficient, but machine generated questions are not natural.
  - **Ours:** Template-based Generation (machine generation and human annotation) + Rephrasing (back translation).



VAERSESQ Dataset



Data	Value
# of tables	3
# of fields/columns in tables[a]	35/8/11
Number of template/natural questions	13,040
Average template question length (in words)	12.13
Average NL question length (in words)	11.52
Average query length (including template keywords)	167.65

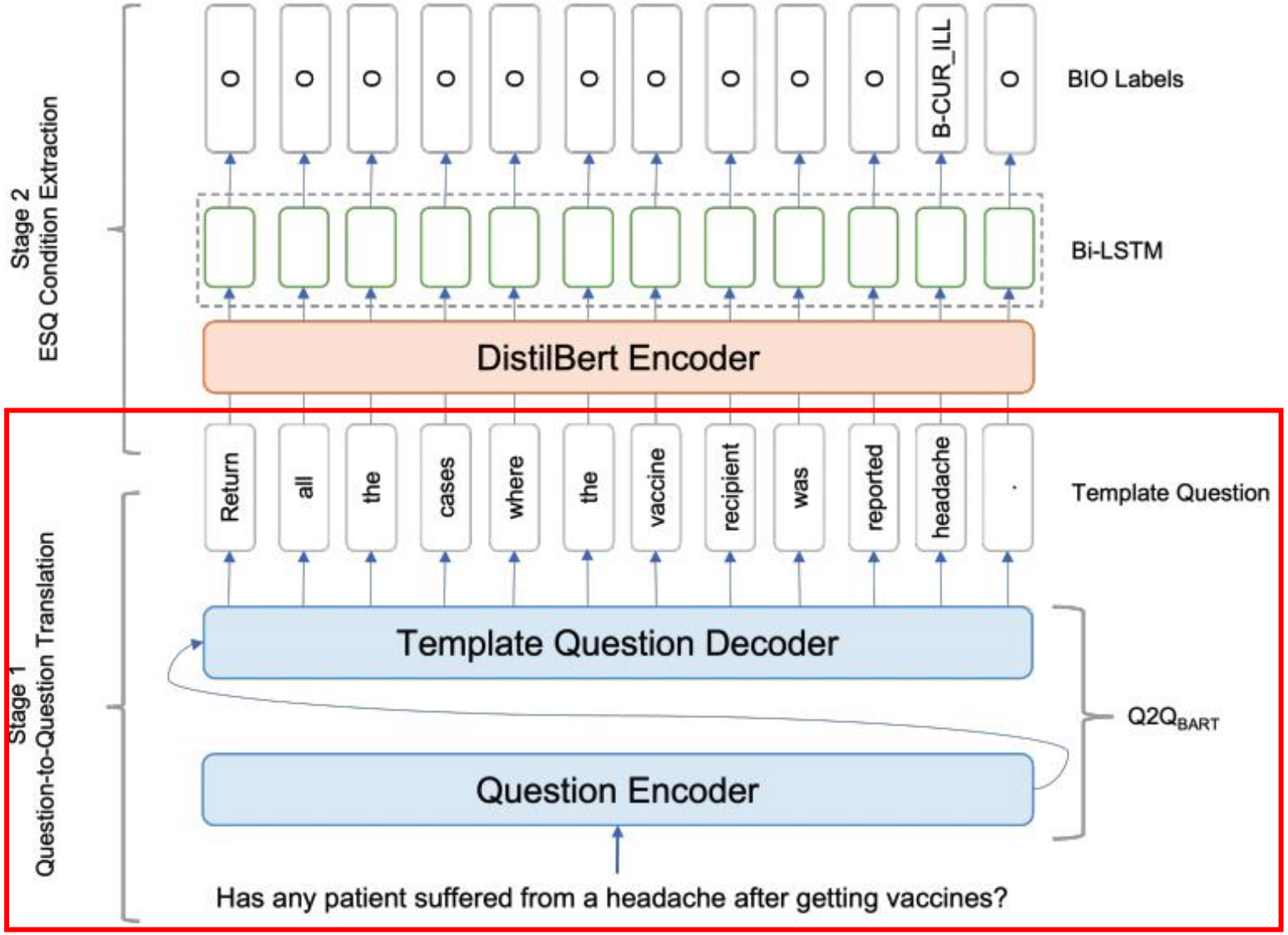
- Transform the structured database into the Elasticsearch database for Text-to-ESQ.
- Provide a reusable paradigm to facilitate the initial exploration Text-to-ESQ.

The VAERSESQ dataset is publicly available at https://github.com/LEAF-Lab-Stevens/Text2ESQ.

Text-to-ESQ: A Two-Stage Controllable Framework

Stage 1: Question-to-question (Q2Q) translation module

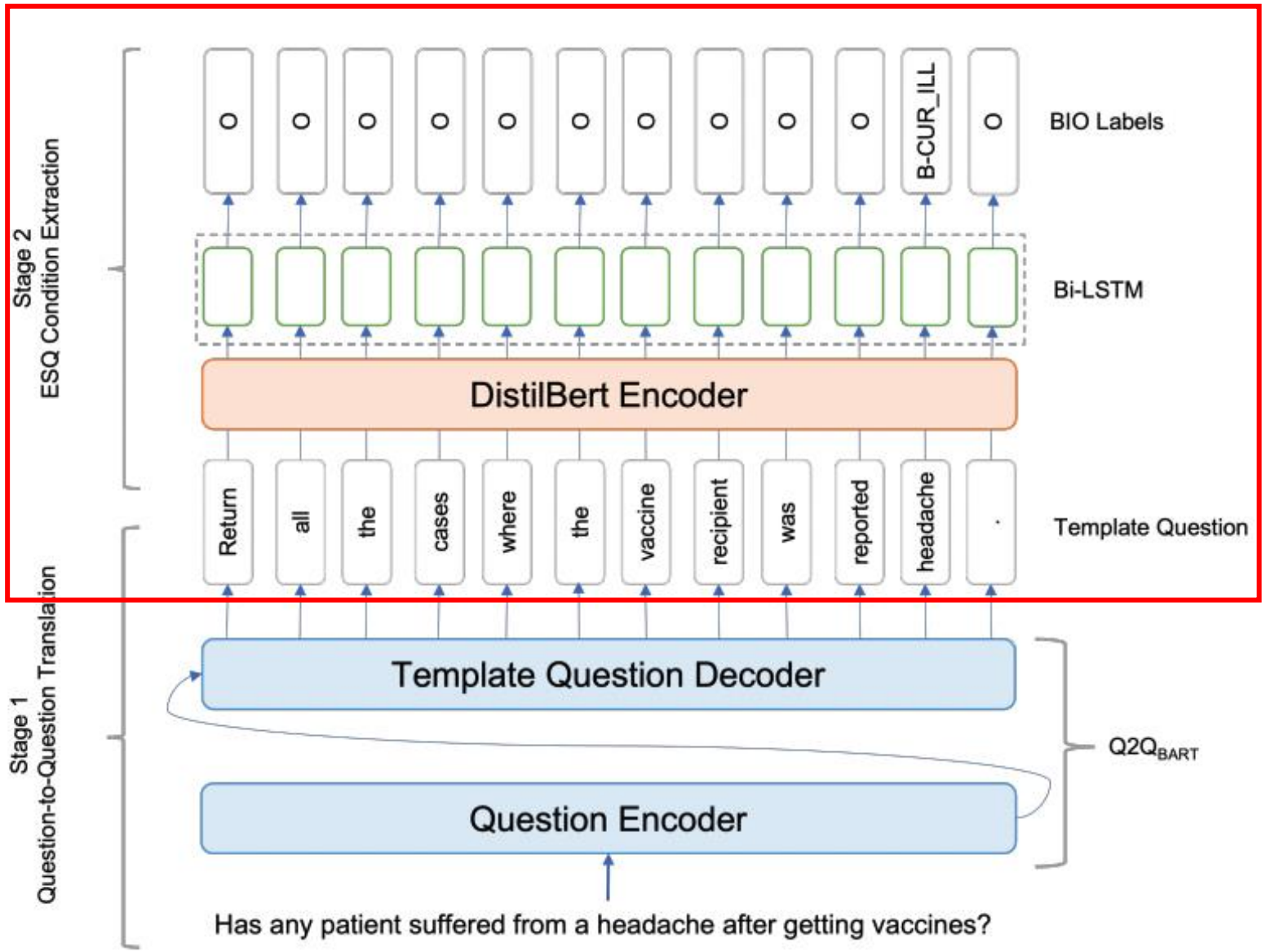
- The questions may not contain terms that exactly match the condition fields and values.
- Designed for translating natural language questions into the corresponding template questions.
- Adopt and finetune the BART model on the natural language and template question pairs.
- The generated question will serve as the input for the second stage.



Text-to-ESQ: A Two-Stage Controllable Framework

Stage 2: ESQ condition extraction (ECE) module

- For parsing name entities about **condition fields** and **values** from template questions for further populating the query templates.
- Formulate as a sequence labeling task:
  - Condition values: relevant tokens
  - Condition fields: labels/categories



Evaluation Metrics

➢ **We leveraged traditional score and query components for comprehensive evaluation.**

- To evaluate the performance of the **Q2Q translation module**:
  - **ROUGE-1**: ROUGE-1 to measure the similarity of the overall original natural language questions and the generated template questions.
  - **Condition**: we also evaluate the specific condition values extract from human intents.

- To evaluate the performance of the **ESQ condition extraction module**:
  - **String matching metric** : we adopt the commonly used string matching metric logic form accuracy

$$A_{CCf+v} = N_{f+v}/N$$

    to evaluate the overall performance on both condition fields and values.
  - **Condition value:** we also evaluate values by calculating $A_{ccv} = N_v/N$.

Performance of Q2Q for Question Translation

Methods	Development		Testing	
	Overall	Value	Overall	Value
Seq2Seq	0.73	0.35	0.70	0.36
M2M	0.92	0.60	0.90	0.63
Q2Q	0.88	0.65	0.85	0.63

- Performance of different methods on translating natural language questions to template questions.

Methods	Question
NLQ	Which type of reaction is most common after a COVID vaccine?
Ground Truth TQ	Which symptom is most common after a COVID-19 vaccine?
Seq2Seq	Which symptom is most common after a ?
M2M	Which symptom is most common after a COVID vaccine?
Q2Q	Which symptom is most common after a COVID-19 vaccine?

- An example of translating the natural language questions (NLQ) into template questions (TQ) with the Q2Q module.

Performance of ECE for Condition Extraction

Type	Method	Development			Testing		
		Overall	Field	Value	Overall	Field	Value
Template	Seq2Seq	0.515	0.646	0.316	0.690	0.740	0.640
	RoBERTa	0.959	0.986	0.991	0.956	0.979	0.986
	RoBERTa+Bi-LSTM	0.967	0.982	0.992	0.967	0.982	0.992
	DistilBERT	0.981	**0.993**	0.995	0.975	**0.989**	0.992
	ECE	**0.982**	0.992	**0.998**	**0.983**	**0.989**	**0.999**
Natural language	Seq2Seq+Seq2Seq	0.351	0.350	0.231	0.301	0.324	0.287
	Seq2Seq+RoBERTa	0.355	0.358	0.357	0.360	0.366	0.362
	Seq2Seq+RoBERTa+Bi-LSTM	0.352	0.357	0.354	0.358	0.360	0.359
	Seq2Seq+DistilBERT	0.343	0.346	0.349	0.342	0.347	0.347
	Seq2Seq+**ECE**	0.343	0.348	0.349	0.348	0.350	0.350
	M2M+Seq2Seq	0.389	0.374	0.291	0.351	0.404	0.307
	M2M+RoBERTa	0.544	0.551	0.552	0.471	0.476	0.477
	M2M+RoBERTa+Bi-LSTM	0.547	0.551	0.551	0.477	0.478	0.479
	M2M+DistillBERT	0.552	0.554	0.554	0.475	0.479	0.478
	M2M+**ECE**	0.553	0.553	0.554	0.476	0.478	0.479
	Q2Q+Seq2Seq	0.469	0.588	0.288	0.473	0.537	0.304
	Q2Q+RoBERTa	0.599	0.612	0.609	0.593	0.601	0.602
	Q2Q+RoBERTa+Bi-LSTM	0.606	0.612	0.610	0.596	0.602	0.604
	Q2Q+DistilBERT	**0.609**	**0.613**	**0.612**	0.598	0.604	0.603
	Q2Q+**ECE**	0.601	0.612	**0.612**	**0.601**	**0.605**	**0.605**

Performance of ECE for Condition Extraction

Type	Method	Development			Testing		
		Overall	Field	Value	Overall	Field	Value
Template	Seq2Seq	0.515	0.646	0.316	0.690	0.740	0.640
	RoBERTa	0.959	0.986	0.991	0.956	0.979	0.986
	RoBERTa+Bi-LSTM	0.967	0.982	0.992	0.967	0.982	0.992
	DistilBERT	0.981	**0.993**	0.995	0.975	**0.989**	0.992
	ECE	**0.982**	0.992	**0.998**	**0.983**	**0.989**	**0.999**
Natural language	Seq2Seq+Seq2Seq	0.351	0.350	0.231	0.301	0.324	0.287
	Seq2Seq+RoBERTa	0.355	0.358	0.357	0.360	0.366	0.362
	Seq2Seq+RoBERTa+Bi-LSTM	0.352	0.357	0.354	0.358	0.360	0.359
	Seq2Seq+DistilBERT	0.343	0.346	0.349	0.342	0.347	0.347
	Seq2Seq+**ECE**	0.343	0.348	0.349	0.348	0.350	0.350
	M2M+Seq2Seq	0.389	0.374	0.291	0.351	0.404	0.307
	M2M+RoBERTa	0.544	0.551	0.552	0.471	0.476	0.477
	M2M+RoBERTa+Bi-LSTM	0.547	0.551	0.551	0.477	0.478	0.479
	M2M+DistillBERT	0.552	0.554	0.554	0.475	0.479	0.478
	M2M+**ECE**	0.553	0.553	0.554	0.476	0.478	0.479
	Q2Q+Seq2Seq	0.469	0.588	0.288	0.473	0.537	0.304
	Q2Q+RoBERTa	0.599	0.612	0.609	0.593	0.601	0.602
	Q2Q+RoBERTa+Bi-LSTM	0.606	0.612	0.610	0.596	0.602	0.604
	Q2Q+DistilBERT	**0.609**	**0.613**	**0.612**	0.598	0.604	0.603
	Q2Q+**ECE**	0.601	0.612	**0.612**	**0.601**	**0.605**	**0.605**

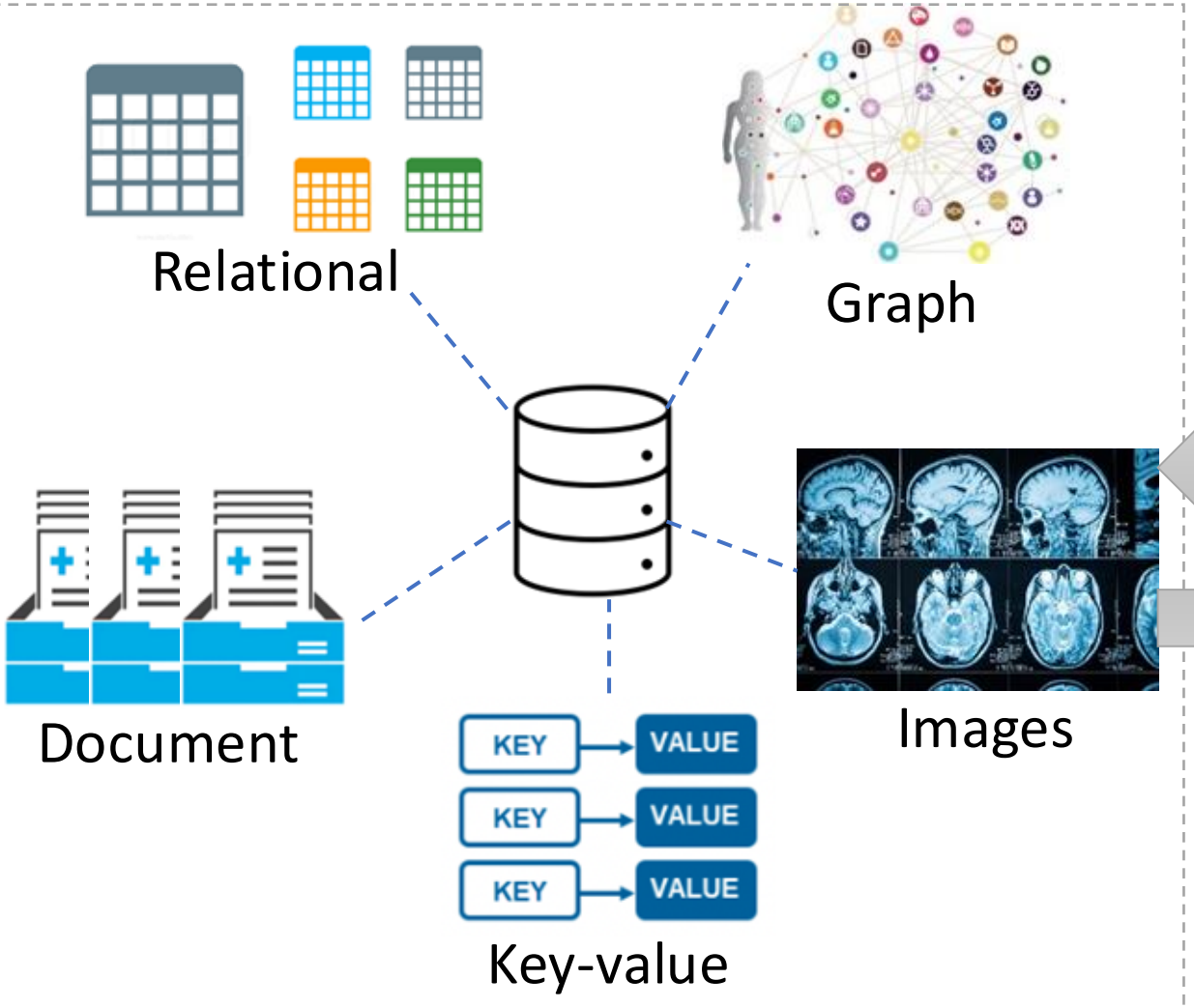
Summary

- **Findings:**
  - Intent understanding is a first-class modeling objective.
  - Decomposition improves reliability.
  - Benchmark construction shapes what models learn.
- **Impacts:**
  - This initial work defined the task, benchmark, and evaluation framework, laying the foundation for later studies.
  - This work established staged task decomposition as a key strategy for improving the reliability of natural language querying.

Natural Language Querying on Domain-Specific NoSQL Databases with Large Language Models

Zhang, Wenlong, Chengyang He, Guanqun Yang, Dipankar Bandyopadhyay, Tian Shi, and Ping Wang. "Natural language querying on domain-specific nosql database with large language models." In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 5174-5181. IEEE, 2024.

LLM-Based NLQ on NoSQL Databases



Various data types in healthcare

Question

Users using human-like language to query information from LLM interface.

LLMs

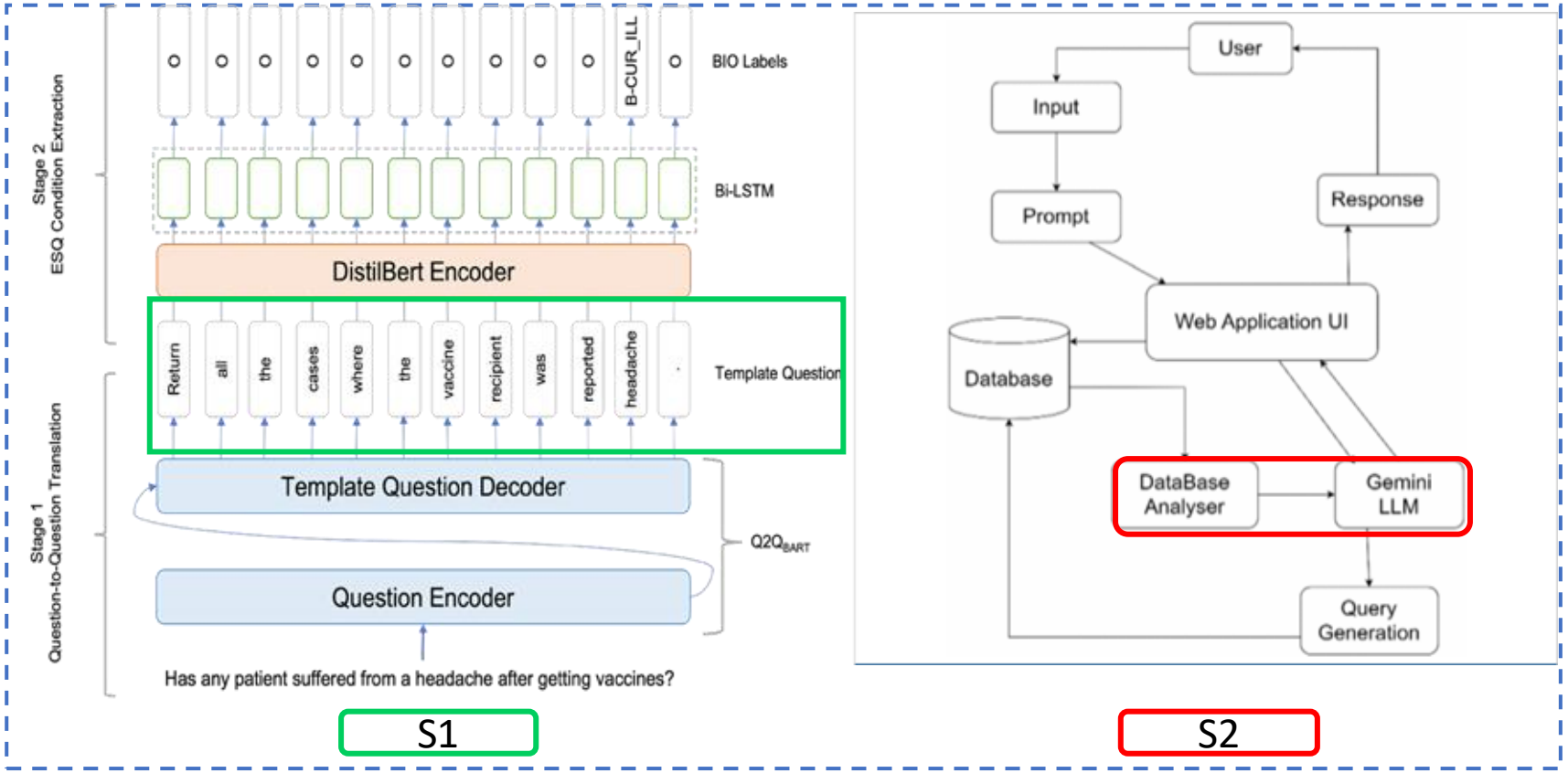
Answer

LLM transform the human-like query into a machine-readable search query or directly return the answers.

A reasoning process

Existing Challenges

Existing Work



Limitation

S1: **Only focuses on template questions.**
S2: **Lack of detailed analysis and evaluation.**

1. Wenlong Zhang, Kangping Zeng, Xinming Yang, Tian Shi, and Ping Wang. 2023. Text-to-ESQ: A Two-Stage Controllable Approach for Efficient Retrieval of Vaccine Adverse Events from NoSQL Database, ACMBCB 2023
2. V. Usha, N. C. Abhinash, S. N. Chowdary, V. Sathya, E. R. Reddy and S. P. S, "Enhanced Database Interaction Using Large Language Models for Improved Data Retrieval and Analysis," 2024 ICoICI

Our Contributions

- **Systematically categorize and examine existing Chain-of-Thought (CoT) prompting.**
  - Instruction-based CoT prompting
  - Explanation-based CoT prompting

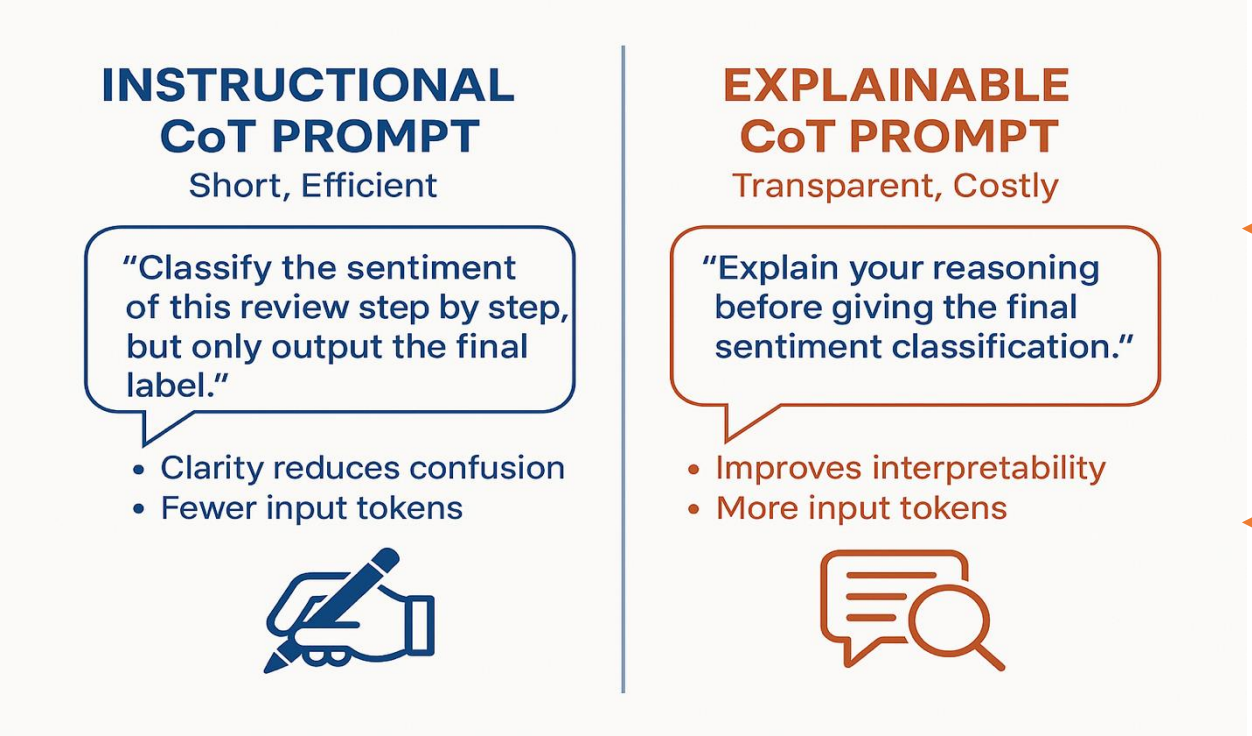
- **Explore a new framework to utilize LLMs for Text-to-ESQ task.**
  - Introduce the InstructEx CoT prompting for an initial exploration of combining diverse prompt strategies for solving the Text-to-ESQ task on NoSQL databases with LLMs.
  - Evaluate nine LLMs and three types of baseline prompting strategies, and conduct an extensive experimental analysis of utilizing LLMs for the Text-to-ESQ task.

Existing Chain-of-Thought Prompting Methods

- **Chain-of-Thought (CoT) prompting methods:**
  - Designed to enhance the reasoning capabilities of LLMs by breaking down complex problems into a series of intermediate steps.
  - We categorize the existing work into two categories: **instruction-based prompts** and **explanation-based prompts.**

Routine or high-volume task →

Production deployment with cost constraint →



← **Domain-specific or safety-critical reasoning**

← **Educational or analytical demos**

Many complex problems, such as Text-to-ESQ, tend to involve both intricate logic judgments and domain-specific language explanation. When handling these tasks with LLMs, it is essential to **combine both advantages**.

InstructEx Chain-of-Thought (CoT) Prompting

Prompt Header

Elasticsearch database with field names:
RECVDATE, STATE, AGE YRS, VAERS ID, SEX, SYMPTOM TEXT, DIED, ER VISIT, L THREAT, HOSPITAL, HOSPDAYS, DISABLE, VAX DATE, LAB DATA, OTHER MEDS, CUR ILL, HISTORY, PRIOR VAX, TODAYS DATE, OFC VISIT, VAX TYPE, VAX MANU, VAX LOT, VAX DOSE SERIES, VAX NAME, ALLERGIES

Prompt Body of Different Prompting Strategies

A. Base Prompt

[*Give me all the patients who reported Blood potassium decreased after the vaccination.*]
Generate the Elasticsearch query for the question.

B. Instructional CoT Prompt

[*Give me all the patients who reported Blood potassium decreased after the vaccination.*]
Identify the filed name first, then extract the specific condition values from the question for this field.
Generate the Elasticsearch query based on the condition and condition value.

C. Explainable CoT Prompt

[*Give me all the patients who reported Blood potassium decreased after the vaccination.*]
Which field name can represent the condition?
The filed name is [ ]
What is the value for this field?
The condition value for this field is [ ]
Generate the Elasticsearch query based on the condition and condition value.

D. InstructEx CoT Prompt

S1: ### [*Give me all the patients who reported Blood potassium decreased after the vaccination.*]

S2:
Identify the condition first.
Which field name can represent the condition?
Extract the specific values from the question for this field.
What is the value for this field?

S3: ### Generate the Elasticsearch query based on the condition and condition value.

LLM

```
SYMPTOM TEXT
Blood potassium decreased
POST _scripts/9
{"script": { "lang": "mustache",
  "source": {"track_total_hits": "true",
   "query": {"bool": {
      "must": [
       { "term": { "{{SYMPTOM TEXT}}": "{{'Blood
potassium decreased'}}" }}}}}}
```

Prompt Header: Include the database schema

Natural Language Question (S1)

Instructions with Stepwise Explanation (S2)

Query Generation Instruction (S3)

This new structured approach blends directive and conversational elements.

Performance of LLMs on Text-to-ESQ

- How do various LLMs perform on the domain-specific Text-to-ESQ task?

Performance measured by BLEU Score for the Text-to-ESQ task.

LLMs	Base	Instructional	Explainable	InstructEx
gpt-4o	32.0	33.4	35.0	**45.4**
gpt-3.5-turbo	35.0	35.5	35.3	42.4
gpt-4	24.0	30.7	34.3	38.9
CodeLlama	26.5	34.0	37.8	39.4
Llama2	10.5	38.0	37.0	38.5
Llama3	30.5	30.0	28.4	33.0
LlamaChat	10.3	24.0	28.4	32.5
StarCoder	7.0	32.0	38.6	39.8
Falcon	17.0	6.0	26.0	17.6

Findings:

1. Generally, the GPT series perform better than other models.
2. We observe that models with more parameters tend to achieve higher accuracy.

Performance of InstructEx CoT Prompting

- Does the proposed InstructEx CoT prompting improve the performance of the Text-to-ESQ task?

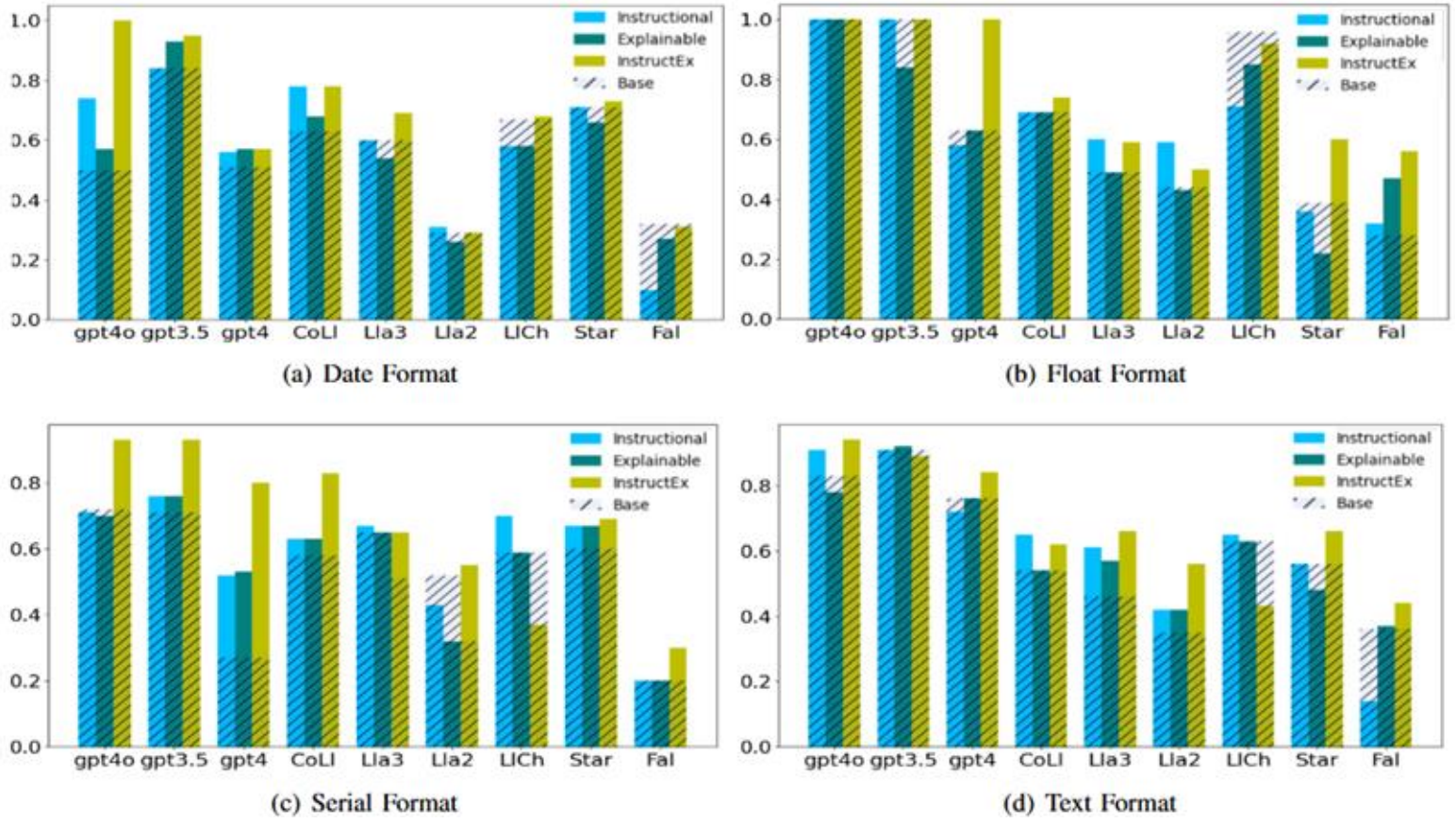
Prompt	Evaluation	gpt-4o	gpt-4	gpt-3.5	Llama2	Llama3	LlamaChat	CodeLlama	Falcon	StarCoder
Base	Frame similarity	0.453	0.328	0.367	0.289	0.257	0.141	0.312	0.234	0.140
	Condition match	0.427	0.345	0.412	0.105	0.213	0.192	0.472	0.235	0.392
	Value match	0.745	0.519	0.898	0.375	0.529	0.670	0.578	0.379	0.592
Instructional	Frame similarity	0.421	0.335	0.351	0.242	0.213	0.172	0.362	0.218	0.180
	Condition match	0.435	0.401	0.420	0.112	0.186	0.202	0.506	0.397	0.400
	Value match	0.867	0.611	0.961	0.407	0.629	0.667	0.788	0.166	0.615
Explainable	Frame similarity	0.454	0.421	0.383	0.313	0.273	0.148	0.356	0.328	0.320
	Condition match	0.381	0.368	0.376	0.093	0.316	0.202	0.376	0.110	0.190
	Value match	0.768	0.509	0.932	0.453	0.590	0.653	0.455	0.344	0.543
InstructEx	Frame similarity	**0.472**	0.382	0.461	0.281	0.287	0.226	0.352	0.341	0.328
	Condition match	0.477	0.422	0.432	0.150	0.396	0.196	**0.520**	0.236	0.303
	Value match	**0.985**	0.704	0.973	0.482	0.623	0.676	0.720	0.406	0.720

Findings:

1. Prompting strategies demonstrated an enhanced performance for all LLMs.
2. Instructional and explainable strategies contribute differently for the model performance.
   - Instructional CoT prompts improved special keywords, notably in the Condition and Value segments.
   - Explainable CoT prompts, improve the code framework, but have negative impact on keywords.

Detailed Performance on Elasticsearch

- What is the performance of different LLMs on various data types in the Elasticsearch queries?



Findings:

- Different strategies show different impact when data type change.
- There remains untapped potential for further exploration.

Summary

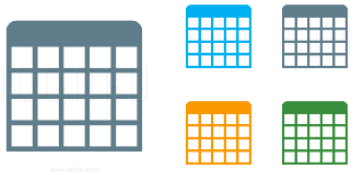
- **Findings:**
  - LLMs show different sensitivities to prompt design.
  - Prompting strategies affect different dimensions of output quality.
  - Their impact varies across query components and data types.
- **Impacts:**
  - It identified a reasoning-guided, verification-oriented query generation pipeline.
  - It benchmarked multiple LLMs on the Text-to-ESQ task and informed the later agent design.

Rethinking Efficient Text-to-NoSQL Query Generation with a Complexity Taxonomy for Lucene Data Structures

Wenlong Zhang, Amrutha Kanakatte Ravishankar, Denghui Zhang, Tian Shi, and Ping Wang. Rethinking Efficient Text-to-NoSQL Query Generation with a Complexity Taxonomy for Lucene Data Structures. In submission.

Taxonomy in Existing NLQ Datasets

SQL Databases



Taxonomy rules

- **Spider 2.0:** SQL queries tokens [1]
- **BIRD:** Tokens, JOINs, Keywords, n-grams [2]

NoSQL Databases

Taxonomy rules

- **TEND:** Transform from SQL database, without specific design for NoSQL [3]
- **BirdES:** Same strategy as used for the Bird dataset [4]

1. Lei, Fangyu, et al. "Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows." ICLR 2025
2. Li, Jinyang.et “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*
3. Lu, Jinwei, et al."Bridging the gap: Enabling natural language queries for nosql databases through text-to-nosql translation." arXiv preprint arXiv:2502.11201 (2025).
4. Dongge Xue et al. “Text-to-ES Bench: A Comprehensive Benchmark for Converting Natural Language to Elasticsearch Query” ACL 2025

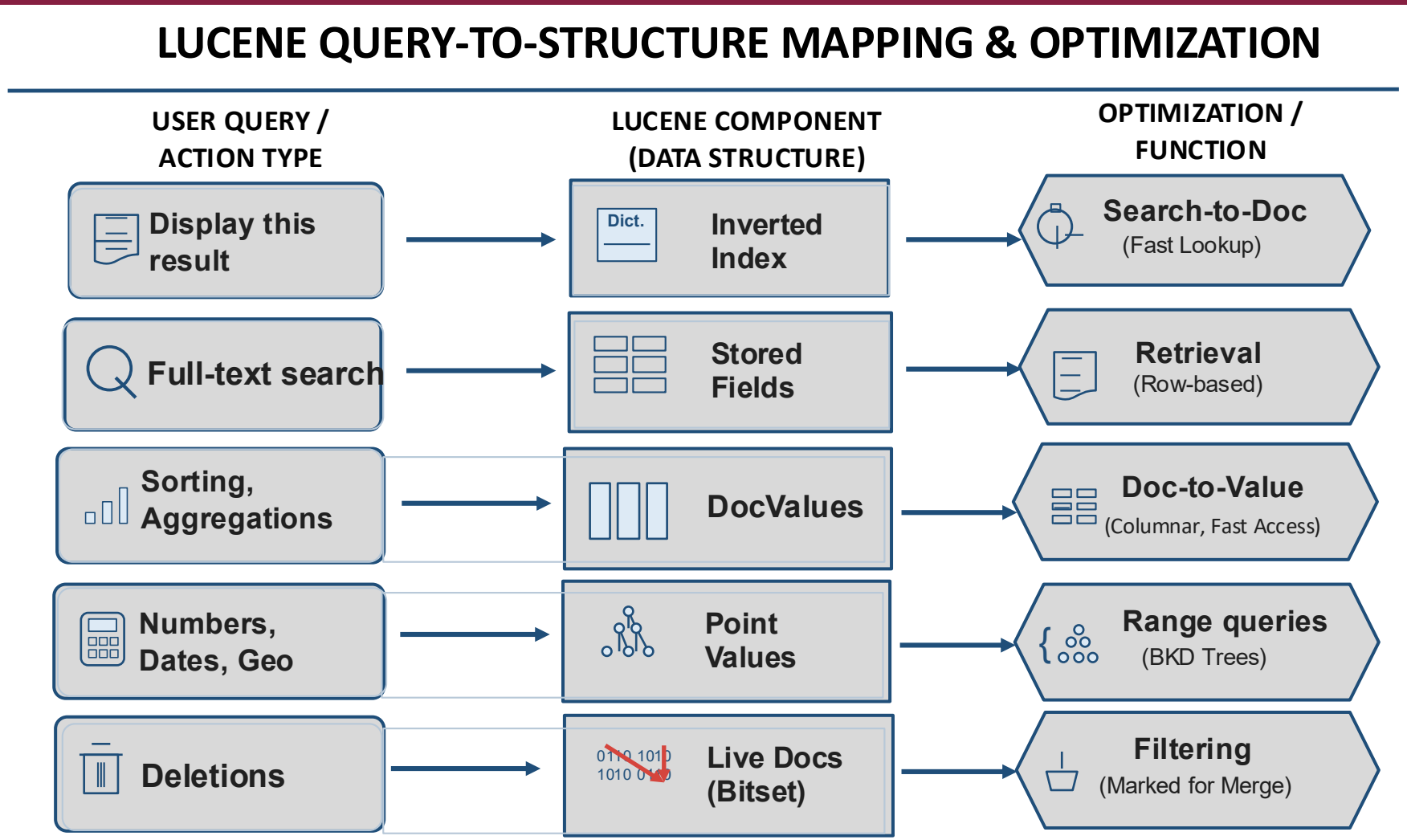
Existing Challenges

- **NL question tokens or SQL query tokens do not explicitly reflect query complexity.**
  - How many **patient** from state NC already took HPV4?
  - How many **teenager** from state NC already took HPV4?

- **Joint tables or any single key operation do not reflect query complexity.**
  - How to determine complexity when multiple tables and operations involved?
  - It is hard to reuse the taxonomy in other datasets.

Our solution: We build a more comprehensive taxonomy based on query time complexity for a broad and systematic use.

Lucene Based Data Structure and Elasticsearch



- **Lucene Data Structure:**
  - Widely deployed in Wikipedia, LinkedIn, and Twitter.
  - How to analyze the whole construction remains challenging.

Elasticsearch is a distributed, open-source search and analytics engine built on top of Apache Lucene, designed to handle massive volumes of data in near real-time.

elastic

Seven-Classes Complexity Protocol of ESQ

Operational reading

Relative query cost (conceptual scale)

Class	Query family	Core complexity	Load
A	**Constant lookup**	O(1)	
B	**Single-term match**	O(df(t))	
C	**Boolean filter**	O(Σ df(t_i))	
D	**Range / prefix†**	O(log N + R)	
E	**Pattern query**	O(W + Σ df)	
F	**Full-text ranking**	O(Σ df + R log K)	
G	**Analytics / sort**	O(Σ df + R + agg)	

Intensity Evaluation

Classes A–B: low intensity
Direct lookup and exact term filters are the safest choices for high-concurrency production.

Classes C–E: medium intensity
Boolean composition, ranges, and pattern expansion become sensitive to document frequency and traversal cost.

Classes F–G: high intensity
Full-text scoring, sorting, and aggregation add ranking or columnar work and dominate latency and memory.

- Classes A–B are generally safe for high-concurrency production.
- Classes C–E require closer monitoring because cost depends on document frequency and pattern expansion.
- Classes F–G are the most resource-intensive because they move from index lookup to scoring, sorting, or aggregation.

MedESQ Dataset

Medical Text-to-Elasticsearch (MedESQ)

ESQ Function Collection (A)

A	B	C	…
_id	ids	should	…
_doc	term	bool	…
⋮	⋮	⋮	⋮
_GET DB	terms	AND	…

Function Composition (B)

Easy	Medium	Hard
value	prefix	multi_match
Term + Terms	range	match_phrase
⋮	⋮	⋮
term + script-filter	constant_score + prefix	match + sort + aggs.terms

VAERS

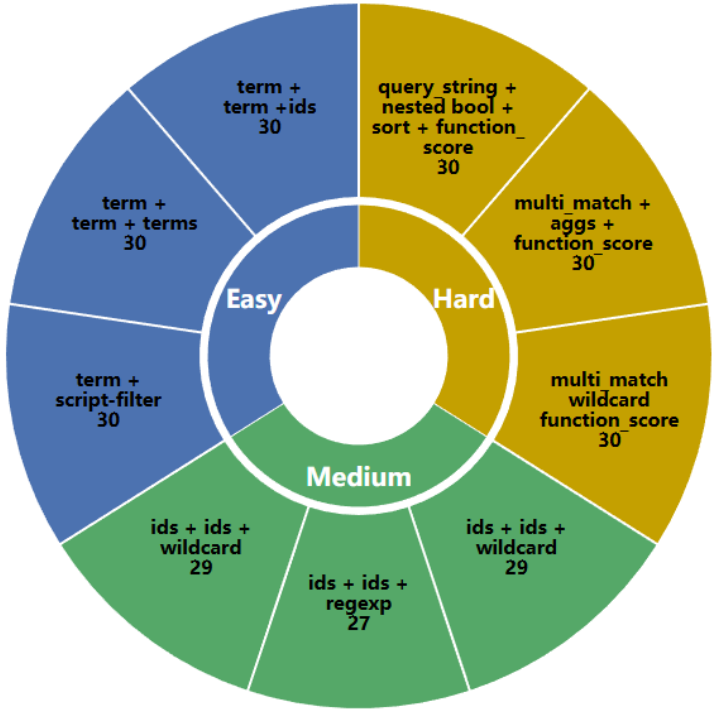
VAERS_ID	STATE	SEX	…
1996873	CA	U	…
1996875	OH	M	…
⋮	⋮	⋮	⋮
1996936	NC	F	…

NL Question Generation (D)

How many patient from state NC already took HPV4, HPV9, TTOX?

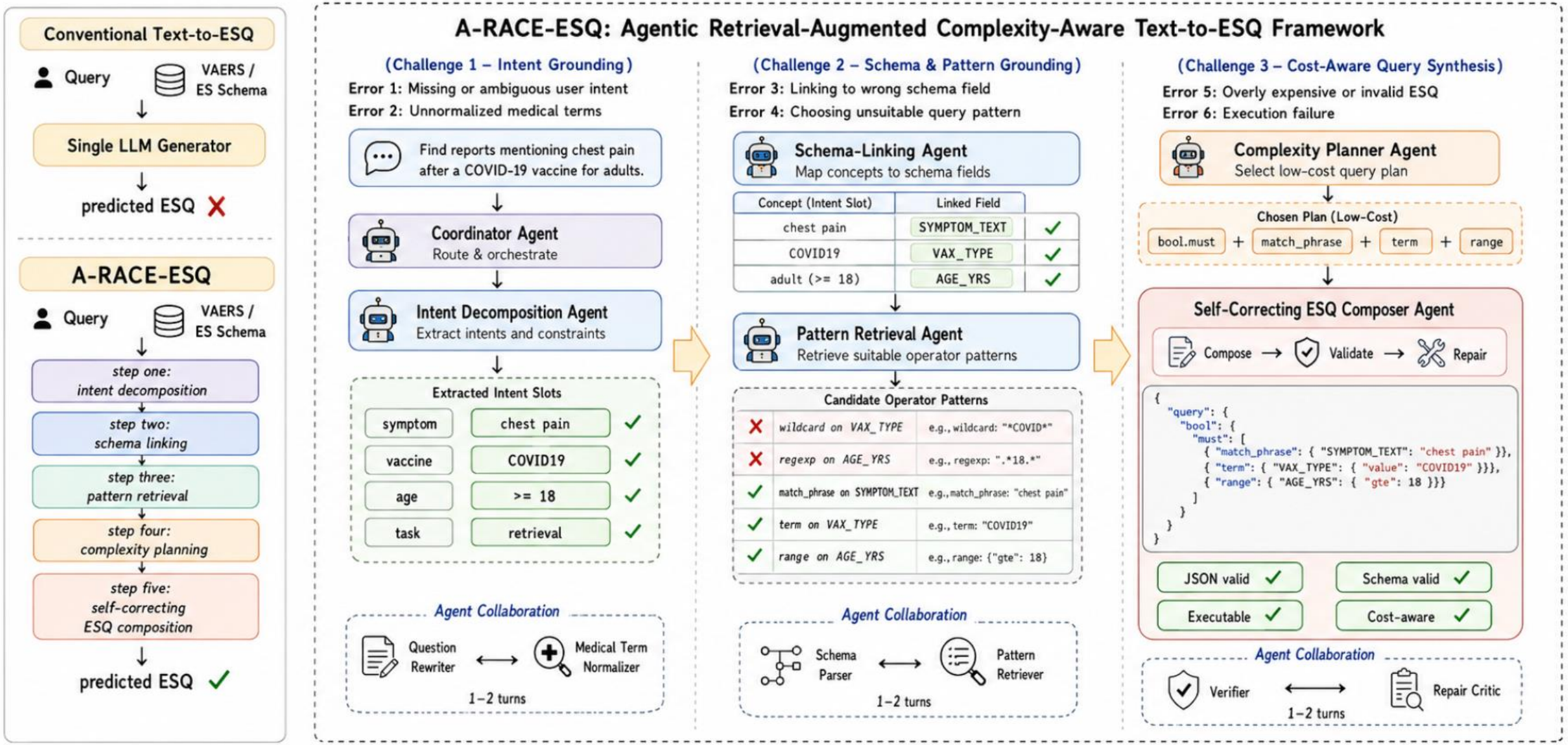
ESQ Query Generation (C)

```
{"query":
  {"bool":
    {"must":
      [{"term":
        {"STATE":{"value": 'NC'}}},
        {"terms":{"VAX_TYPE":
              ["HPV4','HPV9','TTOX'
]}
      }]
    }
  }
}
```



Distribution of the top three frequent function combinations in different difficulty levels in MedESQ.

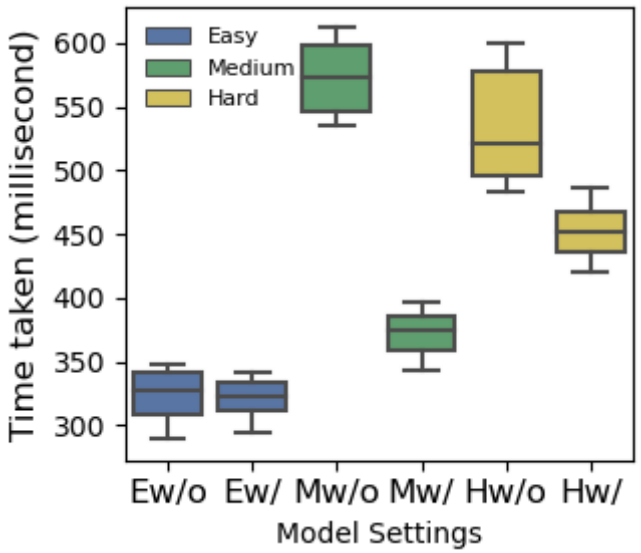
Agentic Retrieval-Augmented Complexity-Aware Text-to-ESQ Framework



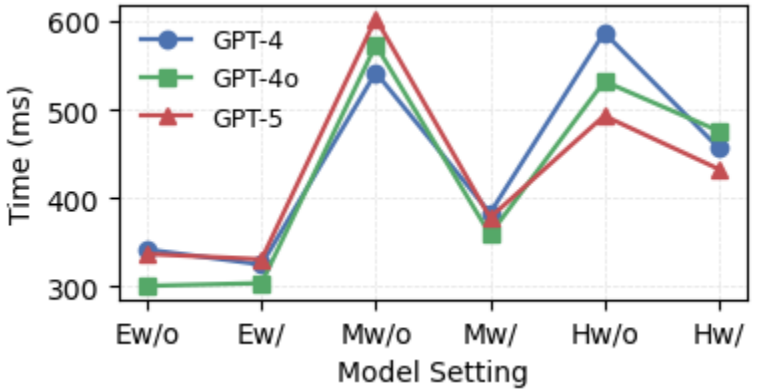
Performance of RACE-ESQ on MedESQ

Table 4: Performance of six models on **MedESQ** dataset. Each score is the median of three runs on 200 sampled queries per level. The best performance is shown in **bold**, and the second best is underlined.

Model	Easy		Medium		Hard	
	ECR	PR	ECR	PR	ECR	PR
GPT4	80%	64%	77%	36%	55%	35%
GPT4o	85%	77%	80%	65%	67%	55%
GPT5	**97%**	87%	80%	67%	75%	63%
Grok4	**97%**	**90%**	72%	69%	71%	61%
GLM4.5	90%	75%	76%	62%	67%	40%
Qwen3	87%	80%	78%	60%	65%	53%
RACE-ESQ	93%	87%	**82%**	**79%**	**79%**	**77%**



Query runtime of GPT5 with/without RACE-ESQ



Query runtime from models with/without RACE-ESQ

Accuracy

Efficiency

Summary

- **Findings:**
  - Explicit efficiency factors, such as operator complexity and query difficulty, are highly useful for guiding efficient query generation.
  - Efficiency-aware modeling provides more informative performance evaluation beyond query correctness.

- **Impacts:**
  - This work introduced efficiency as a new evaluation dimension for Text-to-ESQ.
  - Improved practical relevance of Text-to-ESQ evaluation and generation.

Summary of NLQ on NoSQL Databases

Accuracy

- **Task formulation and benchmark**
- Text-to-ESQ task formalization; VAERSESQ benchmark; Two-stage controllable model

Reasoning

- **LLM-based Reasoning augmentation**
- InstructEx prompting schema-aware reasoning; Executable LLM generation

Efficiency

- **Efficiency-aware execution**
- Complexity protocol; MedESQ benchmark; RACE-ESQ optimization

Thank you!

Q & A

Email: ping.wang@stevens.edu

Web: https://leafnlp.org/ping

The LEAF Lab @Stevens: https://leafnlp.org/