PP02 JSON Data Generation: Linking Statistical Analysis with Large Language Models

Changhong Shi , Merck & Co., Inc., Rahway, NJ, USA (changhong_shi@merck.com)
Jaime Yan, Merck & Co., Inc., Rahway, NJ, USA (mingyu.yan1@merck.com)

A framework for link structure data with LLM via extrating schema with JSON format

Background

Large Language Models (LLMs) excel at natural language tasks but struggle to interact with structured data. Conversely, structured clinical datasets, such as those following the CDISC ADaM standard, are essential for comprehensive data analysis, yet their complex schema poses challenges for LLMs. To bridge this gap, we present a transformative approach that seamlessly integrates LLMs with the generation and utilization of synthetic clinical datasets.

Advantages of JSON and Graph Databases

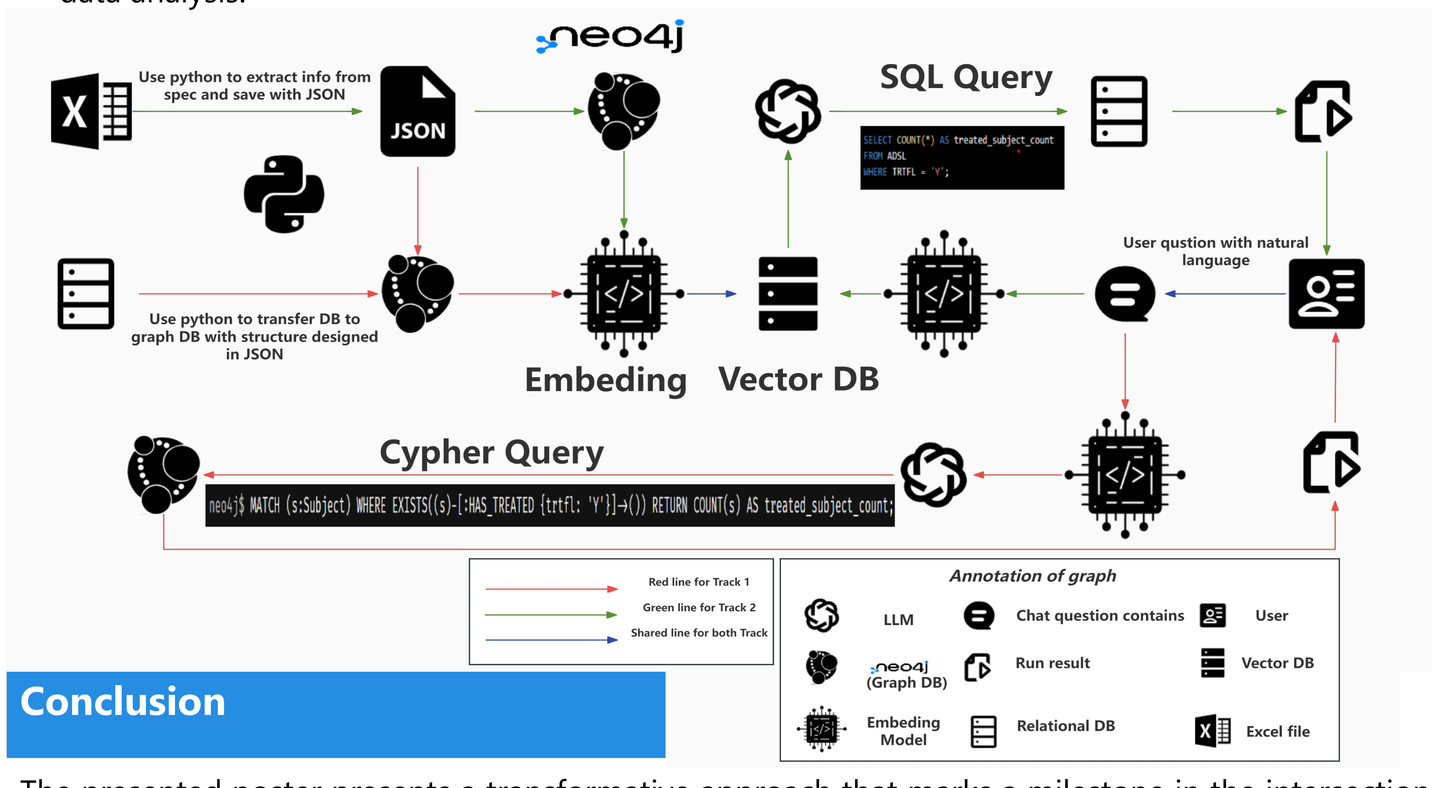
1. Flexible schema representation in JSON, with efficient relationship extraction and storage in graph databases.
2. Seamless integration of structured data with Large Language Models through graph-based modeling.
3. Comprehensive framework combining the strengths of JSON and graph databases for effective clinical data analysis.

Motivation

- **Overcoming LLMs' Limitations with Structured Data:** LLMs excel at natural language tasks but often struggle with directly processing and analyzing structured data. By leveraging the power of LLMs to comprehend and manipulate schemas, we aim to enable these models to work effectively with structured clinical datasets.
- **Bridging the Gap between Structured Data and Analytical Prowess:** Structured clinical datasets contain valuable information, but their intricate nature can limit the analytical capabilities of LLMs. Our approach seeks to design synthetic datasets that bridge this gap, allowing LLMs to harness their full potential in clinical data analysis.
- **Enhancing Efficiency and Automation in Healthcare Insights:** The current clinical data analysis workflows are often time-consuming and labor-intensive. By integrating LLMs with the generation and processing of synthetic clinical data, we aim to significantly improve the efficiency and automation potential of these processes, ultimately driving faster and more impactful healthcare insights.

Outcome and Benefits

- Develop a robust and scalable method to extract the schema information from structured clinical datasets, such as those following the CDISC ADaM standard.
- Establish a comprehensive framework for representing the schema and relationships of structured clinical data using JSON and graph-based databases.
- Seamlessly integrate LLMs with the designed graph-based representations of structured clinical data, enabling efficient and automated data analysis workflows.
- Showcase the practical applications and benefits of this approach through detailed case studies, demonstrating its potential to advance healthcare insights and decision-making:
- **Track 1:** Transfer original data to graph DB via the designed graph based on extracted JSON file. LLM will provide Cypher query based on the retrived info from Vector DB.
- **Track 2:** Only extract the schema from spec and save with JSON format, LLM will provide SQL query based on the retrived variable chain from Vector DB.



Leverage the schema information from the CDISC ADaM standard or other relevant structured clinical data specifications to extract the schema details into a JSON format.

Extract Schema from Structured Clinical Datasets:

Perform graph embedding on the designed database to generate vector representations that can be efficiently queried and integrated with LLMs.

Embed Graph Database:

01 02 03 04

Design Graph-based Database:

Utilize the JSON schema representation to design the structure and relationships of a graph-based database, such as Neo4j or a similar graph technology.

Figure 1. Snapshot of ADAE dataset schema with Json format

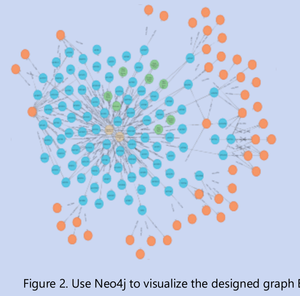


Figure 2. Use Neo4j to visualize the designed graph BD based on the schema with Json format

LLM Interaction with Structured Clinical Data:

1. Enable LLMs to interact with the structured clinical data by:
   a. Retrieving the relevant variables and relationships from the graph database based on the user's natural language query.
   b. Generating tailored code, such as SQL or Cypher, to query the underlying structured clinical data sources.
   c. Providing the results back to the user, leveraging the structure and relationships captured in the graph-based representation.

Conclusion

The presented poster presents a transformative approach that marks a milestone in the intersection of artificial intelligence and structured clinical data analysis. By seamlessly integrating LLMs with the representation and processing of structured clinical data, we have unlocked new possibilities for efficient, automated, and insightful data analysis, paving the way for groundbreaking advancements in healthcare.

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the author by scan the right QR code.