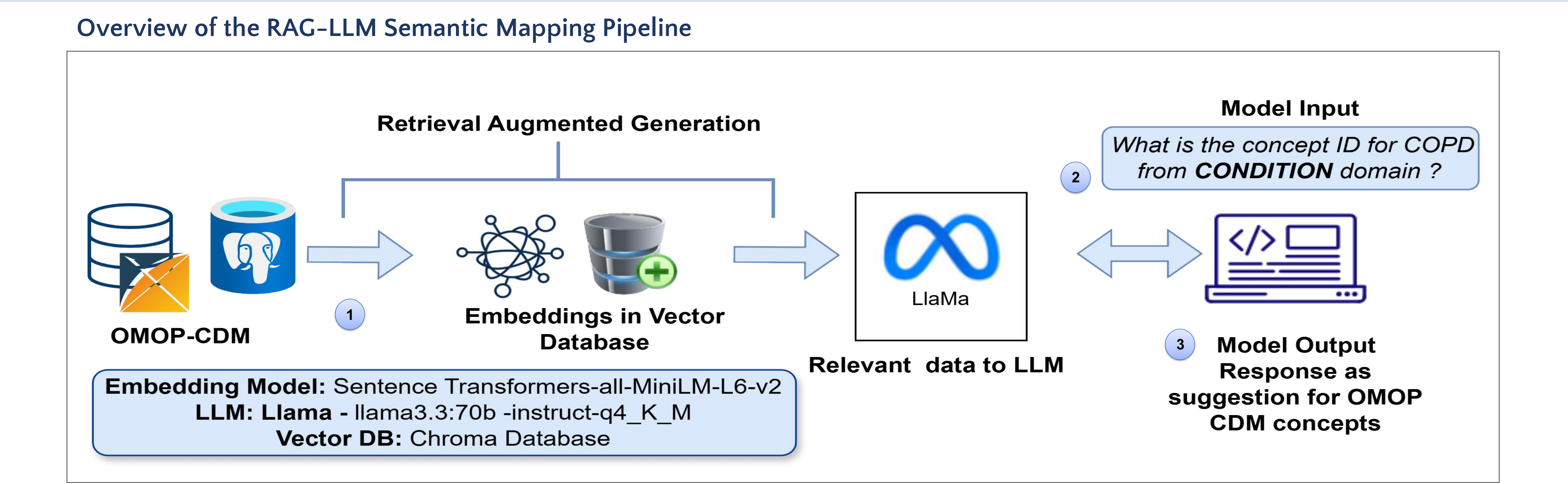


# Accelerating Feature Extraction with AI-Powered RAG-LLM: Automated Concept Mapping to OMOP-CDM Vocabulary.

*Title: RAG-Enhanced LLM Pipeline for Semantic Mapping of Context-based Features to OMOP Vocabulary*

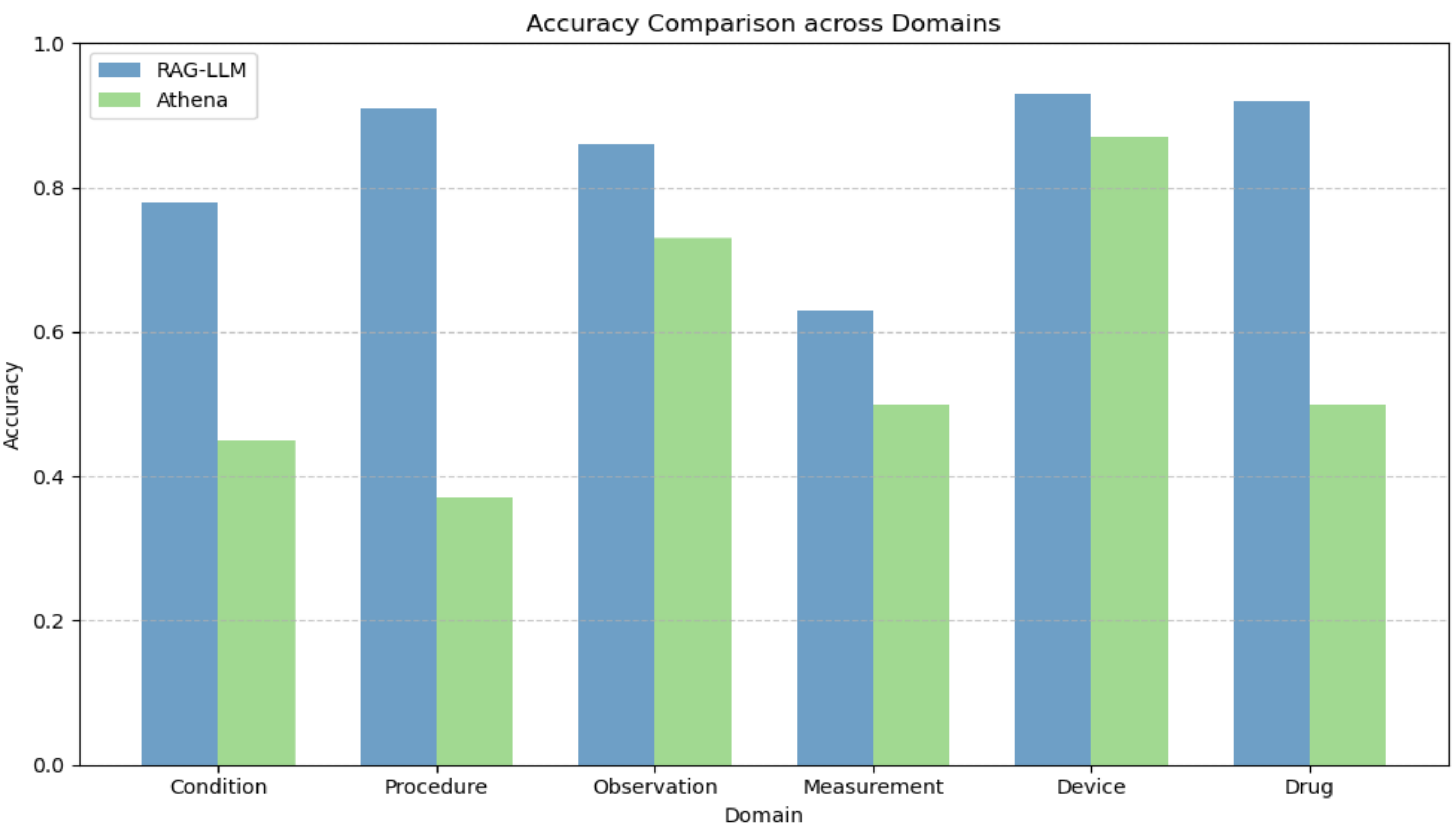
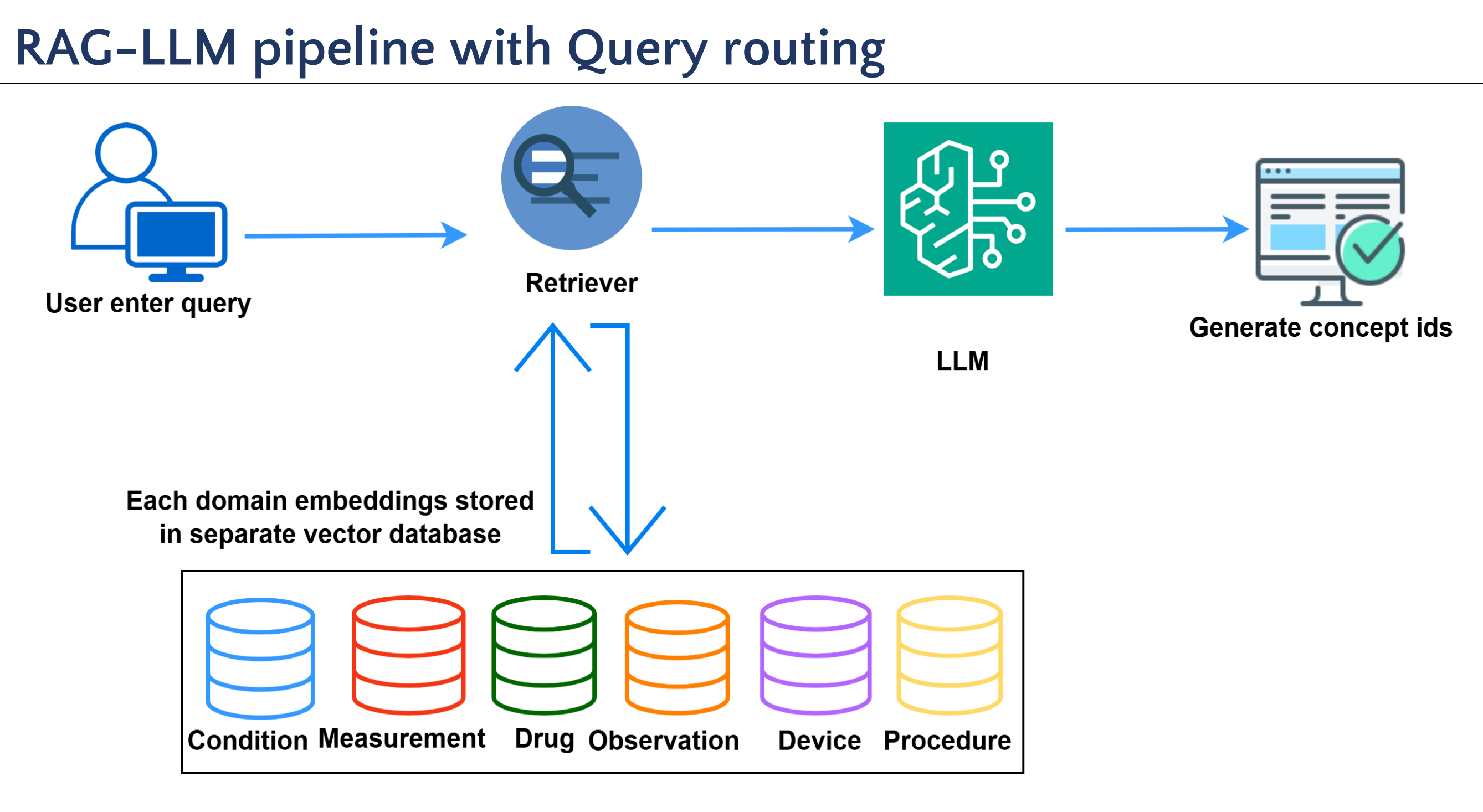
**Background:** Observational health data are often standardized to the commonly used OMOP-CDM standards. This enables us to carry out efficient analyses that can generate reliable evidence. However, understanding these standards and vocabulary terms requires medical knowledge, along with OMOP-CDM expertise. This makes feature extraction crucial, particularly for users without domain expertise.



**Method:** In the end-to-end pipeline: (1) OMOP concepts stored in a vector database. (2) User input is encoded and compared against pre-generated embeddings (3) The top-k most semantically similar matches are retrieved and LLM generates context-aware concepts as suggestions.

**Results:** The pipeline achieved improved performance over the OHDSI tool Athena. With the proposed approach in addition to the suggestions on matching concepts we get explanation on which concepts are more appropriate as per the input query.

**Conclusion:** The proposed tool mainly focuses on aiding the AI model developers to evaluate their software with a focus on safety, efficacy, and usability, for the direct benefit of patients and healthcare practitioners.



Mapping Accuracy for RAG-LLM pipeline versus Athena at top k = 10 searches



Sariga Kakkamani<sup>1</sup>, Frédéric Jung<sup>2</sup>, Joeri Verbiest<sup>1,3,4</sup>, Liesbet Peeters<sup>1,3,4</sup>

<sup>1</sup>Data Science Institute (DSI), Hasselt University, Diepenbeek, Belgium  
<sup>2</sup>VITO, Vlaamse Instelling voor Technologisch Onderzoek, Mol, Belgium  
<sup>3</sup>Biomedical Research Institute, Hasselt University, Diepenbeek, Belgium  
<sup>4</sup>University MS Center (UMSC), Hasselt-Pelt, Belgium



Funded by  
the European Union

