Model Metadata Chatbot with GenAI

DATA SCIENCE INSTITUTE COLUMBIA UNIVERSITY

Tina Cao, Lu Liu, Wanxin Luo, Xinwei Qiao, Yao Xie
Mentors: Sydney Son, Thomas Covella, Chengwei Wang, Sining Chen

**Data Science Capstone Project with**

KPMG

Introduction

As businesses increasingly adopt models, they face challenges in navigating the complex landscape of interconnected reports and models, each with significant upstream and downstream impacts. Consequently, the risks associated with these models also amplify.

This project aims to develop a cutting-edge solution that leverages generative AI technology for effortless incorporation into daily business workflows and efficient handling of metadata management.

Workflow
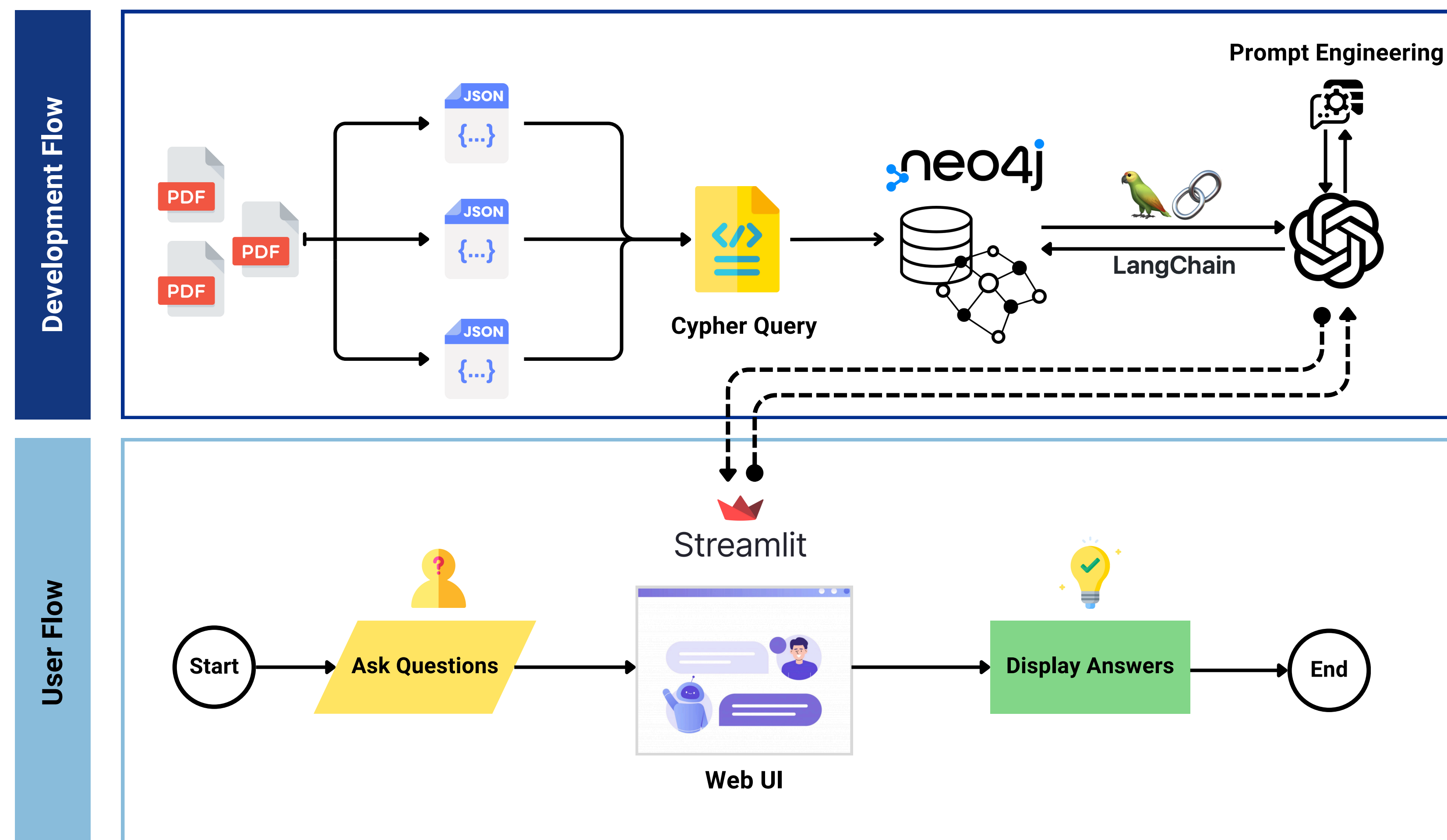
The chatbot development begins by converting PDFs to Cypher queries based on JSON schema files for integration into a Neo4j graph database, forming an interconnected knowledge network. Through LangChain and prompt engineering, the chatbot's capability to understand and generate accurate answers is refined. Finally, the system is deployed into a web interface using Streamlit, completing the end-to-end development flow.

Neo4j
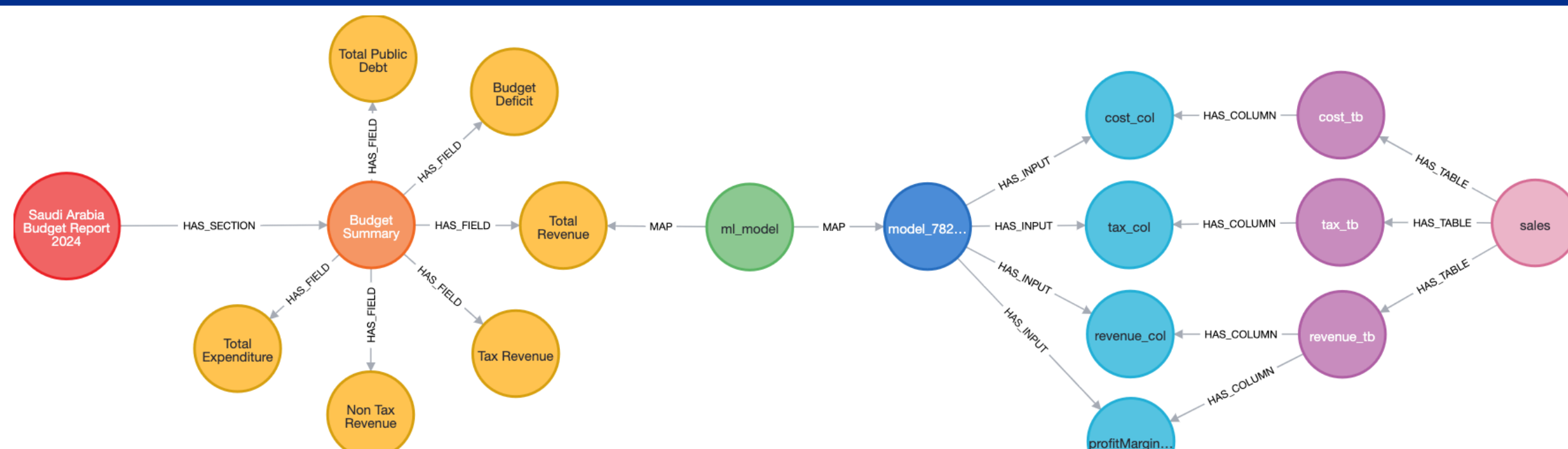
The graph as a whole represents the flow and relationship of data from a high-level report down to the specific tables in a database that feed into an ML model, delineating the hierarchy from report nodes, to sections, to models, to input columns, and finally to the tables and the encompassing database.

Chatbot Logics & User Interface
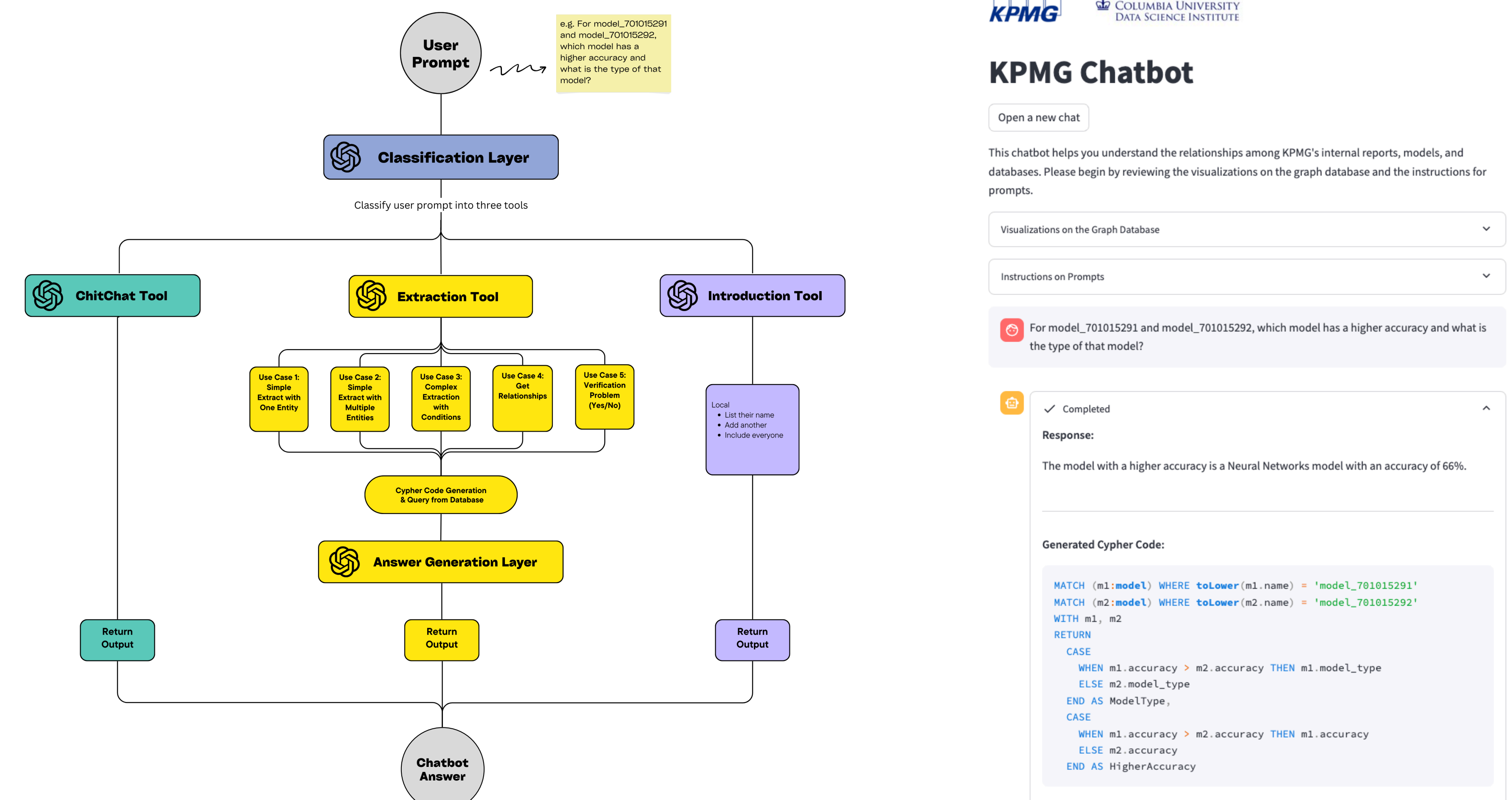
We designed a sophisticated chatbot system with a classification layer that assigns user prompts to one of three specialized tools: the ChitChat Tool for casual conversations, the Introduction Tool for chatbot usage guidelines, and the Extraction Tool for complex data retrieval from our Neo4j graph database. These tools, driven by the OpenAI API, employ various prompt engineering techniques with guardrails to avoid AI hallucinations. The Extraction Tool is trained on five use cases, with 20 sets of user prompts and the corresponding Cypher codes. Once the classification process is completed, the output is transformed into a front-end UI/UX, providing clear and concise responses to user questions.

Performance
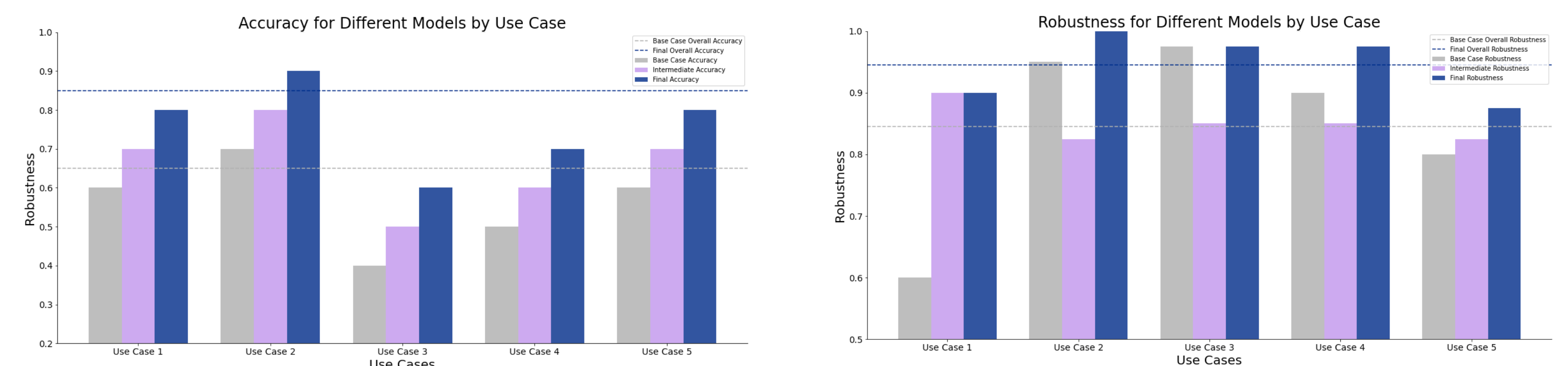
We developed an automated testing pipeline to assess the semantic consistency between the ground truth and the answers produced by our model. Accuracy is measured by the average agreement between the generated answers and the underlying facts, while robustness is evaluated based on the consistency of the answers—whether correct or not—over three trials. The plots reveal that our model excels in accuracy and robustness, surpassing the baseline model, particularly in the complex Use Cases 3, 4, and 5.

Conclusion

Our model introduces an advanced GenAI-powered chatbot designed to streamline data analysis for scientists, enhance system integrity for engineers, and provide new graduates with a user-friendly onboarding experience — all aimed at efficiently delivering business value to KPMG.