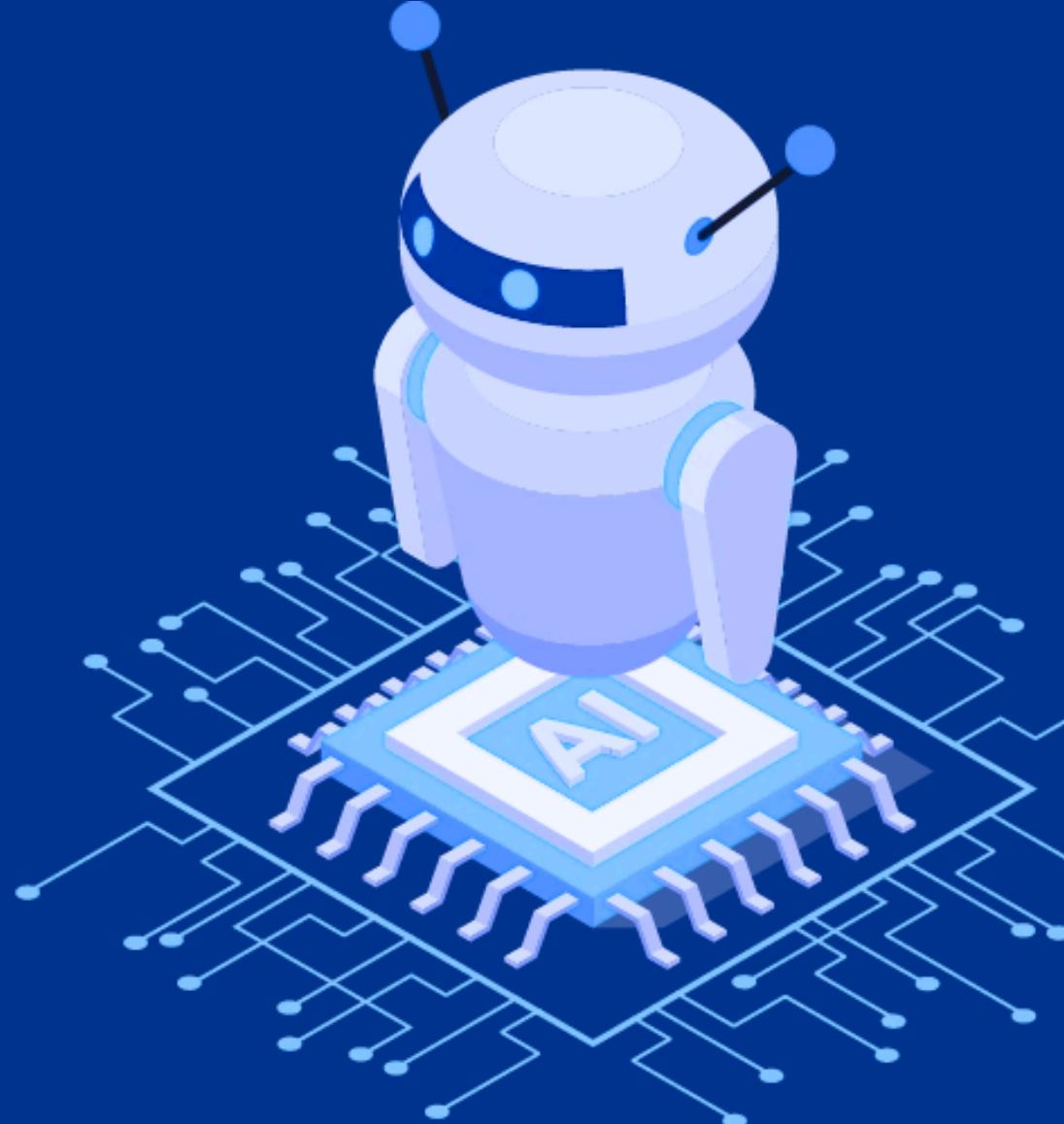




# Model Metadata Chatbot with GenAI

Tina Cao, Lu Liu, Wanxin Luo, Xinwei Qiao, Yao Xie

**Mentors:** Sydney Son, Thomas Covella, Chengwei Wang



- 1 Introduction
- 2 Our Approach
- 3 Business Case Implementation & Demo
- 4 Conclusion & Future Improvements
- 5 Acknowledgement & References

- 
- 1 Introduction
  - 2 Our Approach
  - 3 Business Case Implementation & Demo
  - 4 Conclusion & Future Improvements
  - 5 Acknowledgement & References

# 1. Introduction



## 1. Our Team

1.1 Our Team

COLUMBIA UNIVERSITY  
DATA SCIENCE INSTITUTE

• First Year Master of Science in Data Science Students  
• Affiliated with the Data Science Institute at the School of Engineering and Applied Science of Columbia University  
• Course highlights: Natural Language Processing, Deep Learning, Machine Learning, Algorithms, Computer Systems, Statistics, etc.

Tina Cao (Chatterbot Development, UI/UX Interface, Mock Data Generation, Mock API Testing, Classification Model) t3333@columbia.edu  
Yuxuan Guo (Mock Data Generation, Mock API Testing, AI-powered Engineering, Mock API Testing) m2222@columbia.edu  
Lu Liu (Mock Data Generation, Chatbot Development, AI-powered Engineering, Mock API Testing) l3333@columbia.edu  
Xirui Qiao (Chatterbot Development, Mock Data Generation, Mock API Testing, Mock API Testing) x3333@columbia.edu  
Yibo Xie (Mock Data Generation, Chatbot Development, Mock API Testing, Mock API Testing) y3333@columbia.edu

5

## 2. Background

1.2 Background

COLUMBIA UNIVERSITY  
DATA SCIENCE INSTITUTE

Context: As companies incorporate AI-powered tools into their day-to-day core operations of business. This significantly increases operation efficiency but also introduces complexity and risks of model management.

Research Question: How to create a smart management system that can leverage AI-powered tools to improve companies' operation efficiency while minimizing model risks?

Project Flow:

- Mock Data Generation & Graph Database - Neo4j
- Chatbot Development
- Iterations
- User Interface
- Business Case Implementation (Giving recommendation scores on optimized business)

5

# 1.1 Our Team

- 1st Year Master of Science in Data Science Students
- Affiliated with the Data Science Institute at the School of Engineering and Applied Science at Columbia University
- Course highlights: Natural Language Processing, Deep Learning, Machine Learning, Algorithms, Computer Systems, Statistics, etc.



**Tina Cao** (*team captain*)  
*Chatbot Development,  
UI/UX Interface,  
Prompt Engineering  
(Classification Tools)*  
[tc3334@columbia.edu](mailto:tc3334@columbia.edu)



**Wanxin Luo**  
*Mock Data Generation,  
Neo4j Visualization,  
Prompt Engineering  
(Cypher Code Generation)*  
[w12930@columbia.edu](mailto:w12930@columbia.edu)



**Lu Liu**  
*Mock Data Generation,  
Chatbot Development,  
Prompt Engineering  
(Answer Layer & Guardrail)*  
[ll3721@columbia.edu](mailto:ll3721@columbia.edu)



**Xinwei Qiao**  
*Chatbot Architecture  
Development,  
Automatic Testing Pipeline  
Development*  
[xq2236@columbia.edu](mailto:xq2236@columbia.edu)



**Yao Xie**  
*Mock Data Generation,  
Chatbot Development,  
Prompt Engineering  
(Cypher Code Improvement)*  
[yx2845@columbia.edu](mailto:yx2845@columbia.edu)

## 1.2 Background

### Context

As companies incorporate AI-powered tools into their day-to-day core operations of business. This significantly increases operation efficiency but also introduces complexity and risks of model management.

### Research Question

How can we create a smart management system that can leverage AI-powered tools to improve companies' operation efficiency while minimizing model risks?

### Project Flow

Mock Data  
Generation &  
Graph Database  
- Neo4j

Chatbot  
Development

Iterations

User  
Interface

Business Case  
Implementation  
(Giving recommendation scores on optimized locations)

- 1 Introduction
- 2 Our Approach
- 3 Business Case Implementation & Demo
- 4 Conclusion & Future Improvements
- 5 Acknowledgement & References

## 2. Our Approach



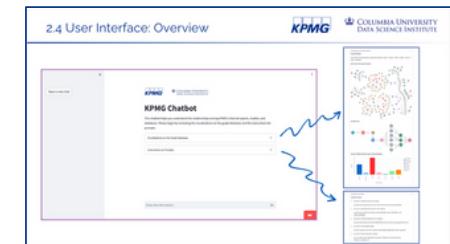
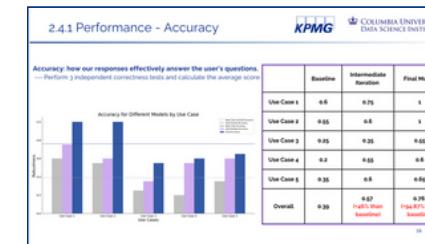
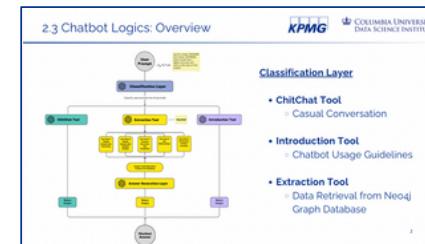
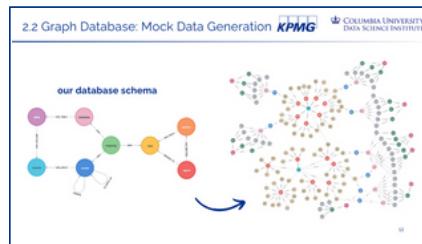
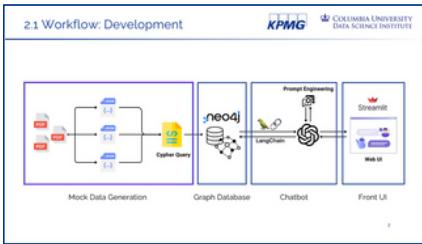
1. Process

2. Graph Database

3. Chatbot Logics

4. Performance

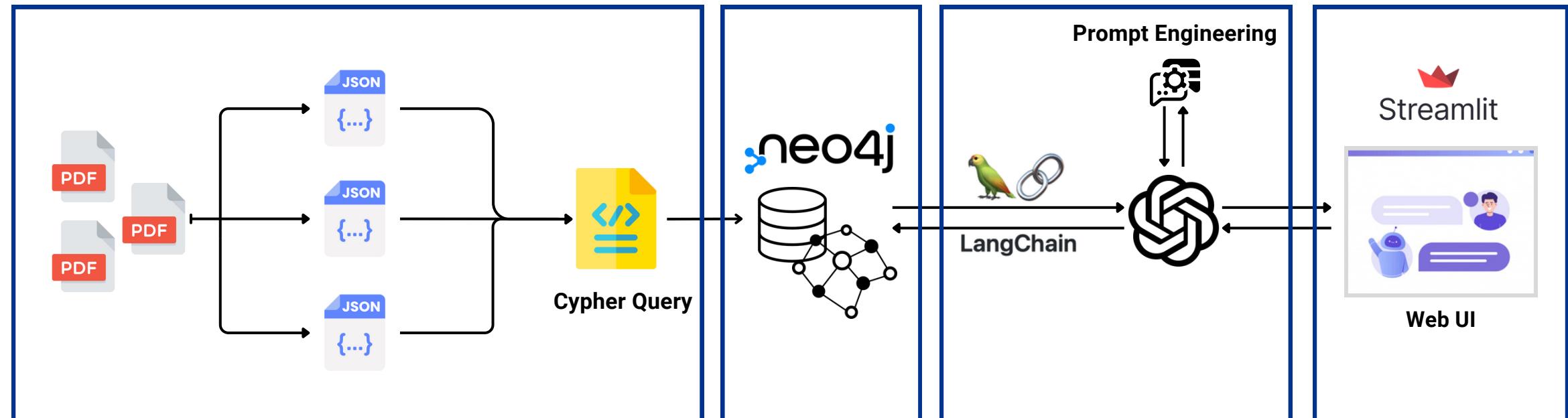
5. User Interface



## 2.1 Process: Technical Flow



This project showcases an end-to-end workflow, developed from mock data generation to a user-friendly front-end UI.



### Mock Data Generation:

The workflow begins by generating mock data from reports in PDF formats to the Cypher queries based on the JSON schema.

### Graph Database:

Using Cypher queries, this data is loaded into a Neo4j graph database.

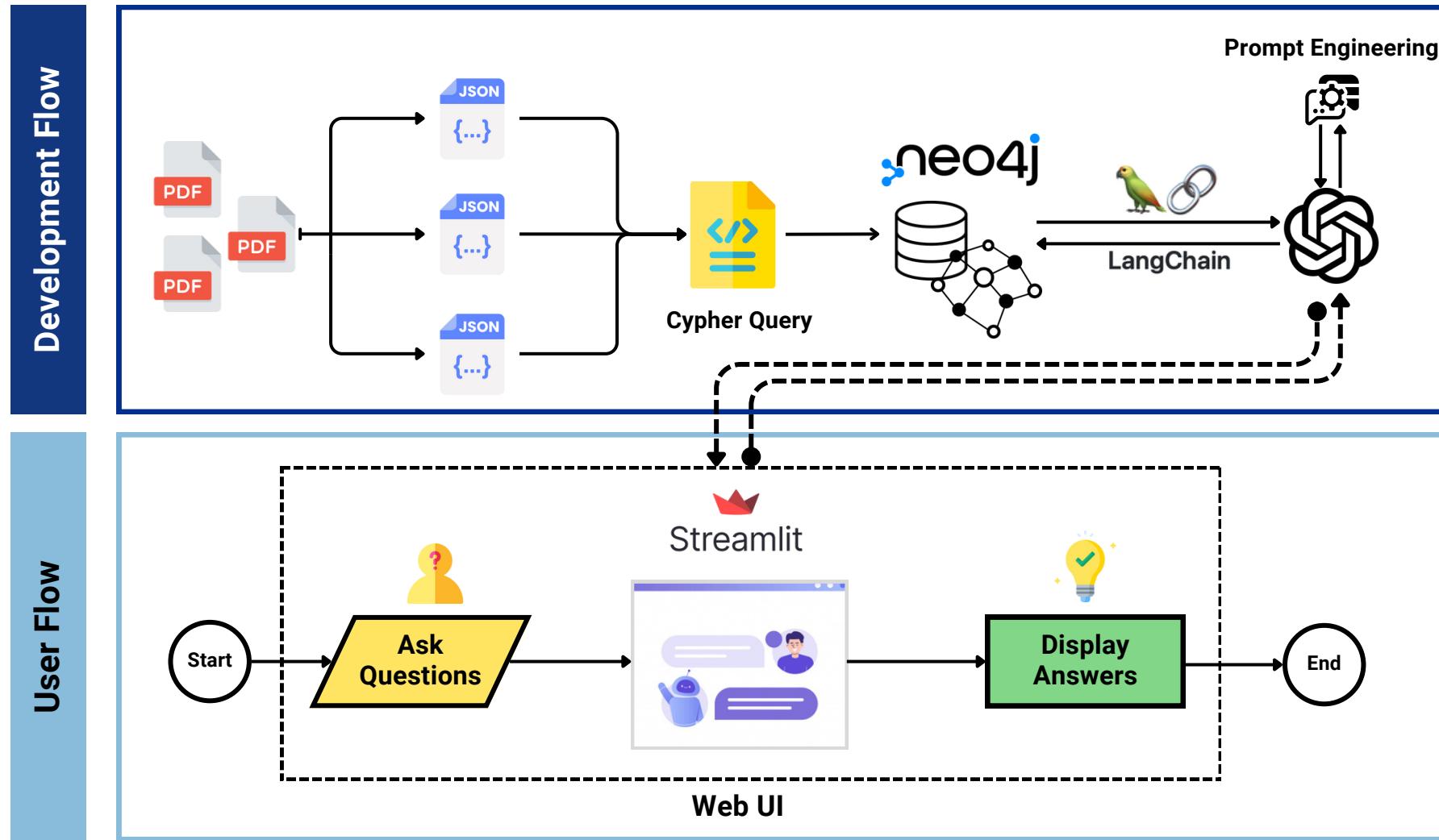
### Chatbot:

The chatbot utilizes LangChain to enable the ChatGPT API to interact with Neo4j.

### Front UI:

Integrated with Streamlit Web UI.

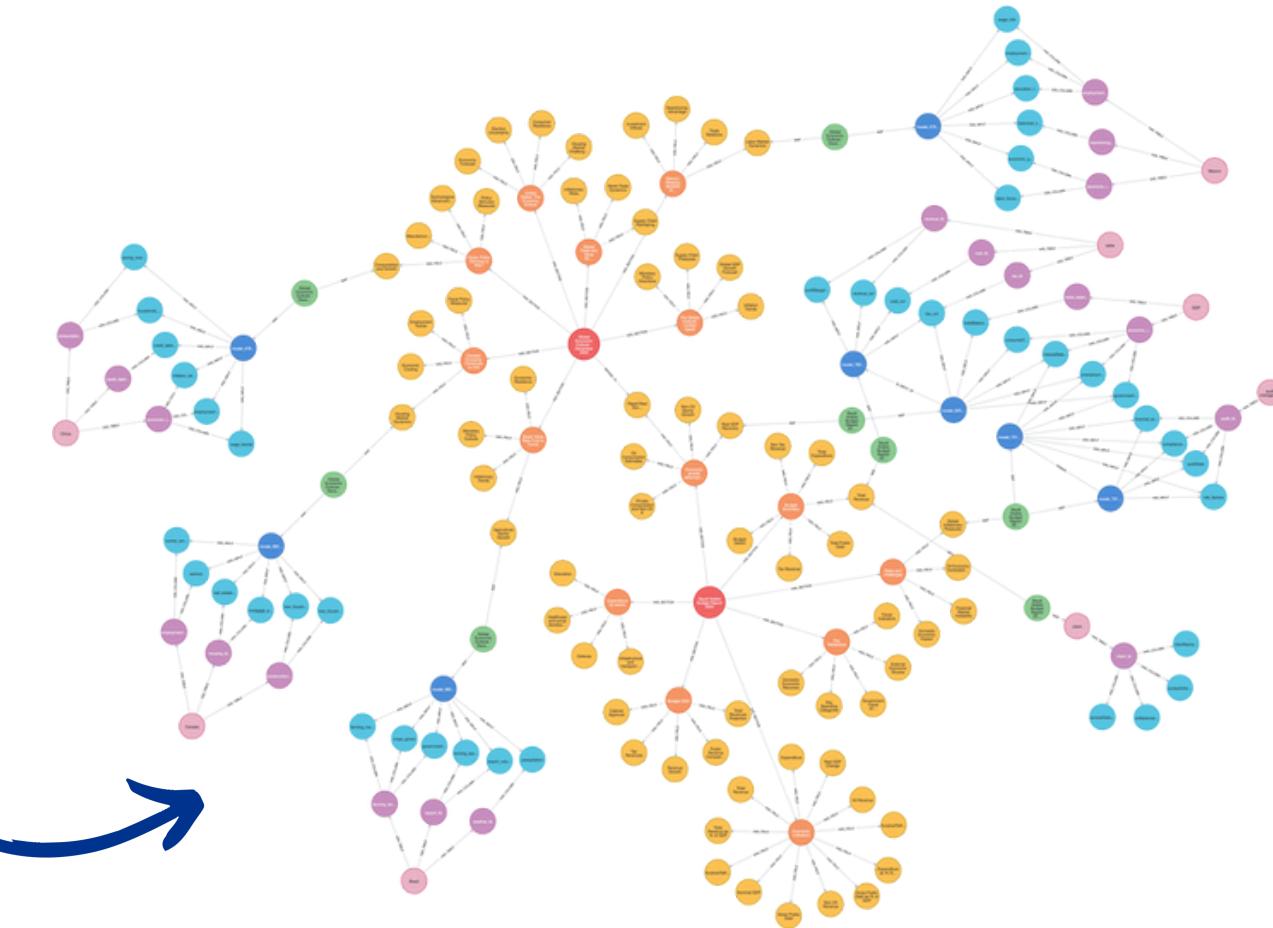
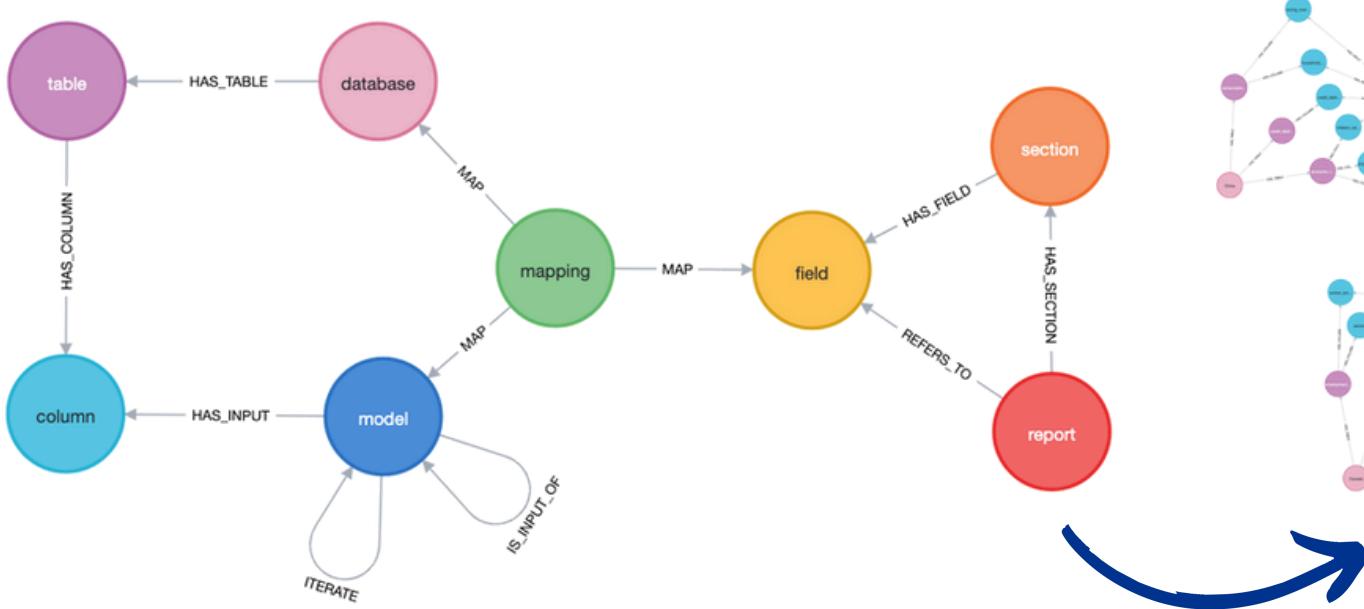
## 2.1 Process: User Flow



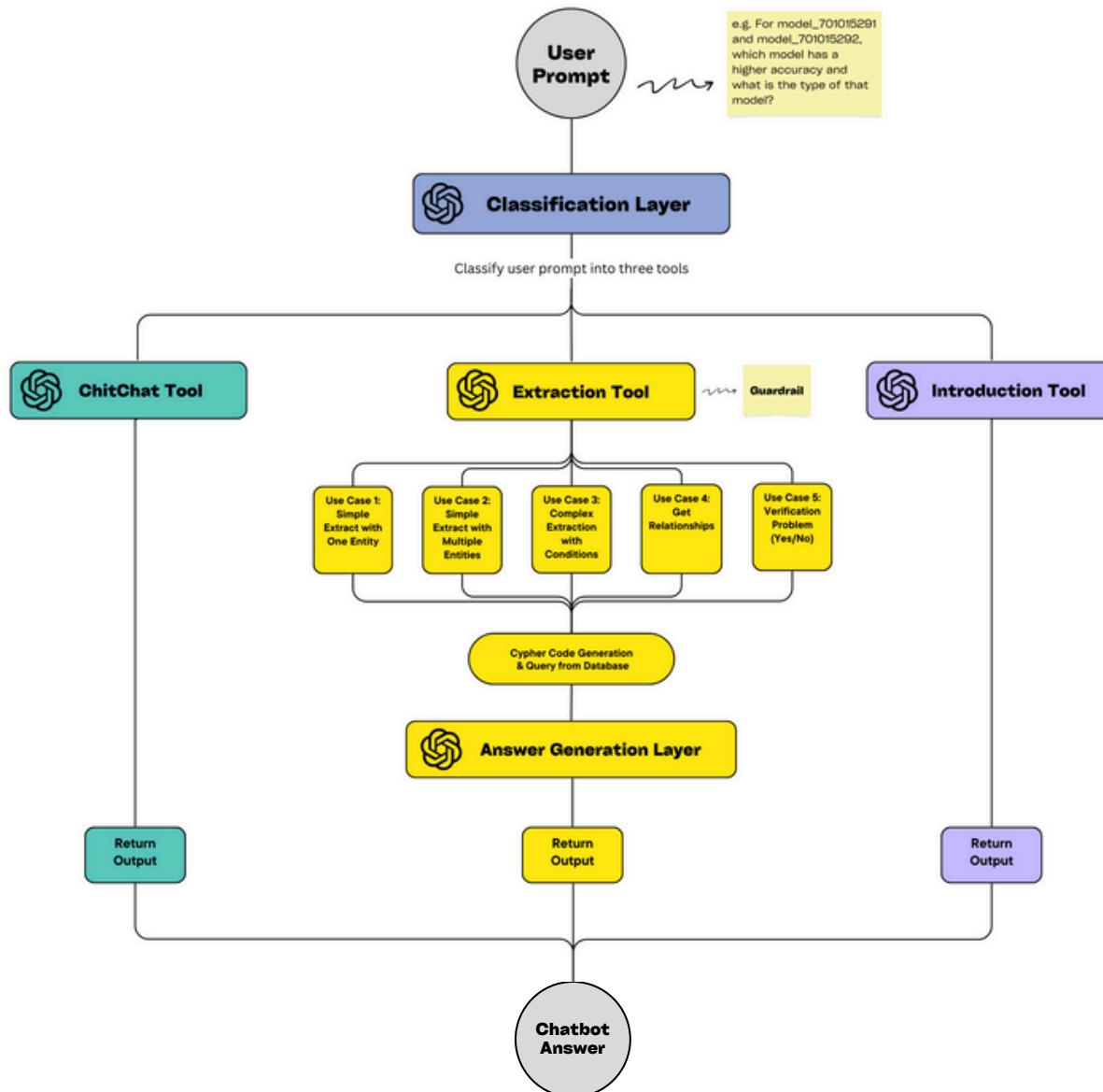
- **Graph database** stores model metadata as nodes (entities) and edges (relationships).
- **Neo4j** is a leading graph database management system.
- **Cypher** is Neo4j's query language that enables us to retrieve data from the graph.
- Enhances visualization & management of data relationships.
- Empowers our chatbot to handle metadata.



### our database schema



## 2.3.1 Overview of the Final Chatbot Logic



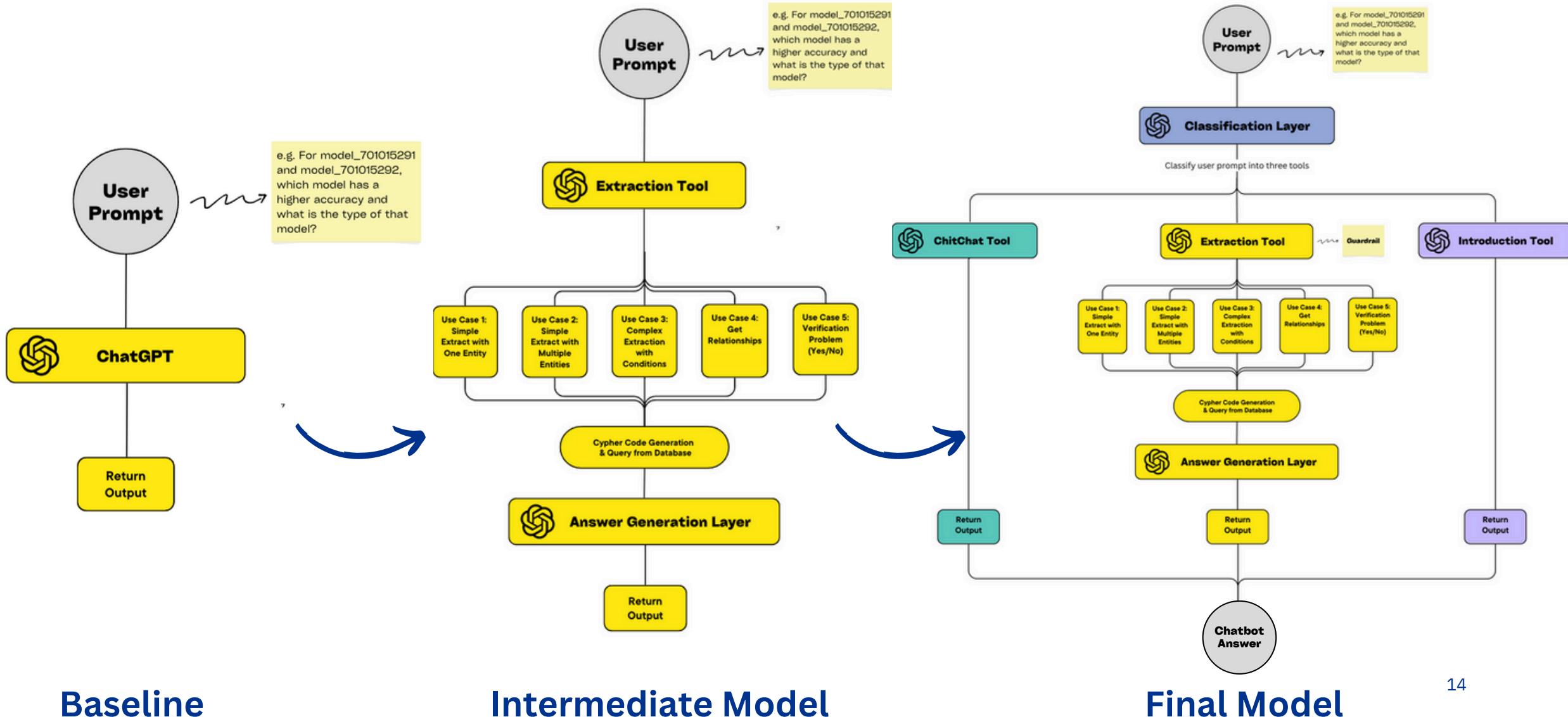
### Chatbot Overview:

- Handles queries related to reports, data, and models.
- Engages users in conversations and provides guidance on chatbot usage.

### Classification Layer to Categorize User Prompts:

- ChitChat Tool:** Casual conversation
  - e.g., "Can you recommend some restaurants in New York?"
- Introduction Tool:** Chatbot usage guidelines
  - e.g., "How do I use this chatbot?"
- Extraction Tool:** Data retrieval from Neo4j graph database
  - e.g., "For model1 and model2, which model has higher accuracy and what is the type of that model?"

## 2.3.2 Structural Iterations



## 2.3.3 Iteration Improvements



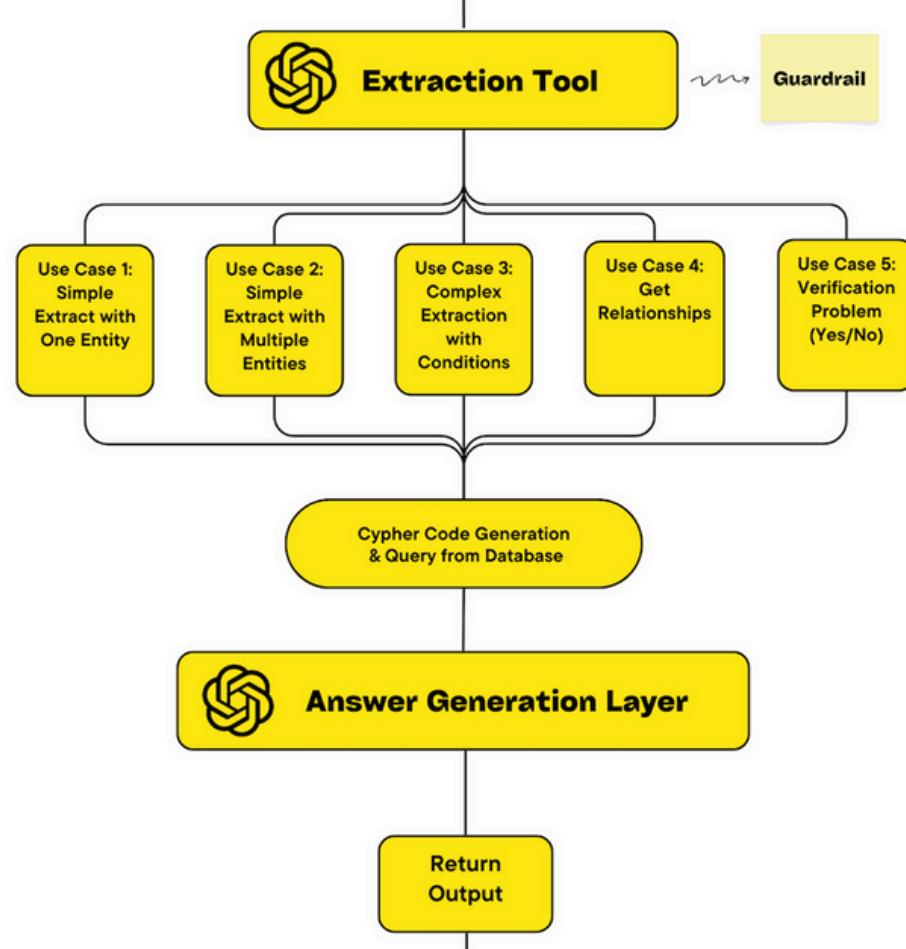
### Baseline --> Intermediate

- **Better Extraction Accuracy**
  - Improved handling and accuracy in data extraction from the database
- **Tailored Use Cases with Small Train Sets**
  - Developed specific use cases with smaller, focused training datasets to refine the bot's responses.

### Intermediate --> Final

- **Enhanced Extraction Accuracy**
  - Further improvements in accuracy for more complex queries.
- **Expanded Training Use Cases**
  - 20 sets of user prompts and their corresponding Cypher codes to cover more scenarios
- **Introduction of Classification Layer**
  - Added ChitChat Tool and Introduction Tool to better categorize user prompts
- **Implementation of Guardtrail**
  - Introduced safety measures to prevent misuse or handling of sensitive topics.
- **Case Insensitivity**
  - Enhanced the system to handle queries irrespective of text case, improving user interaction.
- **Synonym Handling**
  - Integrated capability to understand and process synonyms, enhancing the system's natural language understanding.

## 2.3.4 Our Core: The Extraction Tool



### The Extraction Tool

#### • Training Phase

- Use Case 1: Simple Extraction with One Entity
  - e.g., "Who is the author of this report?"
- Use Case 2: Simple Extraction with Multiple Entities
  - e.g., "What are the output columns of model1, model2, and model3?"
- Use Case 3: Complex Extraction with Conditions
  - e.g., "How many models in the 'Consumption and Growth' field have an R-squared higher than 0.8?"
- Use Case 4: Get Relationships
  - e.g., "What report does model1 depend on?"
- Use Case 5: Verification Problem (Yes/No)
  - e.g., "Does model1 predict GDP?"

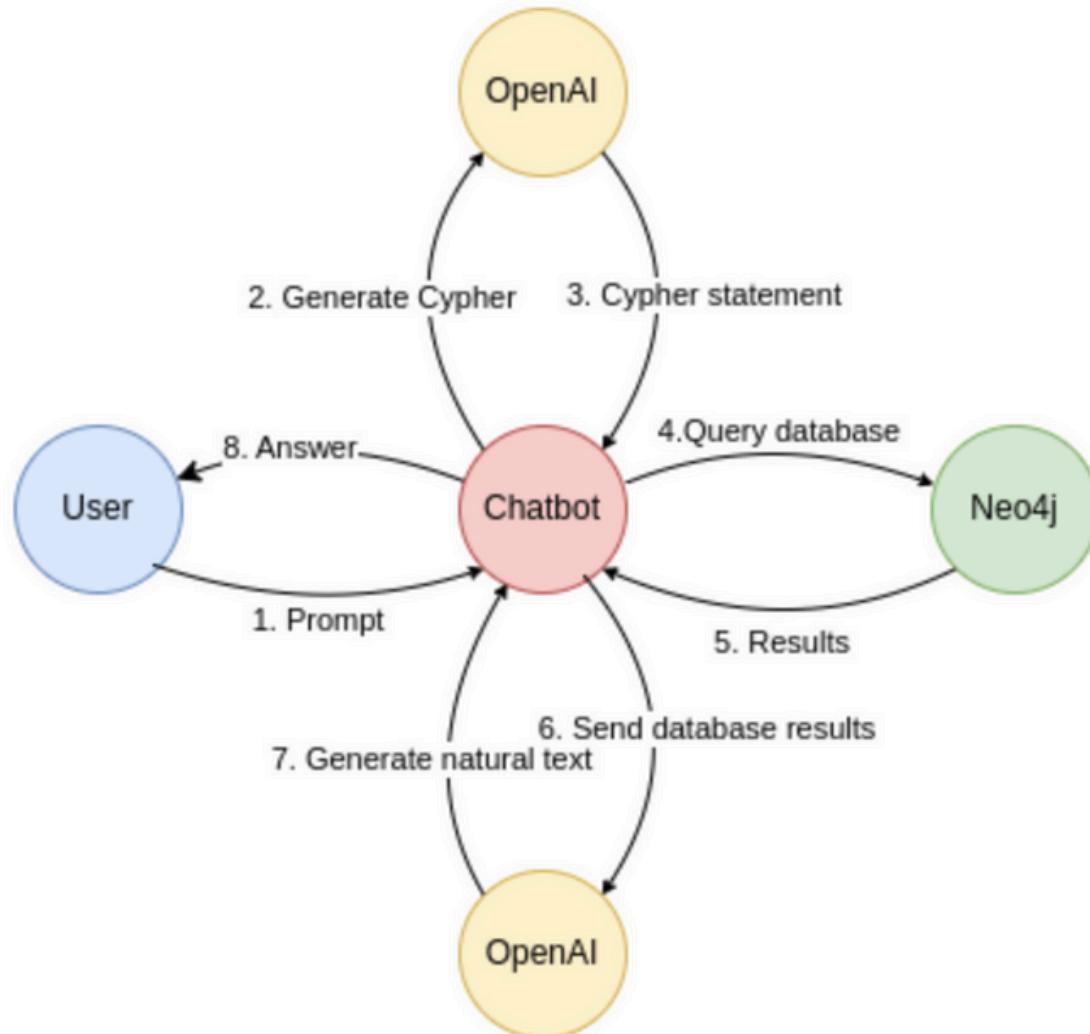
#### • Cypher Generation Phase

- Generating Cypher Queries to Extract Information from the Database

#### • Answer Generation Phase

- Crafting Sentences from the Extracted Information

## 2.3.5 Connection to the Graph Database



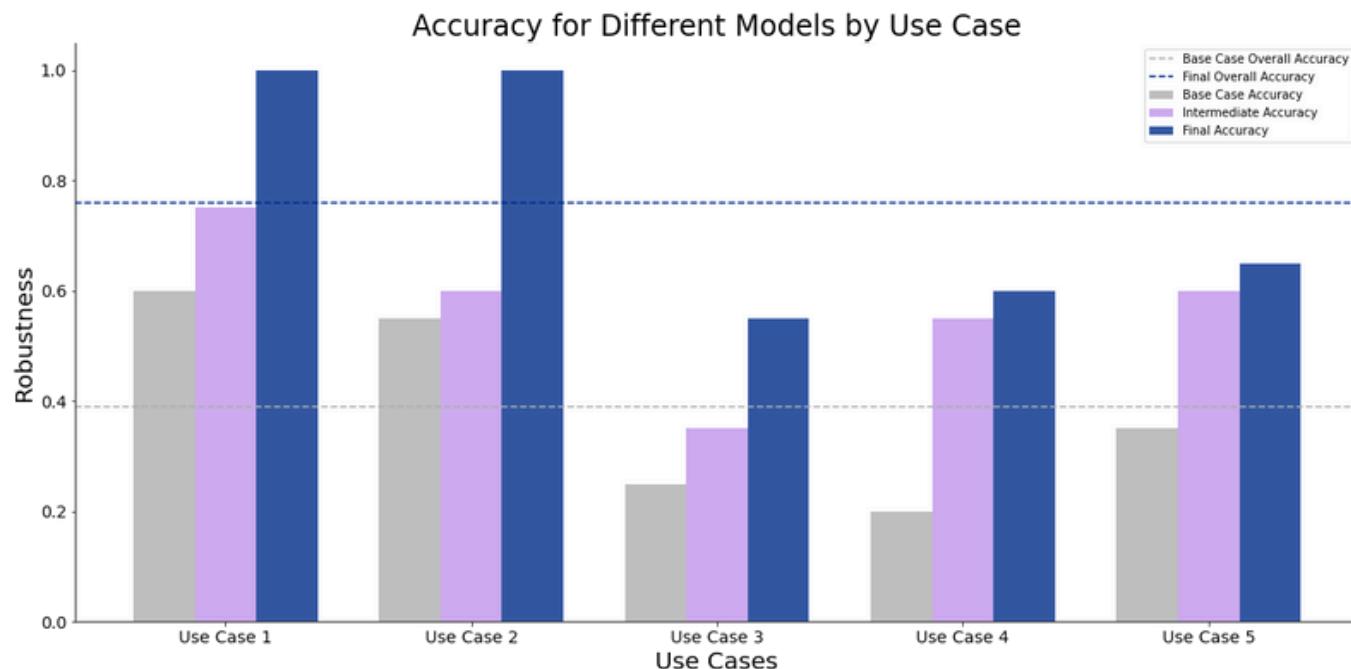
### Chatbot Interaction Flow

1. **Prompt:** The user initiates interaction by providing a prompt/question.
- 2&3. **Generate Cypher & Cypher Statement:** The prompt is sent to GPT-3.5. GPT-3.5 generates a Cypher query based on a trained dataset.
4. **Query Database:** The generated Cypher query is used to retrieve information from the Neo4j Graph Database.
5. **Results Retrieval:** The results are obtained from the database.
6. **Send Database Results:** The results are sent back to GPT-3.5.
7. **Generate Natural Text:** GPT-3.5 processes the results to generate a natural language response.
8. **Answer:** The final answer is delivered to the user.

## 2.4.1 Performance - Accuracy

**Accuracy: how our responses effectively answer the user's questions.**

--- Perform 3 independent correctness tests and calculate the average score.

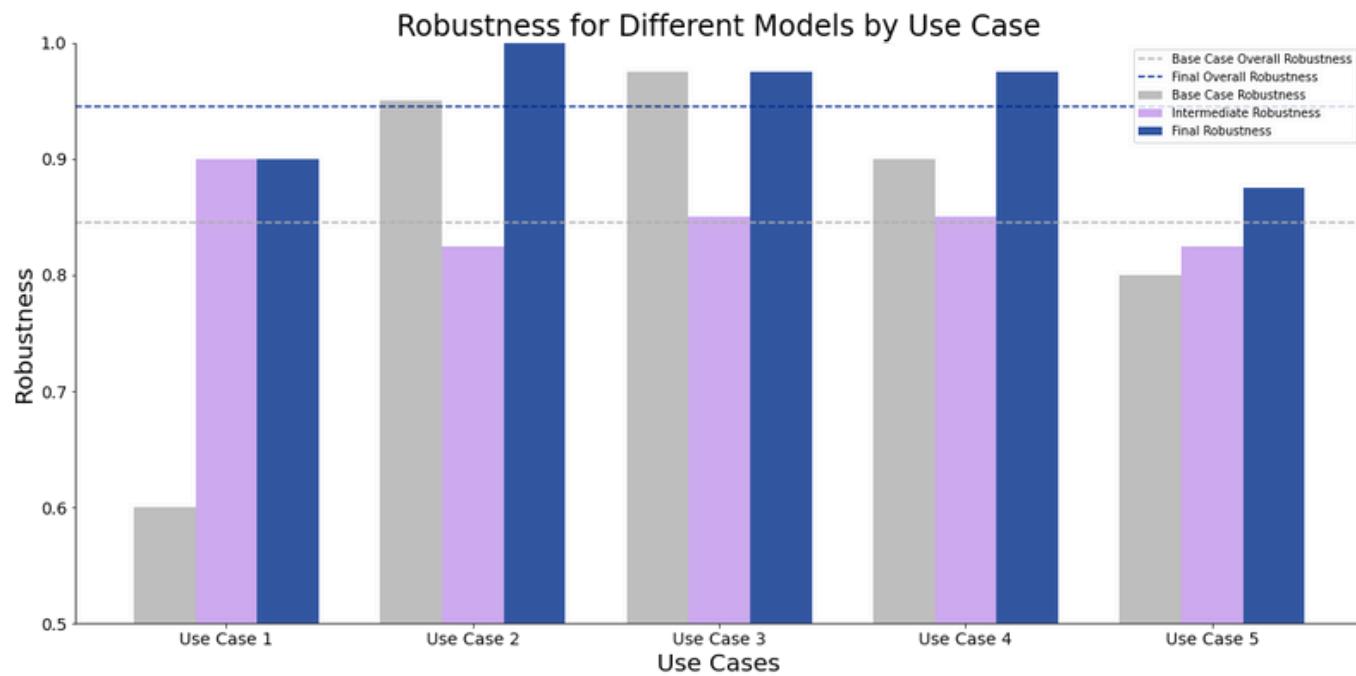


	Complexity	Baseline	Intermediate Iteration	Final Model
Use Case 1	Low	0.6	0.75	1
Use Case 2	Low	0.55	0.6	1
Use Case 3	High	0.25	0.35	0.55
Use Case 4	High	0.2	0.55	0.6
Use Case 5	High	0.35	0.6	0.65
Simple Average Accuracy	--	0.39	0.57 (+46% than baseline)	0.76 (+94.87% than baseline)

## 2.4.2 Performance - Robustness

**Robustness:** the model gives stable answers based on questions, no matter whether the answer is correct or not.

--- Comparison of results across 3 trials and assess the consistency.

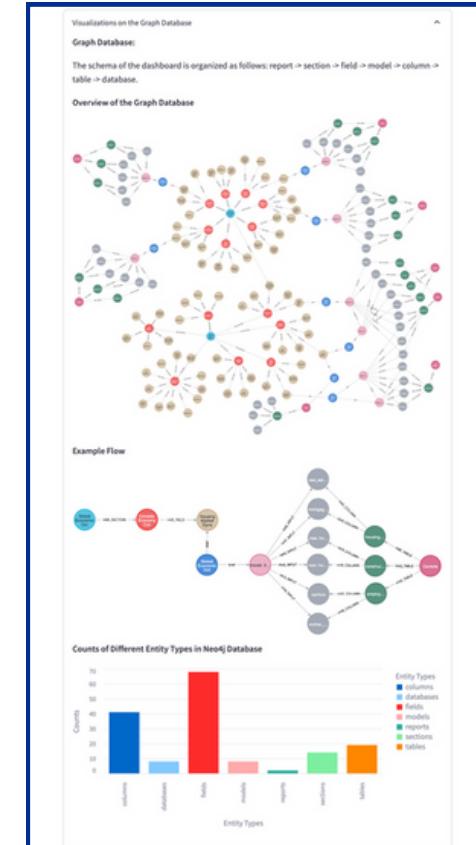


	Complexity	Baseline	Intermediate Iteration	Final Model
Use Case 1	Low	0.6	0.9	0.9
Use Case 2	Low	0.95	0.825	1
Use Case 3	High	0.975	0.85	0.975
Use Case 4	High	0.9	0.85	0.975
Use Case 5	High	0.8	0.825	0.875
Simple Average Robustness	--	0.845	0.85 (+0.59% than baseline)	0.945 (+11.83% than baseline)

## 2.5 User Interface: Overview



The screenshot shows the KPMG Chatbot interface. On the left, there's a vertical purple sidebar with a white header containing a red 'X' icon and a button labeled 'Open a new chat'. Below this is a white content area with the KPMG logo and Columbia University Data Science Institute logo. The main title is 'KPMG Chatbot'. A sub-instruction reads: 'This chatbot helps you understand the relationships among KPMG's internal reports, models, and databases. Please begin by reviewing the visualizations on the graph database and the instructions for prompts.' Two dropdown menus are visible: 'Visualizations on the Graph Database' and 'Instructions on Prompts'. At the bottom, there's a text input field with placeholder text 'Enter the information:' and a red 'Send' button with a crown icon. A blue wavy arrow points from the right side of the purple sidebar towards the right-hand visualization panel.



This screenshot shows the 'Instructions on Prompts' section. It lists five sample prompts for different use cases:

- Use Case 1: Simple Extraction (one entity)  
e.g. What are the performance metrics of the model named model\_925814632?
- Use Case 2: Simple Extraction (two or more entities)  
e.g. Who are the authors of models model\_701015292, model\_701015291, and model\_925814632?
- Use Case 3: Complex Extraction with if condition  
e.g. How many models in the field of Real GDP Recovery have an accuracy greater than 0.9?
- Use Case 4: Finding Relationships  
e.g. Which tables has the report named "Saudi Arabia Budget Report 2024" employed?
- Use Case 5: Yes/No Verification Problem  
e.g. Do models model\_925814632 and model\_701015291 use the table named economic\_indicators\_tb?

## 2.5 User Interface: Example



The screenshot shows a chatbot interface for the KPMG Chatbot. At the top, there's a header with the KPMG logo, Columbia University Data Science Institute logo, and a 'Deploy' button. A sidebar on the left has an 'Open a new chat' button. The main area is titled 'KPMG Chatbot' and contains instructions about understanding relationships in KPMG's internal reports, models, and databases. Below this are two dropdown menus: 'Visualizations on the Graph Database' and 'Instructions on Prompts'. A user message is shown: 'For model\_701015291 and model\_701015292, which model has a higher accuracy and what is the type of that model?'. A response is provided: 'Model\_701015292 has a higher accuracy of 66% and it is a Neural Networks model.' This response is marked as 'Completed'. Below the response, the 'Generated Cypher Code:' section displays the following code:

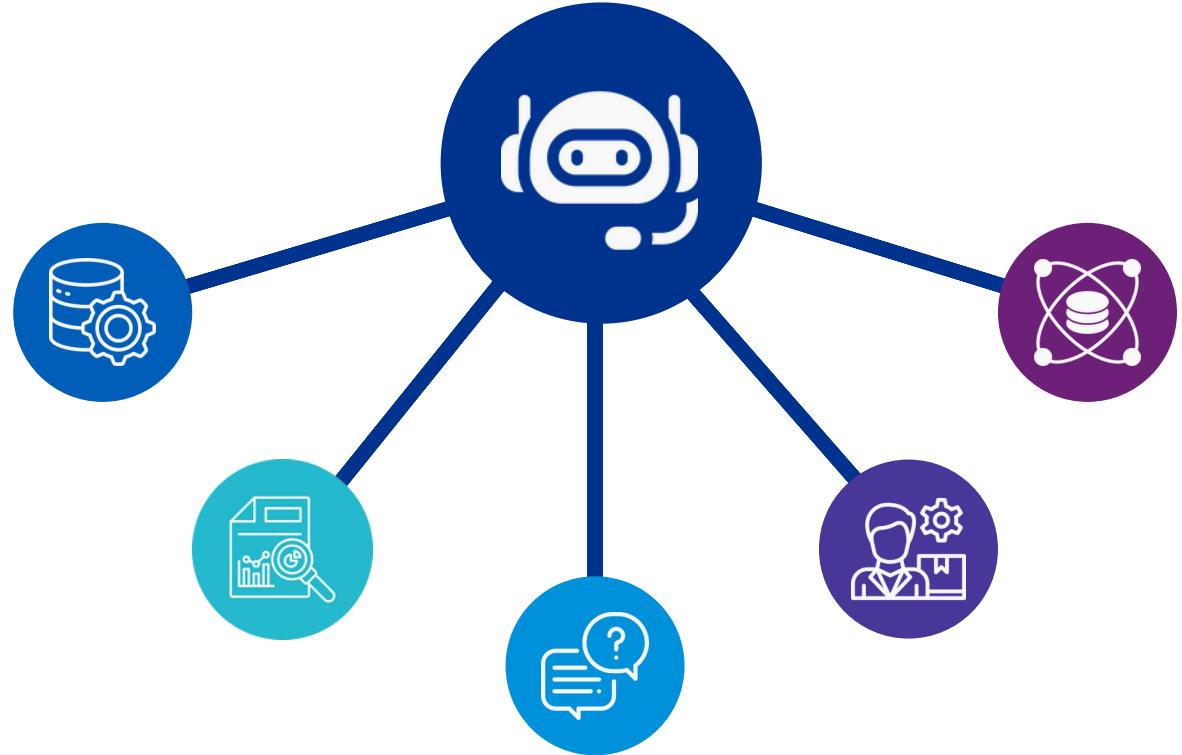
```
MATCH (m1:model) WHERE toLower(m1.name) = 'model_701015291'
MATCH (m2:model) WHERE toLower(m2.name) = 'model_701015292'
WITH m1, m2
RETURN
CASE
WHEN m1.accuracy > m2.accuracy THEN m1.model_type
ELSE m2.model_type
END AS ModelType,
CASE
WHEN m1.accuracy > m2.accuracy THEN m1.accuracy
ELSE m2.accuracy
END AS HigherAccuracy
```

At the bottom, there's an input field 'Enter the information:' with a right-pointing arrow button.

- 1 Introduction
- 2 Our Approach
- 3 Business Case Implementation & Demo
- 4 Conclusion & Future Improvements
- 5 Acknowledgement & References

### 3.1 Potential Use Cases

The development of our GenAI-powered chatbot for KPMG signifies a pivotal advancement in the effective management of model metadata tailored to enhance the functionalities across various business roles within the organization. This chatbot serves distinct use cases for diverse professional personas, illustrating its versatility and utility.



# 3.1 Potential Use Cases



## New Employee

- New employees can use this chatbot for general assistance to ensure a smooth onboarding by asking questions about how to use this chatbot or finding nearby restaurants.



## Business Analyst

- Business analysts can use this chatbot for market research and decision support. The chatbot can help them quickly find the market trends by providing relevant reports.



## Data Scientist

- Data Scientist can ask chatbot to provide detailed information about the machine learning models like the model lineage, their performance and their types.



## Data Engineer

- The chatbot can assist in tracing the lineage of data used by the machine learning model, identifying the source columns, tables and databases involved in generating the model inputs.



## Product Manager

- The chatbot can help product manager to retrieve the profile of the specific employee and the job description associated with their role within the company.

## 3.2 Demo



[Link: https://final-kpmg-chatbot.streamlit.app/](https://final-kpmg-chatbot.streamlit.app/)

- 1 Introduction
- 2 Our Approach
- 3 Business Case Implementation & Demo
- 4 Conclusion & Future Improvements
- 5 Acknowledgement & References

### Conclusion

- **GenAI chatbot for Metadata Management System:** Developed a system incorporating a GenAI-powered chatbot to manage information efficiently.
- **Operational Efficiency:** Streamline operations and minimize model risks, improving metadata management and delivering business value for KPMG
- **Real-world Application:** The chatbot can be used across various scenarios, such as enhancing data analysis for scientists, ensuring system integrity for engineers, and providing a user-friendly onboarding experience.

### Future Improvements

- **Reduce Response Latency.**
- **Experiment with different LLM models and compare performance.**

# Reflections



I'm grateful for the chance to develop a chatbot from the backend to the frontend—a project that has been both educational and enriching. Delving into areas like graph databases, prompt engineering, chatbot development, and UI design has provided me with hands-on experience that goes beyond classroom learning. Learning from industry mentors has been particularly enlightening, fostering a shift from a student to an industry mindset. I'm excited about the potential of our final chatbot to significantly enhance KPMG's metadata management and make a meaningful impact on the industry.

I'm truly grateful for the opportunity to collaborate with KPMG mentors on this capstone project, which allowed us to explore cutting-edge topics like LLM and chatbots. Integrating the OpenAI API and developing a chatbot using Streamlit has been a highly educational experience, pushing the boundaries of what I thought was possible in an academic project. This experience with graph databases also opened our eyes to their unique potential, particularly in handling metadata. Working on this project has not only broadened my technical skills but has also deepened my understanding of how these technologies can be applied in real-world scenarios.

The capstone experience was instrumental in teaching me the significance of metadata management. I learned to execute end-to-end projects, from mock data generation, importing data into a structured graph database, utilizing Langchain, and prompt engineering to craft a chatbot, to final UI design. This journey provided hands-on exposure to modern data and AI tools. Meanwhile, the guidance from mentors was invaluable, offering insights into presentation skills, project management, and more. Their advice bridged the gap between academic knowledge and industry practice, equipping me with skills that are invaluable for my career!

The capstone project has been incredibly valuable in deepening my understanding of the applications of large language models in the industry. Throughout this project, I gained a comprehensive understanding of the end-to-end process involved in designing and implementing AI solutions. This included generating synthetic data, constructing robust model architectures, and designing intuitive front-end UI interfaces. Working on this project provided hands-on experience with each stage of development. I look forward to leveraging more AI technologies to build more cutting edge solutions to industry problems.

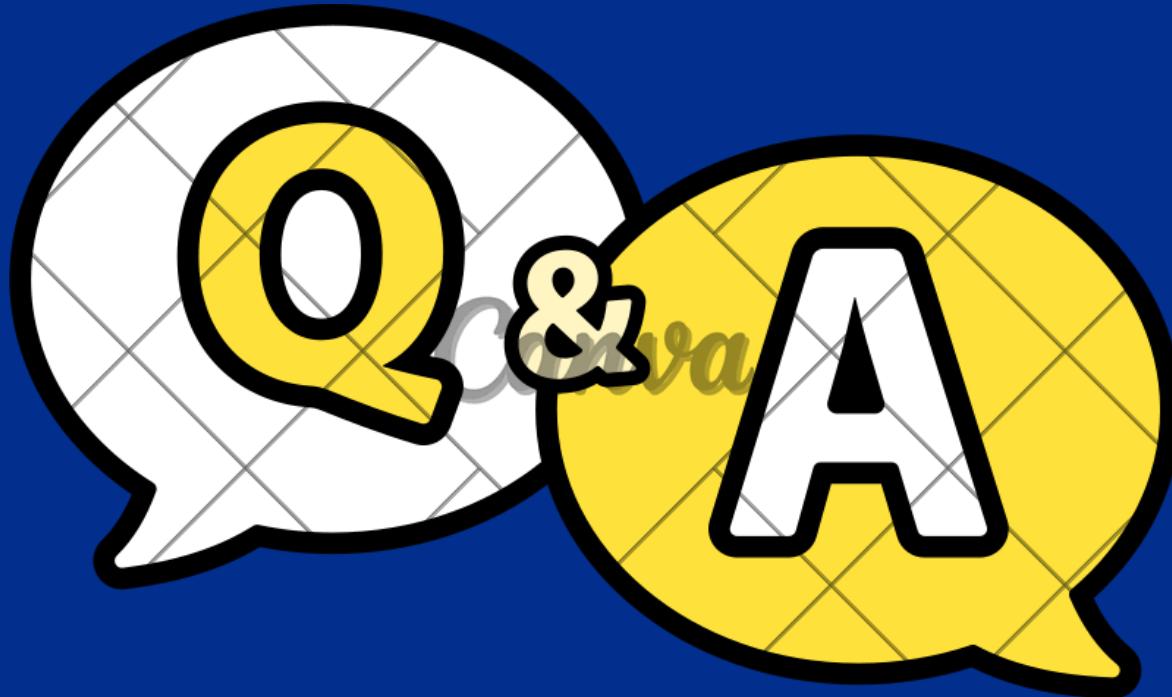
This capstone project with KPMG this semester has been incredibly rewarding. Exploring tools like Langchain, Neo4j, and Streamlit expanded my technical skill set significantly. While facing challenges in debugging, performance optimization, and UI refinement, each obstacle served as a unique learning opportunity. I'm proud of our tangible solution addressing real needs within the organization. This experience has taught me valuable lessons in project management, collaboration, and innovation. I look forward to applying these skills in future endeavors, grateful for the growth and relationships forged along the way.

- 
- 1 Introduction
  - 2 Our Approach
  - 3 Business Case Implementation & Demo
  - 4 Conclusion & Future Improvements
  - 5 Acknowledgement & References

# THANK YOU!

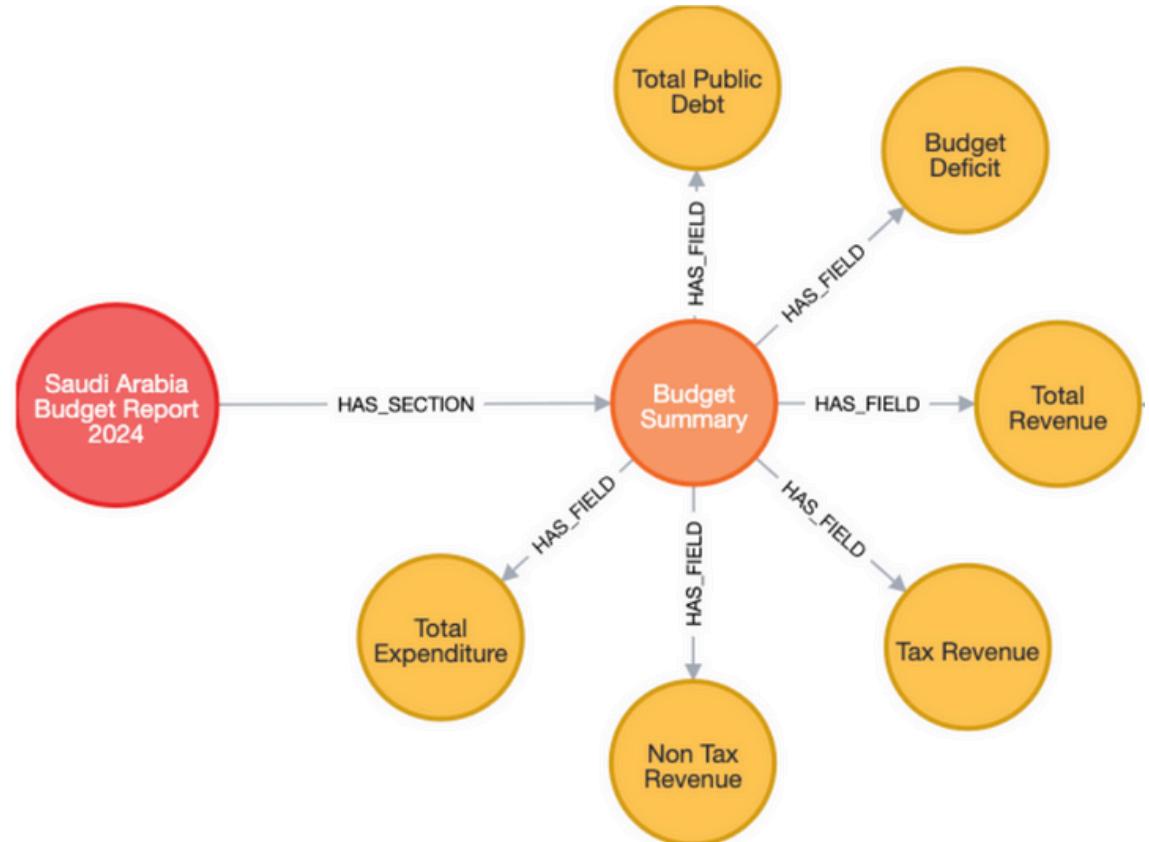
KPMG Mentors: Sydney Son, Thomas Covella, Chengwei Wang

DSI Mentors: Sining Chen



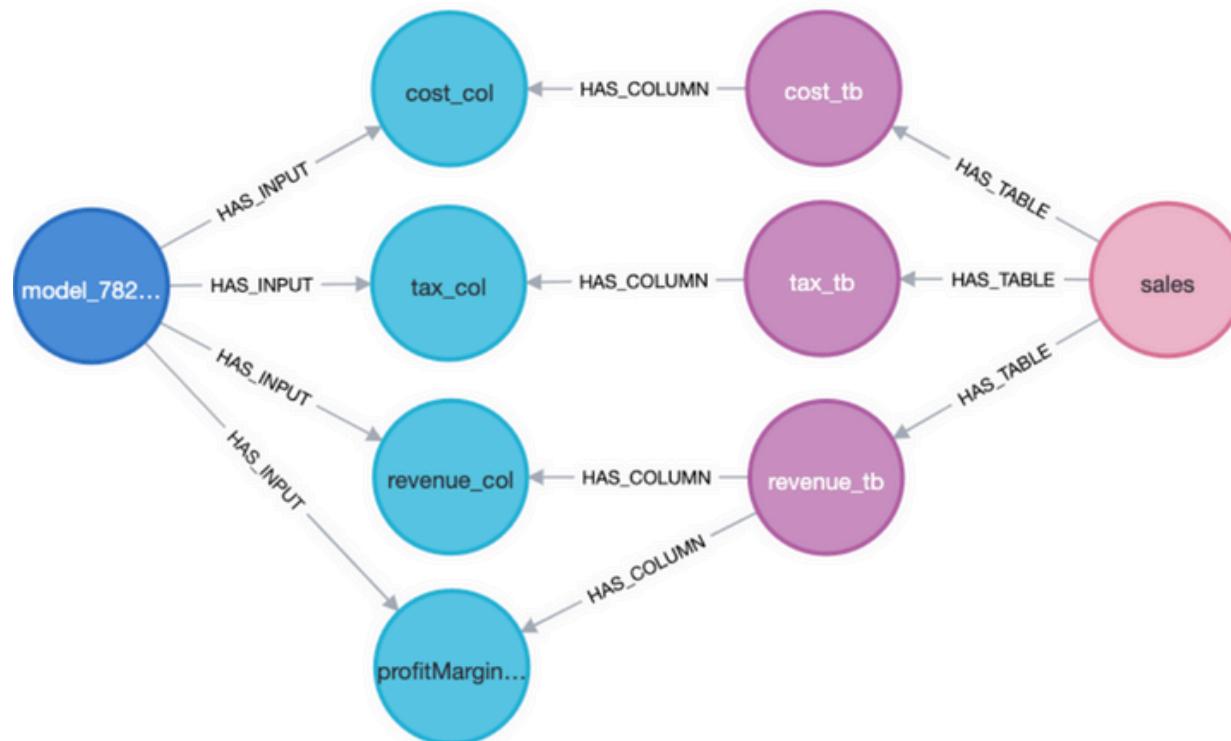
# Appendix

## 2.2 Neo4j: Structure of Graph Database



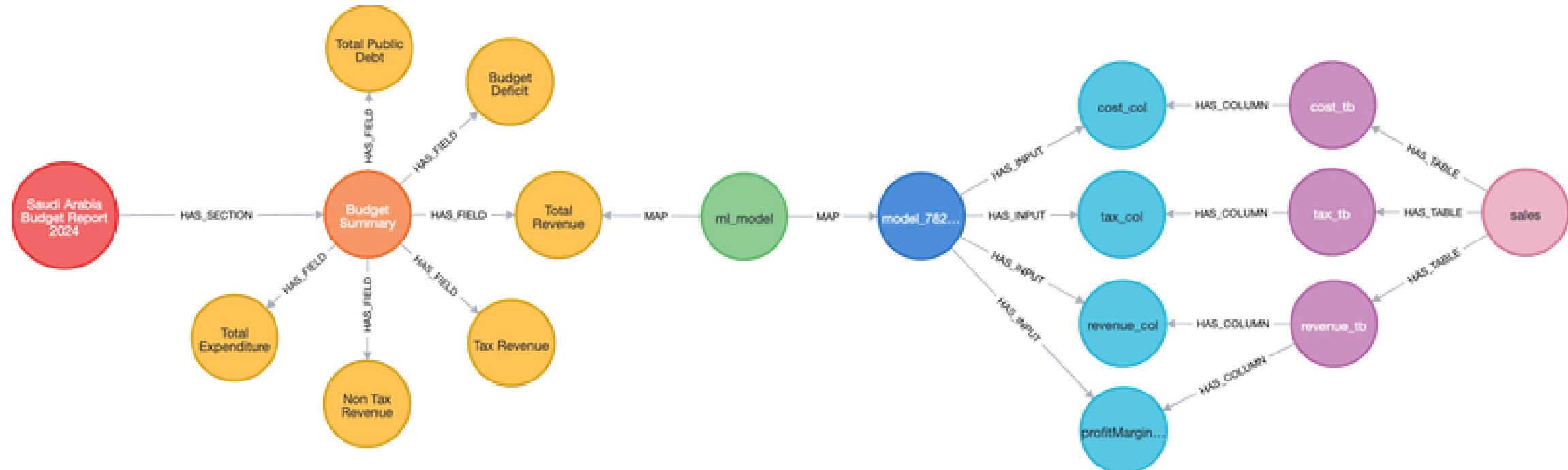
- Report {
  - name
  - entitlements
  - report format
  - author
  - business group
- Section {
  - name
- Field {
  - name
  - description

## 2.2 Neo4j: Structure of Graph Database



- Model** { name, creator, model type, input columns, performance metric, version, creation date }
- Database** { name }
- Column** { name }
- Table** { name }

## 2.2 Neo4j: Structure of Graph Database



### Node labels

\* (168)   field (68)   model (8)   mapping (8)   database (8)  
table (19)   column (41)   report (2)   section (14)

### Relationship types

\* (207)   HAS\_FIELD (68)   MAP (17)   HAS\_TABLE (19)  
HAS\_COLUMN (41)   HAS\_INPUT (45)   HAS\_SECTION (14)  
REFERS\_TO (1)   ITERATE (1)   IS\_INPUT\_OF (1)

## 2.4 Test result



Link: [https://docs.google.com/spreadsheets/d/13pAL-CoYBonbjFMU1apB\\_iwk-cMWQX1V1ZSlv6zS8ms/edit#gid=1330825756](https://docs.google.com/spreadsheets/d/13pAL-CoYBonbjFMU1apB_iwk-cMWQX1V1ZSlv6zS8ms/edit#gid=1330825756)

Sample Test result:

Use Case	Prompts	Ground Truth	True Answer	Base GPT Answer	Base GPT T/F	Intermediate GPT A	Intermediate GPT I	Final GPT Answer	Final GPT T/F	Base GPT Answer	Intermediate GPT A	Final GPT Answer
1	What are the author	Rain	The author of the md	The author of the md	1	The author of the md	1	The author of the md	1	0.898095608	0.898095608	0.898095608
1	What are the author	rainp@kpmg.com	The author_email of	rainp@kpmg.com	1	The author_email of	1	The author_email of	1	0.575640798	0.84532696	0.899365246
1	What are the target	labor_market_readiness_level	The target of model	I don't know the ans	0	The target of model	0	The target of model	1	0.291751325	0.572404027	0.823134899
1	What are the input	["real_estate_sales", "mortgage_approval"]	The input columns o	The input columns o	1	The input columns o	1	The input columns o	1	0.877209902	0.8783172312	0.827617049
1	What are the output	housing_market_trend	The output columns	The output columns	0	The output columns	1	The output column o	1	0.728826251	0.778846979	0.765443206
1	What is the model type	Category Boosting	The model type of m	I don't know the ans	0	The model type of m	0	The model type of m	1	0.303236783	0.742925286	0.804296994
1	Who is the author of	rmayor@kpmg.com	The author of the re	I don't know the ans	0	The author of the re	0	The author of the re	1	0.266056222	0.661485851	0.949511766
1	Who is the developer	Harry	The developer of the	Harry is the developer	1	The developer of the	1	Harry is the developer	1	0.872261405	0.894287229	0.872261405
1	What are the predictors	auditDate,financial_score,compliance_soc	The predictors of md	The predictors of md	1	The predictors of md	1	The predictors of md	1	0.870406747	0.816437125	0.870406747
1	What is the format of	PDF Document	The format of the re	The format of the re	1	The format of the re	1	The Saudi Arabia Bu	1	0.879118502	0.879118502	0.84517616
1	What is the model type	Neural Networks	The model type of m	The model type of m	1	The model type of th	0	The model type of m	1	0.865059376	0.621051073	0.865059376
1	Who are the authors	Alex, Martin	The authors of the m	Alex, Martin are the	1	The authors of the m	1	The authors of the m	1	0.855400324	0.88457038	0.88457038
1	What is the created	2024-02-12T20:59:17Z	The created_at date	The created_at date	1	The created_at date	1	The created_at date	1	0.898291111	0.85879457	0.85879457
1	What is the business	KPMG International	The business group	KPMG International	1	The business group	1	The business group	1	0.450438231	0.914160132	0.914160132
1	What is the version	1	The version of the m	I don't know the ans	0	The version of the m	1	The version of the m	1	0.363843203	0.877933621	0.877933621
2	What are the business	KPMG International, KPMG Professional	The business group	KPMG Professional	1	The business group	1	The business group	1	0.895038807	0.858906448	0.908971906
2	Who are the owners	rmayor@kpmg.com, pbannink@kpmg.com	The owners of the m	pbannink@kpmg.com	1	The owners of the m	0	The owners of the m	1	0.554428101	0.771530986	0.832327004
2	What are the author	gabvd@kpmg.com, rainp@kpmg.com	The author_emails fo	I don't know the ans	0	The author_emails fo	1	The author_emails fo	1	0.352024674	0.889727294	0.903252602
2	Who are the authors	(Alex,Martin), (Rain)	The authors of mode	Alex, Martin, and Ra	1	The authors of mode	1	The authors of mode	1	0.787176967	0.851846218	0.851846218
2	What are the targets	revenue, housing_market_trend	The targets of mode	I don't know the ans	0	The targets of mode	0	The targets of mode	1	0.281412721	0.684633255	0.737843454
2	What are the report	PDF Document, PDF Document	The report_format of	The report formats o	1	The report format of	1	The report format of	1	0.883329988	0.872165442	0.872165442

## 3.1.1 KPMG New Employee



How can I use this chatbot?

✓ Completed

**Response:**

To use this chatbot effectively, you can follow these steps:

- Explore Graph Database:** You can start by exploring the graph database visualizations provided. Each visualization represents a different aspect of KPMG's internal reports, models, or databases. You can ask for more details about a specific visualization or its components.
- Ask Questions:** Feel free to ask any questions related to the data, reports, or models within KPMG. The chatbot is here to provide you with information and guidance.
- Follow Prompts:** The chatbot will provide prompts and suggestions based on your interactions. Follow these prompts to navigate through the information effectively.
- Request Assistance:** If you need help with a specific topic or have trouble understanding something, don't hesitate to ask for assistance. The chatbot is here to help you.
- Feedback:** If you have any feedback or suggestions on how we can improve the chatbot or the information provided, feel free to share it.

By following these steps, you can make the most out of the chatbot and access the information you need efficiently. Let me know if you have any specific questions or if you'd like to start exploring the graph database visualizations.

**Base Case Response:**

I don't know the answer.

Where're the good places around KPMG New York to grab lunch? Any recommendations?

✓ Completed

**Response:**

There are plenty of great places around KPMG New York to grab lunch! Some popular recommendations include Shake Shack for delicious burgers and shakes, Sweetgreen for fresh salads and bowls, Xi'an Famous Foods for flavorful Chinese cuisine, and Joe's Pizza for a classic New York slice. If you're in the mood for something different, you can also check out Eataly for a variety of Italian options or Le Pain Quotidien for tasty sandwiches and pastries. Enjoy your lunch!

Remember, I'm here to help you understand the relationship between KPMG's reports, data, and models.

---

**Base Case Response:**

I don't know the answer.

## 3.1.2 Business Analyst



⌚ Hello, I'm analyzing market trends and need specific information. Can you provide the names of the reports that include sections titled with the word 'China'?

✓ Completed

**Response:**

The report that includes a section titled with the word 'China' is "Global Economic Outlook - December 2023".

**Generated Cypher Code:**

```
MATCH (s:section)
WHERE toLower(s.name) CONTAINS 'china'
MATCH (r:report)-[:HAS_SECTION]-(s)
RETURN DISTINCT r.name AS ReportName
```

**Base Case Response:**

The reports that include sections titled with the word 'china' are not provided in the schema of the Neo4j database.

⌚ I'm currently analyzing fiscal policies for a project. What are the fields included in the Saudi Arabia Budget Report 2024?

✓ Completed

**Response:**

The fields included in the Saudi Arabia budget report for 2024 are Total Revenue, Total Expenditure, Budget Deficit, Total Public Debt, Tax Revenue, Non Tax Revenue, and various other categories such as Defense, Education, and Infrastructure.

**Generated Cypher Code:**

```
MATCH (r:report)-[:HAS_SECTION]-(s:section)-[:HAS_FIELD]-(f:field)
WHERE toLower(r.name) = 'saudi arabia budget report 2024'
RETURN f.name AS FieldName
```

**Base Case Response:**

The fields included in the Saudi Arabia budget report 2024 are name, entitlements, business\_group, report\_format, and author.

### 3.1.3 Data Scientist



what models are predicting risk level in the Saudi Arabia Budget Report 2024?

✓ Completed

**Response:**

Two models are predicting the risk level in the Saudi Arabia budget report for 2024:  
model\_701015291 and model\_701015292.

**Generated Cypher Code:**

```
MATCH (r:report)-[:HAS_SECTION]->(:section)-[:HAS_FIELD]->(:field)<-[M: WHERE toLower(r.name) = 'saudi arabia budget report 2024' AND toLower(m.name) AS ModelPredictingRiskLevel
```

**Base Case Response:**

The models predicting risk level in the Saudi Arabia budget report 2024 are not specified in the provided information.

For model\_701015291 and model\_701015292, which model has the higher accuracy and what is the type of that model?

✓ Completed

**Response:**

The model with the higher accuracy is a Neural Networks model with an accuracy of 0.66.

**Generated Cypher Code:**

```
MATCH (m1:model) WHERE toLower(m1.name) = 'model_701015291'  
MATCH (m2:model) WHERE toLower(m2.name) = 'model_701015292'  
WITH m1, m2  
RETURN  
CASE  
WHEN m1.accuracy > m2.accuracy THEN m1.model_type  
ELSE m2.model_type  
END AS ModelType,  
CASE  
WHEN m1.accuracy > m2.accuracy THEN m1.accuracy  
ELSE m2.accuracy  
END AS HigherAccuracy
```

**Base Case Response:**

For model\_701015291 and model\_701015292, the model with the higher accuracy is model\_701015291 and its type is regression.

## 3.1.4 Data Engineer



Could you tell me which specific tables and databases are utilized by model\_701015291?



✓ Completed

Response:

The model\_701015291 utilizes the "audit\_tb" table in the "audit management" database and the "economic\_indicators\_tb" table in the "GDP" database.

Generated Cypher Code:

```
MATCH (m:model {name: 'model_701015291'})-[:HAS_INPUT]->(c:column)<-[:H  
RETURN DISTINCT t.name AS TableName, d.name AS DatabaseName
```

Base Case Response:

The specific tables and databases utilized by model\_701015291 are not provided in the schema representation of the Neo4j database.



What tables and columns are in the "GDP" database?



✓ Completed

Response:

The "gdp" database contains two tables: "economic\_indicators\_tb" with columns for unemployment rate, consumer confidence, government spending, and interest rates, and "trade\_balance\_tb" with a column for trade balance.

Generated Cypher Code:

```
MATCH (d:database)-[:HAS_TABLE]->(t:table)-[:HAS_COLUMN]->(c:column)  
WHERE toLower(d.name) = 'gdp'  
RETURN t.name AS TableName, COLLECT(c.name) AS ColumnNames
```

Base Case Response:

The "gdp" database contains tables and columns.

## 3.1.5 Product Manager



What kind of work does Luke do?

✓ Completed

**Response:**

Luke works in the field of agriculture production, specifically focusing on analyzing factors such as crops grown, farming methods, government subsidies, export volumes, farming sector, and precipitation to improve agricultural outcomes.

**Generated Cypher Code:**

```
MATCH (m:model {author: 'Luke'}) RETURN m.name AS ModelName, m.author AS AuthorEmail
```

**Base Case Response:**

I don't know the answer.

Could you give me Luke's email

✓ Completed

**Response:**

Luke's email is [luke@kpmg.com](mailto:luke@kpmg.com).

**Generated Cypher Code:**

```
MATCH (m:model {author: 'Luke'}) RETURN m.author_email AS AuthorEmail
```

**Base Case Response:**

I don't know the answer.