pmsa
PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

ELEVATING ANALYTICS THROUGH HUMAN & ARTIFICIAL INTELLIGENCE

Knowledge Graph Infused Natural Language to SQL with Domain Adapted LLM for Patient Data

Bharath Bommakanti, Manager – Data Science
Sameer Saraf, Principal Data Scientist

Agenda

- Introduction
  - Current challenges with data democratization
  - Challenges with traditional LLMs / RAG
- McKesson Compile's NL to SQL approach
- NL to SQL showcase
- Results and Metrics
- Conclusion

Current Challenges in Data Democratization

- Expertise limitations
  - Analyst required to query data using SQL
- Domain-specific complexities
  - Specialized nomenclature
  - Multiple diverse coding systems
  - Dilution of context from business user to analyst to SQL query
  - Diverse healthcare elements hosted on complex database models
- Limitations of traditional LLMs / RAG
  - Complex schematic understanding
  - Non-standard table and schema nomenclature
  - Low-precision retrieval due to lack of semantic knowledge

Challenges with Traditional LLMs / RAG

- Complex schematic understanding
- Non-standard table and schema nomenclature
- Low-precision/ Low-recall retrieval due to lack of semantic knowledge
- Lengthy prompts to guide through the intricacies of healthcare data analytics
  - Financial implications
- Ground-truth or the retrieved context do not capture the relationships in the data
- Co-pilots and task specific LLMs (NL to SQL) lack the required context of the domain and data elements

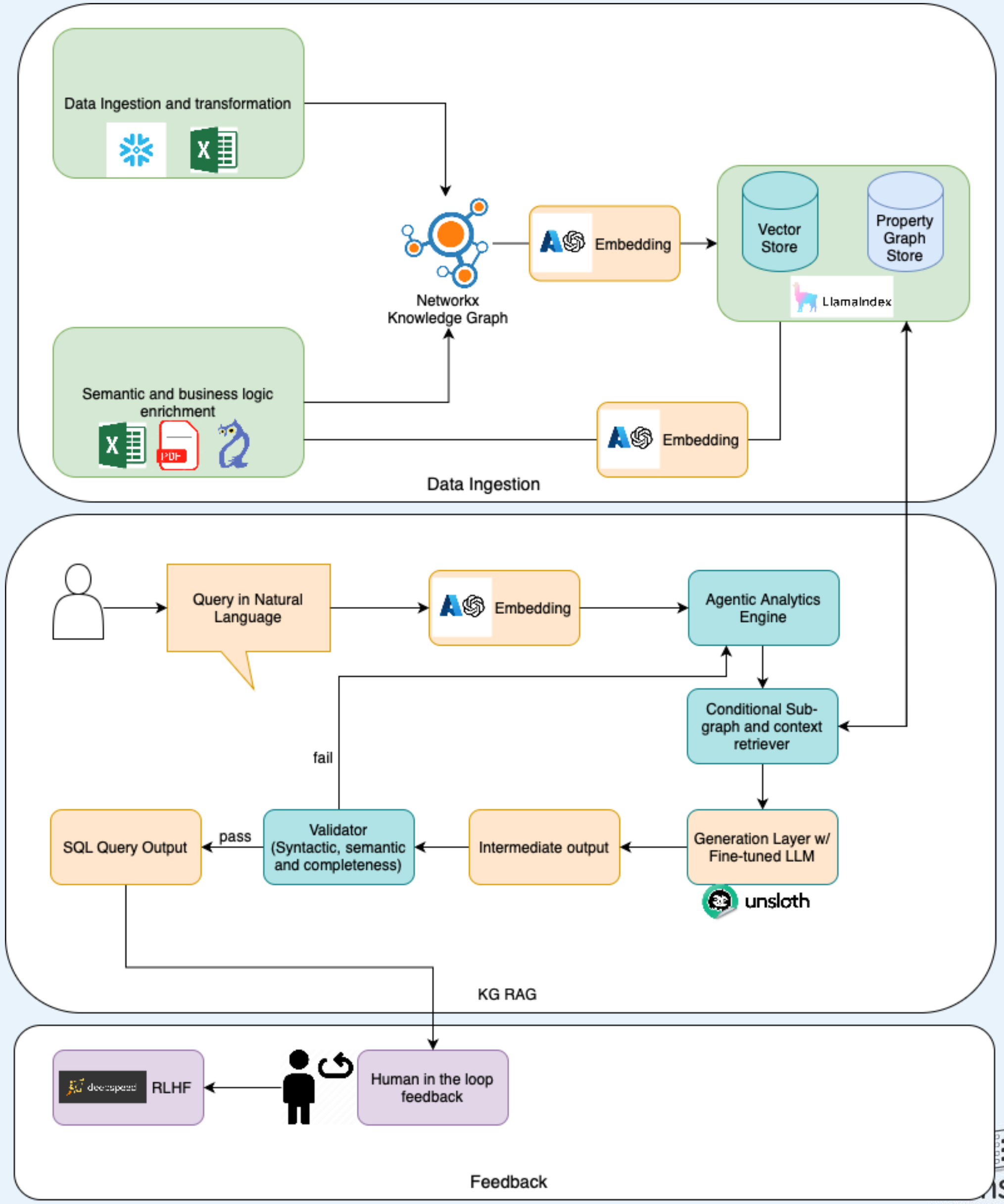
McKesson Compile's NL to SQL approach

The approach is broadly divided into 3 parts:

1. Data ingestion
- Semantic and schematic knowledge ingestion

2. Knowledge graph (KG) RAG
- Proprietary retrieval mechanism from Graph and Vector stores
- Generational layer with domain adapted LLM

3. Feedback loop
- Human feedback capture mechanism for further model alignment with RLHF

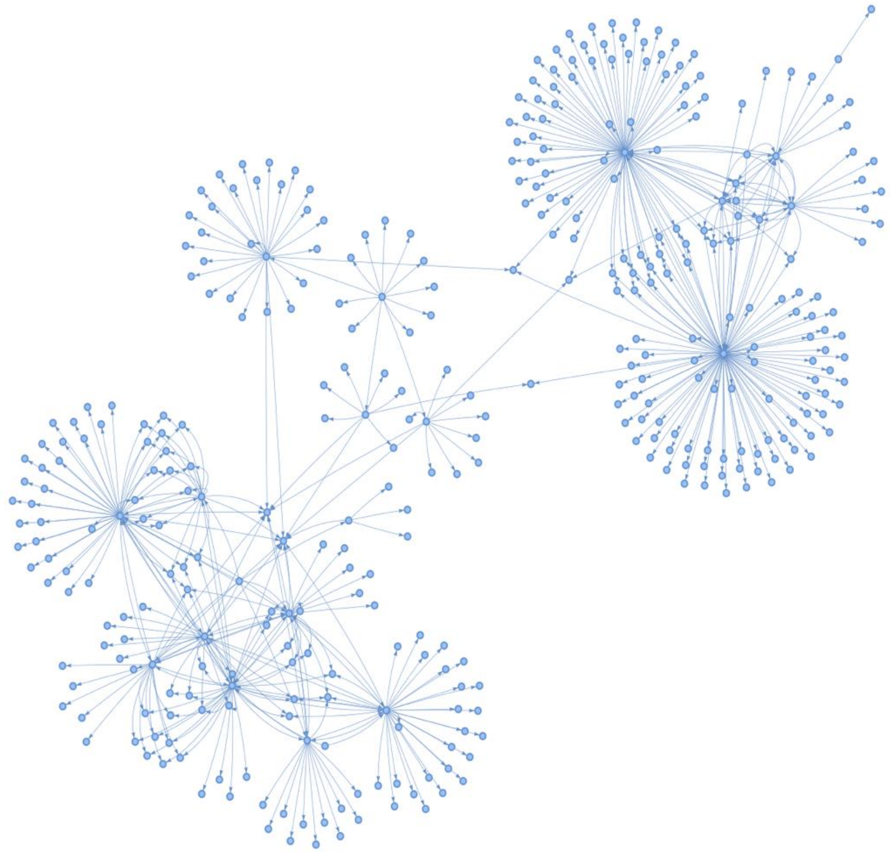


Data | Data Understanding

Schematic data	Semantic data
• Schema – table – column relationships • Table – table relationships	• Description of the node types • Description of the node values • Understanding the data types and data distribution • Semantic tagging using the diverse node values and aliases • Business logic

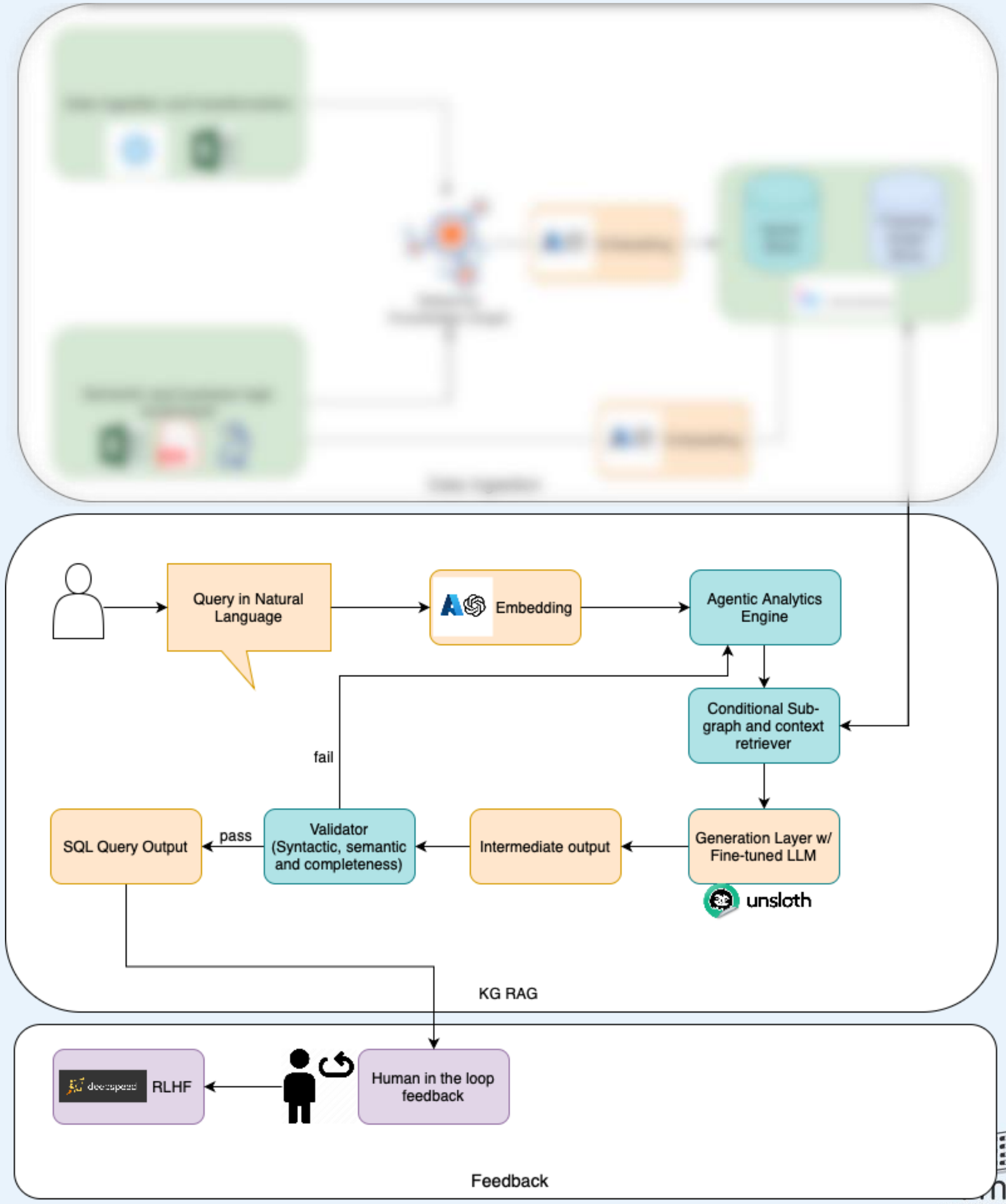
Embedded Knowledge Graph (KG)

- Semantic knowledge embedding with schematic relationships
- Relationship prediction using semantic tagging
  - Useful for relevant sub-graph / community extraction
- High-precision retrieval and generation
  - Constrained sub-graph retrieval from KG
  - Less guidance required for complex problems through structured prompting
  - Fine-tuned LLM leveraging gisting approach for output generation



Query Synthesis & Continual Learning

- Agentic SQL generation
  - Iterative retrieval agent to ensure completeness in the answer through conditional retrieval
  - SQL Validation engine
    - Agent that ensures syntactical accuracy
    - Contextual relevance and completeness
  - Fine-tuned LLM leveraging gisting approach for output generation
- Feedback loop
  - Quick feedback through bucketized granular options
  - Specific feedback through free flow text
  - Maintaining output – feedback pairs for further fine-tuning of the LLM through RLHF



NL to SQL showcase: Example 1

Q: Extract Top 10 HCPs (based on patient volume) that are diagnosing patients with Lung Cancer disease (Diagnosis code: C34) for the year of 2023

Filters and transformations are thoughtfully applied to deliver precise and concise answers to the user's question

Leverages CTEs to structure multiple operations

Joins the CTEs to generate final summary

Identifies right tables and columns of reference

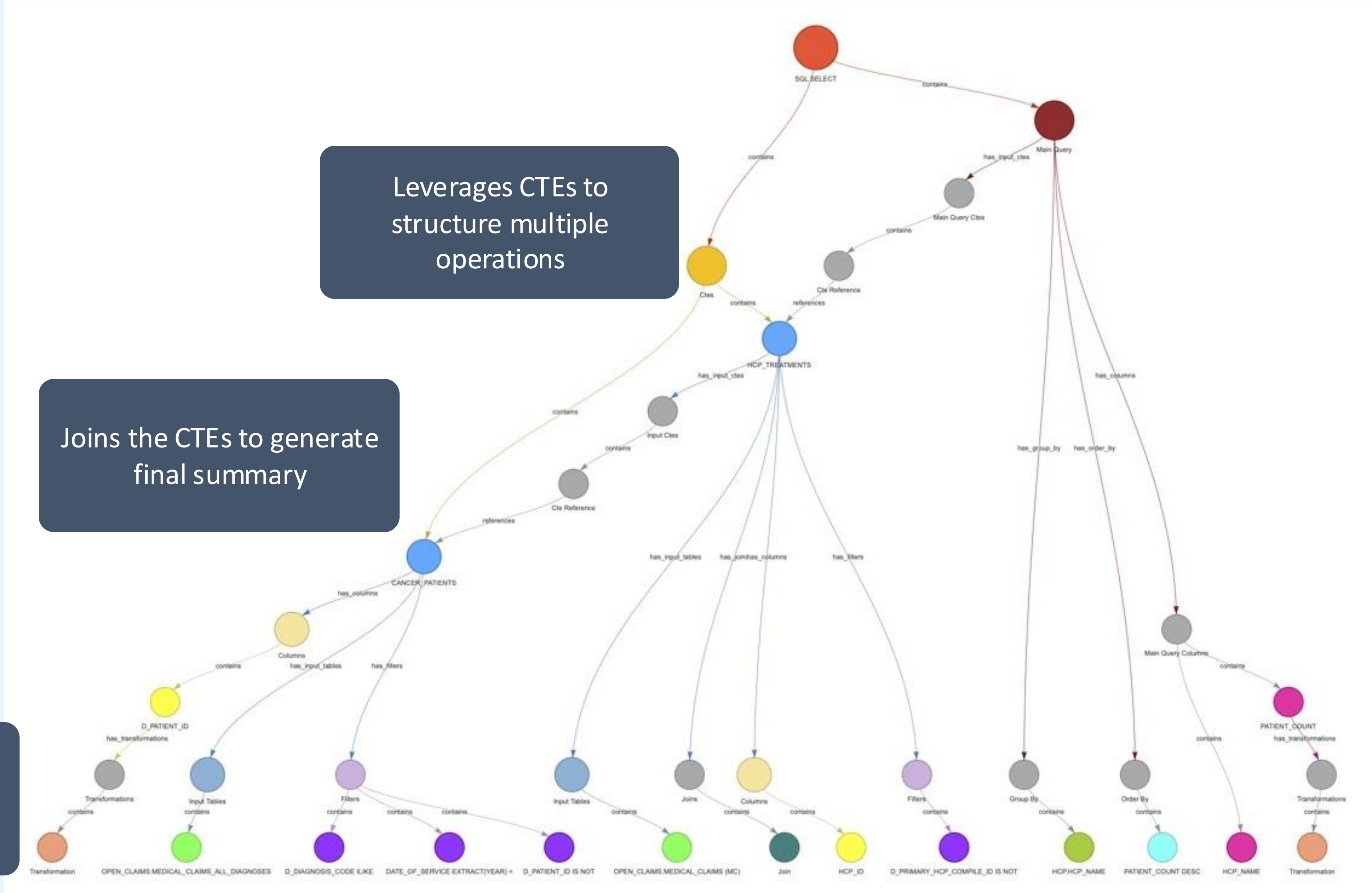


Figure: Visualization of SQL query generated by NL to SQL

NL to SQL showcase: Example 1 Output

Q: Extract Top 10 HCPs (based on patient volume) that are diagnosing patients with Lung Cancer disease (Diagnosis code: C34) for the year of 2023

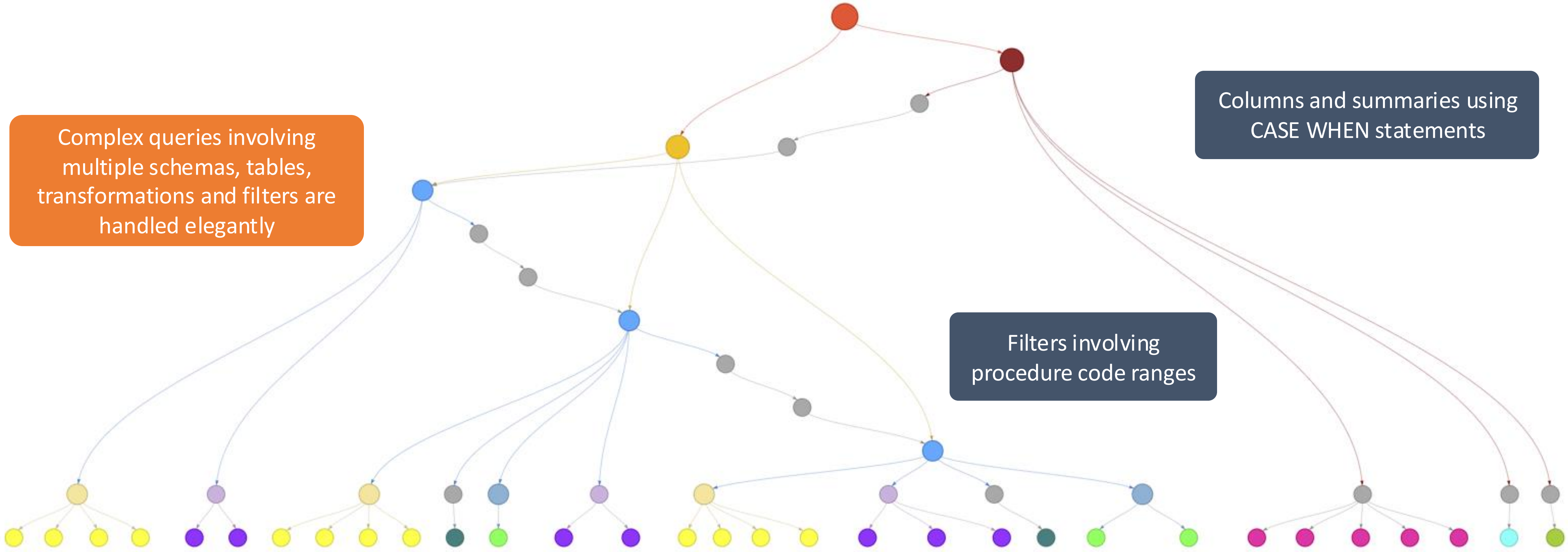
HCP	PATIENT_COUNT
HCP 1	3,123
HCP 2	2,513
HCP 3	2,441
HCP 4	2,158
HCP 5	1,578
HCP 6	1,555
HCP 7	1,503
HCP 8	1,496
HCP 9	1,460
HCP 10	1,431

Note: Data shown is for illustrative purposes only

NL to SQL showcase: Example 2

Q: Provide feasibility estimates (patient counts) by year from 2020 to 2024 in 25 year age group buckets for patients aged ≥18 years with a non-inpatient, non-diagnostic* claim and having an hMPV diagnosis.

Non-diagnostic claims are claims without any procedures for Radiology, Venipuncture Pathology and Laboratory.



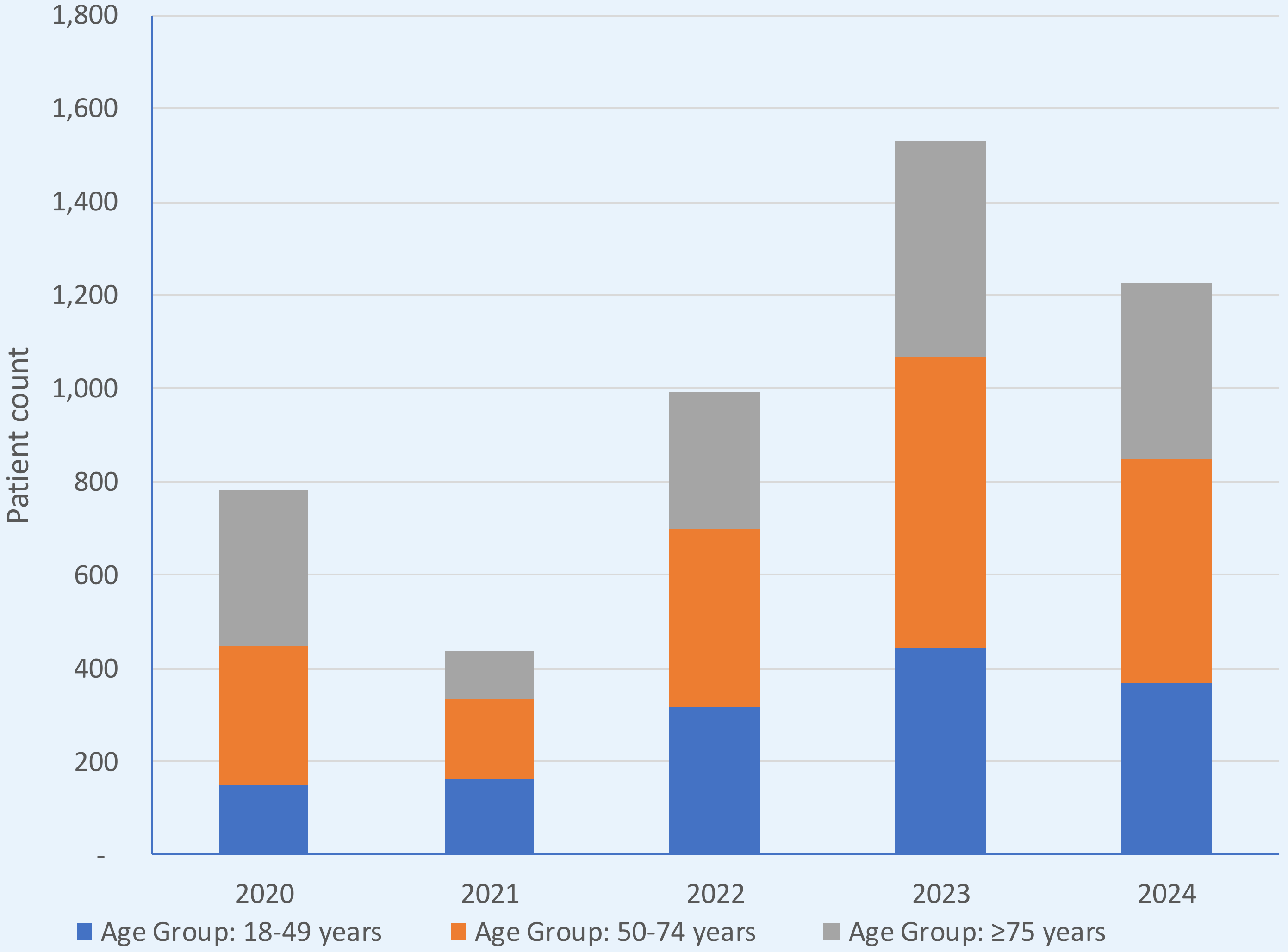
Complex queries involving multiple schemas, tables, transformations and filters are handled elegantly

Columns and summaries using CASE WHEN statements

Filters involving procedure code ranges

NL to SQL showcase: Example 2 Outputs

METRIC	SERVICE_YEAR	PATIENT_COUNT
Age Group: 18-49 years	2020	149
Age Group: 18-49 years	2021	162
Age Group: 18-49 years	2022	318
Age Group: 18-49 years	2023	443
Age Group: 18-49 years	2024	368
Age Group: 50-74 years	2020	299
Age Group: 50-74 years	2021	169
Age Group: 50-74 years	2022	379
Age Group: 50-74 years	2023	623
Age Group: 50-74 years	2024	479
Age Group: ≥75 years	2020	334
Age Group: ≥75 years	2021	106
Age Group: ≥75 years	2022	295
Age Group: ≥75 years	2023	466
Age Group: ≥75 years	2024	379



Note: Data shown is for illustrative purposes only

Results Demonstrate a Reduction in 3 Key Areas

⬇ 95%
Non-executable queries

- Schema misidentification & incorrect syntax
- Attributed to
  - SQL Validation agent
  - KG relationships

⬇ 5x
Hallucinations

- Semantic / contextual relevance & completeness
- Attributed to
  - Agent orchestrator
  - Conditional retrieval agent

⬇ 15x
Token count

- Attributed to
  - Fine-tuned LLM with gisting
  - Condensed context through KG relationships

Conclusions and Future Directions

- Real-world applications
  - Clinical decision intelligence
  - Analytical applications like feasibility analysis, cohort analysis, etc.
  - Advanced provider network analytics
  - Operational improvements
- Improved precision and recall in data democratization
- Cost-efficient GenAI application
- Strong foundation for NL to insights

PMSA
PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

Thank you for your attention

Questions?

Visit McKesson Compile at booth 212