



GROUP PROJECT — MIDTERM PRESENTATION

LumenAI: A Conversational Analytics Assistant for Intelligent Database Analysis

Mohamed Kassem, Erich Gunsenheimer | 27 MARCH 2025

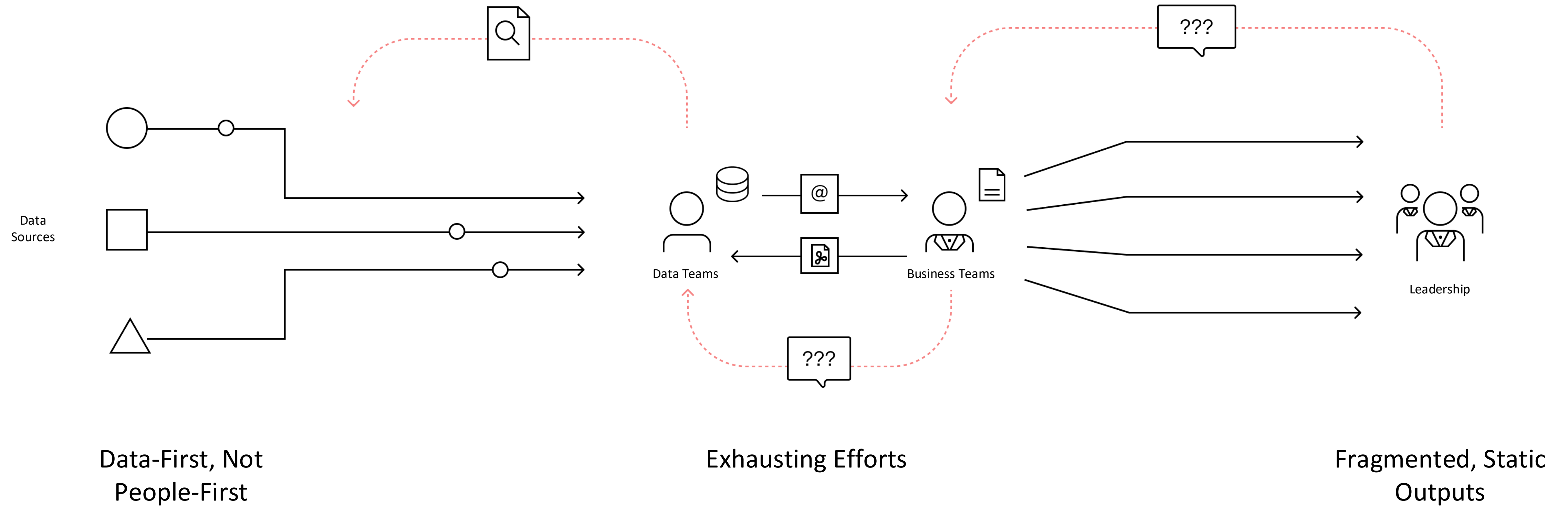
MLOps & LLMOps: Production AI Systems - E115

SPRING 2025

Harvard Extension School



Overflowing Data, Elusive Clarity.





Introducing LumenAI

A context-rich ecosystem AI Agent that feeds from the data you provide it and provides business users access to flexible content management and reasoning



LumenAI

Show me sales trends for Dubai in the last quarter.

Sales Trends for Dubai (Last Quarter: Nov 2023 - Jan 2024):

- November: \$145,000
- December: \$180,000 (highest)
- January: \$160,000

Analysis:

Dubai saw a steady increase in sales in December, driven by seasonal promotions. Sales slightly decreased in January but remained strong compared to historical averages. Product A and Product C were top performers during this quarter.

Show me sales trends for Dubai in the last quarter





Introducing LumenAI

A context-rich ecosystem AI Agent that feeds from the data you provide it and provides business users access to flexible content management and reasoning

LumenAI can:

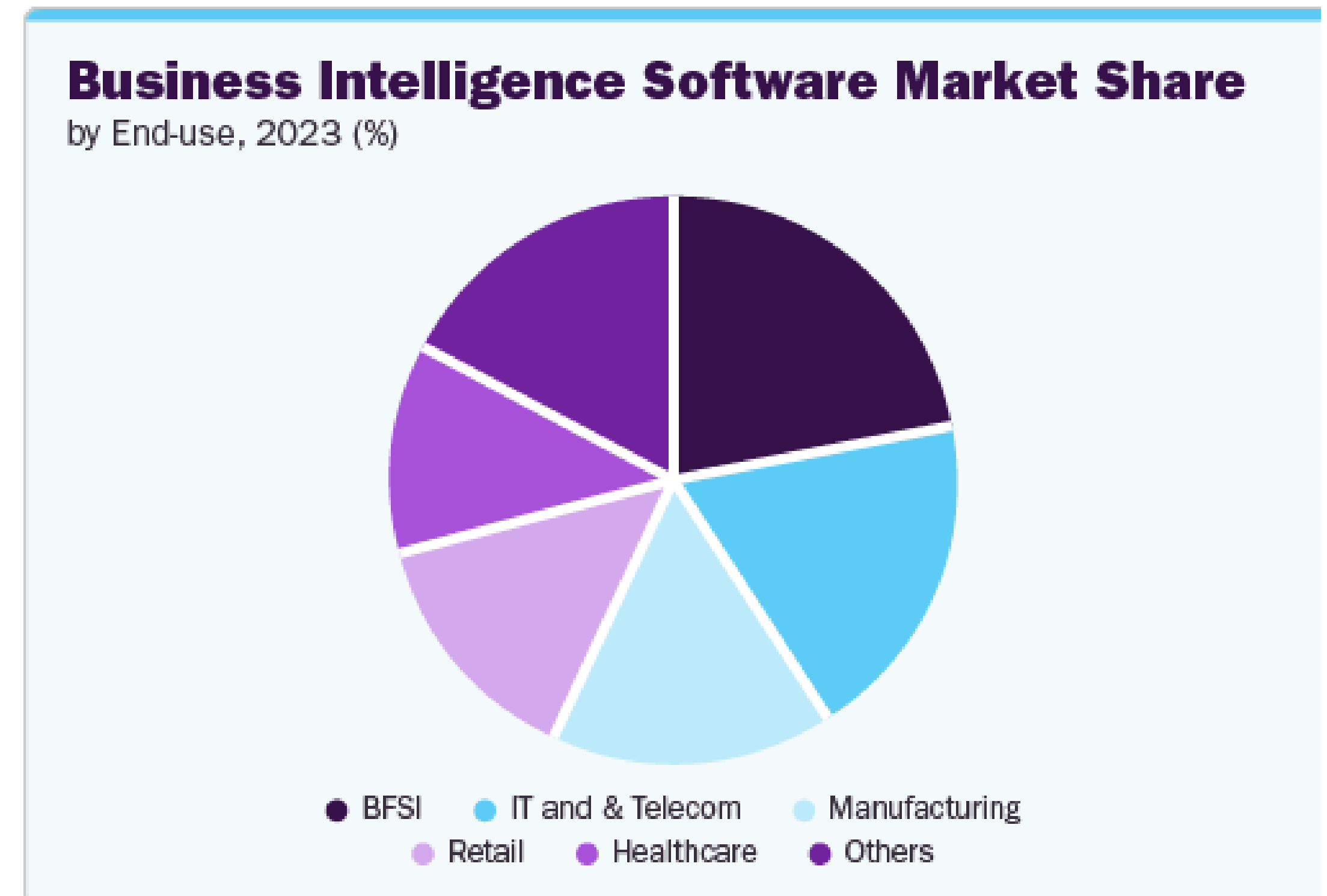
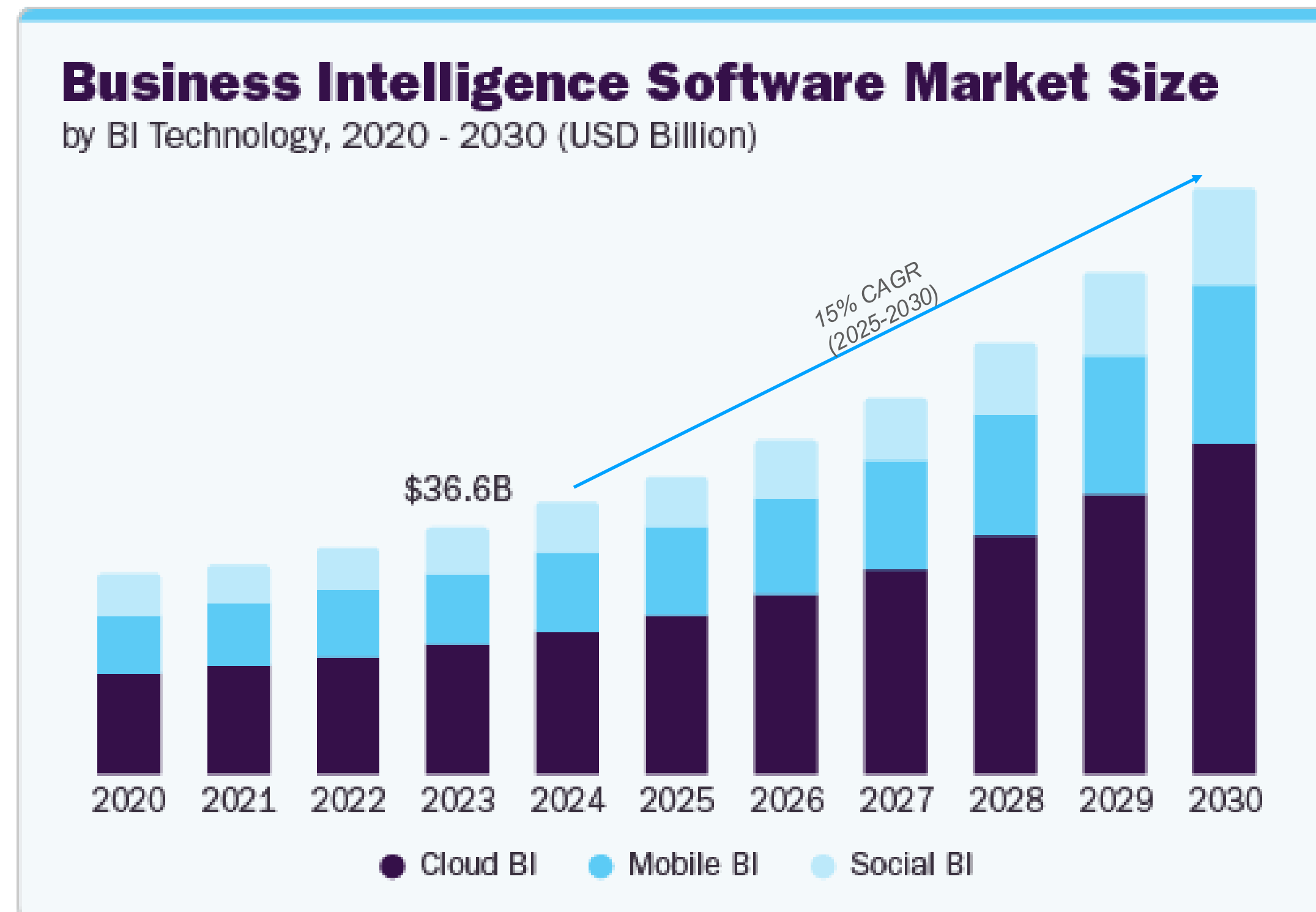
- Use Organization DB or Adhoc Datasets
- Incorporate previous presentations or reports
- Incorporate user context, role, and goals
- Initiate SQL DB and generate queries
- Use validation rules, reasoning and tool selection.

LumenAI offers to:

- Provide LLM summaries for schema and columns
- Extract data analytics based on text-based files
- Administer user access and roles
- Respond to specially-designed reporting prompts.
- Offer relevant accurate and information that can incorporate historical context



Market

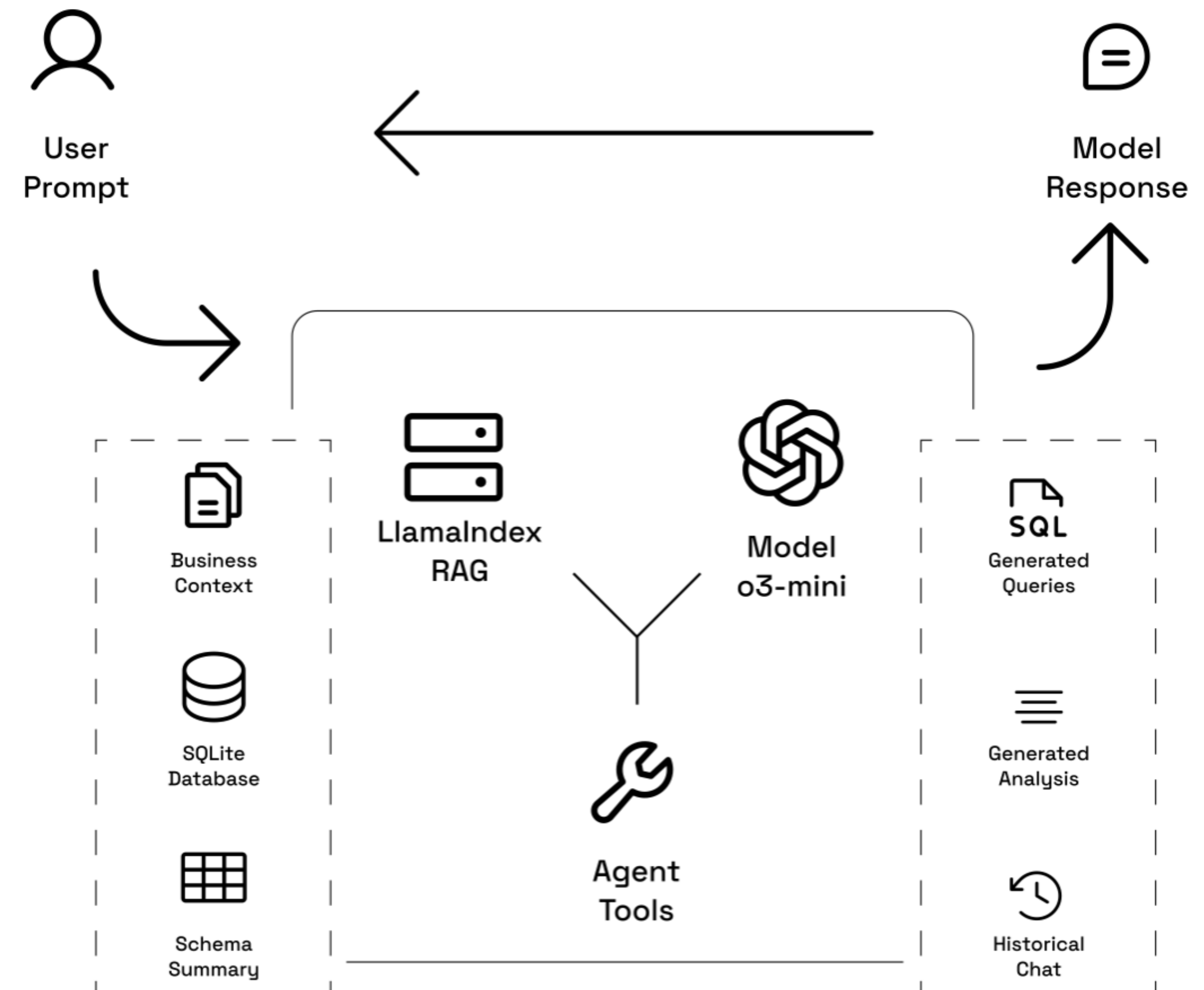


Source: Grand View Research 2025



Under the Hood

- Built on OpenAI o3-mini Model APIs
- Uses Embeddings from HuggingFace Embedding Model for vector embeddings
- Uses LlamaIndex for creating vector store and enabling semantic search over database schema
- Integrated with python functions
- Initiates SQLite DB through Python



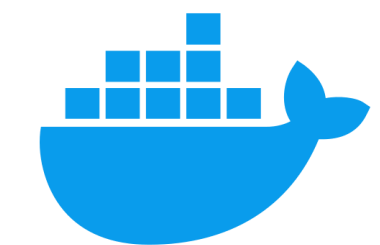


Scalability & Efficiency

- Tested Multiple Model structure including locally run deepseek 3B Model and API-based OpenAI o3-mini model
- App is supported by dockershell, integrated with Git Repo, and runs on GCP



OpenAI



Google Cloud Platform





Scalability & Efficiency

Factors	#1	# 2	# 3	# 4	# 5
Model Used	Deepseek	o3-mini	o3-mini	o3-mini	o3-mini
Model Run	Locally	OpenAI	OpenAI	OpenAI	OpenAI
Context					
Schema				LLM	LLM
Data				All — LLM	50 Rows — LLM
Statistics				LLM	LLM
Historical Prompts and Response					5 Records
Cost	Low	Low	Reasonable	Very High	Reasonable
Quality	Low	Low	Average	High	High



Future Development / Application

- Connect to User Friendly Interface
- Build ad-hoc document uploads
- Enhance the need-for-query validation workflow
- Enhance Error-Handling Workflow and Reasoning
- Enhance Report Prompting
- Optimize running cost



Future Development / Growth

- Introduce Domain-Refined Versions
- Optimize locally hosted models for more secured environments
- Integrate with Data Management systems for deeper context on lineage, usage, sources, and others
- Enhance KPI Suggestion
- Build a unified pool of queries/governance on an organization level



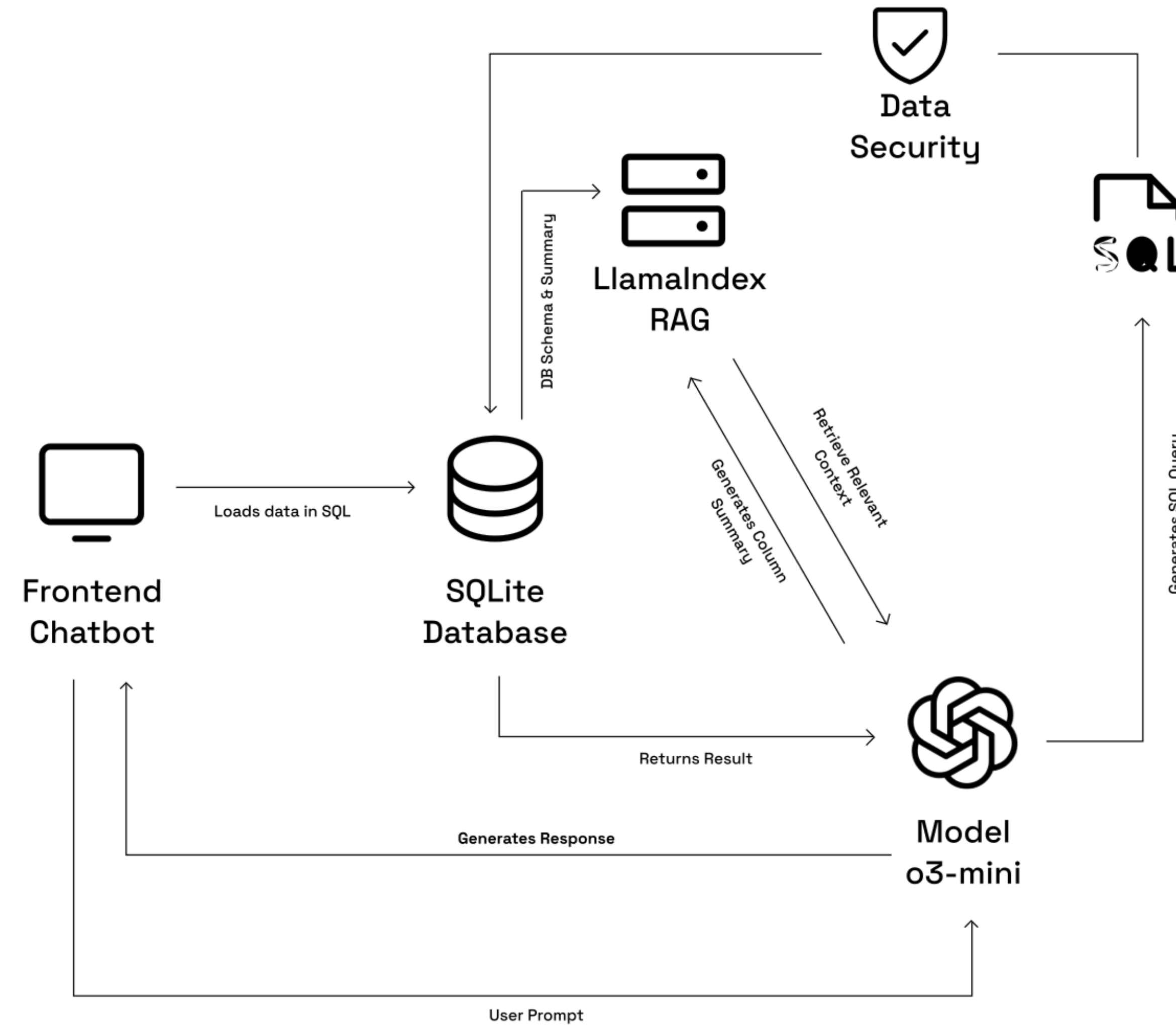
Thank You





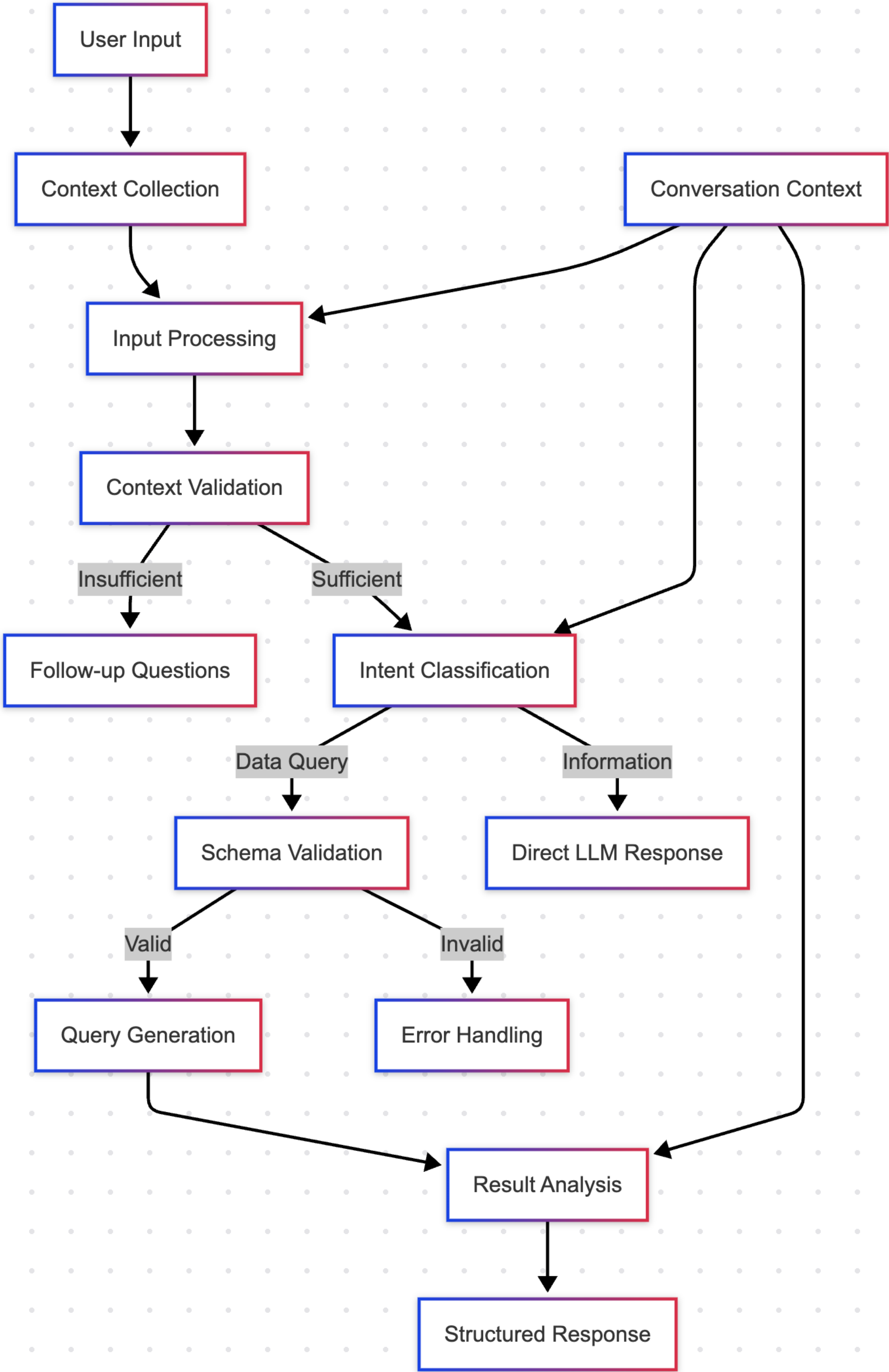
Architecture

LumenAI Architecture



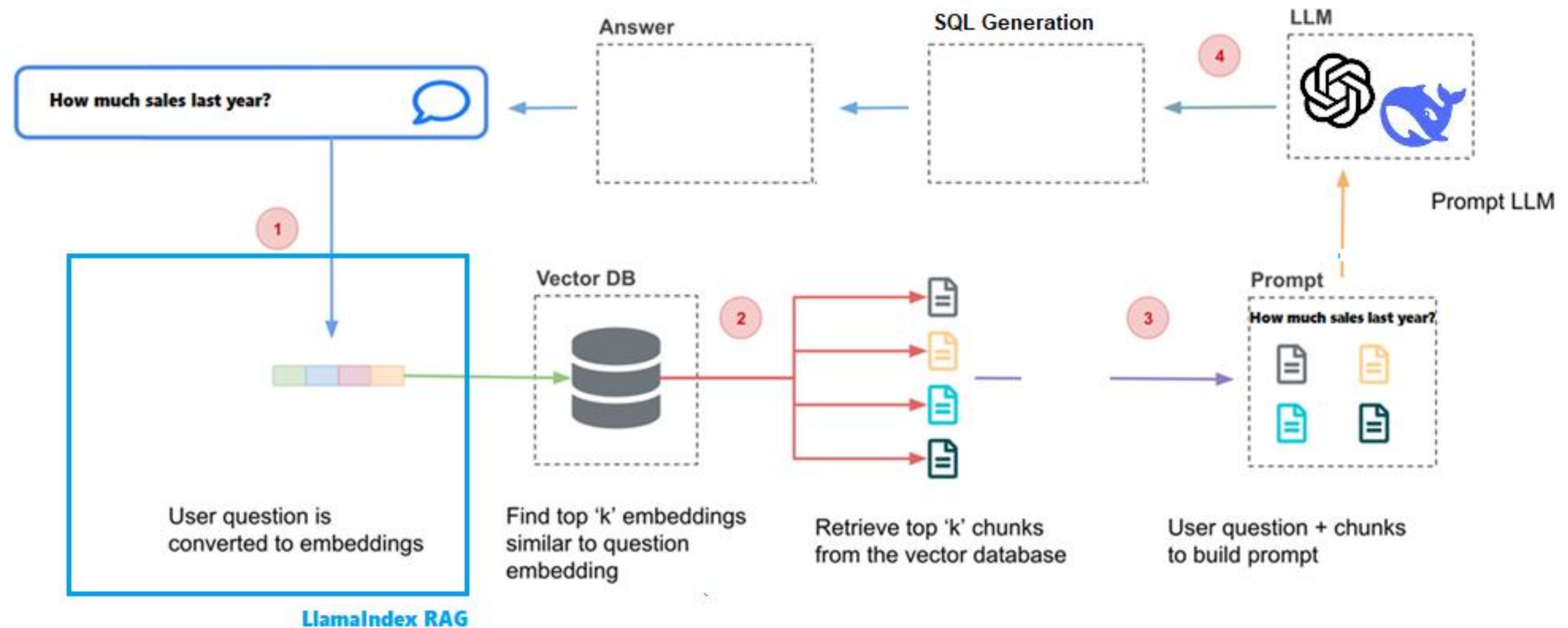


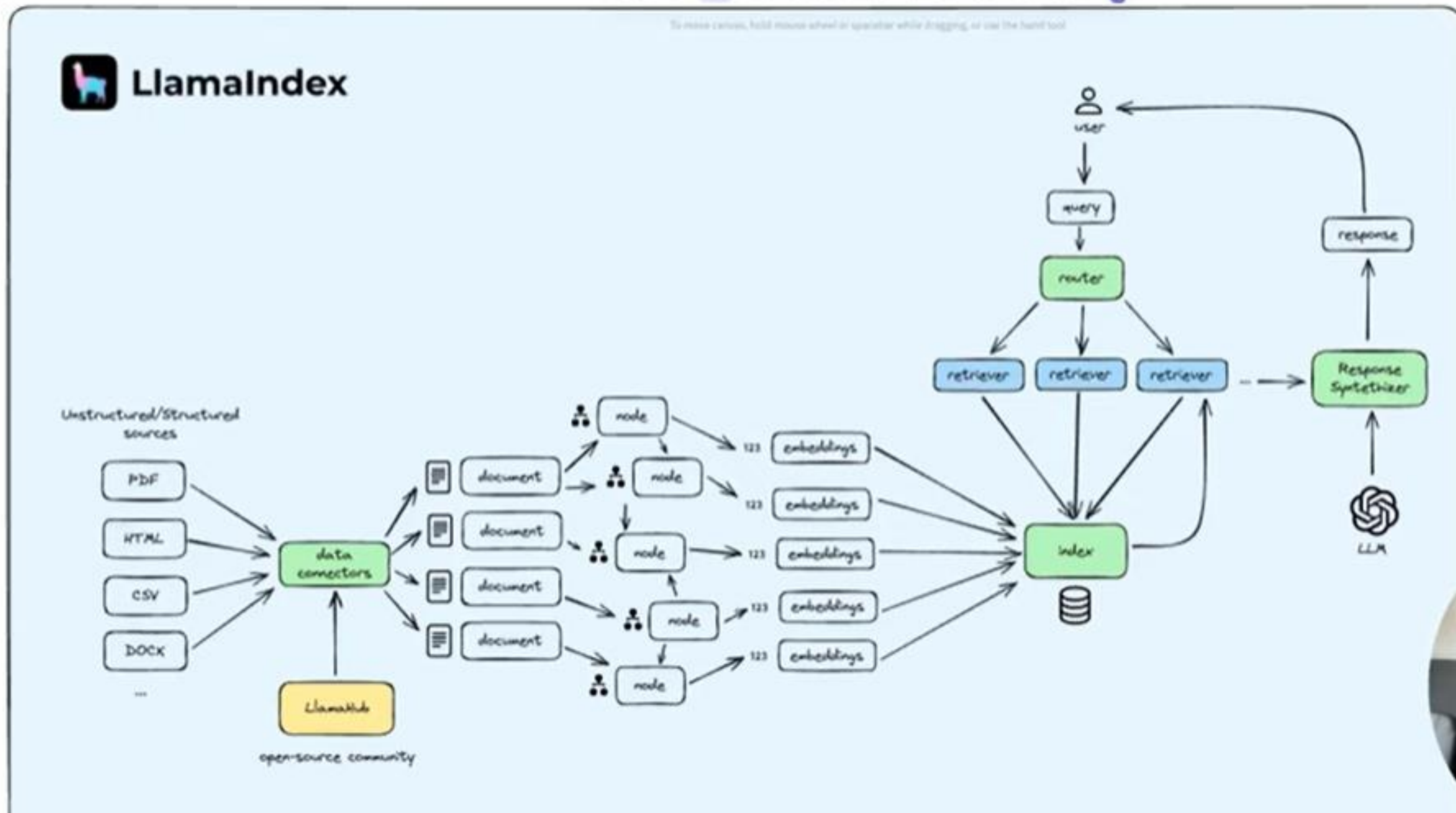
Validation Workflow





LLM - RAG







LLM - RAG

