Recap and Next Steps

Elements of Data Integration (CS 329E)
First Edition

April 26, 2024

Phase 1: Data modeling and transformations

- What we learned:
  - Entity decomposition, merging, and linking
  - Referential integrity
  - Change data capture
  - Slowly changing dimensions
- Next steps:
  - Dimensional modeling
  - dbt (or data build tool)
- Suggested resources:
  - Agile Data Warehouse Design (book)
  - Data Engineering: dbt for SQL (LinkedIn course)

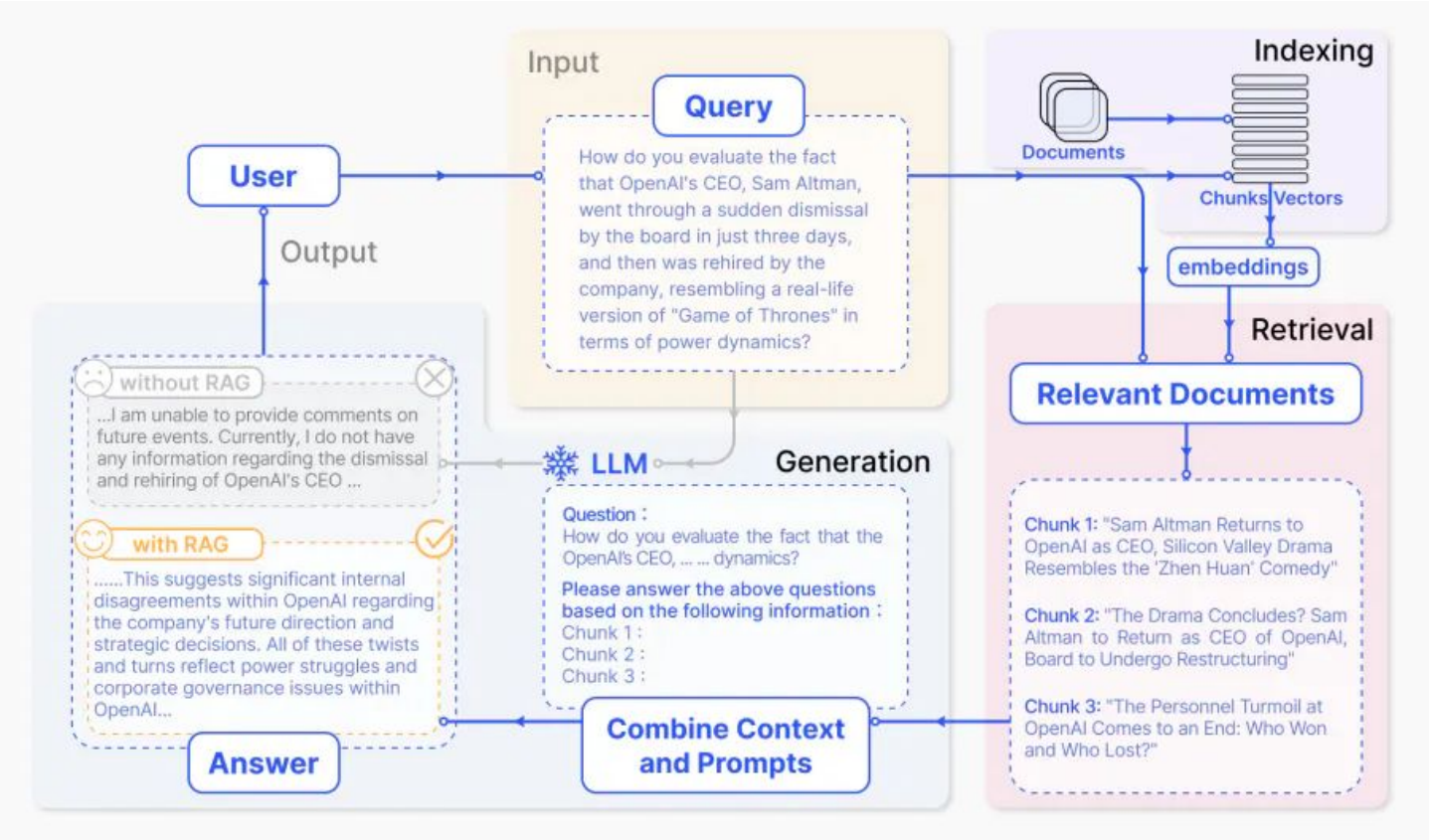
Phase 2: Data pipeline orchestration

- What we learned:
  - Airflow tasks, dependencies, triggers
  - PythonOperator, BigQueryInsertJobOperator, and TriggerDagRunOperator
  - Composer as an Airflow managed service
- Next steps:
  - XCom, templating, sensors, trigger rules, custom operators and hooks
  - Event sourcing and stream processing
- Suggested resources:
  - Data Pipelines with Apache Airflow (book)
  - Grokking streaming systems (book and video)

Phase 3: Data enrichment

- What we learned:
  - Attribute generation with Gemini Pro and BigQuery's ML.GENERATE_TEXT function
  - Basic prompting techniques (one shot, few shot, chain-of-thought)
  - Parsing predictions with BigQuery's JSON functions
- Next steps:
  - Prompting with a document retrieval system (see RAG example on next slide)
  - Prompting with tool use (API calls)
  - Prompt orchestration with LangChain
  - LLM fine-tuning with BigQuery
- Suggested resources:
  - Prompt Engineering Guide (open-source guide) and Gemini-specific prompting (tech talk)
  - The Complete LangChain and LLMs Guide (8-hour masterclass)
  - Gorilla: Large Language Model Connected with Massive APIs (research paper)
  - Introducing LLM fine-tuning and evaluation in BigQuery (blog post)

RAG (Retrieval Augmented Generation) Example



Source: [Retrieval-Augmented Generation for Large Language Models: A Survey]

Additional Resources

- Intro to Machine Learning (Kaggle Learn course)
- AI Crash Course (YouTube series)
- LLM Overview (LinkedIn Learning course)
- Natural Language Processing with Attention Models (Coursera course)
- GPT Under the Hood (YouTube, 2 hour video)