

# Finance AI Agent

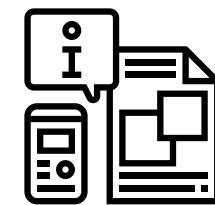
---

- Yingxuan Bian, Xinxin Liu, Wenjun Song, April Yang

# Problem

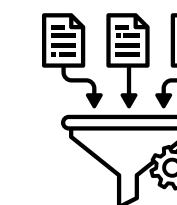


How can we **efficiently extract actionable insights** from 10-K filings to **support better financial decision-making?**



## Information Overload

10-Ks are long and complex, and often hard to parse quickly



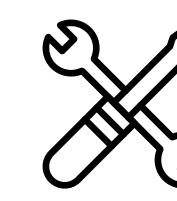
## Unstructured Data

10-K text lacks structure, limiting traditional analysis



## Missed Opportunities

Hidden signals often go unnoticed, leading to reactive rather than proactive decision-making



## Lack of Context Tools

Current tools can't capture relationships or context well

# Solution

A scalable **GraphRAG system** that transforms unstructured 10-K filings into a structured knowledge graph, enabling **AI-powered answers** to user questions and context-aware insights through intelligent retrieval and language generation.

## Organized Information

- Extracts and structures key sections from 10-K filings
- Save time and effort by making information easy to search and compare

## Connected Insights

- Builds a network of relationships between risks, strategies, and financial topics.
- Helps uncover patterns that are often hidden

## Smart Answers

- Allows users to ask questions in plain English.
- Delivers clear, AI-powered answers backed by filing content.

## Faster & Better Decisions

- Empowers investors with instant, actionable insights.
- Turns complex regulatory documents into strategic decision-making tools.

# Data

Existing data from **10-K filings**, retrieved via SEC EDGAR and company websites using CrewAI + Serper tools.

## \* Design:

- Queried SEC EDGAR using **CIK** to fetch 10-K **accession numbers by year**
- Cleaned and parsed filings into **structured JSON format**
- Extracted key sections for **NLP and graph-based analysis**

## \* Outcome and Target:

- Context-aware **answers to user financial queries** using our AI agent
- In later stages, potential **trading signal generation** based on filing insights

## \* Explanatory Features:

Our cleaned 10-K JSON files include **8 structured elements**:

- **company\_name, cik, cusip\_number, source**
- **item1** (Business Overview)
- **item1a** (Risk Factors)
- **item7** (Management Discussion & Analysis)
- **item7a** (Quantitative & Market Risk)

## \* Automation Cadence:

- **Data is refreshed annually** to ensure up-to-date 10-K filings
- Script supports **flexible parameters** for form type and date range
- Uses an `update_log` to **track progress and resume** if interrupted

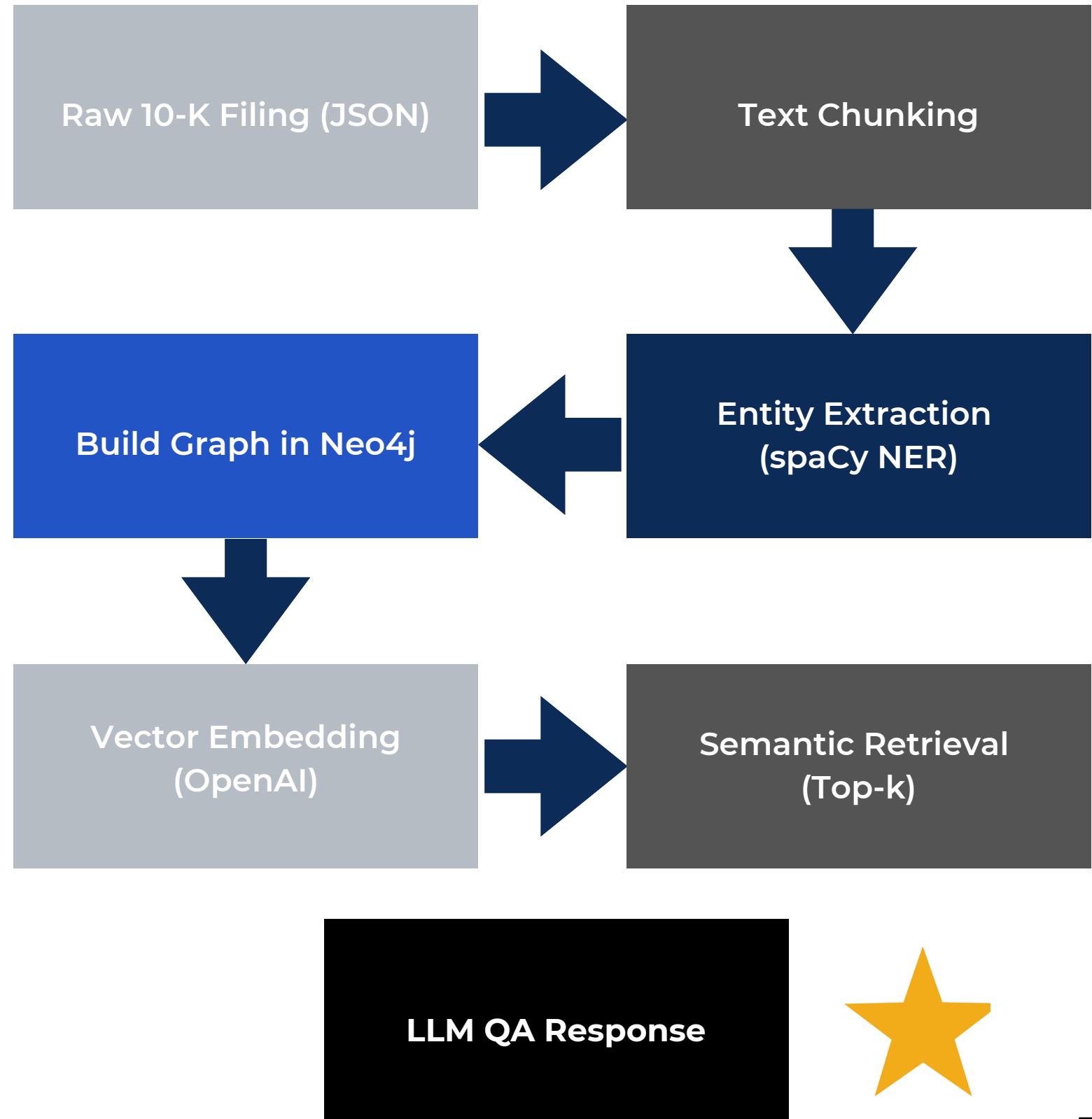


# Methodology

## Overview of the Pipeline

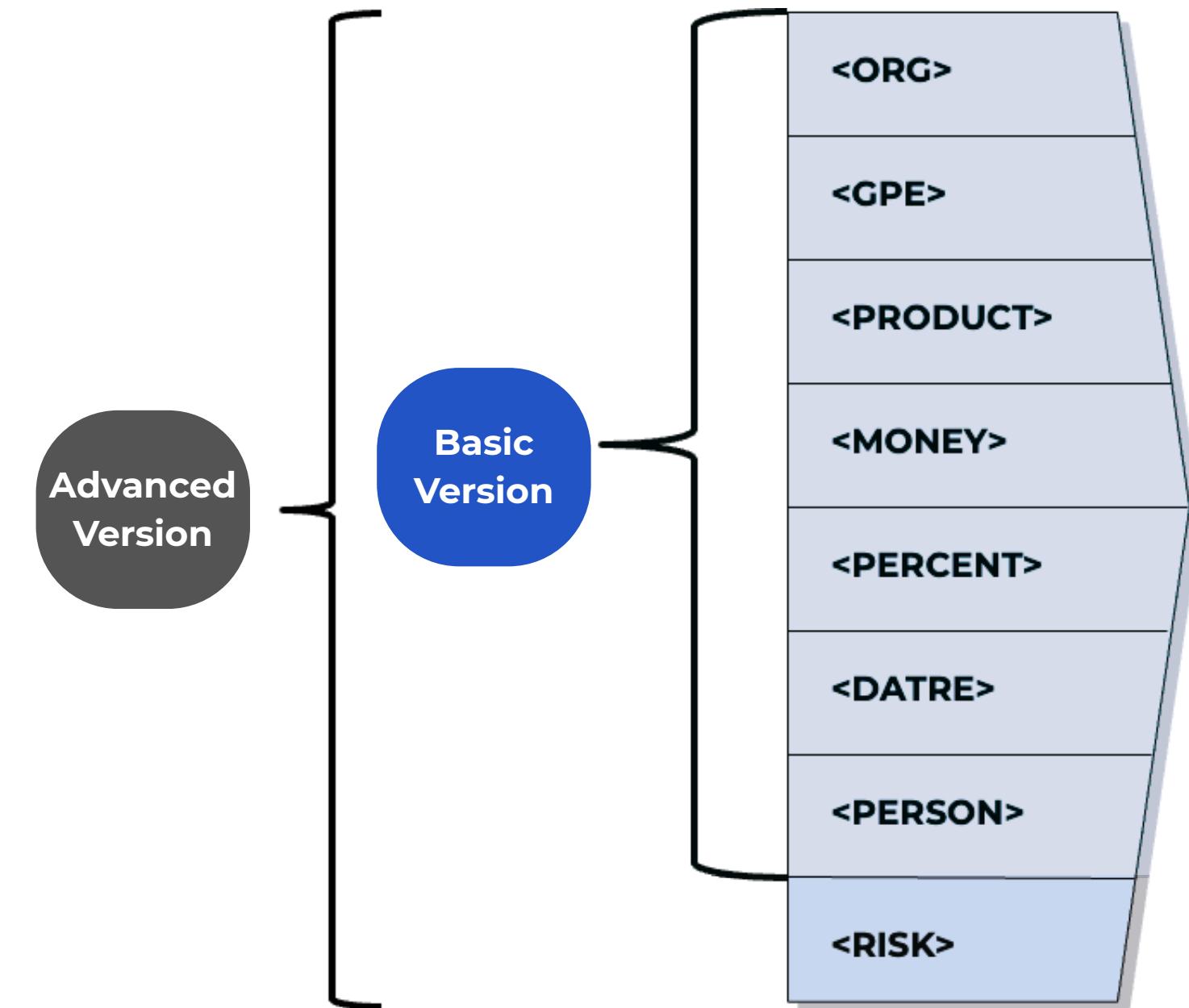
### GraphRAG System Architecture

- **Paragraph-level chunking**  
improves granularity for retrieval.
- **spaCy** used for extracting **financial entities**: ORG, GPE, PRODUCT, etc.
- **Neo4j graph schema** includes  
TextChunk, Entity, and MENTIONS,  
CO\_OCCURS\_WITH relationships.



# Entity Extraction: From Basic NER to Risk-Aware Enrichment

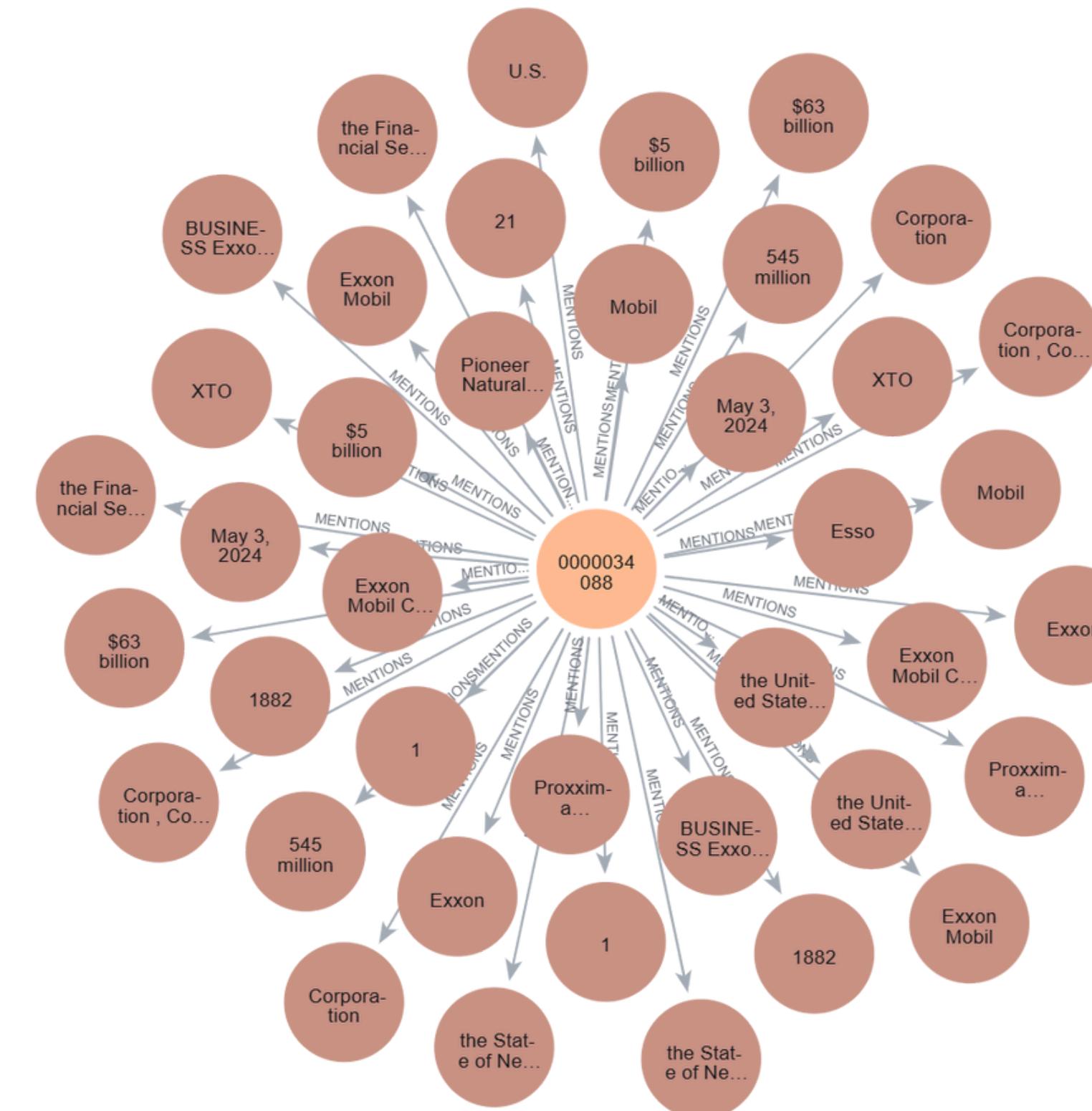
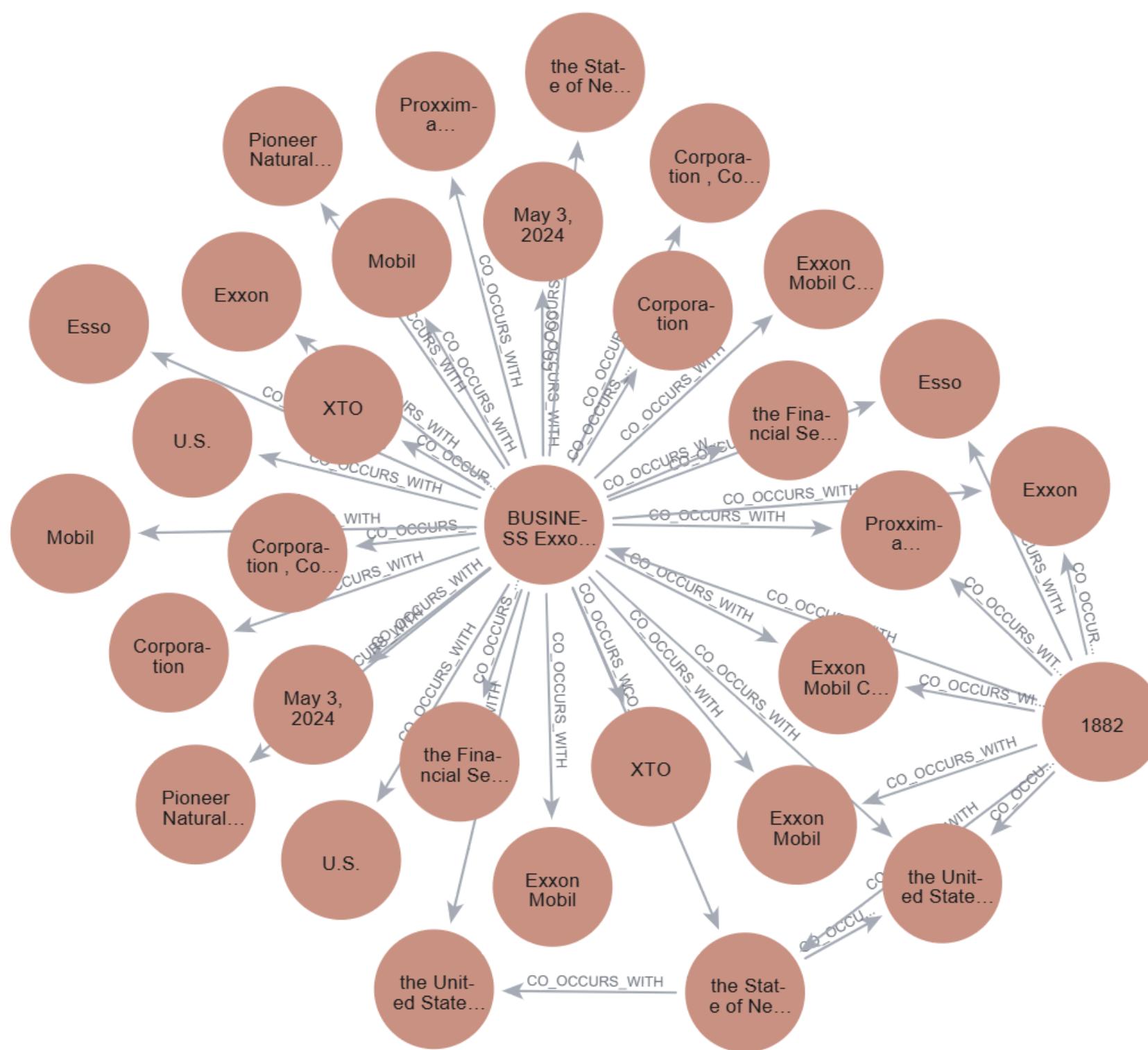
Aspect	Basic Version	Advanced Version
NLP Method	spaCy NER	spaCy NER + PhraseMatcher
Entity Coverage	Generic entities	Generic + Risk-specific phrases
Write Strategy	One-by-one MERGE	Batched UNWIND MERGE
Performance	Slower on large data	Optimized for bulk graph creation
Use Case	Initial proof of concept	Scalable, production-grade graph enrichment



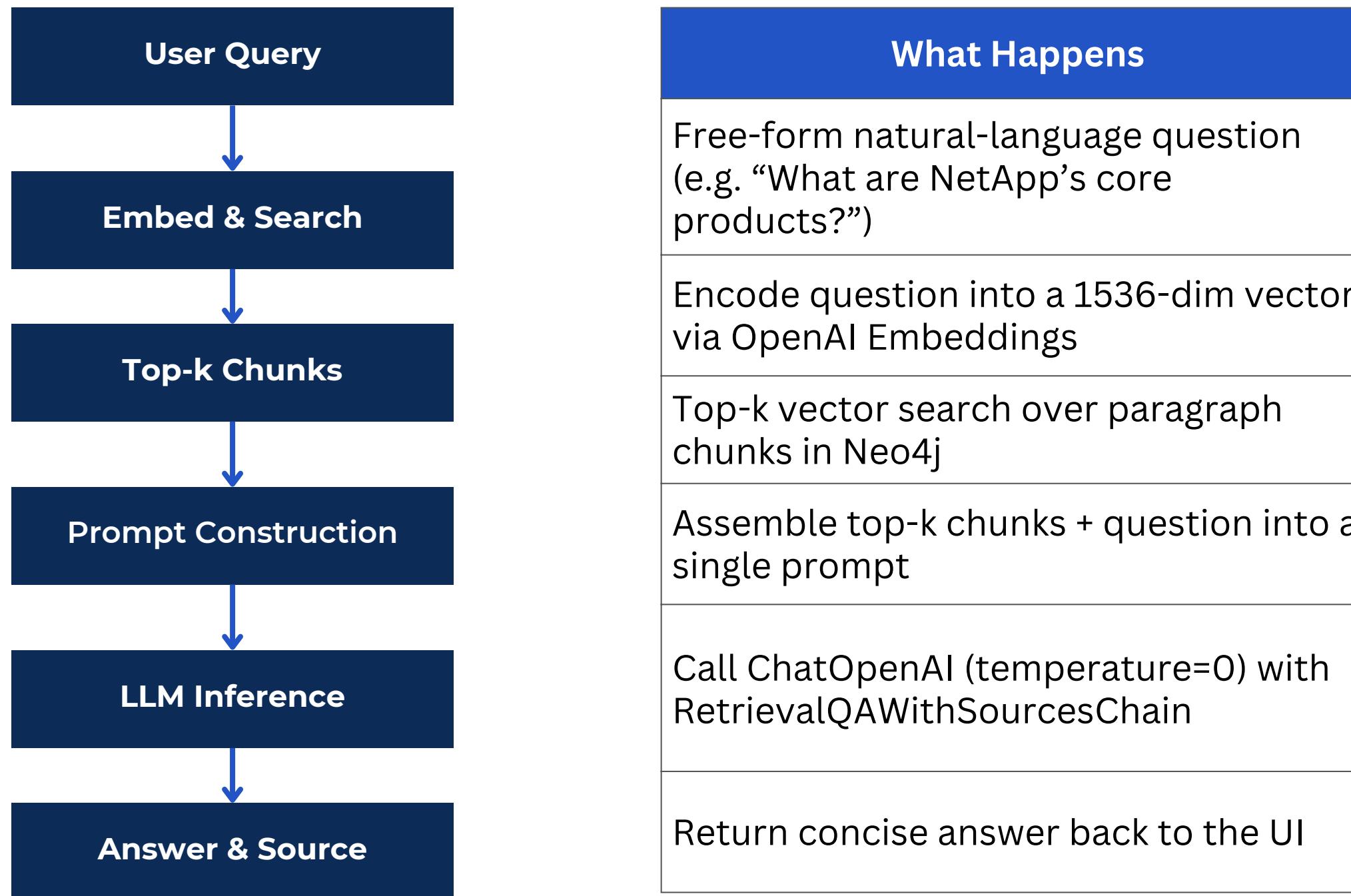
\* Risk: "supply chain disruption", "macroeconomic downturn", "foreign exchange risk", "component shortage", "regulatory changes", "inflation risk", "pandemic impact", "political instability", "competitive pressure", "litigation risk"

# Graph Relationship built with Neo4j

Two major relationship: CO\_OCCUR\_WITH & MENTIONS



# LLM QA Response: Empowered by Vector Embedding and Semantic Retrieval



# Result

## Query Answers about 10k-fillings with the LangChain



### Financial Document Q&A System

This application allows you to ask questions about companies based on their SEC 10-K filings. The system uses a knowledge graph built from multiple companies' filings to provide accurate and contextual answers.

Enter your question about any company:

What are the top risks mentioned in Johnson & Johnson's 10-K?

#### Example Questions:

- What is Netflix's primary business?
- Where is Apple headquartered?
- What are the top risks mentioned in Johnson & Johnson's 10-K?
- Where are the primary suppliers for Tesla?
- How is ExxonMobil addressing climate change and the energy transition?

#### Answer:

The top risks mentioned in Johnson & Johnson's 10-K include legal proceedings, government regulations, intellectual property rights challenges, and product safety concerns.

# Conclusion

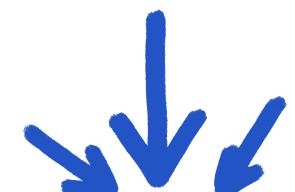
## Summary



We built a **GraphRAG-powered Finance AI Agent** that extracts insights from 10-K filings.

Our pipeline integrates **data automation**, **NLP-based entity extraction**, and **graph-based retrieval** for context-aware querying.

Users can now **interactively ask financial questions** via our deployed **Streamlit demo**.

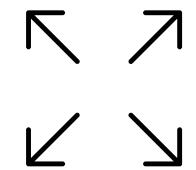


**This validates our end-to-end pipeline  
From raw data to intelligent, grounded answers**

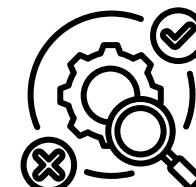
## Next Steps



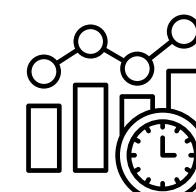
**Refine LLM + Graph integration** for better query precision and robustness.



**Scale to more companies and filing types** with optimized data pipelines and storage.



**Evaluate system performance** through both domain experts and benchmark tools.



Explore potential for **trading signal generation** and **agent deployment in real-time**.

# Thank You

---

- Yingxuan Bian, Xinxin Liu, Wenjun Song, April Yang