Finance AI Agent Final Report

Your Smart Assistant for Insights on Listing Companies

Yingxuan Bian, Xinxin Liu, Wenjun Song, April Yang
BAX423 Group 13

Table of Contents

1. Business Objective

Our top business priority is to empower financial professionals with rapid, accurate and actionable insights extracted from complex regulatory documents like 10-K filings. These documents contain critical strategic and risk information, yet their length and unstructured nature make them difficult to process efficiently, often leading to overlooked signals and delayed decisions.

Our objective is to turn these filings into a competitive advantage by making them easily searchable, interpretable, and directly usable in financial decision-making. Through a scalable AI-powered system, we aim to reduce information overload, increase transparency, and significantly improve the speed and quality of financial decision-making.

2. Key Actionable Business Initiative

We are considering several business initiatives to bring this objective to life: (1) launching an AI-powered finance assistant platform that provides natural language Q&A over regulatory filings; (2) developing a red-flag alert system that automatically detects and notifies users of emerging risks in new filings; and (3) offering API-based integration for institutional tools such as Bloomberg Terminal, Power BI, and Excel to embed our insights into daily workflows.

Among these, the most impactful initiative is launching a public-facing AI-powered finance assistant platform as a beta product targeting investors and analysts. This initiative is specific, actionable, and directly tied to our current system capabilities.

To execute this initiative, we would begin with clearly articulating the platform's core value proposition: helping financial professionals extract and interpret key insights from complex filings with minimal effort. We would start with a focused MVP targeting a few key industries to ensure precision and relevance. This MVP would include core features such as semantic search, financial metric extraction, and natural-language Q&A.

To validate and improve the product, we would release it as a public beta with a managed waitlist to build early interest and control onboarding. We would embed feedback mechanisms within the platform and actively engage early users to identify pain points, measure utility, and guide product refinement. This phased approach ensures we deliver meaningful value from day one while creating a scalable foundation for long-term growth.

3. Metrics of Success

3.1 User Adoption & Engagement

This can be broken down into Number of registered β-users on the AI-powered finance assistant platform, and Volume of unique queries submitted per week. This matters because higher engagement signals that users find the platform immediately valuable.

Hypothesis we have is within 3 months of public β launch, we will onboard 500 active users. And those users will submit an average of 20 unique queries per user per month, indicating sustained engagement.

3.2 Answer Accuracy & Relevance

This can be broken down into the percentage of user queries for which the system's AI-generated response is rated "accurate" (vs. "incomplete" or "incorrect") by a small panel of domain experts, and the average relevance score (on a 1–5 scale) collected via follow-up user surveys asking, "How well did the answer address your question?" This matters because high precision is essential to building trust—if the AI hallucinates or misses key points, professional users will quickly abandon the tool—and relevance scores help ensure we're surfacing the right 10-K passages and interpreting them correctly.

Hypothesis is that by the end of Month 1 of β testing, we will achieve ≥ 85 % accuracy on a standardized basket of 50 test queries, and by Month 2 our average user-reported relevance score will reach ≥ 4.0/5.

3.3 Time-to-Insight Reduction

This can be measured as the average time (in minutes) required for a user to locate and interpret a specific financial insight (for example, "What is Company X's risk exposure to foreign exchange?") using our platform versus performing a manual search through a 10-K document. This matters because our core value proposition is speed in terms that if financial professionals spend significantly less time finding the same information, we deliver quantifiable ROI and make regulatory filings a genuine competitive advantage.

Hypothesis is that at public β launch, our system will reduce average time-to-insight by 50 percent (for instance, from 20 minutes manually to 10 minutes via AI), and within two iterative product cycles (by

Month 4) that reduction will improve to 70 percent (e.g., from 20 minutes to 6 minutes) as we optimize embeddings and fine-tune retrieval pipelines.

4. Role of Analytics

4.1 Enable the Business Initiative

Data Pipeline Health Monitoring: It can track the completeness and freshness of 10-K JSON ingestion (e.g., percentage of filings processed per quarter). Monitor embedding generation throughput (e.g., chunks processed per hour) to maintain updates for new filings.

Retrieval Performance Metrics: It can continuously measure vector-search recall and precision on a hold-out set of "ground-truth" answers, ensuring that semantic retrieval is delivering the correct chunks.

4.2 Ideate & Refine the Business Initiative

Query Pattern Analysis: It can (1) Analyze top user queries (e.g., "Show me cash-flow trends over the past 5 years") to identify missing features or data types worth surfacing; And (2) Detect recurring failure modes (e.g., if many users ask "What is Company X's guidance?", but our system cannot parse "guidance," we know to add a specialized parser for MD&A forward-looking statements).

A/B Testing of Prompt Templates & UX Flows: It can (1) Deploy alternative prompt-engineering variants (e.g., "Extract risk factors" vs. "Summarize top 3 risks") and measure their relative "helpfulness" scores via user feedback; (2) Iterate on the interface copy (e.g., "Ask a question" vs. "Search 10-K") and track click-through rates to optimize clarity and reduce friction.

4.3 Evaluate the Success of the Business Initiative

KPI Dashboards & Reporting: It can (1) Build a real-time dashboard showing all three success metrics (user adoption, accuracy, time-to-insight) to share with stakeholders weekly; (2) Incorporate cohort analyses (e.g., Month-1 β registrants vs. Month-2 β registrants) to compare retention and "stickiness."

Net Promoter Score (NPS) & Qualitative Surveys: It can (1) Embed short post-session surveys ("Was this answer useful?") that feeds into an NPS calculation; (2) Collect anonymized feedback on edge-case

failures to pinpoint limitations in our NER or graph schema (e.g., “I asked about Deferred Revenue but got an unrelated risk summary”).

Business Impact Modeling: It can (1) Estimate dollar-value savings by calculating the average hourly rate of a financial analyst and multiplying by “minutes saved per query” aggregated over active users. Compare this to our operating costs to validate ROI; (2)Track any downstream conversions: for β users who upgrade to a paid subscription (once launched), attribute revenue lift to specific analytic enhancement.

5. Thinking Through the Analytics

5.1 Data

In this project, we rely on existing public financial filings—specifically, each company’s most recent 10-K report—as our primary data source. These filings are downloaded directly from the SEC EDGAR database, ensuring data consistency and official quality.

Our outcome variable is the accuracy and relevance of answers generated by our system to user queries about the 10-K. Our explanatory variables/features include:

- Text content from 10-K sections (e.g., Item 1, 1A, 7, 7A)
- Named entities extracted from the text (e.g., companies, products, risks, locations)
- Embedding vectors from OpenAI models (representing semantic meaning)
- Graph structure features from Neo4j (e.g., centrality, co-occurrence)

These features vary across companies and over time due to differences in writing style, risk disclosure, business models, and structure of the filings. This variation allows us to assess how well our retrieval and QA pipeline adapts to different corporate contexts.

5.2 Type of Analytics

Our project primarily involves exploratory analytics, with elements of causal reasoning considered for future iterations.

Our exploratory analysis helped us understand how companies disclose different types of information. We examined how risk factors are framed, how business models are described, and how financial strategies

are discussed across firms. Since the source is public regulatory filings, we have a clear understanding of how the data is generated and what it captures.

We also explored which types of named entities and sections tend to yield more accurate answers when retrieved for a given query. This guided our graph construction and retrieval strategy.

Although we do not explicitly make causal claims, we recognize that understanding relationships between sections (e.g., how risk factors relate to business performance) may suggest underlying causal patterns. These interpretations remain exploratory unless backed by further experimental design.

5.3 Impediments

A key challenge we face is entity noise and irrelevance in the knowledge graph. Not all extracted entities are meaningful or useful for downstream QA, and some may even degrade performance. We are currently working on:

- Filtering entities based on frequency, type, and graph position.
- Incorporating confidence scores from entity recognition to discard low-quality results.

Another limitation is the lack of ground-truth answers for evaluation. Without labeled question-answer pairs, we rely on manual inspection. Future work may include building a labeled evaluation dataset or using external QA benchmarks for fine-tuning.

6. Executing the Analytics

The analytics work in our project has been carried out collaboratively by our student team, with specific roles assigned to ensure efficient execution.

Data Collection was primarily handled by Wenjun Song and April Yang, who developed Python-based scripts to automate the downloading and preprocessing of each company's most recent 10-K filings from the SEC EDGAR database. This included extracting relevant sections and parsing financial metadata.

Model Development and Execution was shared across the team. Selina Bian focused on building the GraphRAG pipeline, integrating named entity recognition and Neo4j graph construction, while Stephanie Liu focused on QA retrieval and response generation using OpenAI's language models. Wenjun Song was responsible for scaling the pipeline, extending the model beyond individual examples to operate

effectively across multiple companies. For evaluation, we designed a small set of manual test queries and used and relied on human judgment to assess answer quality.

Defining Metrics and Evaluation Criteria was a joint effort involving all team members. We held regular meetings to discuss how to assess the accuracy and relevance of responses, as well as how to filter out irrelevant or noisy graph entities. As part of this process, we consulted with our project advisor for feedback on what metrics would be most meaningful in an applied setting.

While we did not have an external organization as a partner, we operated as an internal team, and each member contributed to thinking critically about the value and validity of our analytics.

7. Implementation

7.1 Decision Impact

Transforming a raw 10-K into a knowledge graph and retrieval-augmented interface alters the sequence of choices that analysts make during a reporting cycle. Prior to this project, the first task was locating relevant text; after deployment, the initial task was interpreting system-ranked paragraphs and deciding whether the disclosed information is material. This shift elevates three practical decisions:

- Pre-ranked paragraphs direct the analyst to the most salient risks, strategy changes, or quantitative disclosures; as a result, marginal time is reallocated from document scanning to valuation analysis.
- Co-occurrence patterns between risk phrases and product lines facilitate earlier identification of issues warranting management attention or inclusion in earnings-call briefs.
- A harmonised graph schema enables defensible cross-company comparisons, e.g., the relative prevalence of “inflation risk” language, thereby informing sector commentary and watch-list updates.

Because these decisions now rest on automatically surfaced evidence, a larger share of analyst time can be devoted to developing forward-looking scenarios rather than to manual information retrieval.

7.2 Workflow Integration

To ensure that this new capability is actually used, the project treats workflow integration as a design requirement rather than an afterthought. The Streamlit front-end launches with a single command and appears in the browser alongside the data terminals and notebooks that analysts already open at the start of the day. Answers can be copied directly into those notebooks, and a built-in export button produces JSON or CSV so that the graph evidence can be merged with valuation models without reformatting. A short video walkthrough and a two-page quick-start guide are distributed with the repository; both emphasise how to trace every generated sentence back to its source paragraphs, which preserves the audit trail expected in professional diligence. Finally, the interface records simple thumbs-up and thumbs-down reactions; that feedback is reviewed during the weekly code session so that problematic queries are fed back into synonym lists or retrieval filters before they erode user confidence.

8. Scale

8.1 Challenges

Scaling the Finance AI Agent beyond a course prototype will confront four tightly–linked constraints.

Data volume will jump once quarterly 10-Q, 8-K, and 20-F filings join the corpus, inflating both storage and embedding costs. System capacity will be stretched, because a single-instance Neo4j and a flat-file vector index cannot guarantee sub-second retrieval when node counts reach the millions. Human resourcing is another choke-point: routine monitoring of scrapers, API quotas, and model drift requires skills not uniformly held by future student cohorts or potential industry users. Finally, cultural acceptance cannot be assumed; analysts unfamiliar with retrieval-augmented generation may reject AI-written summaries unless provenance is transparent and validation effortless.

8.2 Mitigation

The mitigation plan begins with data optimisation. The paragraph-level deduplication paired with incremental batch processing ensures that only genuinely new content is embedded, keeping marginal cost predictable. On the systems side, the entire pipeline is container-ready, so migrating the graph to a lightweight managed service and substituting an approximate-nearest-neighbour index (e.g., FAISS or a hosted alternative) can be achieved by changing environment variables rather than rewriting code. Skills risk is managed through a version-controlled run-book and automated CI tasks. These approaches allow

the next cohort or an external research-tech partner to assume maintenance with minimal ramp-up. To build trust, the interface will retain paragraph citations by default and any wider rollout will run in ‘shadow mode’ for one reporting cycle, letting users compare AI output with legacy notes before making it the primary source.

8.3 Further Improvement

The initiative is designed to remain a living analytical asset, not a one-shot submission. Extraction templates, risk lexicons, and model checkpoints are all tracked in Git; a small, manually curated benchmark of question–answer pairs is rerun every month. Pull requests—whether they introduce a new document template or a more compact embedding model—are merged only when they deliver measurable gains in precision, recall, or latency on that benchmark. User feedback captured through the interface feeds the same backlog, ensuring that practical pain-points drive the next development sprint. These governance steps allow the Finance AI Agent to grow with the expanding disclosure universe and the evolving expectations of its users while preserving the transparency and reproducibility required in an academic or professional setting.