DBA5102 Business Analytics Capstone Project

Roy Yeo Fu Qiang (A0280541L)

Investment Context Engine (ICE): A Lightweight Graph-Based Context-Aware AI Backbone for Hedge Fund Workflows

1. Executive Summary

Investment professionals are inundated with information – from portfolio data and watchlists to a flood of sell-side research emails and market news. Lean, traditional hedge funds face a disadvantage when competing against larger investment firms with greater resources. The proliferation of third-party AI tools, combined with democratization of cutting-edge large language models (LLMs) present a clear opportunity: lean hedge funds must become AI-ready to tap onto these tools to harness its potential and level the playing field with industry giants.

The Investment Context Engine (ICE) is a modular, lightweight AI system designed to serve as the cognitive backbone of a hedge fund's core workflows—spanning idea generation, equity research, portfolio monitoring, risk management, and investor communications.

ICE integrates and encodes both external financial data (e.g., filings, news, earnings transcripts) and internal firm sources (e.g., research notes, emails, portfolio holdings) into a unified, compounding knowledge graph (KG) that continuously encodes the hedge fund's DNA, mental models and investing philosophy. This KG structure enables multi-hop reasoning and long-term memory which is queried through a graph-aware retrieval-augmented generation (Graph-RAG) (Edge et al., 2024) pipeline, processed by a monolithic but extensible agentic layer that is designed to be compatible with external Model Context Protocol (MCP)-based tools, databases, and models.

By combining structural graph context with semantic retrieval, ICE grounds large language models (LLMs) in the most relevant and complete firm-specific knowledge, allowing the investment teams to ask high-leverage questions such as:

- "What KPIs drive Uber's stock performance?" or
- "Which companies in the portfolio holdings are exposed to the AI infrastructure buildout theme?"

Answering these queries would showcase ICE's ability to synthesize fragmented data into structured, traceable and actionable insights – which accelerates and improves investment reasoning, aligns decision-making across teams, and builds a compounding knowledge asset that delivers sharper insights and more informed decision-making across different investment functions.

2. Problem Statement & Background

Modern equity research and portfolio management depend on synthesizing insights from an ever-expanding range of fragmented sources: earnings calls, SEC filings, market news, broker reports, internal research notes, emails, and price action. For lean boutique hedge funds, they face a structural disadvantage (Alpha FMC, 2020) with this process remains largely manual and disjointed—heavily reliant on human memory, speed, and judgment—placing them at a competitive disadvantage to larger funds with dedicated teams, advanced infrastructure, and vast resources.

This fragmentation creates four critical pain points for the lean investment fund:

1. **Delayed signal capture:** Analysts may miss soft signals or early narrative shifts buried in transcripts, filings, or news flows. PMs and traders often react to events with latency or incomplete context.
2. **Low insight reusability:** Investment theses, observations, and research remain siloed in decks, chats, or emails—rarely rediscovered or built upon across time, teams and use cases.

3. **Inconsistent decision context:** Different team members interpret or prioritize information inconsistently, leading to fragmented understanding and reactive, uncoordinated decisions.
4. **Manual triage bottlenecks:** Stitching together context from disparate data sources remains a fully manual task, limiting both speed and scale.

While AI tools are proliferating in finance, most remain generic—surfacing market-consensus views rather than a fund's unique mental models, thesis structures, or interpretations of risk. Their outputs are often static, disconnected, and unlinked to institutional memory or compounding insight. Frontier models like ChatGPT are powerful and widely accessible, but raw capability is no longer the true edge; the differentiator is *context engineering*—crafting rich, current, and fund-specific context that aligns with the firm's proprietary knowledge, investment logic, and the exact query. This engineered context grounds state-of-the-art LLMs, enabling them to reason across structured and unstructured data and deliver bespoke, explainable, traceable insights.

Large fund managers solve this with dedicated functions, well-funded teams, generous budget and advanced enterprise infrastructure. Small investment firms, however, face structural constraints (Alternative Investment Management Association AIMA, 2024) in talent, bandwidth, and integration capacity—yet they are equally exposed to the same information deluge. For these small firms, the build-or-buy dilemma must be approached strategically: build the *spine*—a proprietary cognitive engine like ICE—while buying the *muscles*—third-party AI tools, databases, and LLMs. Building isolated in-house AI tools ('muscles') risks wasting already limited resources on commoditized capabilities, while outsourcing the cognitive core ('spine') risks leaking proprietary information and surrendering control over how investment logic is encoded and maintained.

ICE resolves this asymmetry by serving not as another tool, but as the fund's systemized memory, reasoning engine, and knowledge flywheel that grows more useful with every query, document or insights added. It is a lightweight, purpose-built in-house system – designed to integrate proprietary knowledge and processes while enabling modular interoperability, via Model Context Protocol (MCP (Anthropic, 2024a), with third-party tools and databases (Model Context Protocol, 2025). This allows small firms to exploit the power of frontier AI without enterprise-scale budgets, retaining strategic control while scaling insight quality, speed, and precision over time.

3. Proposed Solution: Investment Context Engine (ICE)

The Investment Context Engine (ICE) is a modular, lightweight AI system that engineers high-quality context by combining fragmented external data (e.g. earnings transcripts, filings, news) with internal firm knowledge (e.g. research notes, investment memos, emails and portfolio holdings information). Functioning as a dedicated context assembler for investment workflows, ICE structures highly relevant context that ground frontier LLM in the firm's proprietary as well as relevant external information. This ensures every generated response by the LLM is not only relevant, comprehensive and up-to-date, but also align with the fund's investment philosophy and processes.

ICE builds context as a structured composition of facts, signals and documents drawn from the knowledge graph as well as semantic and lexical vector stores, tailored to the query and the firm's decision logic.

Context assemble involves two components:

1. **Short-term context:** user's current query, recent chat history (optional), market snapshots, and transient session memory.
2. **Long-term context:** path-aware subgraph expansion (e.g., NVDA → TSMC → China risk), mapping causal dependencies between entities (e.g. tickers, events, themes, drivers), enriched by semantic embeddings of relevant retrieved documents. Future iterations will extend this to full knowledge-graph retrieval.

ICE's modular design enables it to dynamically assemble the most relevant knowledge slice—from ad hoc company questions to peer comparisons or thematic risk queries (e.g., "Which holdings are most exposed to a China slowdown?").

This composability means ICE is not just a query-answering backend—it also powers higher-level workflows such as the automated generation of daily portfolio briefings, summarizing material changes and soft signals across all positions and watchlist names, anchored in evidence and causal reasoning. In effect, the graph scaffolding used in ICE becomes the connective tissue between fragmented data sources and evolving investment needs, delivering structured, decision-ready insights for the investment team.

3.1. System Architecture Overview & Pipeline

At its core, the Investment Context Engine (ICE) is a domain-specific knowledge graph (KG) that integrates both external financial data (e.g., earnings transcripts, filings, news) with internal firm knowledge (e.g., research notes, investment memos, portfolio holdings). This graph encodes key relationships — such as narrative linkages, KPI drivers, causal chains, and soft signals — aligned with the fund's unique mental models and investment logic.

Queries to ICE—such as *"Why did stock price of ABC fall post-earnings?"* or *"What companies are exposed to the EV subsidy narrative?"*—are handled through a graph-aware retrieval-augmented generation (Graph-RAG) pipeline. ICE retrieves relevant nodes, paths, and documents from the graph and vector database, then passes that structured context to a monolithic agentic layer that simulates the triage logic of a human analyst: extract → reason → synthesize. The result is an explainable, traceable response grounded in the firm's own mental models and context. Unlike SaaS tools or LLM wrappers that focus on access or summarization, ICE is purpose-built to reflect the fund's unique investment processes and cognitive lens. It is MCP-compatible by design, enabling plug-and-play interoperability with future third-party tools, models, or internal systems.

Over time, ICE becomes the fund's cognitive operating layer—a reusable infrastructure that encodes, evolves, and surfaces how the firm reasons about information and data. It transforms every research cycle into a contribution to institutional memory and enables faster, more informed decision-making with each use.

Lazy Graph Retrieval Augmented Generation (Lazy Graph-RAG)

Traditional KGs, while powerful, are often too rigid, labor-intensive, and schema-heavy to be reasonable built and supported by the under-resourced lean investment fund. They often require exhaustive upfront modeling and continuous manual curation—an impractical burden when market narratives shift rapidly. ICE instead adopts a **Lazy Graph-RAG** approach (Edge, Trinh, & Larson, 2024), which retrieves sparse, high-signal subgraphs on demand using a few key edge types (e.g., *depends_on*, *exposed_to*), enabling dynamic reasoning paths without the overhead of a fully materialized graph. This balances precision, explainability, and build efficiency—ideal for answering analyst and PM queries in real time. These subgraphs are retained, audited, and periodically merged into a persistent graph via batch LLM-assisted extraction and schema expansion – balancing immediate utility with long-term KG growth.

Key MVP Features of Lazy Graph-RAG

For the MVP rollout, edges are selected for high business value, low construction cost, and strong interoperability, directly answering recurring PM/analyst questions on theme exposure, KPI sensitivity, and causal drivers. Each edge can be feasibly built using readily available data sources – such as news articles and headings, earnings transcripts, and internal portfolio holdings information. MVP Lazy Graph-RAG implementation priorities include:

- Temporal Edge Metadata – Timestamped evidence for recency and trend tracking
- Reverse Traversal – Bidirectional queries without storing duplicate edges
- Multi-Path Reasoning – Multiple causal paths per query for richer insights
- Edge Confidence Scoring – Ranking by evidence quality and recency
- Source-Aware Attribution – Every edge tied to its originating document

A detailed breakdown of the MVP feature set is provided in **Appendix A.1 (**MVP Lazy Graph-RAG Feature Set & Implementation Priorities) and **Appendix A.2 (**Catalog of Typed Edge Templates).

Appendix A.3 (Limitations of Generic AI Tools and How ICE Addresses Them) elaborates on how off-the-shelf LLMs and generic AI search tools often fail to meet the accuracy, context, and timeliness requirements of an investment workflow — and how ICE's domain-tuned architecture overcomes these gaps with curated edges, proprietary data integration, and analyst-approved reasoning paths.

These synergistic edges naturally connect events → KPIs → themes → holdings, amplifying insight when combined, enabling the KG to surface rich, explainable insights from minimal inputs. This composability not only boosts the utility of individual edges but also amplifies their value when combined. For the MVP rollout, we will implement the knowledge graph in the priority order of the edge patterns listed, establishing a lightweight yet powerful foundation for investment reasoning and decision support.

These priorities are directly tied to the questions analysts and PMs ask in practice. **Appendix A.4** (Example user Queries and Corresponding Graph Edge Activations) provides concrete examples, linking representative queries (e.g., *"What names are exposed to the obesity/GLP-1 drug theme?"*, *"How might China risk impact KPIs across our books?"*) to the specific edge types and reasoning chains used to answer them. This mapping illustrates why these MVP edges were chosen and how they deliver immediate, explainable value from day one.

Hybrid RAG: HyDE + Vector-based Search + Keyword Search + Lazy Graph RAG

Building on the Lazy Graph-RAG foundation, ICE employs a hybrid retrieval architecture in its information retrieval that integrates four complementary retrieval strategies:

- **Semantic vector search** – Captures meaning-aligned documents
- **Keyword search** – High-precision symbol and phrase matching
- **HyDE (Hypothetical Document Embedding)** - Expands vague prompts into more informatively rich pseudo-queries, enabling richer retrieval targets (Gao et al., 2023).
- **Lazy Graph-RAG** – Surfaces structured reasoning paths and entity relationships that align with the firm's cognitive lens

Each layer contributes distinct retrieval capabilities—together, they seek to form a high-recall, high-precision retrieval system that can assemble rich, traceable context bundles tailored to each query.

ICE orchestrates this architecture through a **monolithic agentic workflow**: a single, structured agent receives the user's query and a curated context bundle, executes a deterministic reasoning pass (`extract → reason → synthesize`), and produces a grounded, audit-ready response. This design ensures consistency, transparency, and rapid inference—essential for real-time decision support in high-stakes investment workflows.

For a detailed step-by-step plan of this architecture (query parsing, hybrid retrieval, lazy graph expansion, claim extraction, path scoring, and UI design), see **Appendix B** (Component-level Architecture & Pipeline).

End-to-end Traceability

ICE enforces source traceability at every stage—from document retrieval to final synthesis. Each extracted fact inherits its source document IDs, timestamps, and evidence snippets, which persist through summarization, graph construction, and reasoning. This ensures that when the LLM delivers a final answer, it can cite exact provenance ("trace paths" and "edge IDs"), giving analysts the ability to audit and trust outputs. Any fact lacking a source is automatically flagged or dropped—mitigating hallucination risk (Deloitte, 2024).

Model Context Protocol (MCP) Formatting

All query plans, context bundles, and reasoning outputs are packaged in MCP-style JSON blocks, ensuring modularity, interoperability (Ehtesham et al., 2025), and downstream compatibility. This structure standardizes both LLM inputs (retrieval plans, reasoning paths, document) and LLM outputs (final answers, citations, summaries)—future-proofing ICE for integration with MCP-compliant tools.

RAG Evaluation

The development and subsequent updates of ICE would incorporate RAG evaluation loops to measure answer completeness, faithfulness and context adherence (groundedness). This safeguards against reasoning drift—verifying that responses remain grounded in retrieved evidence and align with the query's intent.

Data Confidentiality (Deferred for MVP)

While not a primary MVP focus, ICE's architecture supports secure handling of sensitive data. For example, company names or proprietary metrics can be anonymized before sending to API-hosted models (e.g., GPT-4) and restored afterward. Alternatively, sensitive processing can be routed through locally hosted open-source LLMs, ensuring IP and regulatory compliance.

Explore: Hypothesis-Driven Lazy Graph Expansion

Rather than only retrieving existing edges, we use financial-specialised LLMs (e.g. FinBERT fine-tuned) to generate plausible "ideal" hypothetical subgraphs based on financial priors which can guide in building the subgraphs:

1. **Hallucinates strategically:** Given the query, the LLM proposes context-aware edge chains (e.g., `AI Regulation → impacts → TSMC → supplies → NVDA`) using financial priors
2. **Validates thoroughly**: For each hypothesized path, run targeted retrieval to find citation-backed spans; do extraction-only parsing; gate with NLI entailment and recency/source-quality scoring to weight edges/paths; prune low scores and stop expansion when marginal gain or token/hop budget is exceeded.
3. **Learns continuously:** Confirmed edges become persistent KG elements with confidence scores, while disproven hypotheses improve future modeling

3.2. Tech Stack and Components

ICE's tech stack is deliberately modular, lightweight, and maintainable by a single quant or developer, ensuring minimal operational overhead while remaining extensible for future scaling. The architecture is designed for developer simplicity, cost efficiency, and strategic upgrade paths – all critical for the fast-moving, resource-conscious environment of a lean hedge fund.

Each component in the stack—from ingestion to retrieval, graph reasoning, and synthesis—has been selected for its high signal-to-effort ratio and ability to integrate into a coherent, auditable pipeline. The design favors open-source, low-cost tools for the MVP, with well-defined transition points to more robust infrastructure as data volume, query complexity, or user demands increase.

A detailed breakdown of all components, their roles, and suggested upgrade paths is provided in **Appendix C** (Detailed Tech Stack Table), which serves as a reference for implementation, maintenance, and future planning.

3.3. Data Infrastructure

The strength of ICE's insight engine comes from its ability to synthesize multi-source, heterogeneous data into a coherent causal knowledge graph—linking companies, suppliers, KPIs, and thematic drivers into a structured reasoning substrate. This process transforms raw, unstructured inputs into decision-grade intelligence by extracting and cross-validating relationships across diverse data streams.

ICE's data strategy deliberately blends:

- **Qualitative sources** – e.g., earnings transcripts, filings, broker research, internal notes.
- **Structured/semi-structured metadata** – e.g., GICS tags, portfolio holdings, market data.
- **Optional alternative data** – e.g., supply chain signals, sentiment feeds, and macro events.

This diversity in data sources is foundational to reasoning fidelity. Cross-source validation reduces single-source bias, strengthens edge confidence, and enables the system to detect hidden drivers and ripple effects that might be invisible in siloed data. The outcome is a high-trust, auditable knowledge framework that empowers analysts and PMs to act faster, with greater confidence, and with traceable evidence paths.

A comprehensive breakdown of the specific data types, their descriptions, sources, cost profiles, and implementation notes is provided in **Appendix D** (Data Infrastructure Table). This serves as both an implementation guide for the MVP and a roadmap for phased data enrichment as ROI is proven.

Explore: Using Financial Corpora

Financial corpora provide curated, labeled datasets that encode domain-specific relationships—such as company-to-KPI links, sentiment-to-earnings effects, and thematic exposures—while capturing the canonical language patterns used by analysts and financial media. In ICE, these are not treated as raw fact sources, but as domain-tuned "edge grammar" coaches for training relation extraction, semantic parsing, and typed edge models in knowledge graph construction.

We adopt a "corpus-as-coach" paradigm: pretrained corpora act as *scaffolds and validators*, guiding edge extraction from real, timestamped documents rather than directly populating the knowledge graph. This preserves ICE's core promise—every edge is traceable, explainable, and grounded in verifiable evidence.

To maximize value while containing costs, corpora are used in four bounded roles (see refined table in **Appendix E:** Financial Corpora Roles, Safeguards and Examples):

1. **Bootstrapping high-confidence edge patterns** – Train and fine-tune NER + RE pipelines using labeled corpora (e.g., FiQA, EDGAR Corpus) to identify common high-precision financial edges.
2. **Pre-filtering and candidate generation** – Leverage corpus-trained embeddings to surface high-likelihood document snippets for Lazy Graph Expansion.
3. **Typed edge validation via distant supervision** – Cross-check extracted edges against corpus-annotated examples to enforce schema consistency and prevent "semantic drift."
4. **Graph densification via semantic matching** – Suggest weakly supported edges based on corpus pattern similarity, flagged as "candidate" until confirmed by new evidence.

Explore: Lazy Graph Expansion with Web Search AI Tools

In ICE, Lazy Graph Expansion leverages web search AI tools such as Perplexity.ai or SERPAPI to overcome knowledge graph sparsity and maintain freshness without the cost of pre-building a fully materialized KG. When a user query requires relationships absent from the current graph—such as linking TSMC to "China risk"—the system formulates a targeted natural language search aligned with the relevant edge pattern (e.g., Company → exposed_to → Theme). Web search results, returned with citations from trusted financial, news, or academic sources, are processed by LLM-based entity and relationship extraction pipelines to yield timestamped, source-attributed edges. These edges are scored for confidence based on source quality, extraction certainty, and recency, then either cached for short-term reuse or temporarily injected into the graph for the duration of the query. This workflow, detailed in **Appendix F** (Lazy Graph Expansion with Web Search Tool) ensures each inference chain benefits from up-to-date, verifiable evidence while controlling latency and API costs through selective triggering. By combining retrieval precision, dynamic edge construction, and temporal caching, Lazy Graph Expansion enables ICE to deliver explainable, high-signal answers even in under-modeled areas of the graph, turning web search into an on-demand extension of institutional memory.

3.4. Build Plan

ICE's build strategy follows a lean-first, value-driven development approach: deliver usable capability early, prove utility quickly, and scale modularly. Each phase is independently deployable, ensuring progress compounds without requiring the full system to be complete before value is realized.

The plan is anchored on three principles:

1. **Leverage existing technologies** – Prioritise adopting proven, readily available technologies over building from scratch. Smithery provides a platform of MCP-compatible solutions and databases that accelerate development while ensuring compatibility and scalability.
2. **Modular development with phased ROI** – Each build phase delivers a self-contained capability, allowing for graceful fallback or pause without rendering prior work obsolete.
3. **Future-proofing with minimal redundancy** – Design extensible foundations to support long-term features like full knowledge graph construction without rework.

For the Hybrid RAG architecture, the initial focus is on establishing the **core workflow**, with advanced extensions like **full KG enrichment** added iteratively.

The roadmap is structured into five sequential phases—from an MVP RAG system through to full batch knowledge graph enrichment. Each phase builds upon the previous, compounding value while minimizing technical debt. For full operational detail, including objectives, deliverables, and example implementations, see **Appendix G.1** (Build Plan Phases & Deliverables) and **Appendix G.2** (Implementation Plan for Each Edge Type)**.**

4. Scope of MVP

The Investment Context Engine (ICE) is a graph-native, purpose-built investment context engineering system designed to unify fragmented external data sources (e.g., earnings transcripts, filings, news) with internal firm knowledge (e.g., research notes, memos, portfolio data) into a structured, verifiable knowledge graph. This graph grounds LLM reasoning, enabling high-fidelity, explainable, and fund-aligned investment insights.

The MVP version of ICE is a lean, high-leverage implementation of a context engineering system – designed to assemble just enough structured, decision-relevant context to power meaningful LLM-driven insights. It is pragmatically scoped to deliver meaningful, decision-relevant insights from day one, while operating within the real-world constraints of time, data access, and headcount of a lean investment fund.

By focusing on a small set of high-signal, typed relationships and relying on lazy, query-triggered graph expansion, the MVP focuses on the highest ROI functionality (i.e., portfolio monitoring and fundamental research support) avoids unnecessary complexity without compromising its business impact. Its pragmatic MVP interface design, as shown in the accompanying mockups in the next section, prioritizes interpretability, traceability, and fast integration with existing workflows—making it a high-leverage context engine that's immediately useful to PMs and analysts, while laying the groundwork for future expansion.

The details of the MVP scope – including focus areas, data sources, edge patterns, processing cadence, and interaction UI design – is presented in **Appendix H** (MVP Scope Table).

MVP Hop Reasoning Chain Retrieval Priorities

The MVP version of ICE will target queries requiring up to three hops in the knowledge graph, covering the majority of portfolio management and analyst use cases while keeping extraction, ranking, and synthesis lightweight. Retrieval will support forward and reverse traversal, temporal scoring, and confidence-weighted ranking—all implemented without a heavy graph database, relying instead on NetworkX + metadata with lazy expansion to contain cost and noise.

Execution approach:

- Explicit query parameters (expand_edges, reverse_lookup, graph_hops) will cap depth per query.
- Queries with different hops will be implemented incrementally, starting with 1-hop (fastest ROI), then 2-hop (portfolio-aware and causal stubs), and finally 3-hop queries (high-signal causal traces).
- Each hop tier provides clear incremental business value, aligning with the phased build plan.

Detailed retrieval patterns, examples, and data sources are provided in **Appendix I** (Hop Retrieval Patterns & Use Cases).

5. User Interface and Mockup

The ICE MVP interface is intentionally lean—built to deliver immediate usability, high-impact insights, and trust—while laying the groundwork for iterative expansion. The design focuses on three pillars:

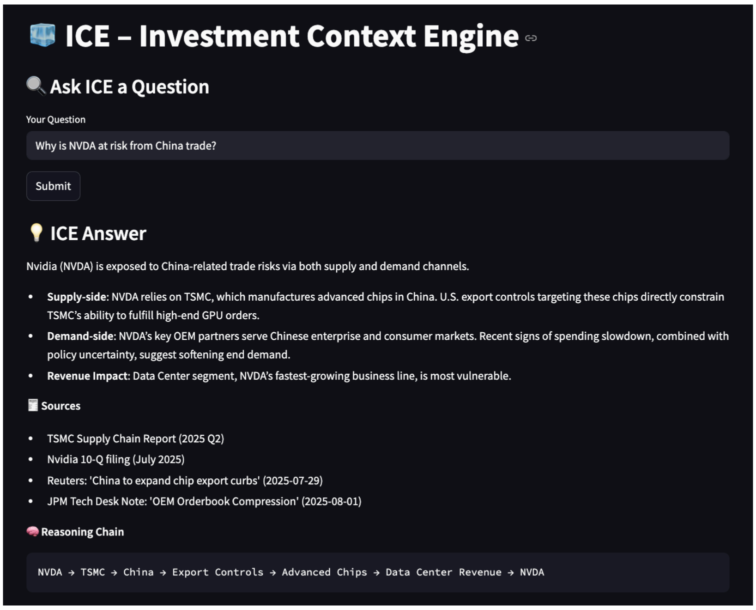
1. **Speed-to-Value** – Instant, causality-grounded answers to complex investment questions without manual data aggregation.
2. **Transparency** – Every output is backed by verifiable evidence (sources, causal paths, confidence scores), fostering trust in AI-assisted insights.
3. **Adaptability** – Modular architecture enables seamless integration of future enhancements (e.g., conversational AI, auto-alerts) without disrupting existing workflows.

The MVP dashboard targets portfolio management, investment research, risk management, and trading workflows for lean equity hedge fund teams. It highlights only the most material drivers of change, explains them in human-readable terms, and grounds every insight in traceable, timestamped evidence.

Below is the phased feature breakdown of the end-user delivery, with the full mockup UI provided in **Appendix J** (ICE MVP User Interface Mockup):

Module 1: Ask ICE a Question (LLM Query Interface)

- **Purpose:** A QA query that accepts natural language queries (e.g., “Why is NVDA at risk from China trade?”) and returns structured, explainable answers.
- **Output:** Key drivers (e.g., supply chain disruptions, demand shifts), a causal reasoning chain, and linked source citations.
- **Impact:** Converts fragmented data into decision-grade insight in seconds.
- **Future:** Evolves into a conversational, multi-turn analysis tool.



🧊 **ICE – Investment Context Engine**

🔍 **Ask ICE a Question**

Your Question

Why is NVDA at risk from China trade?

💡 **ICE Answer**

Nvidia (NVDA) is exposed to China-related trade risks via both supply and demand channels.

- **Supply-side**: NVDA relies on TSMC, which manufactures advanced chips in China. U.S. export controls targeting these chips directly constrain TSMC's ability to fulfill high-end GPU orders.
- **Demand-side**: NVDA's key OEM partners serve Chinese enterprise and consumer markets. Recent signs of spending slowdown, combined with policy uncertainty, suggest softening end demand.
- **Revenue Impact**: Data Center segment, NVDA's fastest-growing business line, is most vulnerable.

🧾 **Sources**

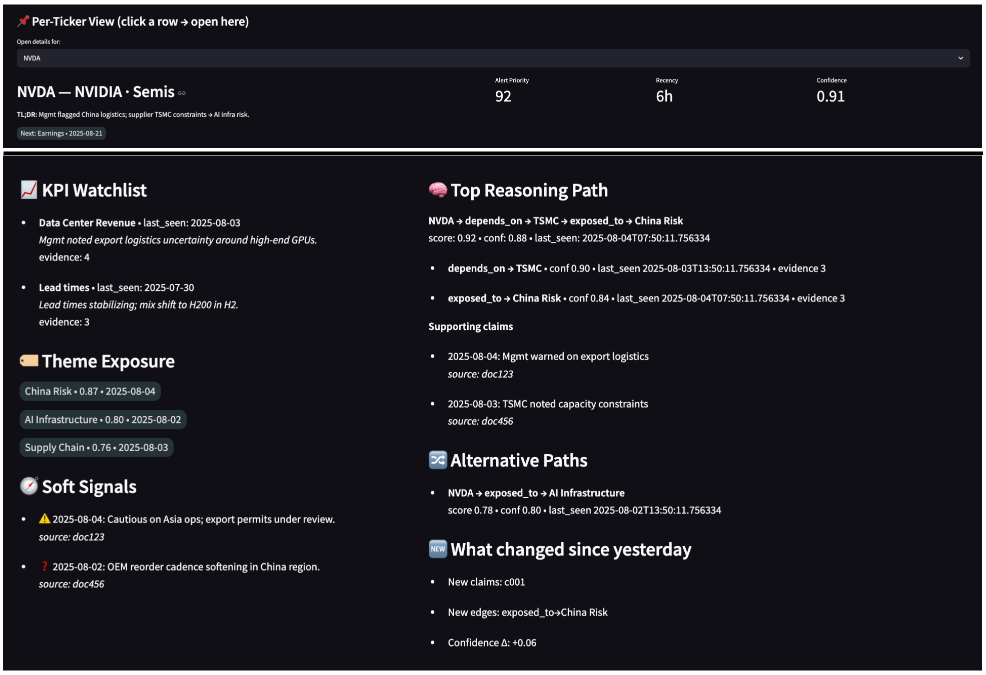
- TSMC Supply Chain Report (2025 Q2)
- Nvidia 10-Q filing (July 2025)
- Reuters: 'China to expand chip export curbs' (2025-07-29)
- JPM Tech Desk Note: 'OEM Orderbook Compression' (2025-08-01)

🧠 **Reasoning Chain**

```
NVDA → TSMC → China → Export Controls → Advanced Chips → Data Center Revenue → NVDA
```

Module 2: Per-Ticker View – Intelligence Panel

- **Purpose:** In-depth drill-down for any company in the portfolio/watchlist or from a query.
- **Features:** TL;DR summary, alert priority, confidence score, recency, KPI drivers, thematic exposures, soft signals, and causal paths.
- **Integration:** Opens subgraph view for visual exploration; "What Changed" tracker highlights shifts in evidence, edges, or confidence.
- **Impact:** Balances compactness and depth—insights at a glance with the option for deeper causal analysis.



📌 **Per-Ticker View (click a row → open here)**

Open details for:

NVDA

NVDA — NVIDIA · Semis

TL;DR: Mgmt flagged China logistics; supplier TSMC constraints → AI infra risk.

Next: Earnings • 2025-08-21

Alert Priority	Recency	Confidence
92	6h	0.91

📈 **KPI Watchlist**

- **Data Center Revenue** • last_seen: 2025-08-03
  Mgmt noted export logistics uncertainty around high-end GPUs.
  evidence: 4
- **Lead times** • last_seen: 2025-07-30
  Lead times stabilizing; mix shift to H200 in H2.
  evidence: 3

🏷️ **Theme Exposure**

China Risk • 0.87 • 2025-08-04

AI Infrastructure • 0.80 • 2025-08-02

Supply Chain • 0.76 • 2025-08-03

🧭 **Soft Signals**

- ⚠️ 2025-08-04: Cautious on Asia ops; export permits under review.
  source: doc123
- ❓ 2025-08-02: OEM reorder cadence softening in China region.
  source: doc456

🧠 **Top Reasoning Path**

NVDA → depends_on → TSMC → exposed_to → China Risk
score: 0.92 • conf: 0.88 • last_seen: 2025-08-04T07:50:11.756334

- **depends_on → TSMC** • conf 0.90 • last_seen 2025-08-03T13:50:11.756334 • evidence 3
- **exposed_to → China Risk** • conf 0.84 • last_seen 2025-08-04T07:50:11.756334 • evidence 3

Supporting claims

- 2025-08-04: Mgmt warned on export logistics
  source: doc123
- 2025-08-03: TSMC noted capacity constraints
  source: doc456

🔀 **Alternative Paths**

- **NVDA → exposed_to → AI Infrastructure**
  score 0.78 • conf 0.80 • last_seen 2025-08-02T13:50:11.756334

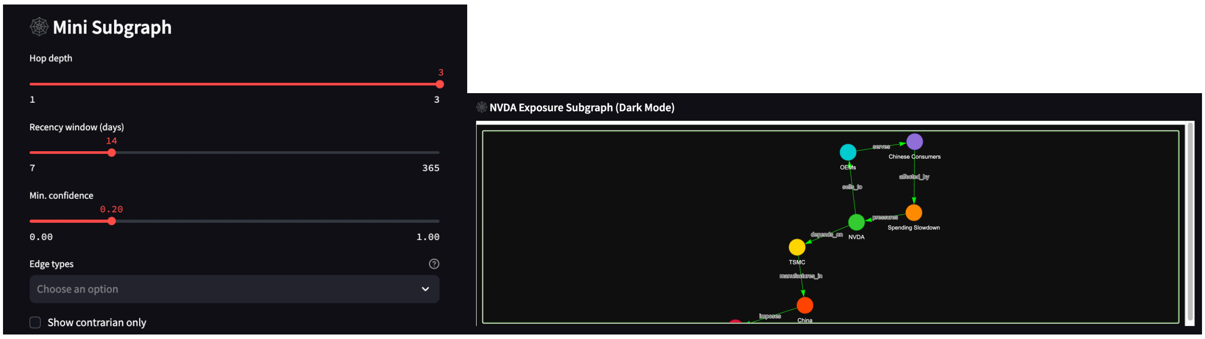
🆕 **What changed since yesterday**

- New claims: c001
- New edges: exposed_to→China Risk
- Confidence Δ: +0.06

Module 3: Mini Subgraph Viewer *(Linked to Per-Ticker View)*

- **Purpose:** Interactive visualization of 1–3 hop relationships around a focal ticker, surfacing links with other entities (e.g. suppliers, KPI drivers, themes, news)
- **Filters for graph:** Hop depth, recency, edge type, confidence, contrarian signals.

- **Impact:** Makes complex interdependencies intuitive—"what's going on and why" in a single, navigable map.



Mini Subgraph

Hop depth: 3 (1 – 3)

Recency window (days): 14 (7 – 365)

Min. confidence: 0.20 (0.00 – 1.00)

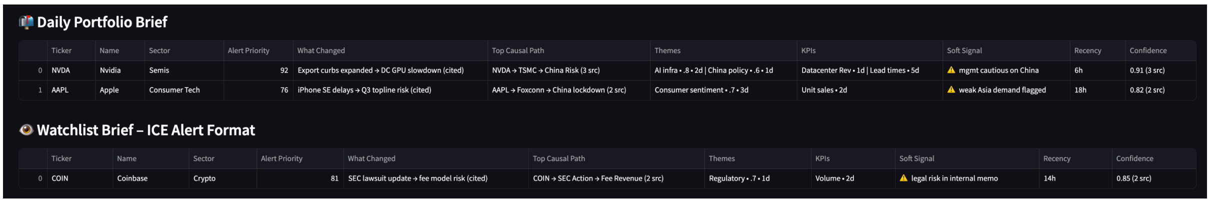
Edge types: Choose an option

□ Show contrarian only

NVDA Exposure Subgraph (Dark Mode)

Module 4: Daily Portfolio / Watchlist Brief Tables

- **Purpose:** High-signal, daily-updating table of emerging risks and opportunities across names in portfolio or watchlist.
- **Columns:** Ticker, "What Changed", top driver, relevant themes/KPIs, soft signals, evidence recency, confidence score.
- **Workflow:** Rapid triage—click on specific ticker to open detailed Per-Ticker View and interactive subgraph.



📫 Daily Portfolio Brief

	Ticker	Name	Sector	Alert Priority	What Changed	Top Causal Path	Themes	KPIs	Soft Signal	Recency	Confidence
0	NVDA	Nvidia	Semis	92	Export curbs expanded → DC GPU slowdown (cited)	NVDA → TSMC → China Risk (3 src)	AI infra • .8 • 2d | China policy • .6 • 1d	Datacenter Rev • 1d | Lead times • 5d	⚠ mgmt cautious on China	6h	0.91 (3 src)
1	AAPL	Apple	Consumer Tech	76	iPhone SE delays → Q3 topline risk (cited)	AAPL → Foxconn → China lockdown (2 src)	Consumer sentiment • .7 • 3d	Unit sales • 2d	⚠ weak Asia demand flagged	18h	0.82 (2 src)

👁 Watchlist Brief – ICE Alert Format

	Ticker	Name	Sector	Alert Priority	What Changed	Top Causal Path	Themes	KPIs	Soft Signal	Recency	Confidence
0	COIN	Coinbase	Crypto	81	SEC lawsuit update → fee model risk (cited)	COIN → SEC Action → Fee Revenue (2 src)	Regulatory • .7 • 1d	Volume • 2d	⚠ legal risk in internal memo	14h	0.85 (2 src)

Optional Visuals

- **Portfolio Sector Exposure (Pie Chart)** – Sectors/Industries breakdown of portfolio holdings.
- **Portfolio Theme Exposure (Bubble Chart)** – Frequency & market value of holdings linked to each theme.

Module 5: Email Module

- **Purpose:** Lightweight delivery of ICE outputs via email—summaries, alerts, or PM digests.
- **Future Extensions**: Auto-push alerts, customizable templates, and periodic strategy reports.

6. Use Cases and Business Value

ICE delivers immediate and tangible business value to the investment firm by accelerating information retrieval and deepening contextual reasoning across the investment workflow. In practical terms, ICE can save analysts and PMs over 30 minutes per day – time reclaimed from manual document parsing, fragmented searches, and redundant analyst queries. More importantly, it enables faster, more confident investment decisions.

In capital markets where even a 5-minute informational lead can generate a multi-basis-point advantage (e.g., +50 bps), this speed translates into more than $1M in incremental alpha for a $200M AUM fund. This is not speculative—recent empirical evidence shows that hedge funds adopting generative AI technologies achieve 3–5% higher annualized abnormal returns compared to peers without such capabilities (Sheng et al., 2024).

While speed and depth of insight are strategic advantages, hallucinated or unverified outputs can be equally costly—potentially erasing millions in capital. For this reason, ICE is not built as an autonomous trading agent. Instead, it operates as a source-grounded reasoning assistant with traceability, transparency and human oversight.

This preserves both trust and compliance alignment, positioning ICE as a force-multiplier for the human investment process rather than a black-box decision-maker.

7. Constraints, Risks, and Feasibility Plan

Building ICE within the constraints of a lean investment fund – namely a 3-month timeline, limited IT resources, lack of database and a single quant analyst as the primary builder – present a range of non-trivial challenges that demand thoughtful planning, rigorous feasibility analysis, target technical choices and strategic scoping to deliver a focused and achievable MVP. To mitigate these risks, we have outlined potential solutions and workarounds for each anticipated challenge and constraint.

Our feasibility assessment addresses:

- **Structural Constraints (Appendix K.1:** Build Constraints & Mitigation Strategies**)** – Time, resource, and budget limits mitigated through a lean, modular stack, open-source reuse, AI-assisted development, and phased delivery.
- **Operational Risks (Appendix K.2:** Key Risks and Mitigation Strategies**)** – Sparse/stale KG, hallucination, conflicting views, siloed retrieval, and token overload managed via lazy graph expansion, evidence grounding, and temporal/confidence scoring.
- **Technical Enablers (Appendix K.3:** Targeted Technical Enablers for Risk Mitigation**)** – Temporal quadruple KGTransformer, entity resolution, intelligent query routing, ColBERT re-ranking, and multi-way recall fusion improve precision and maintain performance on lean infrastructure.

By acknowledging these constraints and identifying potential risks, we can address them by incorporating the corresponding mitigation strategies and technical enablers from the outset, thereby significantly increasing the probability of delivering the MVP successfully. Detailed mappings of constraints, risks, and enablers are provided in **Appendix K** as mentioned before.

8. Future Plans / Extensions

The ICE architecture is intentionally designed for scalable extensibility, enabling it to mature from a focused MVP into the firm's core investment intelligence layer. Post-MVP development (**Appendix L** – Future Extensions) will concentrate on:

- **Deepening reasoning capabilities** – expanding from lean causal paths to richer, multi-hop, explainable narratives.
- **Broadening data coverage** – incorporating structured financials, alternative datasets, and internal workflows.
- **Evolving into intelligent, agent-based operations** – enabling autonomous but auditable analysis loops.

These enhancements are aimed at compounding the firm's proprietary knowledge base, strengthening alpha-generation, and positioning the fund for AI-native operational scale.

9. Conclusion

The competitive gap between large and small investment firms is expanding, driven by unequal access to AI and the growing fragmentation and volume of financial data (Sheng et al., 2024). Resource-rich firms can leverage AI at scale, leaving lean funds at a structural disadvantage in speed, analytical depth, and precision. ICE is designed to close this gap. Its foundation is a compounding, domain-specific knowledge graph that functions as the fund's context-engineering system, embedding state-of-the-art AI models within the firm's unique investment logic. By providing institutional-grade intelligence without the need for enterprise-scale infrastructure or large teams, ICE democratizes advanced AI capabilities—allowing smaller firms to compete directly on insight quality, agility, and accuracy. Over time, it compounds institutional knowledge, accelerates decision-making, and equips lean funds to scale and remain competitive in an AI-driven future.

References

Alpha FMC. (2020, June 2). *From fragmented to organised: What are the latest data strategies in asset management?* Alpha FMC. Retrieved from https://alphafmc.com/blog/2020/06/02/data-strategies-2020/

Sheng, J., Sun, Z., Yang, B., & Zhang, A. L. (2024, August 1). Generative AI and asset management (SSRN Scholarly Paper No. 4786575). SSRN. https://ssrn.com/abstract=4786575

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2024). *From local to global: A Graph RAG approach to query-focused summarization* (arXiv:2404.16130). arXiv. https://doi.org/10.48550/arXiv.2404.16130

Alternative Investment Management Association. (2024, February 1). Getting in pole position: How hedge funds are leveraging Gen AI to get ahead [Press release]. AIMA. https://www.aima.org/article/press-release-getting-in-pole-position-how-hedge-funds-are-leveraging-gen-ai-to-get-ahead.html

Edge, D., Trinh, H., & Larson, J. (2024, November 25). *LazyGraphRAG: Setting a new standard for quality and cost*. Microsoft Research Blog. https://www.microsoft.com/en-us/research/blog/lazygraphrag-setting-a-new-standard-for-quality-and-cost/

Gao, L., Ma, X., Lin, J., & Callan, J. (2023). *Precise zero-shot dense retrieval without relevance labels*. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1762–1777). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.99

Ehtesham, A., Singh, A., Gupta, G. K., & Kumar, S. (2025, May 4). *A survey of agent interoperability protocols: Model Context Protocol (MCP), Agent Communication Protocol (ACP), Agent-to-Agent Protocol (A2A), and Agent Network Protocol (ANP)* (arXiv:2505.02279). arXiv. https://doi.org/10.48550/arXiv.2505.02279

Anthropic. (2024, November 25). *Introducing the Model Context Protocol* [Blog post]. Retrieved from https://www.anthropic.com/news/model-context-protocol

Model Context Protocol. (2025, June 18). *Specification* [Protocol document]. Retrieved from https://modelcontextprotocol.io/specification/2025-06-18

Deloitte. (2024, October 7). 2025 investment management industry outlook. Deloitte Insights. Retrieved from https://www.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-outlooks/investment-management-industry-outlook.html

Appendices

Appendix A.1 – MVP Lazy Graph-RAG Feature Set and Implementation Priorities

MVP	Feature	Description

✅	**Temporal Edge Metadata**	Add timestamped metadata (`first_seen`, `last_seen`, `evidence_count`) to existing edges, e.g.:
		`{"edge_type": "exposed_to",`
		`"source": "META",`
		`"target": "AI Infrastructure",`
		`"last_seen": "2024-07-25",`
		`"first_seen": "2023-12-01",`
		`"evidence_count": 6 }`
		Enables trend tracking, recency filtering, and simple updates (refresh timestamps or increment counts).

✅	**Reversed Traversal**	Enable bidirectional queries without storing reverse edges.
		Examples:
		• “Companies exposed to [Theme]?”:
		`Company ← exposed_to ← Theme`
		• “Holdings dependent on Supplier X?”:
		`Company ← depends_on ← Supplier ← holds ← Portfolio`

✅	**Multi-path Reasoning**	Answer a query via multiple reasoning paths, e.g., “Why is $NVDA at risk from China?”: 1. `NVDA→exposed_to→China risk` 2. `NVDA→depends_on→Supplier/Customer→China risk` 3. `NVDA→has_soft_signal→Event` (“Mgmt cautious on Asia Ops”)
✅	**Edge Confidence Score**	Add confidence/evidence counts per edge, based on: • Mentions • Recency • Source quality Rank or filter reasoning chains by confidence for better precision, explainability, and prioritization.
✅	**Source-Aware Edge Attribution**	Store source document/chunk for every edge to ensure trust, auditability, and analyst adoption.
❌	**Edge Composition Chains (Graph Memory)**	Persist reasoning paths from complex queries (e.g., “China Risk via Supply Chain”) as reusable templates. Builds institutional AI memory and speeds future queries.
❌	**Query-Type-Aware Retrieval Profiles**	Define retrieval templates per query type: • Earnings Q → prioritize `key_driver`, `has_surprise` • Thematic scan → prioritize `exposed_to`, `has_soft_signal` Adjust expansion and summarization accordingly.

Appendix A.2 – Catalog of Typed Edge Templates

#	Edge Name / Type	Purpose	Edge Template	Example Edges	Data Sources	Business Value	Example Queries

1	**Thematic Membership/ Exposure Edge / Narrative Cluster Edge**
	(Directed Thematic Edge)
		Group companies into evolving themes or active market narratives	`[Company] → (part_of_narrative / exposed_to) → [Theme/Narrative]`	`[Nvidia] → (part_of_narrative) → [AI Infrastructure]`
				`[TSMC]→ (exposed_to_theme) → [China Supply Chain Risk]`
					News, analysts reports, earnings call commentary, internal notes, brokers notes, LLM-tagged news, analysts commentary	Group companies into themes; thematic exposure; macro exposure mapping	Which companies benefit from deglobalization trends? What names are exposed to the obesity/GLP-1 drug theme?
							Who is exposed to AI infra? What China-risk names do we hold?

2	**Valuation Driver Edge / Anchor KPI Edge** (Valuation Driver Edge)	Identify core KPIs that drive company valuation/stock performance	`[Company] → (key_driver) → [KPI or Metric]`	`[Meta] → (key_driver) → [DAUs]`	Earnings transcripts, financials, broker models, internal notes	Valuation anchoring and KPI tracking	Which firms are sensitive to ad CPM trends? What KPIs drive Uber's valuation?
3	**Early Signal Edge** / Soft Signal Edge (Early Signal Edge)	Detect early signals before hard data emerges	`[Company] → (has_soft_signal) → [Event/Soft Signal]`	`[AMD] → (has_soft_signal) → [Chip launch delay]`	News, press releases, earnings calls	Qualitative foresight and early signal capture	Which firms are undergoing restructuring? What soft signals precede earnings miss for semis?
4	**Ownership Context Edge** (Portfolio Holding Edge) (Watchlist Edge)	Encode firm-specific positions and exposures	`[Portfolio/Watchlist] → (holds)→ [Stock/Company]`	`[Portfolio A] → (holds)→ [TSLA]`	Internal portfolio data, OMS, watchlists	Position tracking and portfolio exposure analysis	What narratives are we exposed to? Which high-beta names are in our book? What is the sector/industry exposures of our portfolio?
5	**Causal Link Edge** (Directed Causal Edge)	Model impact of events on KPIs/Companies; Forward scenario modelling	`[Event] → (likely_impacts/causes)→ [Company/Metric]`	`[Fed Hike] → (likely_impacts)→ [Mortgage volume]`	Brokers reports, macro notes, news	Event impact modeling and risk propagation	Which companies are impacted if oil prices spike? What metrics are affected by China's deflationary environment?
6	**Business Model / Peer Similarity Edge** (Symmetric Similarity Edge)	Group companies by revenue model, monetization, or operating logic	`[Company A] →(similar business model) → [Company B]`	`[Shopify] →(similar business model) → [BigCommerce]`	Company descriptions (10-K), Yahoo, broker notes	Identify better peer comps; thematic clusters	Who has similar biz model to Roblox? Peers for Snowflake?

7	**Revenue Composition Edge**
	(Segment exposure edge)
		Break down company revenue by geography, product line, customer	`[Company] →(generates_revenue_from) → [Segment/Product/Region/Customer]`	`[Apple] →(generates_revenue_from) → [China (35%)]`
				`[Microsoft] → (generates_revenue_from) → [Azure]`
					Segment data in 10-K, FactSet, broker models	Model geopolitical exposure or revenue dependency	Which firms earn >50% in EU? Cloud revenue breakdown for MSFT?

8	**Functional Exposure Edge/** Operating Domain Edge
	(Directed Product/Role Map)
		Map companies to what problems they solve or tools they are	`[Company] → (powers/enables/supports) → [Function/Use Case/Problem Domain]`	`[Snowflake] → (powers) → [Enterprise Data Infra]`
				`[Palantir] → (enables) → [Government Surveillance Analytics]`
					Product docs, broker reports, company websites	Use-case thematic baskets and overlap avoidance	Which companies offer ML model serving? DevOps infra providers?

9	**Value Chain Edge /** Supplier or Customer or Platform Dependency
	(Directed Relational Edge)
		Model upstream/downstream relationships and dependencies between firms—e.g. supplier, customer, platform	`[Supplier] → (supplies) → [Customer]`
			`[Platform] → (enables) → [User/Client]`
			`[Company A] → (depends_on) → [Company B]`
				`[TSMC] → (supplies) → [Apple]`
				`[Shopify] → (depends_on) → [Stripe]`
					10-K risk disclosures, broker reports, product docs, S&P CapIQ, news	Enables knock-on impact reasoning, supplier risk hedging, thematic propagation;
						Supply chain risk analysis, disruption modeling, input-output propagation
							Which companies depend on Nvidia GPUs?
							Who are Apple's key suppliers?
							What are the downstream customers of TSMC?

10	**Sector & Industry Classification Edge** (Directed Taxonomic Edge)	Categorize firms into canonical or custom industry/sector clusters	`[Company] → (classified_as) → [Sector/Industry]`	`[ASML] → (classified_as) → [Semiconductor Equipment][Uber] → (classified_as) → [Mobility Platform]`	GICS codes (public), FactSet/CapIQ, analyst commentary, internal mapping	Enables sector exposure analysis, comp grouping, risk aggregation; Portfolio construction, sector risk attribution, filtering & screening;	Which companies in our book are software infra? Show all holdings classified under semiconductors Which firms are in fintech but not classified by GICS?

Appendix A.3 – Limitations of Generic AI Tools and how ICE addresses it

#	Edge Name / Type	What the question really requires	Why generic LLM tools fail	How ICE addresses it
1	**Thematic Membership / Narrative Cluster Edge** `Company → part_of_narrative / exposed_to → Theme`	Accurate, **fund-specific** mapping of companies to evolving narratives /themes, reflecting internal taxonomy & overrides, scoped to portfolio.	Public LLMs guess from generic news (maybe outdated); no notion of your custom theme definitions or analyst overrides; cannot overlay portfolio holdings.	Internal theme taxonomy with analyst curation; timestamped edges; portfolio overlay filter in every query.
2	**Valuation Driver / Anchor KPI Edge** `Company → key_driver → KPI`	Pinpoint **one or two anchor KPIs** agreed by analysts as stock movers, tied to forecast model columns.	Lists many generic metrics; unaware of fund-agreed anchor KPIs; no link to internal models.	Curated KPI dictionary linked to forecast model schema; edges carry evidence source and recency.
3	**Early Signal Edge / Soft Signal Edge** `Company → has_soft_signal → Event`	Capture faint, pre-consensus signals (exec exits, product delays) with timestamps, source, and confidence.	Generic LLMs surface only widely reported events; miss local/niche data; no freshness or confidence scoring.	Multi-source ingestion (local news, press releases, niche feeds); weak-signal classifiers; edges store `confidence` + temporal metadata.
4	**Ownership Context Edge / Portfolio Holding Edge** `Portfolio → holds → Company`	Ensure all answers reflect **our current book & watchlist**.	Cannot access internal portfolio positions which is highly confidential.	Direct OMS/CSV ingestion; edges refreshed daily; query filters for portfolio scope.
5	**Causal Link Edge** `Event → likely_impacts → Company/Metric`	Analyst-approved **cause-effect priors** for forward scenario propagation.	LLM may hallucinate counterintuitive links; no guardrails or link scoring.	Curated causal edges with documented rationale; multi-path scoring ranks most plausible impacts.
6	**Business Model / Peer Similarity Edge** `Company ↔similar_business_model ↔ Company`	Peers grouped by **revenue mechanics**, not just GICS/SIC.	Defaults to mega-cap peers; no access to your embeddings or factor tags.	Embedding + rules capture monetization model; analyst seed lists refine peer groups.
7	**Revenue Composition Edge** `Company → generates_revenue_from → Segment/Region/Customer`	Accurate, dated % revenue splits by product/geography/customer.	Uses stale or inconsistent data; fiscal/calendar confusion.	Latest filing parsing; % and period stored on edges; source attribution included.
8	**Functional Exposure / Domain Edge** `Company → powers/enables /supports → Function`	Map vendors to **specific use cases** per your taxonomy.	Over-generalizes from marketing copy; misses nuanced roles.	LLM+taxonomy tagging; analyst QA; multi-label edges with provenance.
9	**Value Chain Edge / Supplier-Customer / Platform Dependency** `Supplier → supplies → Customer`	Timestamped, validated upstream/downstream dependencies.	Guesses common pairs; misses niche B2B links; no temporal accuracy.	Multi-source extraction (10-K, CapIQ, news) with entity resolution, timestamps, and evidence count.
10	**Sector & Industry Classification Edge** `Company → classified_as → Sector/Industry`	Your **bespoke taxonomy** with overrides; consistent tagging for new tickers.	Defaults to public GICS/NAICS; can't honor internal overrides.	Internal mapping table + analyst override process; edges stored in KG for reuse.

Appendix A.4 – Example user Queries and Corresponding Graph Edge Activations

#	User's Query	Delivery Method	Why it matters?	Edges Used	Data Sources	Business Value	Failure Risk / Build Complexity
1	What names are exposed to the obesity/GLP-1 drug theme? / What is driving the performance of Uber recently?	Chat interface	Identifies specific companies impacted by a high-momentum market narrative; supports thematic trades and narrative-driven alpha generation.	1-hop: `Company → exposed_to → Theme / News; 2-hop: Theme / News → likely_impacts → KPI/Price`	Newswire feeds, earnings transcripts, analyst reports, public filings	Alpha generation from thematic plays	Low – data is widely available and edges are simple
2	Theme exposure of our portfolio **or** What companies in our portfolio are exposed to China Risk narrative?	Streamlit Dashboard/Table	Provides PM with a high-level view of portfolio's thematic exposures, enabling proactive positioning and risk management.	2-hop: `Portfolio → holds → Company → exposed_to → Theme`	Portfolio holdings CSV, news feeds, internal research notes	Portfolio risk management & thematic positioning	Low – edges are straightforward and data mapping is simple
2b	Sector exposure for our portfolio	Streamlit Dashboard/Table	Supports investor relations and compliance reporting; enables clear communication of sector allocation.	2-hop: `Portfolio → holds → Company → classified_as → Sector`	Portfolio holdings CSV, industry classification datasets (e.g., GICS)	Investor confidence & compliance	Very Low – industry classification data is standardized
3	What KPI drives $Meta share performance or valuation?	Chat interface	Identifies core performance drivers for a company; supports KPI-linked monitoring and predictive modeling.	1-hop: `Company → key_driver → KPI`	Earnings transcripts, KPI datasets, broker research	Alpha through KPI-linked predictive trading	Medium – KPI extraction requires semantic accuracy
4	Is TSMC exposed to China risk? / What is the theme driving the performance of TSMC recently?	Chat interface	Detects company-specific geopolitical or macro risk exposures; informs risk-adjusted positioning.	1-hop: `Company → exposed_to → Theme`	Newswire feeds, geopolitical risk datasets, transcripts	Downside risk mitigation	Low – edges are simple; risk depends on theme tagging accuracy
5	How might China risk impact KPIs across our books?	Chat interface	Links macro risk to operational and financial metrics; enables portfolio stress-testing.	3-hop: `Portfolio → holds → Company → exposed_to → Theme → likely_impacts → KPI`	Portfolio holdings CSV, news feeds, KPI datasets	Forward-looking risk management	Medium – requires multi-hop reasoning and KPI linkage accuracy
6	Which companies are exposed to [Theme]?	Chat interface	Supports screening and idea generation across coverage universe.	1-hop (reversed): `Theme → exposed_to → Company`	Newswire feeds, thematic datasets	Alpha generation via thematic screening	Low – straightforward query reversal

Appendix B – Component-level Architecture & Pipeline

	Components	Description

1	**Query Input, Understanding and Framing**
	
	Hypothetical Document Embedding (HyDE)
		User Query Intake
		Natural language query from:
		- A PM/analyst (e.g. "What is NVDA at risk from China trade?")
		- Or a scheduled batch process trigger (e.g. daily ticker loop of portfolio holding) from portfolio.csv or watchlist.csv.
		
		Query Understanding
		Parse the intent and structure of the user's query.
		Use a small LLM call (or regex+NER fallback) to extract:
		- Entities: tickers, themes, KPIs
		- query_type: one of a fixed set (e.g. earnings_reasoning, theme_exposure, others)
		- intent_flags: e.g. requires causal reasoning? Soft signal scan?
		
		Output (*query_entity_intent* JSON):
		`{`
		`  "query": "Why is NVDA at risk from China trade?",`
		`  "entities": {`
		`    "tickers": ["NVDA"],`
		`    "themes": ["China Risk"]`
		`  },`
		`  "query_type": "THEME_RISK_SCAN"`
		`}`
		Purpose: Used by downstream processes, Hybrid Retriever, Lazy Graph Expansion, Summariser agent
		
		Query Plan Construction
		Build a structured retrieval plan to guide downstream LazyGraphRAG.
		Based on query_type, define:
		- What edge types to expand
		- Which documents to be included?
		- Whether reverse traversal is allowed
		- Graph hop depth (e.g. 1-, 2- or 3-hop)
		- Any metadata filters
		
		Output (*retrieval_plan* JSON), to use for Lazy Graph Expansion:
		`{`
		`  "expand_edges": ["exposed_to", "has_soft_signal"],`
		`  "reverse_lookup": true,`
		`  "graph_hops": 2,`

		"prioritize_recent": true,
		"prioritize_confidence": true,
		"document_sources": [“earnings_transcripts”, “broker_notes”, “high-relevance_news”],
		}
		Purpose: JSON gives explicit instructions to the graph traversal and RAG engine on how to guide reasoning expansion:
		- Decide which typed edges to follow (e.g., exposed_to)
		- Whether to allow reverse edge traversal
		- How deep to expand (e.g. 1-, 2- or 3-hop reasoning path)
		- Whether to filter/score nodes/paths by metadata (confidence, recency)
		HyDE Embedding Construction
		Use a strong LLM to generate a hypothetical “ideal answer” for better semantic recall during retrieval
		Prompt template:
		`“You are a financial analyst. Based on the query ‘{query}’, what is the most complete, plausible answer grounded in known financial reasoning?”`
		Example output (HyDE text):
		`“NVDA may be exposed to China trade risks due to its reliance on TSMC and exposure to AI hardware demand in Asia…”`
		Generate two embeddings:
		1. Original user’s query
		2. HyDE “ideal response” output
		Output (*embeddings* JSON):
		{
		"embedding_raw_query": [0.31, -0.12, ...],
		"embedding_hyde_answer": [0.42, 0.03, ...],
		"hyde_text": "NVDA may be exposed to..."
		}
		Purpose: Provides a semantic representation of the query (and HyDE embedding) for Hybrid Retriever, Re-ranker (optional) and RAG Evaluation.

2.	**Hybrid Retrieval** **(Retrieval & Expansion)**	Retrieves the most relevant documents (and document metadata) for a given query using both **semantic similarity** and **keyword precision**. **Semantic Vector Retrieval** Perform vector search using both (separately): 1. embedding_raw_query 2. embedding_hyde_answer Each returns top_k_semantic docs (e.g. 15-20 docs) **Lexical Keyword Retrieval** Perform keyword search for: • Extracted tickers (“$NVDA”)

- Themes or KPI mentions (“China Risk”, “TSMC”)

Expand with synonyms / sector tags (optional)

Return top_k_lexical docs (e.g. 15-20)

Note: lexical retrieval is performed using key words found in the original user’s query only, and not from the HyDE “ideal answer”.

Contrarian Retrieval (Optional enhancement)

Use programmatically altered queries to retrieve counter-narratives. Generate a contrarian prompt, run lexical or semantic search on this contrarian query, prioritizing results with negative terms, merge results into main document pool (tag these as “retrieval mode”: ‘contrarian”)

e.g. `“What are the risks of NVDA exposure to China?"` or `"What could go wrong with NVDA?"`

Merge + Re-rank

Combine results from both searches (2 sets from semantic retrieval + 1 set from lexical retrieval)

- Deduplicate by doc ID.
- Score by:
  - Source type match (from *retrieval_plan.document_sources*)
  - Recency
  - Exact keyword match
  - Vector similarity (optional)

Return top_k (e.g. 20-40) documents + metadata + tags

Output (*retrieved_document_package* JSON):

```
{
  "retrieved_documents": [
    {
      "doc_id": "doc123",
      "title": "NVDA Q2 Earnings Call Transcript",
      "source_type": "earnings_transcript",
      "date": "2024-07-26",
      "score": 0.92,
      "entities": ["NVDA", "TSMC", "China Risk"],
      "content_snippet": "Mgmt flagged supply chain risk tied to Taiwan operations..."
    },
    {
      "doc_id": "doc456",
      "title": "Broker Note: AI Infra Outlook",
      "source_type": "broker_note",
      "score": 0.87,
      "date": "2024-07-24",
      "entities": ["AI Infra", "NVDA"],
      "content_snippet": "AI chip demand could face headwinds if China sanctions escalate..."
```

		`    }`
		`  ],`
		`  "total_candidates": 38`
		`}`
		Optional: explicitly include some contrarian retrieval – for instance, if all top documents seem to be saying one thing (e.g. analysts bullish on stock”), have the system also search for the opposite sentiment or risk factors (perhaps by programmatically tweaking the query to include terms like “downside” or “risk”. This can be done via keyword search on negative terms or by using another round of HyDE (“What could go wrong with X?” as a prompt).

3.	**Lazy Graph Expansion**
	(Subgraph Construction)
		To build a query-focused, typed, directed subgraph from retrieved documents, aligned with the query intent — and expand it just far enough to support high-quality reasoning.
		<u>Extract Entities & Candidate Edges from Documents</u>
		For each retrieved document (from the retrieval step):
		• Use LLM or rule-based extraction to find:
		- Nodes: Tickers, KPIs, Themes, Events, Companies
		- Candidate typed edges:
		- Company → exposed_to → Theme
		- Company → has_soft_signal → Event
		- Company → key_driver → KPI
		- Company → depends_on → Supplier
		• Each extracted edge is annotated with:
		- confidence_score
		- source_doc_id
		- date_last_seen
		- evidence_snippet
		Note: Only extract edge types listed in *retrieval_plan.expand_edges*
		<u>Build Local Subgraph</u>
		Construct a directed graph (NetworkX DiGraph) with:
		• Nodes: Companies, KPIs, Themes, Events
		• Edges: From extracted data, with types and metadata
		Example of Edge Object (JSON):
		`{`
		`  "source": "NVDA",`
		`  "target": "China Risk",`
		`  "edge_type": "exposed_to",`
		`  "confidence": 0.87,`
		`  "source_doc": "doc123",`
		`  "last_seen": "2024-07-26",`
		`  "snippet": "Mgmt noted rising concerns over US-China trade..."`
		`}`
		<u>Expand Neighbor Nodes via Typed Edge Traversal</u>

From each root entity (e.g., “$NVDA”), traverse:

- Only edges allowed in *expand_edges*
- Up to *graph_hops* depth (e.g. 1-3 hops)
- Reverse traversal allowed if *reverse_lookup=true*

This creates a just-in-time local reasoning graph tailored to the query.

Filter Expanded Subgraph

Apply filters (from retrieval plan):

- Drop edges with confidence_score < 0.6
- Drop edges older than 30 days if prioritise_recent=True
- Rank or tag edges based on:
    - Source priority
    - Frequency of edge occurrence
    - Sentiment or uncertainty markers (optional)

Output (*expanded_subgraph* JSON):

```
{
  "nodes": ["NVDA", "China Risk", "TSMC", "AI Hardware"],
  "edges": [
    {
      "source": "NVDA",
      "target": "China Risk",
      "edge_type": "exposed_to",
      "confidence": 0.87,
      "last_seen": "2024-07-26",
      "source_doc": "doc123"
    },
    {
      "source": "NVDA",
      "target": "TSMC",
      "edge_type": "depends_on",
      "confidence": 0.9,
      "last_seen": "2024-07-25",
      "source_doc": "doc456"
    },
    {
      "source": "TSMC",
      "target": "China Risk",
      "edge_type": "exposed_to",
      "confidence": 0.8,
      "last_seen": "2024-07-20",
      "source_doc": "doc789"
    }
  ]
```

		}
		This lazy retrieval of neighbors enriches the context with relevant but not obviously similar info that might be missed by the vector search and lexical keyword search. Notably, this is done without precomputing all possible links – we simply use metadata like “Company: X” or “Topic: supply chain” as edges. This yields an expanded document set that includes the initial hits plus any additional documents linked via the graph relationships (common entities, citations, etc.).
		The result of this step is one or more clusters or “communities” of documents centered around the query’s key themes. Each cluster might represent a distinct narrative thread or causal pathway relevant to the query.
		These clusters provide semantically and structurally coherent “reasoning bundles” that can be independently summarized or scored. This design allows the system to reason over multi-hop, multi-perspective angles without relying on pure text similarity alone — enabling the discovery of latent risks, counterfactuals, and hidden dependencies. Critically, because this expansion is driven by typed, directed, and query-relevant edges, the process remains lightweight, interpretable, and tailored to the user’s intent.
		Operationally, consolidate outputs into a single bundled JSON:
		`{`
		`  "retrieved_documents": [...], // enriched doc pool`
		`  "expanded_subgraph": {...},  // typed, confident edges`
		`  "query_type": "THEME_RISK_SCAN",`
		`  "root_entities": ["NVDA", "China Risk"]`
		`}`

4.	**Claim Extraction & Evidence Tagging**	Extract investment-relevant facts from each retrieved document, and tag each fact with its source, associated entities, and candidate typed edge(s).
		This is to create structured, source-traceable claims that can later support graph-based reasoning.
		For each retrieved document:
		• Use LLM or NLP pipeline to extract:
		- Facts (metrics, quotes, forward guidance, sentiment, risks)
		- Associated metadata:
		o date, source_type, confidence
		o Entities involved
		o Optional sentiment score or uncertainty flag
		• Tag each claim with one or more linked_edges from the expanded_subgraph.
		Output (*claim_package* JSON):
		`{`
		`  "claims": [`
		`    {`
		`      "claim_id": "c001",`
		`      "text": "Mgmt warned of potential delays in China shipments",`
		`      "linked_edge": ["NVDA", "exposed_to", "China Risk"],`
		`      "confidence": 0.87,`

		"doc_id": "doc123", "source_type": "transcript", "date": "2024-07-26", "risk_signal": true }, ...] }

5.	**Path Enumeration & Claim Alignment**	Enumerate all valid reasoning chains (1–3 hops) through the subgraph, and match each path with the claims that support it.
		To derive a causal logic frame (reasoning paths), and attach supporting evidence (claims) to each.
		• Traverse the expanded_subgraph using root_entities.
		• Enumerate all unique, typed 1-3 hop paths.
		• For each path, match claims using:
		- Exact or partial match on linked_edge
		- Shared entities or KG node overlap
		Output (*reasoning_paths* JSON):
		{
		"paths": [
		{
		"path_id": "p001",
		"nodes": ["NVDA", "depends_on", "TSMC", "exposed_to", "China Risk"],
		"depth": 3,
		"supporting_claims": ["c001", "c002"]
		},
		...
		]
		}

6.	Claim Scoring	Score all extracted claims based on relevance to the user's query, metadata quality, and salience signals. To identify globally high-value insights across the entire claim pool, regardless of path. For each **claim**: • Compute semantic similarity to: - Embedding_raw_query - Embedding_hyde_answer • Combined with: - Recency - Source quality - Confidence - Risk keywords / uncertainty signals.

		Output a normalised relevance score ∈ [0, 1]
		Output (*ranked_claims* JSON):
		{
		"claim_scores": [
		{
		"claim_id": "c001",
		"score": 0.91
		},
		{
		"claim_id": "c002",
		"score": 0.85
		}
		]
		}

7.	**Path Scoring**	Compute an overall reasoning quality score for each path by combining its structural soundness with the relevance of its supporting claims.
		To prioritize the most plausible and evidence-backed causal explanations for downstream synthesis.
		For each path:
		• Compute:
		- Avg/min edge confidence
		- Most recent edge or claim
		- Highest source quality (among claims)
		- Avg relevance score of claims (from claim_ranker)
		- Query alignment (do path nodes match query entities?)
		- Path depth penalty (shorter preferred)
		- Optional: presence of risk words in claims
		• Compute aggregate path_score
		Output (*ranked_paths* JSON):
		{
		"ranked_paths": [
		{
		"path_id": "p001",
		"score": 0.92,
		"supporting_claims": ["c001", "c002"],
		"coverage_tags": ["Supplier", "Theme"]
		}
		]
		}

8.	**Narrative Trace Construction**	“Convert each top-ranked reasoning path into a natural-language, source-backed causal explanation.”
		To provide human-readable reasoning units the LLM can synthesize without hallucination.
		For each top-ranked path:
		• Retrieve the highest-ranked claims linked to that path
		• Assemble a short narrative:
		- Follow the chain of nodes + edges
		- Use factual language from claims
		- Cite sources, dates
		- Preserve tone (e.g., cautious optimism)
		Output (*narrative_traces* JSON):
		`{`
		`  "traces": [`
		`    {`
		`      "path_id": "p001",`
		`      "narrative": "NVDA is exposed to China Risk via its supplier TSMC, which recently flagged production concerns due to Taiwan logistics constraints (Q2 call, July 26).",`
		`      "sources": ["doc123", "doc456"]`
		`    }`
		`  ]`
		`}`

9.	**Final Selection & Packaging**	Select the top-N reasoning paths and claims, enforce coverage diversity, and output the final LLM-ready payload.
		To prepare a compact, high-quality knowledge pack for synthesis or reporting.
		• Select top K paths with highest path_score
		• Ensure diverse tags (e.g. theme + supplier + KPI)
		• Select top M individual claims
		• Package everything into a final JSON bundle
		Output (*llm_context_package* JSON):
		`{`
		`  "query": "Why is NVDA at risk from China trade?",`
		`  "ranked_reasoning_paths": [...],    // includes path, narrative, sources`
		`  "top_claims": [...],                 // 1-2 line alerts`
		`  "query_embeddings": {...},           // optional for MCP`
		`  "query_plan": {...}`
		`}`

10.	**LLM Answer Synthesis/ Generation**	Generate the final answer by prompting a high-quality LLM with a curated bundle of reasoning paths and supporting claims. Using the llm_context_package JSON, generate a clear, actionable, investment-grade response suitable for e.g.: • Portfolio alerts

- Analyst Q&A
- Research summaries
- Decision support

Optionally, generates:

- TL;DR
- Risk flags
- Citations
- Sources

The LLM (e.g. GPT-3.5/4 via API or an open-source model like Llama2) will compose a final answer, using the provided facts to ground its response.

Return formatted answer with citations, trace paths, edge, source attributes.

Prompt Template:

```
"You are a financial analyst generating a concise reasoning memo based on structured reasoning data.

Query: {{query}}

Here are 2–4 causal reasoning paths explaining the query. Each is supported by evidence from transcripts, broker notes, and news.

Please:

1. Summarize the key reasoning threads, including causal links and source-backed statements.

2. Retain all dates, KPIs, and uncertainty language (e.g., "might", "warned", "expects").

3. Optionally return a 2-line TL;DR if requested.

4. Use clear, investment-grade language.


---


Paths:

{{ranked_reasoning_paths[narrative + sources]}}

Top extracted claims:

{{top_claims}}

Answer:

"
```

Graph-augmented retrieval ensures even information that wasn’t obviously related via pure semantics is included if it’s relevant by relationships.

In a batch use-case, these answers might be compiled into a daily report or alert for the investment team.

Output (*final_response* JSON):

```
{

  "final_answer": "NVDA is at risk from China trade primarily through its dependence on TSMC, which flagged geopolitical yield challenges in Taiwan. Management, during the Q2 call on July 26, acknowledged uncertainty around export logistics. In addition, China represents a significant demand driver for NVDA's AI chips, and escalating sanctions could dampen that growth.
⚠️ TL;DR: NVDA’s supply chain and AI chip sales face
exposure to China-related risks via TSMC and export
limitations.",
  "citations": ["doc123", "doc456", "doc789"]
}
```

12. **User Interface / Output Renderer**

Streamlit UI Design:

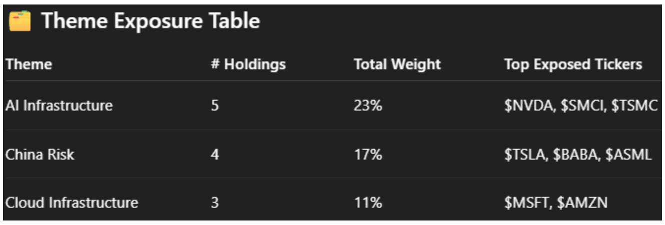
The UI for ICE is deliberately minimalist, designed to maximize usability while minimizing build effort and deployment complexity. At its core is a single-page Streamlit interface that offers a clean “Ask a question” box for ad-hoc equity queries—returning structured answers with source citations and a preview of the reasoning graph (e.g., how Company A’s outlook links to macro, sector, or peer-level signals). The UI also displays the daily markdown brief for all watch-list and portfolio tickers, allowing PMs and analysts to skim changes and drill down only when needed. This lightweight interface runs locally or via simple web-hosting tools (e.g., Streamlit Cloud)mand can optionally push notifications or summary highlights to the email. Crucially, it maintains full traceability—each answer shows which documents and knowledge paths were used, building trust without sacrificing speed. The UI prioritizes signal over noise: clean answers, fast load, zero distractions—perfect for high-pressure, time-boxed morning workflows.

An interactive subgraph viewer for the selected ticker

Works from either the query box input or selection from brief table (portfolio or watchlist)

Lightweight, using network for graph logic and pvis for interactive rendering.

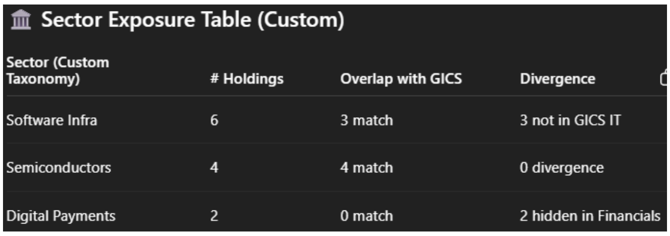
1. Real-Time Alert Feed (Table/List)
   - Timestamp, Ticker, Event Summary, Source, Confidence Score
2. Context Narrative Viewer
   - Click event to expand narrative with sources and timeline overlays
3. Knowledge Graph Visualiser
   - Interactive graph of current portfolio + connected entities (e.g. customers/suppliers/themes/news co-mention/events)
4. Search Bar
   - Ask: “Why did ABC drop today?” or “What’s the latest on XYZ’s supplier risk?”
5. Simple Script (or Streamlit dashboard) that outputs THEME Exposure



📁 Theme Exposure Table

Theme	# Holdings	Total Weight	Top Exposed Tickers
AI Infrastructure	5	23%	$NVDA, $SMCI, $TSMC
China Risk	4	17%	$TSLA, $BABA, $ASML
Cloud Infrastructure	3	11%	$MSFT, $AMZN

6. SECTOR Exposure



🏛 Sector Exposure Table (Custom)

Sector (Custom Taxonomy)	# Holdings	Overlap with GICS	Divergence
Software Infra	6	3 match	3 not in GICS IT
Semiconductors	4	4 match	0 divergence
Digital Payments	2	0 match	2 hidden in Financials

Alternatives: CLI, Email, PDF.

Appendix C – Detailed Tech Stack Table

#	Component	Description
1	**Data Sources & Ingestion**	Unified Python ingestion pipeline (LangChain loaders + PyMuPDF/pdfplumber for PDFs + spaCy NER). Parses earnings transcripts, filings, internal notes, and broker emails. Auto-tags with metadata: tickers, themes, KPIs, dates. **Suggestion:** Add basic ingestion validation (e.g., entity presence, doc length) to prevent garbage-in retrieval.
2	**Vector Store (Knowledge Base)**	**MVP:** FAISS in-memory for zero-infra ultra-fast retrieval. **Upgrade path:** Chroma, Qdrant, or Weaviate for persistent or hybrid search. All support metadata-based filtering for graph expansion.
3	**Embedding Model**	**Default:** Sentence-BERT (all-MiniLM) for local, zero-cost embeddings. **Upgrade:** OpenAI text-embedding-3-small for noisy or domain-specific text. **Tip:** Keep embedding model consistent across retrieval, reranking, and graph node similarity to avoid semantic drift.
4	**HyDE Generator**	Low-cost LLM (e.g., GPT-3.5-turbo, Mixtral) generates hypothetical ideal answers for ambiguous queries, improving recall. One-shot inference; negligible cost.
5	**Hybrid Retriever**	Combines: (1) vector similarity search, (2) keyword/inverted index search, (3) HyDE embeddings. Merges, deduplicates, and re-ranks results by recency, score, and metadata match. **Lean option:** Custom Python merge/rerank; **Future:** Use Weaviate/Milvus native hybrid query.
6	**Lazy Graph Traversal (On-Demand KG)**	On-the-fly subgraph construction from retrieved docs using entity metadata. **MVP:** NetworkX DiGraph; **Upgrade:** Neo4j/AWS GraphRAG Toolkit if edge density >1M. Stores confidence, temporal metadata, and source citations for each edge.
7	**Intermediate Summarization & NLP**	LangChain/LlamaIndex extracts 1–2 query-relevant bullets per doc cluster. **Default:** GPT-3.5-turbo; **Reserve:** GPT-4 for mission-critical; **Fallback:** LLaMA-3/Mistral local. Bullets ranked by cosine similarity or LLM reranker to cut prompt cost.
8	**Generative Model (Answer Synthesis)**	GPT-3.5-turbo for most cases; GPT-4o for complex reasoning. Local fallback with quantized LLaMA-3-8B-Q4. Prompts include citations, reasoning paths, MCP JSON structuring.
9	**Knowledge Graph Store**	MVP: NetworkX pickle for persistence; optional Neo4j/Graphiti for visualization and richer queries. Edges typed and timestamped with confidence scores.
10	**Orchestration & Scheduling**	Python + cron for batch jobs. Optional LangChain RunnableGraph for modular pipelines and retry logic. **Suggestion:** Add simple logging + failure alerts for ingestion and retrieval stages.
11	**Front-end / Interface**	MVP: Streamlit or Jupyter notebook; batch mode via CLI/email. Later: Streamlit dashboard or minimal Flask app with query box, per-ticker views, and KG explorer.

Appendix D – Data Infrastructure table

Phase	Data Type	Description	Source(s)	Cost	Notes / Impact
MVP	**SEC Filings**	Core financial disclosures incl. 10-K, 10-Q, 8-K, DEF 14A, insider trades. Structured + unstructured text for KPI anchors & governance edges.	SEC EDGAR API, sec-edgar-downloader, XBRL feeds	Free	Rich, reliable signal; directly parsed for entity/KPI extraction
MVP	**Earnings Transcripts & Press Releases**	Official announcements, forward guidance, Q&A; tone and sentiment indicators.	Company IR sites, PRNewswire, GlobeNewswire, AlphaSense	Free → $$$	Soft-signal edge (guidance, delays). Start free, add premium feeds later for coverage
Phase 2	**Investor Presentations**	Visual summaries of earnings, strategy, KPIs.	Company IR sites	Free	Supports OCR/PDF parsing for context enrichment
MVP	**Proxy & Special Situation Filings**	Proxy statements, corporate action filings, M&A announcements.	SEC EDGAR	Free	Detects control shifts, governance changes, and shareholder activism
MVP	**News & Events**	Financial news feeds, press releases, trending tickers, event alerts.	Yahoo Finance, Google Finance, Reuters, RSS feeds, NewsAPI, Finnhub	Free → $	Captures "what changed" triggers; use free feeds + light scraping initially.
Phase 2	**Supply Chain Data**	Supplier/customer mapping, shipping, trade flows, industry stats.	Filings, FactSet Revere, Panjiva, ImportGenius, WSTS, EIA	$ → $$$	High-value for risk propagation analysis; integrate selectively
MVP	**Internal Firm Data**	Proprietary research notes, portfolio holdings, trade logs, internal emails.	Firm systems, CSV/XLS exports	Free	Defensible edge; prioritize structured (portfolio) before unstructured
MVP	**Sell-side Analyst Reports**	Broker notes with ratings, targets, qualitative insights.	Broker emails, Bloomberg, FactSet, ResearchPool	Free → $$$	Valuable for KPI/theme triangulation
Phase 2	**Third-party Research**	Independent analyst and investor reports.	SeekingAlpha, SmartKarma, Tegus, SumZero, Substack	$$$	Adds long-tail insight; post-MVP.
Phase 2	**Market & Pricing Data**	Price, volume, consensus estimates, corporate actions.	Yahoo Finance API, IEX Cloud, FactSet	Free → $$$	Contextual grounding; not core for MVP RAG
Phase 2	**Financial Statement Data**	Standardized fundamentals for KPI validation.	Yahoo Finance API, Alpha Vantage, FMP, SEC	$–$$	Improves numerical accuracy and KPI tracking
Phase 2	**Alt-Data**	Web traffic, hiring trends, social sentiment, geospatial data.	SimilarWeb, LinkedIn, SafeGraph, Thinknum	$–$$$	For differentiated edge once ROI proven
MVP	**NLP/ML Corpora**	Pre-trained finance-specific models/datasets.	FinBERT, Financial PhraseBank, FiQA, Reuters archives	Free	Bootstrap domain adaptation & improve model relevance
Phase 2	**Web-search AI Tools**	External AI search assistants.	Perplexity.AI	$$ → $$$	Use sparingly; verify outputs for accuracy
Phase 2	**Regulatory & Gov Data**	Regulatory press releases, macroeconomic indicators.	SEC, FDA, FCC, EPA, FRED API	Free	Timely policy impact signals
	Crowdsourced & Sentiment Data	Earnings estimates, hedge fund holdings, investor sentiment.	Estimize, WhaleWisdom, AAII, Kalshi	$–$$	Useful for sentiment-weighted thesis building

Appendix E – Financial Corpora Roles, Safeguards and Examples

Roles and safeguards of Financial Corpora

#	Role	Description	Safeguard
1	Bootstrapping High-Confidence Edge Patterns	Train/finetune NER + RE pipelines on labeled corpora (e.g., FiQA, EDGAR) for common edge types such as Company → key_driver → KPI.	Accept only if match confidence > threshold **and** supported by retrieved snippet.
2	Pre-filtering & Candidate Generation	Use corpus-trained embeddings to identify high-likelihood snippets for Lazy Graph Expansion.	LLM extracts facts **only** from retrieved docs; no prompt hallucination.
3	Typed Edge Validation via Distant Supervision	Cross-check proposed edges against corpus examples to ensure schema consistency and prevent "semantic drift."	Schema-locked; no new edge types without approval.
4	Graph Densification via Semantic Matching	Suggest weakly supported edges based on similarity to corpus patterns (e.g., supplier risk propagation).	Mark as "candidate edge" until confirmed by further evidence.

Example Edge Types & Corpora

Edge Type	Corpus Role	Example Corpus
Company → key_driver → KPI	Sentence classification + RE	FiQA, FinBERT
Company → exposed_to → Theme	Named entity + similarity match	EDGAR Corpus
Company → has_soft_signal → Event	Tone/sentiment + QA	Financial PhraseBank
Supplier → exposed_to → Theme	Propagation patterns	EDGAR, broker notes
Portfolio → holds → Company	Not corpus-derived (internal)	N/A

Appendix F – System Design: Lazy Graph Expansion with Web Search Tool

Step	Description
1. Trigger & Edge Gap Detection	During traversal, detect a missing-but-critical edge for the current query (e.g., Company → exposed_to → Theme). Only trigger web search if the path is required by the retrieval plan and confidence can't be met from local stores.
2. Structured Query Decomposition	Convert the gap into a typed question constrained by the edge template (e.g., "Is **TSMC** exposed to **China risk**?" → Company → exposed_to → Theme). Include time window and synonyms from ontologies to tighten recall/precision.
3. Precision Retrieval (Perplexity/SERP API)	Issue targeted searches with recency filters, finance-domain allowlists, and snippet length caps. Require citation-backed passages. Run multiple, parallel variants (baseline + contrarian) and dedupe by URL/canonical source.
4. Evidence → Edge Extraction	Apply NER + relation extraction (LLM templates or lightweight RE models) to produce typed edges with metadata: {edge_type, source, snippet, first_seen, last_seen, extraction_confidence}. Canonicalize entities (ticker map) and normalize relation verbs to your schema.
5. Validation, Scoring & Conflict Handling	Compute an edge confidence from source quality, recency, snippet specificity, and model certainty. Drop below-threshold edges; flag contradictions with existing edges for review. Merge duplicates and increment evidence_count.
6. Temporary Augmentation & Caching	Inject validated edges into a query-scoped subgraph and a short-TTL cache (e.g., 1–24h). Tag as provisional=true. Promote to the persistent KG only after repeated confirmations (e.g., ≥2 independent sources or analyst approval). Track API usage and latency; fall back to local-only reasoning if budget/latency limits are hit.

Appendix G.1 – Build Plan Phases & Deliverables

Phase	Objective / Description
1 – Basic RAG MVP	Stand up a minimal but functional Retrieval-Augmented Generation (RAG) system to prove concept value without graph enhancements. • **Data Ingestion & Embeddings:** Load representative document samples (e.g., selected news, reports), chunk for indexing, generate embeddings (e.g., sentence-transformer) with metadata (source, date, entities). • **Simple Retrieval & QA:** Implement baseline RAG flow — vector similarity search (top-k docs) → LLM answer generation. • **Validation:** Test with sample investment queries, checking grounding in retrieved docs. • **Tech Decisions Finalized:** Select core vector store, embedding model, and initial LLM (e.g., GPT-3.5).
2 – Enhanced Retrieval (Hybrid + Lazy Graph Links)	Boost recall and breadth by adding hybrid keyword + vector search and on-the-fly graph expansion. • **Keyword Search:** Add full-text/regex search to catch exact matches missed by vectors; merge & deduplicate results. • **Lazy Graph Expansion:** Enrich results by pulling related documents via entity tags and metadata filters. • **Clustering (Optional):** Group by entities/themes for downstream summarization. • **Prototype & Evaluate:** Ensure expanded results capture indirect but relevant context. • **Performance Tuning:** Keep latency reasonable; scope expansion to high-value entities.
3 – Context Processing & Answer Generation Refinement	Transform raw retrieval results into concise, high-quality answers. • **Summarization / Fact Extraction:** Condense clusters or docs into key points relevant to the query. • **Ranking:** Prioritize most relevant, high-impact facts for final output. • **Prompt Finalization:** Design final answer template for clarity and factuality. • **Source Linking (Optional):** Retain citations for traceability. • **User Acceptance Testing:** Validate with real-world investment scenarios; adjust retrieval, summarization, and prompts.
4 – Deployment	Operationalize and integrate the system into analyst workflows. • **Batch Automation:** Schedule recurring queries (e.g., pre-market briefings) with logging and error handling. • **Lightweight UI:** Optional Streamlit or Slack integration for ad-hoc queries. • **Scaling & Optimization:** Parallelize steps, swap faster models for intermediates, cache results.
5 – Full Knowledge Graph Construction (Batch Enrichment)	Periodic, large-scale KG enrichment to strengthen long-term reasoning. • **Graph Maintenance:** Deduplicate edges, drop stale links, normalize node aliases. • **Backfill & Expansion:** Mine latent relations from historical data; add multi-hop edges. • **Graph Structuring:** Build thematic hierarchies, summary nodes, edge weights, and community clusters.

Appendix G.2 – Implementation Plan for Each Edge Type

Step	1. Thematic Exposure Edge	2. Valuation Driver Edge	3. Early Signal Edge	4. Ownership Context Edge	5. Causal Link Edge	6. Business Model Similarity Edge	7. Revenue Composition Edge	8. Functional Exposure Edge	9. Value Chain Edge	10. Sector /Industry ClassificationEdge
How to build?	Extract from headlines and semantic clusters using sentence-transformer + LLM for edge typing; validate with manual seed list; LLM + rule-based tagging + clustering	Use LLM + prompt templates to extract KPI mentions and causal relationships from docs; optionally use dictionary of KPI terms	Use LLM with weak supervision rules to extract corporate actions like layoffs, leadership changes, or delays. Use news scrapping pipelines + regex pre-filters	Direct from holdings tables; edge creation is trivial. Can be refreshed daily/weekly from OMS exports or static files	Named entity recognition (NER) + LLM extraction of causality; rule-based event tagging	Text embedding (e.g., all-MiniLM) + cosine sim threshold	Regex + LLM parsing from 10-K segment tables. SEC 10-Ks/10-Qs (especially segment tables), FactSet/CapIQ segment breakdowns, broker models	LLM classification against taxonomy. Extract functional roles or client use cases from descriptions.	LLM + rule-based extraction from 10-Ks and analyst notes; use templates like “X supplies Y” or “Y relies on X”.	Use GICS for standard mapping; optionally enrich with LLM + analyst-informed category taxonomy.
Source	Broker reports, LLM-tagged news, analyst commentary	Earnings transcripts, broker models, financials	News, press releases, earnings calls	Internal OMS, watchlists, 13F filings	News, macro notes, economic calendar	10-Ks, company descriptions, broker notes	SEC 10-Ks/10-Qs (segment tables), FactSet, broker models	Product sites, investor decks, broker notes	10-Ks, broker notes, CapIQ supply chain, news	GICS codes, broker reports, internal sector mapping
Ingestion	LLM preprocessed news/document pipeline	Transcript loader + chunker	News/article scrapers or RSS + document loader	Position table ingestion from OMS	Docs/news fetcher, calendar scrapers	Company doc parser + web scraper	SEC Edgar scraper, XLS/CSV parsers	Website/document crawler, HTML extractors	10-K/Risk section parser + broker model parser	Static mapping (GICS/CapIQ); internal mapping table
Preprocess	Focus on theme/sector-related sentence clusters	Identify KPI mentions + causality context	Filter for early signal keywords or “soft” tone events	Normalize fund/position names; dedupe	Identify economic/corporate events; locate cause-effect windows	Identify monetization models, pricing terms, descriptors	Extract tables + commentary sections	Focus on product descriptions & use cases	Focus on “supplies”, “relies on”, “depends on” etc.	Normalize industry names; handle overlap with other categories
LLM/ Parsing Strategy	Use LLM to tag themes (e.g. “China Supply Chain Risk”)	Prompt-based KPI extraction with weak supervision	Weak signal classifiers or regex + LLM templates	None required (structured data)	Event window extraction + LLM-based causal labeling	Embedding-based similarity + business model classifier	Regex/table parsing + LLMs to label segments (e.g. “Cloud”)	Use LLM to extract functions (e.g., “ML Ops”, “ERP backend”)	Use LLM rules or templates: “X is supplier of Y”, “Y depends on X”	LLM optional; rules/classifier to map firm → sector
Extraction Strategy	Use sentence-transformer embeddings (e.g., all-MiniLM) to cluster semantically similar news/headlines → LLMs label clusters with themes → match companies via co-mentions and contextual proximity	Prompt-based extraction of KPI mentions from earnings transcripts using LLMs → filtered by common valuation indicators (DAU, GM%, AOV, etc.) → validate with KPI dictionaries	Heuristic pre-filtering (e.g., regex for “delays”, “layoffs”, “pause”) + weakly supervised LLM templates over headlines → tag early warning corporate events	Extract directly from OMS/13F/position tables (structured) — no LLM needed; transform rows into fund → holds → ticker edges	Event detection via news scrapers → sentence window extraction → LLM templates for causal relationship detection (e.g., “X will impact Y”) + NER to identify targets	Text embedding model (e.g., all-MiniLM) on company descriptions + cosine similarity scoring → optionally filtered using NAICS/SIC or monetization keyword overlap	Use LLMs or table parsers to extract product/segment/geographic revenue breakdowns from earnings reports or filings. Map each revenue source as a node linked to the company, optionally tagged with % or YoY figures.	Use LLMs to extract functional roles (e.g., "enables payments") from company descriptions, IR, or broker notes. Map companies to standardized functional tags using sentence-level parsing and embedding-based classification.	Extract supplier/customer/platform links via rule-based patterns (e.g., “X supplies Y”, “Y relies on X”) in 10-Ks, broker decks, CapIQ → validate with co-occurrence + entity linking	Use static mapping table (e.g., GICS codes), or run firm descriptions through LLMs to assign fine-grained sectors — map to standard taxonomy tree
Edge Formation	Company → (exposed_to) → Theme	Company → (key_driver) → KPI	Company → (has_soft_signal) → Event	Fund → (holds) → Company	Event → (likely_impacts) → Metric/Company	Company A → (similar_biz_model) → Company B	Company → (generates_revenue_from) → Segment/Region (with % attr)	Company → (powers/enables) → Functional Unit	Supplier → (supplies) → Customer, Customer → (depends_on) → Platform	Company → (classified_as) → Industry/Sector
Validation	Backtest theme membership across news sets	Analyst curation or KPI dictionary	Manual sampling + pattern QA	Direct match from OMS/13F	Human check of impact pairs or scenario examples	Cosine similarity thresholds, seed list validation	Cross-check across multiple filings/models	Taxonomy alignment for functional categories	Cross-source check: CapIQ vs 10-K vs broker model	Validate against sector ETF constituents, analyst sector decks

Appendix H – MVP Scope Table

Focus	Description	Why This Scope for MVP?
Asset Class	Public equities	Aligns with the fund's core investment focus and expertise.
Stock Universe	S&P 500 constituents	Maximizes coverage and minimizes data acquisition costs.
Investment Workflow Focus	Priority order: 1) Portfolio management, 2) Investment research, 3) Risk management, 4) Trading	Targets high-trust, high-impact workflows where AI augments—not replaces—analyst judgment; balances business value with delivery feasibility.
Data	Publicly available qualitative external data (e.g., filings, earnings transcripts, news)	Leverages LLM strengths in unstructured data; allows for future layering of quantitative data.
Typed Edge Patterns	Minimum 3 high-value edge types for Lazy Graph RAG, in priority order: 1) `Portfolio/Watchlist → holds → Company`, 2) `Company → exposed_to → Theme`, 3) `Company → key_driver → KPI`, with optional expansion to: 4) `Company → sector → GICS_sector`, 5) `Company → industry → GICS_industry`, 6) `Company A → supplier/customer → Company B`, 7) `Company → has_soft_signal → Event`	Maximizes decision-relevance with minimal engineering lift; prioritization reflects business impact vs. complexity trade-offs.
Processing Frequency	Batch (daily or intraday)	Simplifies implementation while meeting early signal detection needs.
Graph Construction Technique	Lazy query-triggered graph expansion	Avoids full upfront build costs; enables targeted, performant enrichment.
RAG Architecture	Up to 3-hop lazy traversal over available edge patterns	Balances richer reasoning with manageable complexity and compute load.
Workflow Design	Monolithic pipeline vs. multi-agent architecture	Reduces system complexity while maintaining reasoning depth; simpler to optimize and debug.
AI Interaction Design	Single QA input box; phased rollout of capabilities	Surfaces highest-value queries early with minimal UX friction; staged development delivers incremental ROI and de-risks project.
Interoperability	MCP-style standardization	Ensures plug-and-play extensibility and long-term maintainability.

Appendix I - Hop Retrieval Patterns and Use Cases

1-hop (Fastest ROI, Trivial to Extract)

Retrieval Pattern	Path	Example Query	Value	Data Sources
Theme exposure for a ticker	`Company → exposed_to → Theme`	"Is TSMC exposed to China risk?" / "What themes drove META this week?"	Immediate narrative context for price action and screening; clean extractions from headlines/transcripts.	Earnings transcripts, news articles/headlines, sell-side reports, 10-K risk factors
KPI driver for a ticker	`Company → key_driver → KPI`	"What KPIs actually drive UBER?"	Focuses research on the 1–2 metrics that move the stock; low-cost extraction from earnings-related docs.	Earnings transcripts, investor decks, earnings presentations, sell-side models, analyst initiation reports, earnings call summaries
Soft-signal scan	`Company → has_soft_signal → Event`	"Any early negative signals on AMD?"	Surfaces pre-consensus drift (tone changes, delays, layoffs); early narrative shifts before financial impact; ideal for alerts.	News headlines/wires, internal team notes/emails, sell-side notes, earnings call Q&A/management caution

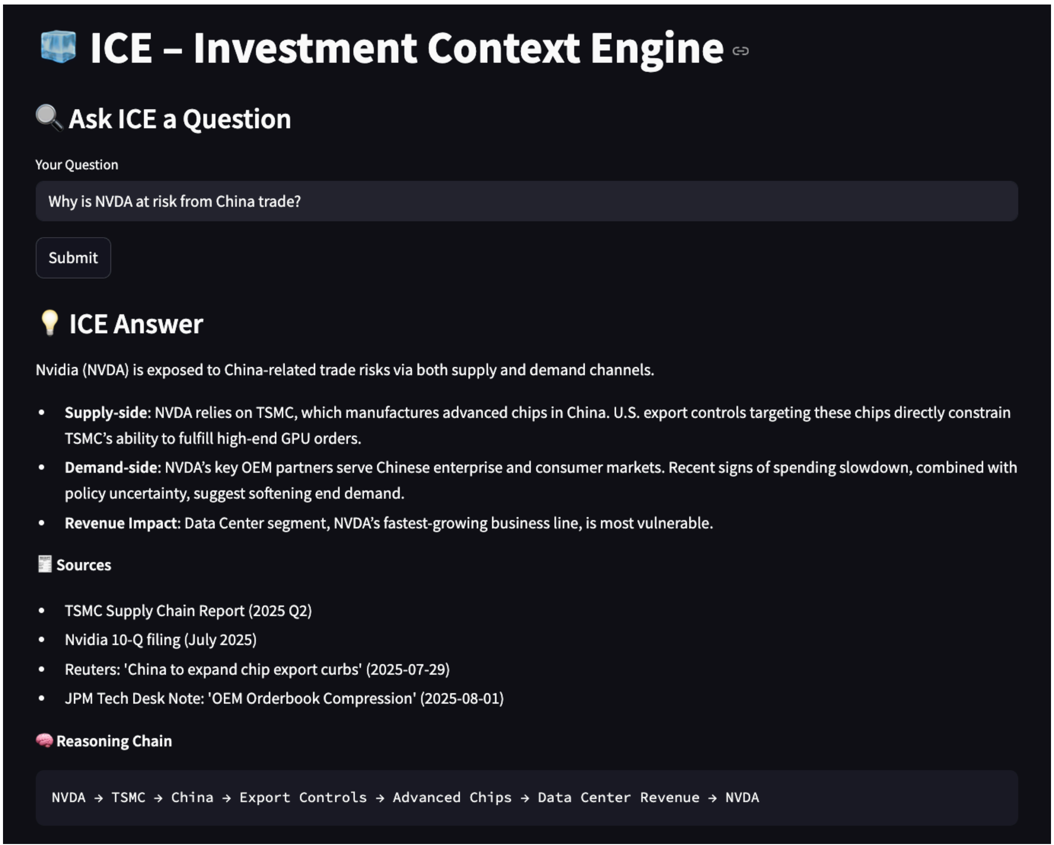
2-hop (Portfolio-aware & Causal Stubs)

Retrieval Pattern	Path	Example Query	Value	Data Sources
Portfolio theme exposure	`Portfolio → holds → Company → exposed_to → Theme`	"Which of our holdings are exposed to GLP-1 / China risk / AI infra?"	Converts themes into book-level views; trivial portfolio edge from CSV.	Portfolio/watchlist CSVs, news/earnings transcripts theme links, LLM-extracted entity-theme co-occurrences
Post-earnings 'why move?'	`Company → key_driver → KPI + Company → has_soft_signal → Event`	"Why did META drop post-earnings?"	Links KPI anchors to fresh management commentary or soft signals to explain direction without modeling magnitude.	Earnings calls, press releases, pre/post-market wires and alerts, soft signals from notes or presentations
Supplier/customer narrative propagation	`Company → depends_on → Supplier/Customer → exposed_to → Theme`	"How does China risk reach NVDA via TSMC?"	Explains second-order impacts with minimal extra steps; easy to score by recency/confidence.	Supply chain disclosures, vendor lists, industry DBs (Factset RBICS, Bloomberg SPLC), filings/news linking supplier to theme
Theme → names reverse lookup	`Theme ← exposed_to — Company (reverse traversal)`	"Show names tied to 'AI infrastructure' with most recent evidence."	Native reverse traversal without storing reverse edges.	LLM-tagged company-theme co-occurrence, earnings transcripts, press releases, news articles

3-hop (High Signal)

Retrieval Pattern	Path	Example Query	Value	Data Sources
Full causal trace for risk	`Company → depends_on → Supplier → exposed_to → Theme ← has_soft_signal — Company`	“How is NVDA exposed to China risk via TSMC, and what’s the freshest management tone?”	Classic risk explanation path; scored by confidence, recency, and evidence count.	Sources from 2-hop + soft-signal sources; temporal metadata & confidence scoring
Portfolio-first risk sweep	`Portfolio → holds → Company → has_soft_signal → Event → Theme`	“Which holdings flashed cautionary signals in the last 30 days, and what themes do they touch?”	Converts soft signals into portfolio-level action items; recency filters remove noise.	Portfolio CSV, soft signals from news/internal notes, event-theme ontology (e.g., layoffs → macro slowdown)
Supplier/customer concentration check	`Portfolio → holds → Company → depends_on → Supplier/Customer`	“Which positions rely on the same supplier?”	Simple consolidation view for risk meetings; uses CSV + supplier edge.	Portfolio CSV, supply chain DB/disclosures, Factset RBICS, Bloomberg SPLC
Theme → KPI lens for a name	`Theme ← exposed_to — Company → key_driver → KPI`	“For GOOG’s AI infra exposure, which KPI should we watch?”	Links narratives to measurable anchors; ideal pre-earnings prep.	Theme extraction from transcripts/press, KPI drivers from investor materials, sell-side notes
Event narrative via supplier	`Company → depends_on → Supplier → has_soft_signal → Event`	“Does a TSMC production warning matter for ASML?”	Tests whether upstream soft signals propagate to relevant names.	Supplier disclosures/inferred links, news/filings with soft signals, event timestamps and tone

Appendix J – ICE MVP User Interface Mockup



ICE – Investment Context Engine

Ask ICE a Question

Your Question

Why is NVDA at risk from China trade?

ICE Answer

Nvidia (NVDA) is exposed to China-related trade risks via both supply and demand channels.

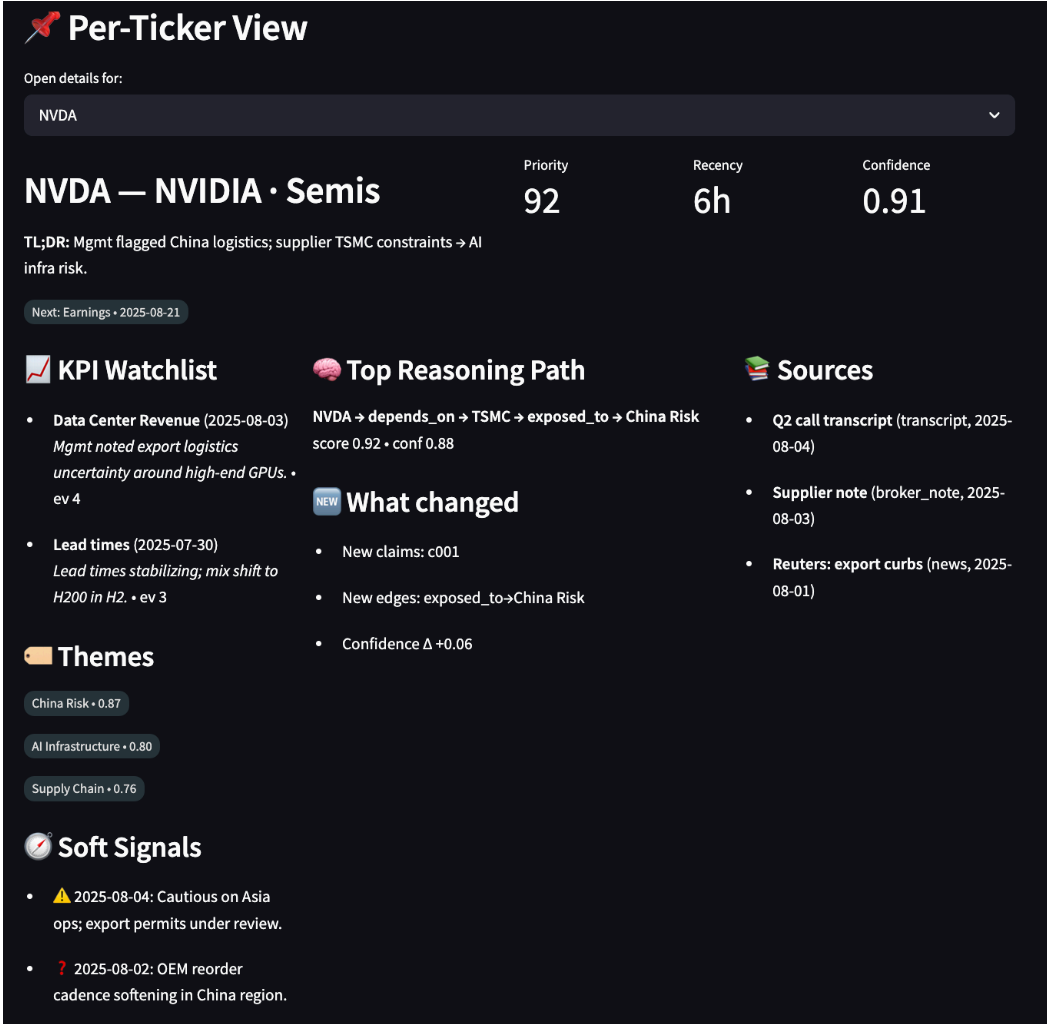
- **Supply-side**: NVDA relies on TSMC, which manufactures advanced chips in China. U.S. export controls targeting these chips directly constrain TSMC's ability to fulfill high-end GPU orders.
- **Demand-side:** NVDA's key OEM partners serve Chinese enterprise and consumer markets. Recent signs of spending slowdown, combined with policy uncertainty, suggest softening end demand.
- **Revenue Impact:** Data Center segment, NVDA's fastest-growing business line, is most vulnerable.

Sources

- TSMC Supply Chain Report (2025 Q2)
- Nvidia 10-Q filing (July 2025)
- Reuters: 'China to expand chip export curbs' (2025-07-29)
- JPM Tech Desk Note: 'OEM Orderbook Compression' (2025-08-01)

Reasoning Chain

```
NVDA → TSMC → China → Export Controls → Advanced Chips → Data Center Revenue → NVDA
```



Per-Ticker View

Open details for:

NVDA

NVDA — NVIDIA · Semis

Priority	Recency	Confidence
92	6h	0.91

TL;DR: Mgmt flagged China logistics; supplier TSMC constraints → AI infra risk.

Next: Earnings • 2025-08-21

KPI Watchlist

- **Data Center Revenue** (2025-08-03) *Mgmt noted export logistics uncertainty around high-end GPUs.* • ev 4
- **Lead times** (2025-07-30) *Lead times stabilizing; mix shift to H200 in H2.* • ev 3

Themes

China Risk • 0.87

AI Infrastructure • 0.80

Supply Chain • 0.76

Soft Signals

- ⚠ 2025-08-04: Cautious on Asia ops; export permits under review.
- ❓ 2025-08-02: OEM reorder cadence softening in China region.

Top Reasoning Path

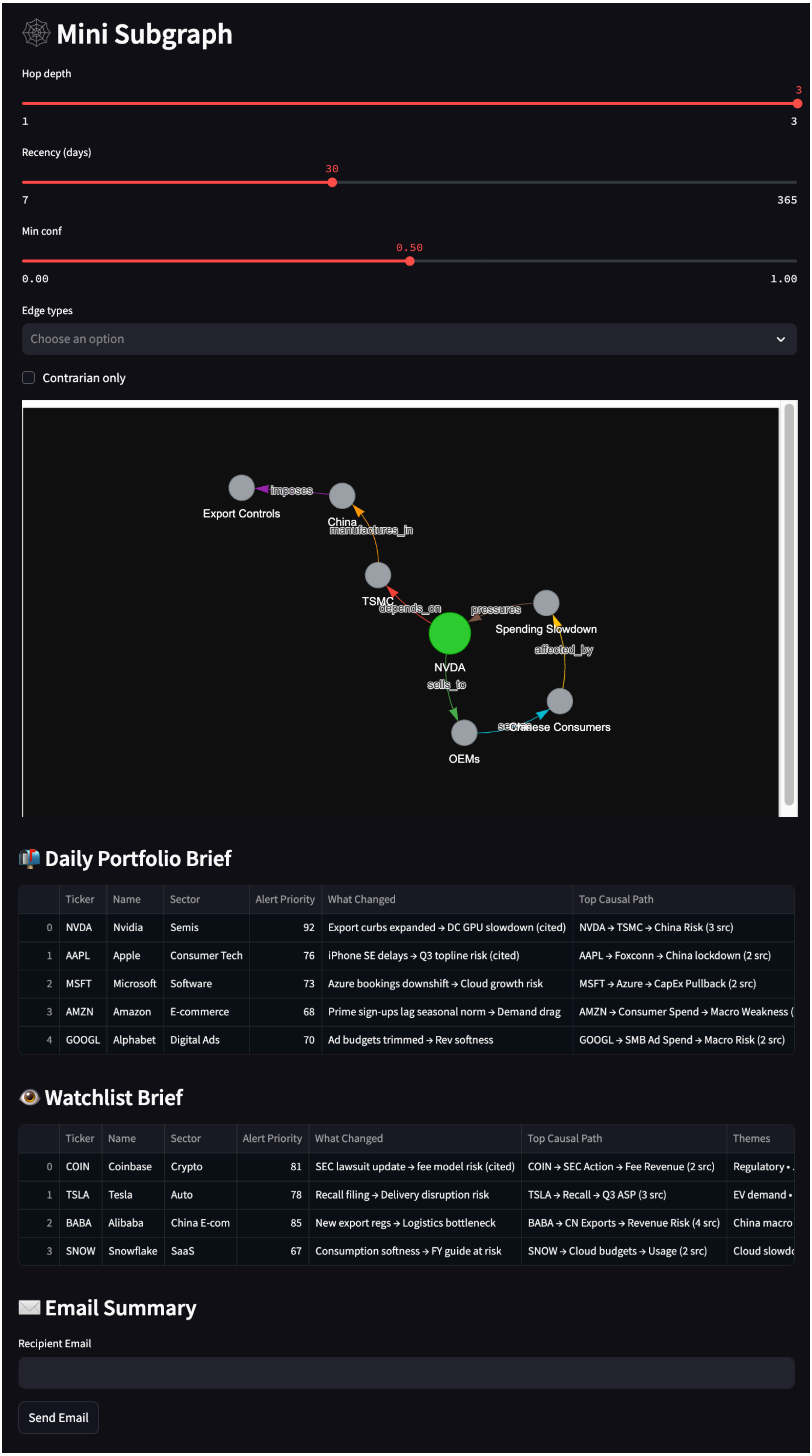
NVDA → depends_on → TSMC → exposed_to → China Risk
score 0.92 • conf 0.88

What changed

- New claims: c001
- New edges: exposed_to→China Risk
- Confidence Δ +0.06

Sources

- **Q2 call transcript** (transcript, 2025-08-04)
- **Supplier note** (broker_note, 2025-08-03)
- **Reuters: export curbs** (news, 2025-08-01)



🕸 Mini Subgraph

Hop depth: 3 (1–3)

Recency (days): 30 (7–365)

Min conf: 0.50 (0.00–1.00)

Edge types: Choose an option

□ Contrarian only

📬 Daily Portfolio Brief

	Ticker	Name	Sector	Alert Priority	What Changed	Top Causal Path
0	NVDA	Nvidia	Semis	92	Export curbs expanded → DC GPU slowdown (cited)	NVDA → TSMC → China Risk (3 src)
1	AAPL	Apple	Consumer Tech	76	iPhone SE delays → Q3 topline risk (cited)	AAPL → Foxconn → China lockdown (2 src)
2	MSFT	Microsoft	Software	73	Azure bookings downshift → Cloud growth risk	MSFT → Azure → CapEx Pullback (2 src)
3	AMZN	Amazon	E-commerce	68	Prime sign-ups lag seasonal norm → Demand drag	AMZN → Consumer Spend → Macro Weakness (
4	GOOGL	Alphabet	Digital Ads	70	Ad budgets trimmed → Rev softness	GOOGL → SMB Ad Spend → Macro Risk (2 src)

👁 Watchlist Brief

	Ticker	Name	Sector	Alert Priority	What Changed	Top Causal Path	Themes
0	COIN	Coinbase	Crypto	81	SEC lawsuit update → fee model risk (cited)	COIN → SEC Action → Fee Revenue (2 src)	Regulatory • .
1	TSLA	Tesla	Auto	78	Recall filing → Delivery disruption risk	TSLA → Recall → Q3 ASP (3 src)	EV demand •
2	BABA	Alibaba	China E-com	85	New export regs → Logistics bottleneck	BABA → CN Exports → Revenue Risk (4 src)	China macro
3	SNOW	Snowflake	SaaS	67	Consumption softness → FY guide at risk	SNOW → Cloud budgets → Usage (2 src)	Cloud slowd

✉ Email Summary

Recipient Email

Send Email

Appendix K.1 – Build Constraints & Mitigation Strategies

Constraint	Details	Potential Mitigations
Time	3-month delivery target for MVP.	Structure as a phased build plan with clear phase-complete fallbacks; optional extension to 6 months if needed; deliver value at each phase to justify continuation.
People	Single quant analyst with limited software engineering depth.	Use lean, proven stack; maximize open-source reuse; rely on AI coding aids (Cursor, Claude Code, ChatGPT); employ templates/utilities over bespoke services; maintain strict scope discipline.
Budget	No dedicated allocation for enterprise AI infra or heavy LLM usage; funding contingent on MVP's demonstrated business value.	Build on open-source Python libs (FAISS, NetworkX, LangChain-like orchestration) running locally; use local embeddings and free/public datasets first; minimize token spend with Lazy GraphRAG + hybrid retrieval/summarization; meter premium LLM calls and reserve for high-value reasoning.
Infrastructure	Limited/no cloud backend; laptop-scale compute only.	Avoid real-time pipelines—prefer batch processing; persist lightweight data stores locally; adopt Lazy GraphRAG instead of full KG construction; zero-ops deployment—all code in a single repo runnable on analyst desktop; no server/cloud creds required.
Data	Public + internal docs; no live OMS/API integration at MVP stage.	Start with public filings, transcripts, newswire, internal notes, CSV holdings; parallel-track short vendor trials (FactSet, Bloomberg) for expansion; adjust use cases based on available feeds.
Architectural Design	Lazy GraphRAG is relatively new and lightly documented in the community.	Maintain a fallback architecture: conventional RAG + light typed relations in metadata; limit MVP scope to reduce complexity; evaluate LLM-assisted KG construction for selective expansion.

Appendix K.2 –Key Risks and Mitigation Strategies

Risk	What can go wrong?	Solution / Workaround (how we control it)
Sparse KG coverage	Early graph lacks edges/nodes, so multi-hop reasoning misses links or collapses to shallow paths.	**Prevent:** Build *on-demand* subgraphs from retrieved docs (Lazy Graph Expansion); seed a small prior graph from high-confidence corpus patterns; allow provisional edges with status=candidate and low weight. **Detect/Respond:** Track coverage metrics (edges per ticker/theme, evidence_count); prompt users to confirm/deny low-confidence edges (active learning); promote candidates only after reconfirmation by new sources.
Hallucination	LLM invents relationships not present in sources, polluting the graph and answers.	**Prevent:** Extraction-only from retrieved, timestamped spans; force inline citations per edge; use constrained prompts and schema validation; apply NLI/entailment check (cross-encoder) that each claim is supported by its source. **Detect/Respond:** Threshold on confidence + retrieval score; quarantine edges without citations; require human review for high-impact alerts.
Conflicting views	Internal notes vs. sell-side/news disagree, producing incoherent inference.	**Prevent:** Store provenance + perspective (source_role = internal/sellside/news) and stance labels where available. **Detect/Respond:** Contradiction detection over edges touching same pair; surface side-by-side narratives rather than auto-resolving; escalate to analyst when conflicting paths change a recommendation.

Siloed vector stores	Different indexes (e.g., FAISS for filings, Chroma for notes) block unified retrieval.	**Prevent:** Prefer a single ANN index with rich metadata; standardize chunking, embedding model, and global doc_id. **Detect/Respond:** If multiple stores are unavoidable, use a federated retriever (BM25+ANN across stores) and a cross-store reranker; normalize metadata and merge top-k before graph expansion.
Token explosion on multi-hop queries	2–3 hop queries pull too many docs/edges and overflow LLM context.	**Prevent:** Enforce a retrieval plan (caps on graph_hops, per-hop top_k, MMR de-dup); perform hierarchical, extractive summaries (sentence-level) before synthesis; maintain a token budget per query. **Detect/Respond:** Drop low-recency/low-confidence edges first; stream/iterate answers (progressive disclosure) when context nears limit.
Stale or spurious edges	Old/incorrect relations bias reasoning (e.g., outdated suppliers/themes).	**Prevent:** Attach first_seen/last_seen, evidence_count, and decay weighting to every edge; treat edges as hypotheses with confidence, not facts. **Detect/Respond:** Weekly re-score batch trims stale edges; version the KG; flag "at-risk" edges whose evidence hasn't refreshed within policy windows.
Market regime drift	Drivers change (e.g., export controls, AI capex cycles), but graph encodes old narratives.	**Prevent:** Tag edges with validity_period / context_epoch; time-aware retrieval windows by default. **Detect/Respond:** Compare path score deltas over time; "What changed" view highlights replaced drivers; fade edges unless reaffirmed by new evidence.
Trust & adoption barrier	PMs/analysts distrust opaque outputs or fear hidden hallucinations.	**Prevent:** UX centers traceability (sources, timestamps, path visual); show confidence bands and edge counts; position ICE as augmented research (not an auto-trader). **Detect/Respond:** Pilot in non-critical workflows; collect feedback on false positives/negatives; make "open source panel" one click away so users can audit every sentence.

Appendix K.3 –Targeted Technical Enablers for Risk Mitigation

Technique	What It Is	Which Risk It Addresses
Temporal quadruple representation + KGTransformer	Encodes (subject, predicate, object, time) to make the KG time-aware; KGTransformer enables temporal reasoning	Stale/spurious edges, market context drift
Senzing-style entity resolution + confidence scoring	Merges duplicate/variant entities and assigns match confidence	Sparse KG, conflicting views
Intelligent query routing (LazyGraphRAG ↔ traditional RAG)	Routes queries to optimal pipeline depending on complexity	Token explosion in multi-hop, sparse KG
ColBERT tensor-based reranking	Token-level embedding interaction improves relevance without full cross-encoder cost	Hallucination risk, retrieval precision
Multi-way recall fusion (dense + sparse + full-text)	Combines multiple retrieval modes before ranking	Sparse KG, siloed vector stores
Modern open-source stack	FAISS/Milvus vector DBs, NetworkX/Graph tooling, high-performance local LLMs	Budget constraints, infra constraints

Appendix L – Future Exte`nsions

Feature / Extension	Description
Full GraphRAG Upgrade	Migrate from Lazy GraphRAG to full graph retrieval-augmented generation, supporting more complex multi-hop causal reasoning and richer, interconnected narratives.
Periodic Full KG Expansion & Maintenance	Supplement query-triggered lazy graph with scheduled full graph construction cycles. Include timestamped evidence, edge decay logic, schema validation, and automated pruning of stale or weak edges to preserve accuracy, semantic integrity, and explainability.
Data Expansion (incl. Quant Data Integration)	Integrate structured financials (e.g., income statement KPIs), alternative datasets, internal trade logs, and expanded qualitative sources for more holistic reasoning.
IT Infrastructure Setup	Establish secure, maintainable infrastructure: zero-ops batch pipelines, standardized data interfaces, and future cloud-readiness.
Secure Document Upload	Enable ingestion of internal PDFs, memos, and proprietary research directly into the graph, ensuring secure handling and tagging.
User Feedback Loop	Implement mechanisms for user-driven corrections, tagging, and relevance feedback, feeding directly into model fine-tuning and retrieval precision improvements.
Edge Traceability & Decay Logic	Enrich every edge with source provenance and recency scoring, applying decay weighting to prevent stale logic from influencing outputs.
Multi-Agent Workflows	Evolve from a monolithic pipeline to specialized agents (Extractor, Reasoner, Synthesizer), improving scalability, explainability, and fault isolation.
Foundation for Agentic Hedge Fund	Lay the groundwork for a fully AI-native research and monitoring stack, with ICE at its core, enabling proactive and personalized investment intelligence.