

Continuous Financial Knowledge Graph Construction with Citation-Grounded Agentic Question Answering

Arnau Rodon Comas
Universitat Pompeu Fabra
Barcelona, Spain
arnau.rodon01@estudiant.upf.edu

Abstract

Financial analysis requires crossing evidence modalities that arrive in incompatible forms — structured market data, reported text (news and regulatory filings), collective expectations from prediction markets, and private portfolio state — which existing systems either handle in isolation or collapse into a single index, destroying the distinction between observed measurements, attributed claims, probabilistic beliefs, and user-specific exposure. We present a continuously running financial NLP system that maintains these modalities natively, linked by canonical entities and time, through continuous knowledge graph construction with per-edge provenance and temporal decay, and answers questions via agentic tool use under a citation contract that is enforced structurally: every citation must resolve to a passage retrieved in the same session, and free-text figures must agree with provider-attributed tool outputs. A learned relevance gate formalises the extraction cost/quality tradeoff, reducing LLM extraction calls while retaining the majority of yield-positive documents. Over 284 benchmark questions executed three times under majority vote, 251 produce non-failing responses (205 strong, 20 pass, 26 weak), 23 fail, and 10 additional questions are data-absent under the evaluation protocol. A structural evaluation framework of eight deterministic pre-judge gates reaches $\kappa = 0.860$ with a human annotator. In a sample evaluation, grounding entity descriptions in retrieved passages cuts fabrication from 1.83 to 0.17 unverifiable claims per description.

1 Introduction

Financial analysis requires synthesising evidence that arrives in fundamentally incompatible forms. A question such as “*Is NVIDIA’s supply chain risk priced into the stock, given recent analyst reports and prediction market sentiment?*” requires (1) graph traversal over extracted supplier relationships with temporal validity, (2) hybrid retrieval over re-

cent news passages, (3) exact lookup of valuation fundamentals, and (4) querying prediction market probabilities — all under a single planning loop in which retrieved evidence remains attributable to the passage or provider record it came from. Among the systems reviewed in Section 2, we did not identify one that addresses all four simultaneously while maintaining citation traceability.

A further critical challenge in financial NLP systems is factual grounding: language models generate plausible-sounding but unverifiable figures, a failure mode particularly dangerous in financial contexts where incorrect values can drive real decisions. A model that reports a price-to-earnings ratio of “approximately 40” when the recorded provider value is 42.3 has produced a fluent answer that no downstream consumer can distinguish from a grounded one. Hallucination mitigation in this setting is therefore not a stylistic preference but a correctness requirement, and it must be enforced structurally rather than left to the generator’s discretion.

The incompatibility between modalities runs deeper than access patterns. Reported text (news, filings) carries attributed claims with reliability profiles — the same fact can be reported differently by different sources, and the disagreement is itself informative. Structured market data carries measurement authority but no semantic context: a P/E ratio of 42.3 is meaningless without a narrative of why the market assigns that multiple. Prediction markets encode market-implied probabilistic expectations about uncertain futures: a 61% probability that TSMC capacity remains sold out is not a fact to be retrieved but a probability to be weighted. Personal portfolio exposure determines which subset of the evidence space is relevant for a given user. Collapsing these into a unified vector index destroys the distinctions that make each modality valuable.

We contribute a deployed financial intelligence

platform that addresses the joint grounding problem with four contributions examined through complementary experiments:

(C1) A live-updated financial KG with per-edge bidirectional links to source passages, a six-class temporal decay taxonomy per predicate type, and contradiction edges that register conflicting evidence rather than overwriting.

(C2) A learned relevance gate formalising when-to-extract as binary classification (gradient boosting; ROC-AUC 0.828 vs. 0.639 heuristic; 25% fewer LLM calls at 6% recall cost, held-out CV estimate; §7).

(C3) Two-level hallucination mitigation: construction-time entity description grounding (fabrication reduced from 1.83 to 0.17 per description) and inference-time citation enforcement via structural evaluation gates.

(C4) A structural evaluation framework of eight deterministic pre-judge gates achieving $\kappa = 0.860$ (Landis–Koch: almost perfect) with a human annotator, catching citation and numeric fabrications that an LLM judge alone rewards. A restricted-access ablation shows the mean judge score falls from 94.8 to 46.4 under text-only retrieval.

2 Related Work

Financial knowledge graphs and temporal modelling. FinReflectKG (Arun et al., 2025) is the closest precedent: LLM triple extraction from SEC 10-K filings with source-chunk linking. We extend it with continuous news-stream ingestion, temporal decay per edge, and integration of structured market data and multimodal tools in the same reasoning loop. Prior financial graphs focus on company–company relations without temporal dynamics; modal-adaptive KG prediction (Ouyang et al., 2024) treats the graph as a fixed predictor input, not a live evidence store; GraphRAG (Shukla et al., 2025) reports mixed gains over vector retrieval on static summaries without per-edge provenance. Standard KG benchmarks do not model temporal decay; prior financial-domain work uses timestamps on triples rather than validity intervals. Our six-class decay taxonomy (permanent through ephemeral) reflects distinct temporal semantics per predicate type. BloombergGPT (Wu et al., 2023) and domain-adapted LLMs encode knowledge parametrically, sacrificing updateability and citation traceability; we treat language models as planners over externally grounded evidence.

Financial NER, QA, and retrieval-augmented generation. Lu and Huo (2025) benchmark LLM-based NER on financial text and find generative models uneven on fine-grained types relative to specialised encoders. We use a generalist encoder with eleven financial entity classes combined with a five-stage resolution cascade that carries the domain burden (turning *Nvidia Corp.* and *NVDA* into one canonical node). HybridRAG (Sarmah et al., 2024) fuses KG and vector retrieval over static collections; FinSearch (Shen et al., 2025) adds temporal search but lacks a persistent graph or structured numeric lookup; hierarchical reranking (Lee and Hong, 2026) addresses document retrieval only. FinAgentBench (Choi et al., 2025), FinQA (Chen et al., 2021), and TAT-QA (Zhu et al., 2021) provide evaluation baselines over static corpora with fixed gold answers; our benchmark targets a live deployed system where the correct answer changes with the data.

Hallucination mitigation and LLM-as-judge evaluation. CHATREPORT (Ni et al., 2023) surfaces passages in ESG disclosures but applies no automatic check linking generated content back to them; FinGPT-based systems (Yang et al., 2023) produce forecasts with no citation traceability. We address hallucination at two levels: construction-time grounding (entity descriptions generated from retrieved passages, reducing fabrication from 1.83 to 0.17 per description in a sample evaluation) and inference-time enforcement via structural gates that check citation resolvability and numerical consistency — a weaker but automatically decidable condition than full semantic entailment (§8). Automated evaluation frameworks RAGAS (Es et al., 2024) and ARES (Saad-Falcon et al., 2024) rely on LLM metrics vulnerable to fluency bias; our eight pre-judge deterministic gates prevent fluency from masking structural failures.

Agentic tool use and extraction cost. Toolformer (Schick et al., 2023) and ReAct (Yao et al., 2023) establish self-supervised tool learning and chain-of-thought with actions, treating tool results as flat text. We extend this with a typed evidence contract and cross-modal chaining, where one modality’s output directly parameterises the next modality’s query. The architecture is a single ReAct-style planning agent with a structured tool manifest — not a multi-agent system in the sense of multiple coordinating agents. FinReflectKG (Arun et al., 2025) and comparable systems

extract unconditionally; our learned relevance gate formalises when-to-extract as a reproducible classification problem.

3 System Architecture

3.1 Four Evidence Modalities

The central design decision is that financial evidence is not a unified corpus but four modality types with distinct truth conditions (Figure 1):

- **Observed state:** prices, fundamentals, earnings calendars. Timestamp-stamped measurements; consulted exactly, never paraphrased.
- **Reported assertions:** news and regulatory filings. Attributed claims with source reliability profiles; stored as canonical documents, embedded chunks, and extracted relations linked to their source passage.
- **Collective expectations:** prediction market propositions with probability time-series. Negotiable beliefs about uncertain futures that resolve over time; not reports of what happened, but bets on what will.
- **Personal exposure:** user portfolio and watchlists. Private state that determines which evidence is relevant for this specific user.

The three shared keys — canonical financial entities, topics, and time — link all four modalities without collapsing them into a single index.

3.2 NLP Construction Pipeline

Text sources — news feeds, SEC EDGAR filings, and prediction market proposition text — enter the NLP pipeline (Figure 2).

Stage 1: Deduplication. A three-layer cascade: exact content hash, normalised-URL match, and

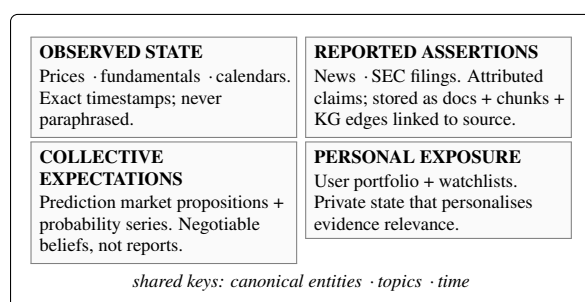


Figure 1: Four native evidence modalities linked by shared canonical keys.

MinHash–LSH for near-duplicate detection. Re-posts are registered as corroboration of the surviving document rather than as independent evidence, preventing syndicated wire copy from inflating the apparent independence of extracted-relation support.

Stage 2: Relevance gate. A gradient boosting classifier using a semantic embedding of the document title and subtitle concatenated with structured signals (document type, source reliability, recency) routes documents to deep extraction or skips it when extraction is unlikely to yield structured artefacts (§6.2).

Stage 3: NER and entity resolution. GLiNER (Zaratiāna et al., 2024) labels eleven entity classes. A five-stage resolution cascade matches mentions to canonical financial entities: (1) exact ticker/ISIN match, (2) fuzzy company name match, (3) alias expansion from a curated synonym table, (4) contextual disambiguation using co-occurring known entities, and (5) provisional entity creation for unresolved mentions, subsequently enriched with grounded LLM-generated descriptions that enable immediate semantic embedding. The eleven classes (organisation, government body, regulatory body, financial institution, person, financial instrument, location, commodity, index, currency, macroeconomic indicator) distinguish policy actors from traded issuers — a distinction generic taggers erase and that downstream supply-chain or policy-exposure traversal requires.

3.3 Structured Data Path

In parallel, a market data service ingests OHLCV price series, financial fundamentals (balance sheet, income statement, key ratios), earnings calendars, and prediction market probability series from external data providers. This path bypasses the NLP pipeline entirely: data is normalised, deduplicated by (symbol, date), and stored in time-indexed relational tables and TimescaleDB hypertables for exact retrieval by symbol and date range. (Prediction market proposition text enters the NLP pipeline as a text source; only the probability time series takes this structured path.)

3.4 Relation Extraction and KG Update

Stage 4: Relation extraction. An LLM with a semi-closed predicate vocabulary (~80 relation types) extracts typed triples with source passage links and temporal metadata. Each extracted triple

carries: subject and object canonical entity IDs, predicate label and confidence score, a validity interval (valid_from: extraction date; valid_to: optional, set only when the text explicitly states the relationship has ended), a system-assigned decay class (one of six: permanent, durable, slow, medium, fast, ephemeral), and the ID of the source RAG chunk.

The ~ 80 predicates span four functional categories: supply-chain (manufactured_by, key_customer), governance (subsidiary_of, ceo_of), financial (analyst_rating, revenue_exposure_to), and event-impact (merger_target, regulatory_action_against). The vocabulary is semi-closed: the extractor may propose free-form predicates, which are promoted on evidence of recurrence across independent documents. Each stored edge carries subject/object entity IDs, predicate, confidence, validity interval, decay class, and the ID of the source RAG chunk.

Stage 5: KG and RAG update. Triples are materialised in the property graph (Apache AGE / openCypher). Crucially, each graph edge carries a bidirectional pointer to its source chunk in the vector index (pgvector HNSW), creating a unified evidence graph where KG traversal and passage retrieval are interoperable. Contradicting evidence does not overwrite existing edges; it is registered as a counter-edge, preserving the controversy in the graph so that a downstream answer can report disagreement rather than silently selecting a side.

3.5 Entity Embeddings and Multitype Representations

Each canonical entity maintains three parallel embedding types updated asynchronously: (1) **definition** over the entity’s description and aliases; (2) **narrative** over recent news passages, providing a temporally sensitive fingerprint; and (3) **fundamentals+market** over structured financial data encoded as text. The retrieval layer selects the embedding type based on query characteristics. Definition embeddings are the point at which construction-time hallucination mitigation applies: descriptions for thinly-covered entities are generated from retrieved passages rather than parametric memory, since a fabricated description poisons the retrieval substrate by becoming part of the vector later queries match against.

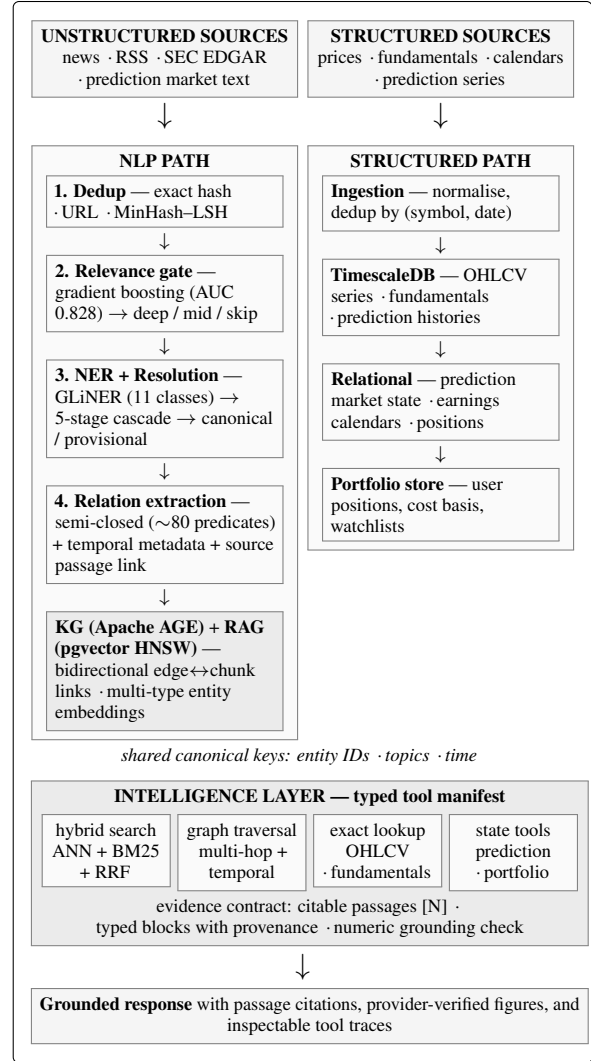


Figure 2: Full system pipeline. NLP and structured paths build parallel evidence stores; the intelligence layer queries them jointly through a typed tool manifest with a cross-modal chaining planner.

3.6 Edge Confidence and Temporal Decay

Each edge carries a confidence score $c = \text{clamp}(S + B - D, 0, 1)$, where each term is:

Support $S = \frac{\sum_i w_i t_i}{\sum_i w_i}$ trust-weighted temporal avg., $t_i = e^{-\alpha \Delta_i}$, Δ_i age in days of source i

Diversity $B = \min(0.20, 0.05k)$; k = distinct qualifying outlet pairs with $t_i \geq 0.10$

Contradiction $D = \min(0.60, \sum_{j=1}^3 s_j e^{-\alpha \delta_j})$ top-3 decayed link strengths (δ_j : age in days of the j -th contradicting evidence item, analogous to Δ_i)

Per-class α (fixed before benchmark execution):

Class	α (d ⁻¹)	$t_{1/2}$
PERMANENT	0.000000	∞
DURABLE	0.000950	730 d
SLOW	0.003851	180 d
MEDIUM	0.011552	60 d
FAST	0.049510	14 d
EPHEMERAL	0.231049	3 d

Assignments follow predicate semantics: acquisitions and HQ location $\rightarrow$ PERMANENT; executive roles $\rightarrow$ DURABLE; legal filings $\rightarrow$ SLOW; earnings guidance $\rightarrow$ MEDIUM; analyst ratings $\rightarrow$ FAST; sentiment $\rightarrow$ EPHEMERAL. Stateful relations hold at full strength while valid and step to a floor on expiry; c is not a calibrated probability.

4 Intelligence Layer

The intelligence layer is where the four evidence modalities converge. Its design solves a fundamental tension: each modality requires different query semantics, yet a single user question may require all four simultaneously.

4.1 Tool Manifest and Query Semantics

The planning agent operates over four evidence-type tool families. **Text retrieval** uses hybrid ANN+BM25+RRF search (Malkov and Yashunin, 2020; Robertson and Zaragoza, 2009; Cormack et al., 2009) with optional HyDE expansion (Gao et al., 2023); results carry chunk IDs and KG entity links. **Graph traversal** executes openCypher queries with temporal filtering, confidence thresholding, and contradiction-edge exposure; each edge includes its source chunk ID so graph facts remain citable. **Structured lookup** performs exact retrieval by canonical entity ID and date range over time-series and relational stores (OHLCV, fundamentals, earnings calendars), returning provider-attributed values rather than model approximations. **State tools** access prediction market probabilities and user portfolio positions, integrating the collective-expectations and personal-exposure modalities.

4.2 Evidence Contract and Grounding

The intelligence layer enforces a typed evidence contract distinguishing two output classes. **Citable passages** [N] resolve to chunk IDs returned by tools in the current session; a structural gate verifies before the response reaches the judge that every [N] resolves to a chunk returned in the current session, failing phantom sources. The gate establishes that the cited passage was retrieved; it does not

check that the surrounding prose is entailed by it. **Typed evidence blocks** render structured values with symbol, date, and provider tag rather than paraphrasing into free text; graph edges are emitted with their confidence score and valid-time interval.

A numeric grounding check compares free-text figures against provider-attributed tool outputs and re-prompts on divergence beyond tolerance. In financial contexts, a confidently wrong figure is more dangerous than a refusal; the contract makes that failure mode structurally detectable.

4.3 Planner Loop and Inter-Tool Reasoning

The planning agent operates in a ReAct-style loop (Yao et al., 2023) — a single agent issuing sequential tool calls, not a multi-agent system: tool results from one modality can directly parameterise queries in another (cross-modal chaining). A graph traversal result can parameterise a structured fundamentals lookup, which in turn scopes a targeted news search. The worked example in Section 5 traces this concretely.

One hard constraint enforces tool coverage: the agent must call at least one tool from each modality family relevant to the question class before generating a response, regardless of accumulated evidence richness.

The planning component is implemented as an interchangeable module: any instruction-tuned language model served through the tool-calling interface can occupy the planning slot without architectural changes. This is a deliberate design choice — these structural guarantees (citation resolvability, numeric grounding) are enforced by the evidence contract and gates, not by any specific model’s grounding discipline. In internal testing across multiple planner configurations, the most common structurally detectable failures were consistent: phantom citation markers ([N] resolving to no retrieved chunk) and numeric claims inconsistent with tool output — both intercepted by the gates before the response reaches the judge.

5 Worked Example

To illustrate the system’s cross-modal reasoning, we trace a representative complex question from the evaluation run (score 97/100, **STRONG**).

Question: “Assess the current state of China-Taiwan geopolitical tension from recent news, and trace through its potential impact on the semiconductor companies in our coverage. Use the supply-

chain graph to identify which companies face direct physical or export-control risk, which face second-order customer risk, and which might benefit from supply-chain diversification moves.”

This question requires three evidence modalities and multi-hop graph reasoning. Figure 3 shows the tool sequence.

```

Step 1 — Export-control news
search_documents(query='China Taiwan export controls semiconductor 2026')
→ [1] "US drafting law to align ally export rules; ASML named as target" (Reuters, 2026-07-14)
→ [2] "ASML: up to 20% of 2026 sales from less-advanced DUV tools" (FT, 2026-07-11)
get_entity_news(entities=['ASML', 'NVDA', 'TSM'])
→ ASML: export-controls clouding outlook; denied selling EUV to China

Step 2 — Knowledge-graph supply-chain traversal
traverse_graph(from='NVDA',
relations=['operates_in_country'], hops=1)
→ NVDA → Taiwan: operates_in_country (conf=0.89)
search_entity_relations(subject='ASML',
predicates=['key_customer'])
→ ASML → TSM: key_customer ("TSMC is ASML's biggest customer", src=[1])

Step 3 — Structured fundamentals and universe screen
get_fundamentals_history_batch(
symbols=[AMD, INTC, MU, NVDA, QCOM], metric='revenue')
→ INTC Q2 2025: $13.65B [3]; revenue histories returned for all five
screen_universe(criteria='US_listed semiconductor') → 11
names in scope

Step 4 — Grounded synthesis
No deterministic gate triggered. All news citations resolve to stored
passages; KG edges carry confidence scores and source provenance.
Verdict: STRONG (97/100).

```

Figure 3: Tool trace for the China-Taiwan geopolitical question. KG edges and news passages jointly enable the three-tier risk map in Figure 4.

6 Evaluation

Table 1 summarises the fixed experimental parameters. Eight deterministic gates pre-filter responses before the LLM judge; a ninth grounding threshold is applied to the numeric-grounding sub-score. Any gate failure sets the score to zero regardless of fluency.

6.1 Entity Description Grounding

As a construction-time hallucination mitigation measure, entity descriptions are generated from retrieved news passages rather than from parametric model memory. To measure the effect, we drew a stratified sample of entity descriptions (weighted toward thinly-covered and provisional entities, where the risk is greatest), generated each description both with and without retrieved-passage grounding, and labeled each output claim as *verifiable* (entailed by the retrieved passage or checkable against a known provider value) or *unverifiable* (not traceable to

System response (abridged):

“Direct risk: The knowledge graph shows **NVIDIA** holds an `operates_in_country` edge to Taiwan (conf 0.89), making it the only coverage name with direct physical/export-control exposure.

Second-order risk: The graph identifies **ASML** as TSMC’s primary equipment vendor [1]. US legislation targeting ASML’s tool exports [1] would constrain the foundry capacity that **AMD**, **Intel**, **Qualcomm**, **Micron**, and **NVIDIA** all depend on — a second-order risk propagated through the KG supply-chain edges.

Diversification beneficiaries: **Intel**, ramping US fabs (\$13.65B revenue, Q2 2025 [3]), is positioned to capture orders from designers seeking non-Taiwan capacity. **GlobalFoundries** and **ON Semiconductor** (US/EU fabs) are natural alternatives.

Summary: direct risk NVDA; second-order risk ASML + all Taiwan-dependent fabless designers; beneficiaries Intel, GFS, ON. All claims are grounded in KG edges, retrieved news, and reported fundamentals.”

Figure 4: Abridged system response (score 97/100). Bracketed numbers ([1], [3]) resolve to stored passages and Intel’s revenue ([3]) is directly retrieved, but **GlobalFoundries** and **ON Semiconductor** are planner-inferred from the `screen_universe` result and carry no citation; they are uncited prose and therefore outside the citation gate’s scope. The answer’s closing sentence (“All claims are grounded. . .”) is the model’s own self-assessment, not a verified property: the gates establish that every citation resolves and that checked figures match tool output, not that every statement is entailed by retrieved evidence. The benchmark judge evaluates the final answer under the grounding rubric.

any retrieved source). Labeling was performed by the system developer; all discards were conservatively marked unverifiable on doubt. This grounding approach reduced the per-description fabrication rate from 1.83 unverifiable claims to 0.17 — a 91% reduction — compared to ungrounded generation. The effect is concentrated on obscure entities: for widely covered issuers the ungrounded and grounded descriptions largely agree, whereas for thinly covered or provisional entities the ungrounded model reliably invents plausible sector, headquarters, and relationship attributes. This is the population that matters most, because those entities are precisely the ones a user cannot verify at a glance.

A methodological note: fabrication rate alone is not a sufficient objective. A candidate generator with near-zero fabrication emitted empty output for most entities at the operating threshold — a degenerate solution an unguarded metric rewards. We therefore score fabrication jointly with a non-empty-output requirement, the same principle behind the empty-response gate in §6.7. Single-annotator labeling is a limitation; the result should be treated as an indicative estimate rather than a calibrated measurement.

Table 1: Experimental setup.

Component	Parameter
<i>Model configuration (generative components via DeepInfra; temperature 0.0)</i>	
Planner	DeepSeek-R1-Distill-Qwen-32B; max 8 tool rounds
Relation extraction	openai/gpt-oss-120b (via DeepInfra)
Entity descriptions	Qwen3-235B-A22B-Instruct-2507 (via DeepInfra)
Judge	DeepSeek-V4-Flash (via DeepInfra)
Query embeddings	BAAI/bge-large-en-v1.5 (1024-dim, top-20 retrieval)
Gate encoder	all-mpnet-base-v2 (768-dim, local inference)
NER	GLiNER (Zarathiana et al., 2024), 11 entity classes (local inference)
Numeric tolerance	$\pm 1\%$ relative; ± 0.01 absolute for ratios < 1
<i>Routing gate</i>	
Cross-validation	5-fold stratified random (no temporal grouping; within-sample generalisation)
Training documents	15,006; threshold $\tau = 0.58$ (Youden-J, in-sample)
<i>LLM judge</i>	
Score bands	STRONG ≥ 90 / PASS ≥ 75 / WEAK ≥ 60 / FAIL < 60
Structural checks	8 pre-judge gates + 1 post-judge threshold; any failure $\Rightarrow$ score 0
<i>Benchmark</i>	
Questions	284 evaluated (300 authored; 15 safety excluded; 1 removed)
Protocol	3 independent runs; majority vote 2-of-3; 50-item calibration set

6.2 Routing Gate Performance

The learned relevance gate uses five-fold stratified CV on 15,006 documents (label: at least one accepted structured artefact — typed relation, claim, or event — from that document). CV uses random document-level stratification without temporal grouping; metrics reflect within-sample generalisation only (see §7 for the live deployment gap). Features: a semantic embedding of title and subtitle concatenated with three structured signals (source reliability, document type, recency).

A gradient boosting classifier improves ROC-AUC from 0.639 to 0.828 ± 0.008 (Brier $0.249 \rightarrow 0.151 \pm 0.003$; prevalence 62.5%). The operating point routes $\sim 75\%$ of documents to deep extraction while retaining $\sim 94\%$ of yield-positive documents — 25% fewer LLM calls at 6% recall cost. Threshold $\tau = 0.58$ (Youden-J, in-sample) is an operational routing parameter, not a generalisation estimate.

Table 2: Quantitative results.

Metric	Result
<i>Relevance gate — 5-fold cross-validation</i>	
Learned gate vs. heuristic baseline	ROC-AUC: 0.639 $\rightarrow$ 0.828 ; Brier: 0.249 $\rightarrow$ 0.151
Operating point (OOF curve)	$\sim 75\%$ docs routed to deep extraction; $\sim 94\%$ of yield-positive documents retained
<i>Response quality — 284 questions $\times$ 3 runs</i>	
Majority-vote verdicts	205 strong · 20 pass · 26 weak · 23 fail (10 no market data)
Majority-vote non-failing rate	251/284 = 88.4% (end-to-end); 251/274 = 91.6% (absent-data questions excluded)
Judge–annotator agreement (50-item calibration set)	$\kappa = 0.860$ (almost perfect, Landis–Koch); raw agreement 94.0%; precision 87.5%, recall 93.3%
<i>Multimodal ablation — 24 cross-modal questions (fixed pre-expansion) $\times$ 3 runs (pooled mean)</i>	
Full system vs. text-retrieval only (restricted access)	94.8 vs. 46.4 (mean judge score 0–100)
Full system vs. –KG (graph tools disabled)	94.8 $\rightarrow$ 68.3
Full system vs. –Structured (numeric lookup disabled)	94.8 $\rightarrow$ 62.9
Full system vs. –State (portfolio / market-state disabled)	94.8 $\rightarrow$ 75.8

6.3 Benchmark Design

The 300-question benchmark spans nine strata: tool coverage (48), retrieval and aggregation (46), cross-modal synthesis (30), date-anchored (29), safety and adversarial (30), deep investigation (39), hypothetical projection (38), prediction market analysis (28), supplementary probes (12). Questions were authored by the system developer to cover failure modes, checked against the deployment-period data snapshot, and executed three independent times without fixing a random seed. A run passes when the judge score is ≥ 60 ; majority vote requires at least 2 of 3 runs to pass. All three runs agree for 198/284 questions (69.7%); disagreement concentrates in hypothetical and long-chain strata.

6.4 Results

Table 2 summarises the full results. Individual run scores exhibit planner non-determinism despite temperature = 0.0 (Table 1), consistent with provider-side batch ordering effects, motivating the multi-run majority-vote protocol.

6.5 Multimodal Ablation Analysis

Table 2 quantifies per-modality tool-access costs. Disabling graph tools ($-KG$) drops the mean score from 94.8 to 68.3; disabling structured numeric lookups ($-Structured$) to 62.9; disabling portfolio and market-state tools ($-State$) to 75.8. The ordering $-Structured < -KG < -State$ held consistently in aggregate across runs. The gap conflates information removal with representation choice; an information-equivalent flattened-evidence baseline is left for future work.

6.6 Failure Analysis

Twenty-three questions fail under majority vote. Failures cluster into five trace-derived categories: **temporal mapping** (5 cases, 2025 dates misclassified as future); **tool routing** (5, wrong tool family or omission); **result misinterpretation** (4); **multi-step abandonment** (4, gives up after one empty call); and **local data unavailability** (5, prediction-market absence producing full refusals). None of the twenty-three failures was caused by an unresolvable citation or by a numeric claim inconsistent with tool output: the structural gates intercept that class upstream. This is evidence about the two properties the gates check, not about semantic entailment of uncited prose, which the gates do not observe; within that scope the residual failures are planning failures rather than citation or numeric grounding failures.

6.7 Evaluation Framework Reliability

Eight pre-judge gates (regex and floating-point comparisons covering response structure, citation provenance, numerical grounding, required tool use, unsupported numeric claims, phantom sources, empty responses after tool calls, and control-token leaks) are evaluated before the LLM judge; a ninth grounding floor threshold is applied to the judge’s numeric sub-score. Any failure sets the score to zero regardless of fluency, preventing the evaluation from rewarding structurally incorrect but well-written responses.

Cohen’s κ over a 50-item calibration set is 0.860 (95% CI 0.71–1.00, asymptotic SE; the lower bound falls in the Landis–Koch *substantial* range, reflecting the limited sample size). The annotator is the system developer; independence is therefore limited. Table 3 shows the confusion matrix: FAIL-as-positive precision 87.5% (14/16), recall 93.3% (14/15). Three conflicts: two false-fails (numeric

Table 3: Judge calibration confusion matrix (human vs. automatic, 50-item set).

	Judge PASS	Judge FAIL
Human PASS	33	2
Human FAIL	1	14

omission) and one false-pass (correct premise with no tool call).

7 Deployment and Operational Findings

The system operated in continuous deployment throughout the evaluation period. *Deployed* means a continuously running public service connected to live ingestion sources; it does not imply multi-user scale or high-availability guarantees. Findings are operational observations relevant to the paper’s claims.

NLP throughput and gate savings. In live deployment, 49% of routed documents are directed to the embedding-only path (no LLM extraction), versus the 25% out-of-fold estimate. The gap is a source-mix effect rather than model drift, and argues for reporting cost-saving claims against live traffic rather than held-out folds.

Knowledge graph composition. The confidence model correctly separates speculative prediction-market extractions from more established relations; the decay-class distribution reflects medium-duration financial relationships, confirming the taxonomy is doing differentiating work. Raw stored-graph precision is not independently evaluated; the dominant observed defect class is co-mention pairs promoted as typed relations, which the confidence model partially suppresses but does not eliminate. The inference-time citation mechanism limits the impact of raw KG errors by requiring cited claims to resolve to retrieved chunks and by checking free-text numerical claims against tool outputs; it does not prevent unsupported uncited prose (§8).

Failure modes in operation. The 23 benchmark failures correspond to identifiable planner limitations: temporal mapping errors (treating 2025 dates as future), tool routing mismatches, result misinterpretation on successful tool calls, multi-step abandonment on empty results, and soft-refusal failures when prediction market data is absent. Manual inspection of failed traces indicates that relevant evidence was present in tool outputs in most cases; the immediate failure occurs during planning or synthesis rather than retrieval.

8 Discussion

On the multimodal architecture. The ablation (94.8 $\rightarrow$ 46.4) shows that questions requiring graph relations, exact market values, or portfolio state score poorly when those data are withheld. The failure mode is not refusal: the restricted system produces fluent, confident responses with plausible but incorrect figures — more dangerous in financial contexts. A flattened-evidence baseline (serialising structured data to text) is a needed follow-on to isolate representation choice from data availability.

On hallucination mitigation. Construction-time grounding (descriptions from retrieved passages) and inference-time citation enforcement are both necessary: an ungrounded substrate makes downstream citations resolve to fabrications; a grounded substrate without enforcement still permits the agent to assert figures no tool returned. The combined two-level approach reflects that hallucination mitigation in financial systems requires both proactive and reactive mechanisms.

On the evaluation framework. The gate framework encodes a claim: a response citing an un-retrieved source or asserting a figure inconsistent with tool output is worse than a refusal. Without structural pre-filtering, the judge alone passed 3 of 8 human-labelled fabrication cases (pre-gate scores 85, 80, 85); the combined pipeline detects all eight of these cases. The gates depend only on a tool-call trace and response, making the framework portable beyond this system.

9 Conclusion

We have presented a deployed financial NLP system that addresses the multimodal grounding problem through continuous KG construction, typed multimodal tool access, and a two-level hallucination mitigation strategy. The four contributions — a continuously updated financial KG with per-edge provenance and temporal decay; a learned relevance gate formalising selective extraction; citation grounding at both construction and inference time; and a deterministic structural evaluation framework — are evaluated through complementary offline, end-to-end, calibration, and ablation experiments.

Two findings are of general interest to the financial NLP community. First, grounding entity descriptions in retrieved passages reduced fabrication from 1.83 to 0.17 unverifiable claims per descrip-

tion, indicating that a substantial share of hallucination in financial KG construction is addressable by construction-time retrieval. Second, the structural evaluation gates reach $\kappa = 0.860$ with a human annotator and catch fabrication cases that the LLM judge alone rates in the 80–85 band, suggesting that financial QA evaluation should treat structural verification as a prerequisite to LLM judging.

The benchmark questions and evaluation scripts are available at <https://github.com/arnaurodondev/WorldView>.

Limitations

The structural gates check two decidable properties — that every citation resolves to a passage retrieved in the same session, and that free-text figures are consistent with provider-attributed tool outputs — and not semantic entailment. A sentence that is fluent, uncited, and unsupported by any retrieved passage passes every gate; the worked example in Figure 4 contains such a sentence. The defensible claim is therefore that all cited passages resolve to retrieved evidence and all checked numerical claims are consistent with provider-attributed tool outputs, not that every generated statement is entailed by the evidence. An entailment-level verifier is the natural next layer. The system is deployed on a single Kubernetes node; multi-user scaling is untested. KG confidence weights are not calibrated against an external annotation set. The benchmark was authored by the system developer, introducing potential coverage bias toward exercised tool paths; independent annotation is planned. Residual planner variance motivates the multi-run majority-vote protocol. The entity-description fabrication comparison is a sample evaluation rather than a full-corpus audit, labeled by a single annotator (the system developer). The routing gate uses stratified random CV without temporal grouping; the 49% live routing rate versus the 25% out-of-fold estimate indicates source-mix shift, and prospective performance on unseen time periods is not characterised. NER quality and entity resolution accuracy are not evaluated on any held-out labelled set. Raw stored-graph precision is not independently evaluated (§7); a flattened-evidence ablation is needed to separate representation choice from data availability in the ablation results.

References

- Abhinav Arun, Fabrizio Dimino, Tejas Prakash Agarwal, Bhaskarjit Sarmah, and Stefano Pasquali. 2025. FinReflectKG: Agentic Construction and Evaluation of Financial Knowledge Graphs. In *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF)*.
- Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of EMNLP*.
- Chanyeol Choi, Jihoon Kwon, Alejandro Lopez-Lira, Chaewoon Kim, Minjae Kim, Juneha Hwang, Jaeseon Ha, Hojun Choi, Suyeol Yun, Yongjin Kim, and Yongjae Lee. 2025. FinAgentBench: A Benchmark Dataset for Agentic Retrieval in Financial Question Answering. In *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF)*.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of SIGIR*.
- Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of ACL*.
- Joohyun Lee and Sungwoo Hong. 2026. Hierarchical Reranking for Scalable Financial RAG System. arXiv:2607.27523.
- Yixin Lu and Yintong Huo. 2025. Financial Named Entity Recognition: How Far Can LLM Go? In *Proceedings of the Joint Workshop on Financial Technology and Natural Language Processing (FinNLP)*.
- Yury A. Malkov and Dmitry A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4).
- Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stambach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools. In *Proceedings of EMNLP: System Demonstrations*.
- Kun Ouyang, Yi Liu, Shicheng Li, Ruihan Bao, Keiko Harimoto, and Xu Sun. 2024. Modal-adaptive Knowledge-enhanced Graph-based Financial Prediction from Monetary Policy Conference Calls with LLM. In *Proceedings of the Joint Workshop on Financial Technology and Natural Language Processing (FinNLP-KDF-ECONLP 2024)*.
- Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4).
- Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Proceedings of NAACL*.
- Bhaskarjit Sarmah, Dhruv Mehta, Brian Hall, Rik Rao, Sunil Patel, and Stefano Pasquali. 2024. HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction. In *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF)*.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yiqing Shen, Jingshu Zhang, Feng Chen, Kaiyuan Yan, and Hongguang Li. 2025. FinSearch: A Temporal-Aware Search Agent Framework for Real-Time Financial Information Retrieval with Large Language Models. In *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF)*.
- Neelesh Kumar Shukla, Prabhat Prabhakar, Sakthivel Thangaraj, Sandeep Singh, Weiye Sun, C. Prasanna Venkatesan, and Viji Krishnamurthy. 2025. GraphRAG Analysis for Financial Narrative Summarization and a Framework for Optimizing Domain Adaptation. In *Proceedings of the Joint Workshop on Financial Technology and Natural Language Processing (FinNLP)*.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjana Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564.
- Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. In *FinLLM Symposium at IJCAI*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of ICLR*.
- Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *Proceedings of NAACL*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In *Proceedings of ACL*.