



Yiming Lu • Yanxin Yang • Supervised by Dr. Ali Hirsra, Will Duan • Sponsored by Wellington Management Company
Columbia Engineering Department of Industrial Engineering and Operations Research

News Agent

The contract both agents obey.

No agent may write a valuation input. Each one emits a *request* against one specific valuation state, identified by a SHA-256 snapshot hash; the sensitivity agent refuses any request whose hash is stale or missing. Every number carries the quote it came from, the page it was on, and the gate decisions it survived.

```
graph LR; VE[Valuation engine] --> S3S[Stage-3 snapshot]; S3S --> AR[Analyst Report Agent]; S3S --> NA[News Agent]; AR --> SA[Sensitivity agent]; NA --> SA; subgraph RuleBox [No agent may change a valuation input]; end
```

Valuation engine
SEC XBRL + investor relations + FRED + Damodaran betas
-> FCFF DCF -> four-stage beta pipeline

Stage-3 snapshot
market D/E, forward levered beta
the one state everything is measured against

Analyst Report Agent
broker PDFs -> filter
-> BM25 (+dense) -> RRF
-> sentence & table readers
-> 3 gates -> request AMZN and XOM

News Agent
as-of control -> entities
-> relevance -> event clusters
-> two-pass LLM features
-> human review -> features AMZN

Sensitivity agent
executes only requests whose expected snapshot hash matches the snapshot
one-way, two-way, scenarios

No agent may change a valuation input

figure 1. Where the two agents attach to the shared valuation engine. Both read the Stage-3 snapshot and both hand their output to the same executor.

- **Fundamentals.** SEC XBRL keyed on period end, IR releases, FRED series, Damodaran betas. AMZN and XOM.
- **Analyst reports.** Four sell-side PDFs, June 2026: two JPMorgan on Amazon, JPMorgan + RBC on ExxonMobil. 85 and 51 chunks.
- **News.** 17 articles, Jan–Aug 2026 — five issuer disclosures, twelve independent-media records, seven publishers.

Everything runs offline from committed files: 444 tests, no network, no key.

- [1] Samuelsen, D., Nyström, K., Mazumdar, S., Hussain, A., & Strange, R. (2026). MimirRAG: A Multi-Agent RAG Framework for Financial Data Retrieval with Metadata Integration. *arXiv:2605.25030*.
- [2] Damodaran, A. *Investment Valuation*, 3rd ed., ch. 14–15 (FCFF valuation; bottom-up betas).
- [3] Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *FTIR*, 3(4), 333–389.
- [4] Cormack, G. V., Clarke, C. L. A., & Büttcher, S. (2009). Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. *SIGIR*.
- [5] Hamada, R. S. (1972). The Effect of the Firm's Capital Structure on the Systematic Risk of Common Stocks. *J. Finance*, 27(2), 435–452.

Reproducibility. Both agents ship as one Code/ folder — `analyst_report_agent` (58 tests, 6 notebooks, AMZN and XOM from one `--ticker` flag) and `news_agent` (386 tests, 5 notebooks). Every gate writes an immutable run directory with a manifest recording its input hashes, its rule file and its counts, so any figure on this poster can be traced back to the raw article or PDF page that produced it. Gate 3 of the news agent ships one complete recorded language-model run, so its behaviour can be re-checked with no API key.

A simplified MimirRAG [1]: the Planner and Writer are removed – the questions are known in advance and the answer must be a validated number – and the retrieval spine is kept (metadata filter, BM25, optional dense, RRF).

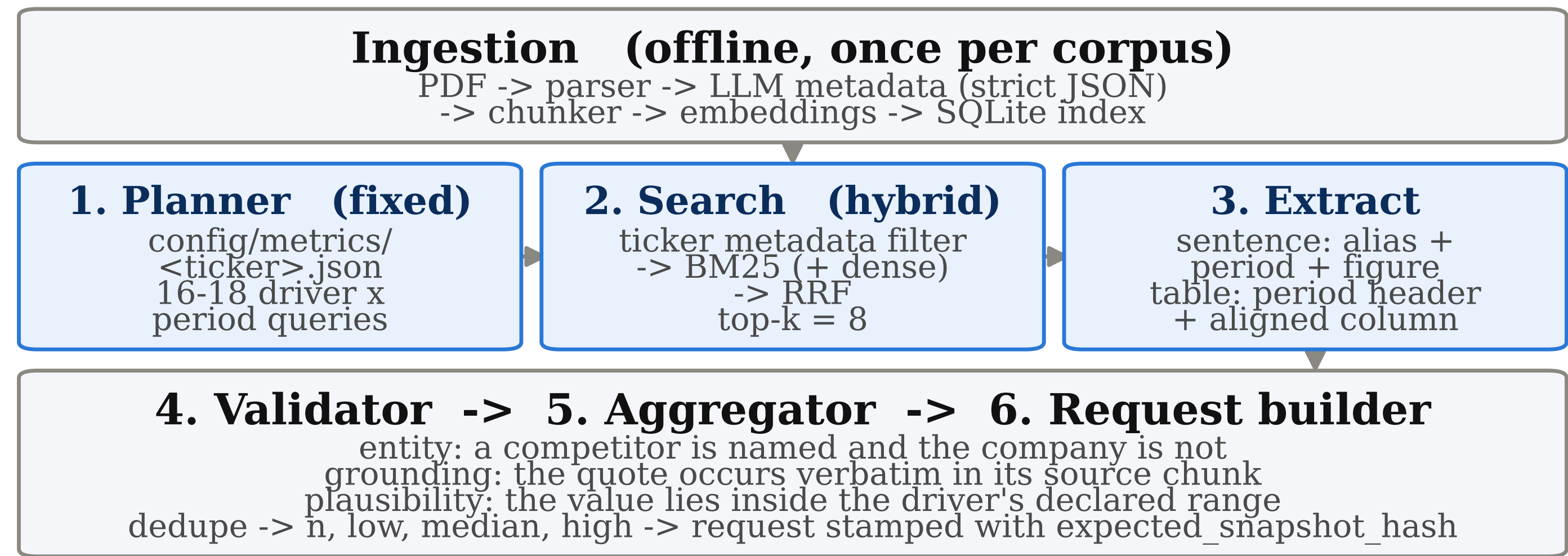


Figure 2. Six stages. Ingestion runs once per corpus; stages 1–6 run per query.

Two extractors, because financial PDFs are not prose. The sentence reader wants an alias, a period token and a figure within 90 characters; the table reader locks onto a period header row and reads column-aligned cells. Three gates then run in local code, after the model: *entity*, *grounding*, *plausibility*.

Amazon

Metric	Value
Chunks in corpus	85
Retrieved by a query	20
Candidate figures	20
Survive all 3 gates	17
Distinct estimates	5
Map to a DCF input	4

2 JPMorgan reports

ExxonMobil

Metric	Value
Chunks in corpus	51
Retrieved by a query	24
Candidate figures	3
Survive all 3 gates	1
Distinct estimates	1
Map to a DCF input	0

JPMorgan + RBC

Extractor ablation

Category	figures found	accepted	drivers evidenced
sentence only	8	5	1
table only	12	12	2
both (shipped)	20	17	3

Figure 3. Evidence funnel for both corpora, and the extraction ablation. Amazon’s 17 accepted figures deduplicate to 5 distinct estimates because the same summary table recurs across pages and reports.

Driver	Period	Value	Maps to
Revenue growth	FY2026	14.9%	initial_revenue_growth
Revenue growth	FY2027	13.4%	initial_revenue_growth
Operating margin	FY2026	15.3%	target_operating_margin
Operating margin	FY2027	17.0%	target_operating_margin
Price target	—	\$330	evidence only

Table 1. Every row is $n=1$ and flagged `single_publisher` – one house’s view, not a consensus.

The finding worth sitting with. Those inputs lift the Stage-3 baseline from \$115.03 to \$139.43 / \$150.11 / \$161.37. So JPMorgan's own forecasts, through our DCF, reach \$161.37 — while the same report carries a \$330 price target and the stock trades at \$240.23. The residual is *not* in the inputs the agent can evidence: our driver ranking puts the equity risk premium (40.9% of baseline value) and levered beta (35.1%) above revenue growth (19.0%), and no analyst publishes an equity risk premium.

Ablation. Sentence extraction alone finds 8 figures covering one driver – the price target, the one the DCF may not consume. The table reader takes it to 20 figures and 3 drivers.

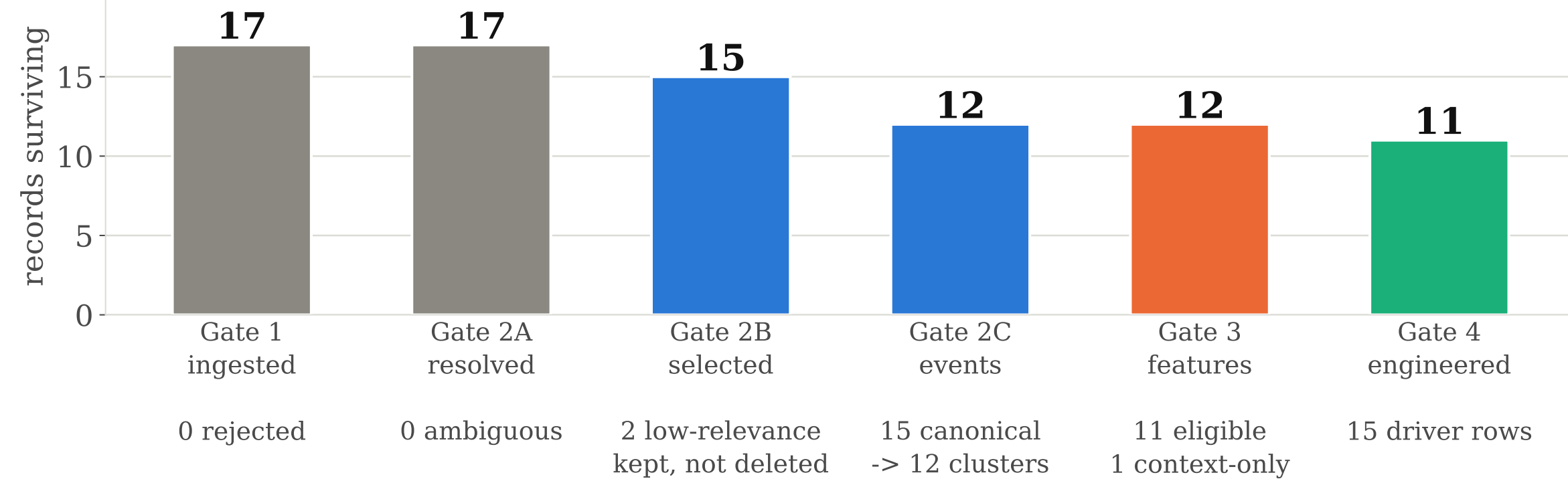


Figure 4. What survives each gate. Nothing is deleted: low-relevance articles and duplicates are kept in audit files. Local code verifies every Gate-3 numeric quote verbatim against its source.

Point in time. Availability is judged on `retrieved_at`, not the publisher's date stamp: the same corpus yields 5 accepted at 23:59:59Z and 0 one second before its collection instant.

One event, one signal. Clustering collapses 15 canonical articles into 12 events before any model sees them so coverage volume cannot masquerade as significance.

Figure 5. Gate 4: a point-in-time daily panel over 205 days. No market returns are used anywhere.

Human review is not optional. Every Gate 3 event carries `review_required=true`. A reviewer signs each row off; applying the sheet writes a new append-only run, and Gate 4 in `approved_only` mode consumes only those.

- **The entity gate never fired.** Zero rejections on both corpora; every rejection came from plausibility. The rule does block 22 of 24 Woodside sentences when applied directly — but an upstream recall limit means those never became candidates.
- **Right answer, wrong reason.** The one Woodside figure to reach a gate, an A\$35 price target, was stopped only because A\$35 is below a USD lower bound — currency filtering by accident.
- **The driver list binds harder than the corpus.** Adding `cost_of_capital` to the XOM spec recovered a 10% discount rate that was there all along.

Next steps

- Broaden recall, *then* re-test the entity gate.
- Resolve currency and units, so a foreign figure is a typed value rather than a failed range check.
- Grow the benchmark past the 12 events the prompts were tuned on.