Mosaic Theory at Scale

A controlled architecture for cross-company equity research with language models

Alistair Smallwood, Head of Applied AI at Primer

ABSTRACT

Public companies disclose far more information than an analyst team can systematically analyse across the market. The bottleneck is the cost of generating, testing and rejecting large numbers of possible cross-company hypotheses, each of which requires tracing a defensible chain of inference from one company's disclosures to another. This paper presents an architecture for using language models to conduct that process at a scale that would be impractical for a conventional analyst team. The system separates evidence retrieval from interpretation and evaluation, requires explicit testing of economic mechanism, timing and comparability, and preserves both supporting and contradictory evidence. Reusable sector maps retain verified relationships and reporting definitions, while matched evaluation compares judgements made from target-company information alone with the same judgements after cross-company evidence is added. Probabilistic outputs make the incremental value of that evidence directly scoreable once outcomes are known.

Initial case studies show both useful and spurious relationships. A physical-flow series for Dover's SWEP business moved ahead of the broader segment recovery; a three-indicator semiconductor signal matched 21 of 22 subsequent Pall disclosures, compared with 18 of 22 using Pall's prior direction alone; and product-level supply-chain evidence provided an independent read on activity at Dover's CPC business. The same semiconductor rule failed to improve predictions for Entegris, illustrating why apparently plausible relationships must be tested rather than merely explained.

These results are preliminary and retrospective. They show how language models can make large-scale mosaic research tractable while preserving the provenance, inference controls and falsification needed to determine whether cross-company evidence actually improves investment judgement.

1 Introduction

Most uses of AI in equity research are designed to help an analyst complete familiar tasks more quickly, whether that means searching filings, summarising earnings calls, updating financial models or preparing a first draft of a research note. These applications reduce the cost of baseline research and leave analysts with more time for judgement, but the shape of the research process remains largely unchanged because the analyst still decides which companies matter, which questions to ask and where to look for evidence.

This paper examines a different use of AI: expanding the set of investigations that an investment firm can conduct in the first place. Language models provide enough reading capacity, memory and analytical attention to examine relationships across thousands of companies, making it practical to search suppliers, customers, competitors, distributors and adjacent markets for evidence that may change the interpretation of a target company.

We describe this approach as mosaic theory at scale. Mosaic theory is the practice of combining public information that is individually incomplete or non-material to form a material investment conclusion; for example, a supplier may report that orders are normalising while a customer says channel inventory has returned to an appropriate level and a competitor becomes more confident about pricing. None of these observations settles the investment question by itself, but together they may reveal a change that is not yet visible in the target company's reported results.

The opportunity is therefore larger than finding more correlations. A useful research system must establish what was actually observed, explain how it could affect the target, test whether the timing and financial definitions are comparable, preserve evidence that supports a different account, and distinguish several independent observations from several companies repeating versions of the same underlying fact.

Those controls become more important as the search space grows, because thousands of companies will produce a large supply of apparent relationships and only some of them will be economically meaningful. Language models are unusually good at turning weak evidence into coherent explanations, so a system that increases the scale of reading without controlling the chain of inference may produce plausible stories at scale rather than better research.

We therefore treat large-scale mosaic research as a system-design and evaluation problem. GPT-6-Astra drives the investigation through versioned research skills that define how evidence should be retrieved, interpreted and challenged; Primer's document tools provide access to company source material and high-quality retrieval; Jev converts the final judgement into a constrained probability distribution that can be stored and scored once the outcome is known; and a machine-readable relational map preserves verified relationships, rejected hypotheses and unresolved questions for later investigations.

The central question is whether adding cross-company evidence improves the judgement that could have been made from the target company's information alone. The evaluation compares two otherwise identical cases, one containing only target-company information and the other adding a frozen packet of cross-company evidence, so that the system can measure whether the mosaic improved the judgement, made no useful difference or encouraged a spurious relationship.

The paper does not claim that this initial work establishes persistent forecasting performance or investment alpha. It sets out an architecture through which those claims can be tested properly and reports initial findings that include both potentially useful cross-company relationships and relationships that appeared plausible but did not survive closer examination.

2 Research question and scope

In investment practice, mosaic theory permits an analyst to form a material conclusion by combining information that is both public and individually non-material.[1] At scale, the same method can combine observations from several companies to revise a view of one target company.

Language models alter the feasible search space. A system can inspect many more filings, calls and company relationships than an analyst could examine manually. That advantage carries a familiar research risk. Once the number of possible comparisons becomes large, coincidences are plentiful and persuasive explanations are easy to construct after the outcome is known.

The initial casework focuses on the global bioprocessing sector. Its value chain is distributed across specialist equipment suppliers, consumables manufacturers, distributors and drug producers, while the businesses involved experience different demand patterns depending on their exposure to capital investment, production volumes, customer inventory and project timing. This makes bioprocessing a useful setting for cross-company research: related signals appear in different companies and at different points in the operating cycle, but cannot be treated as interchangeable. We use the sector as a demanding initial test of the method, rather than assuming that the findings automatically generalise to other industries.

3 System architecture and control points

The architecture reflects the distinct requirements of the research process. The first is reliable access to source material, so we use Primer's document tools to search a structured universe of company filings and transcripts directly rather than relying on open-web search. Document Search identifies the relevant company documents, while Document Reduce locates and compares the passages required to answer a specific question.

The retrieved evidence is then passed through versioned research skills, which specify how different types of evidence could be interpreted and compared. These skills encode guidance and guardrails that a general-purpose model might otherwise apply inconsistently, including how to preserve publication dates, test whether financial measures are comparable and retain evidence that contradicts the proposed relationship.

The reasoning is performed by GPT-6-Astra through the Codex harness. Astra was selected for its ability to reason across large volumes of mixed financial evidence; Codex provides reliable tool calling, computer use and web search when external evidence is required.

A separate evaluation layer compares the judgement made from the target company alone with the judgement made after adding cross-company evidence. This is necessary because an economically plausible relationship may still add no useful information. Finally, every investigation updates a machine-readable relational map containing the relationships, definitions, sources and unresolved questions uncovered during the work. That accumulated knowledge can be supplied as context to later investigations, allowing the system to build on previous research while keeping the evidence and status of each relationship visible.

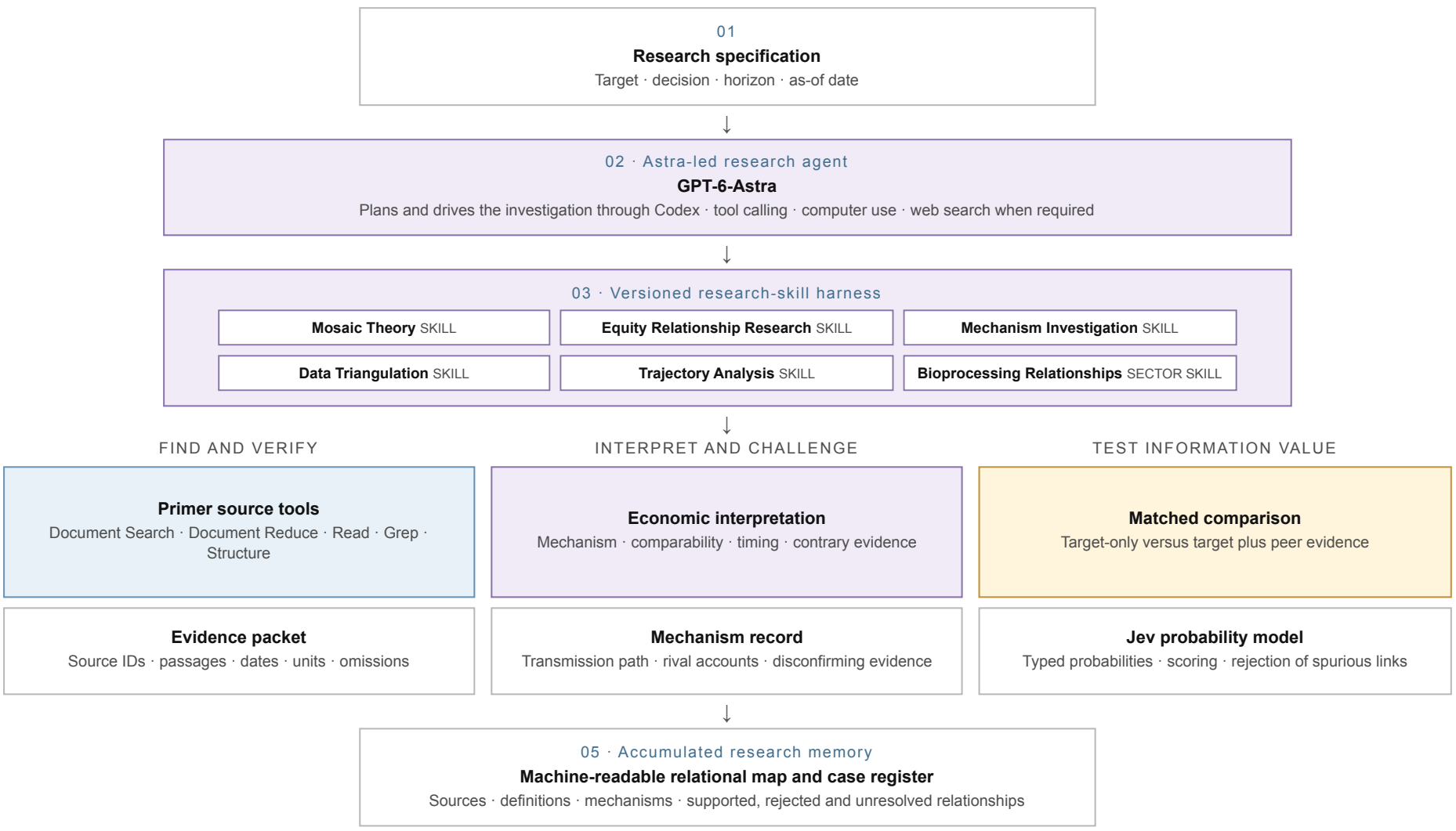


Figure 1. Proposed architecture. GPT-6-Astra drives retrieval, interpretation and evaluation. A separate versioned research-skill layer supplies the method and guardrails. Each stage leaves a distinct record, while the relational map and dated case register preserve reusable knowledge, including rejected relationships.

4 Evidence acquisition and provenance

The following sections examine the architecture in sequence: source-bound evidence acquisition; the skills and research-memory layers; and matched evaluation.

Incorrect retrieval is a potential source of system error in this setting. A passage from the wrong filing can remain semantically plausible. A later report may restate information that was unavailable on the historical decision date. Year-to-date and standalone quarterly values are also easy to confuse when the units and period definitions are separated from the number.

Primer's retrieval tools search a defined universe of company filings and transcripts, surface the relevant sources and compare the passages needed to answer a question. They preserve the original passage, table and surrounding definition, while the resulting evidence packet records the source identity, publication date, period, unit and retrieval path.

Evaluation	Questions	Metric	Result
Primer internal FinRetrieval run	500	Exact-answer accuracy	**100%**

Table 1. Primer's reported result on FinRetrieval.[2] Each of the 500 questions required retrieval of an exact numerical value, its unit and the correct positive or negative sign, making the benchmark a direct test of whether the right financial datapoint was recovered. Primer answered all 500 correctly under the tested configuration.

Correct retrieval establishes what was reported, but not whether it matters to the target. The next step maps the observation to a product or workflow, locates it within the target's reporting boundary, identifies whether it reflects orders, inventory or revenue, and states what should change and when. If any link is assumed rather than supported by evidence, the relationship remains provisional.

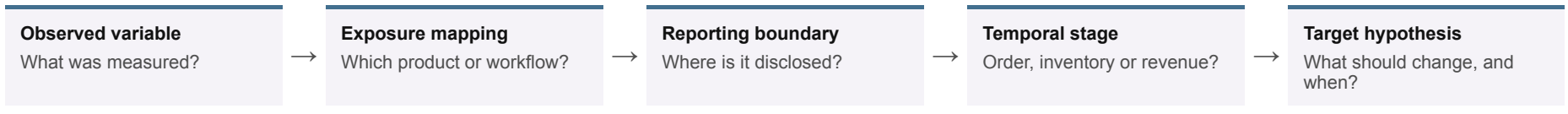


Figure 2. Minimum provenance chain used to connect a retrieved observation to a target-company hypothesis.

5 Method encoding and accumulated research capital

The skills layer is built as a set of version-controlled research protocols rather than one large instruction block. Each skill has a defined scope, required inputs, an ordered procedure, rules for transformations that are not allowed and a specified output record, making the method inspectable and allowing changes to the research procedure to be traced over time. The instructions are operational rather than descriptive: a skill may prohibit converting cumulative percentage growth into a standalone quarterly rate, require the original publication date for each historical fact and force contrary periods to remain in the comparison instead of allowing the model to select only the observations that fit.

Separate skills are used because the analysis can fail at different points, and each failure requires a different control. The mosaic skill asks which uncertainty an observation could reduce, whether it adds independent evidence and which rival explanation might also fit; relationship research then requires an explicit economic path from the observed company to the target, while mechanism investigation identifies the product, workflow, customer or channel through which that path could operate.

Data triangulation checks what a datapoint measures and whether its period, unit and accounting perimeter are comparable, while trajectory analysis preserves quarter-level divergence and prevents the system from choosing a lag simply because it fits the outcome. A sector overlay adds domain-specific questions, including the distinction between capital equipment, recurring consumables, customer inventory and production demand in bioprocessing. This modular design makes omissions visible, so a convincing narrative cannot substitute for a missing source, an unverified exposure or an unexplained timing mismatch.

Skills contain method rather than company conclusions, while company-specific facts remain in dated case records. This separation allows the research procedure to improve across investigations without turning an earlier result into a hidden premise for the next one.

The sector relational map is costly to establish because products, reporting segments, customers and operating stages must be mapped by hand or verified from source documents. Operationally, it is a typed graph in which each edge carries a source, effective date, proposed transmission mechanism, status and condition for reassessment. Later investigations can reuse those edges and definitions. The gain is lower setup cost and a larger set of plausible candidates. Evidential confidence remains case-specific.

Figure 3. The relational map creates a reusable stock of definitions and tested links. Its value can accumulate after the initial mapping work, although the validity of any particular edge remains open to revision.

6 Evaluation design and choice of a probabilistic language model

To test whether cross-company evidence adds value, the model makes the same judgement twice. In the first run, it sees only information about the target company that was public on the decision date. In the second, it receives the same material plus a fixed packet of evidence from related companies. The eventual outcome remains hidden until both forecasts have been recorded. Comparing the two scores then shows whether the additional evidence improved the judgement, made no difference or made it worse.

At this stage, the model must return probabilities for three possible states: improving, stable and deteriorating. We selected Jev 1.13.0 because its TypeSafe Choice interface guarantees a valid distribution across those states. The probabilities can be stored and scored once the outcome is known, whereas free-form prose is harder to compare and can leave the model's confidence unclear.

The model has no role in evidence retrieval or candidate selection during this experiment. Its task is conditional judgement given a fixed packet. The realised-state probability and multiclass Brier loss are recorded for each arm.[3] Directional accuracy is reported separately because a correct class can still be poorly calibrated.

Retrospective model knowledge is an unresolved source of contamination. For that reason, historical runs are treated as development evidence. Prospective, preregistered tests are required before the design can support a claim about forecasting performance.

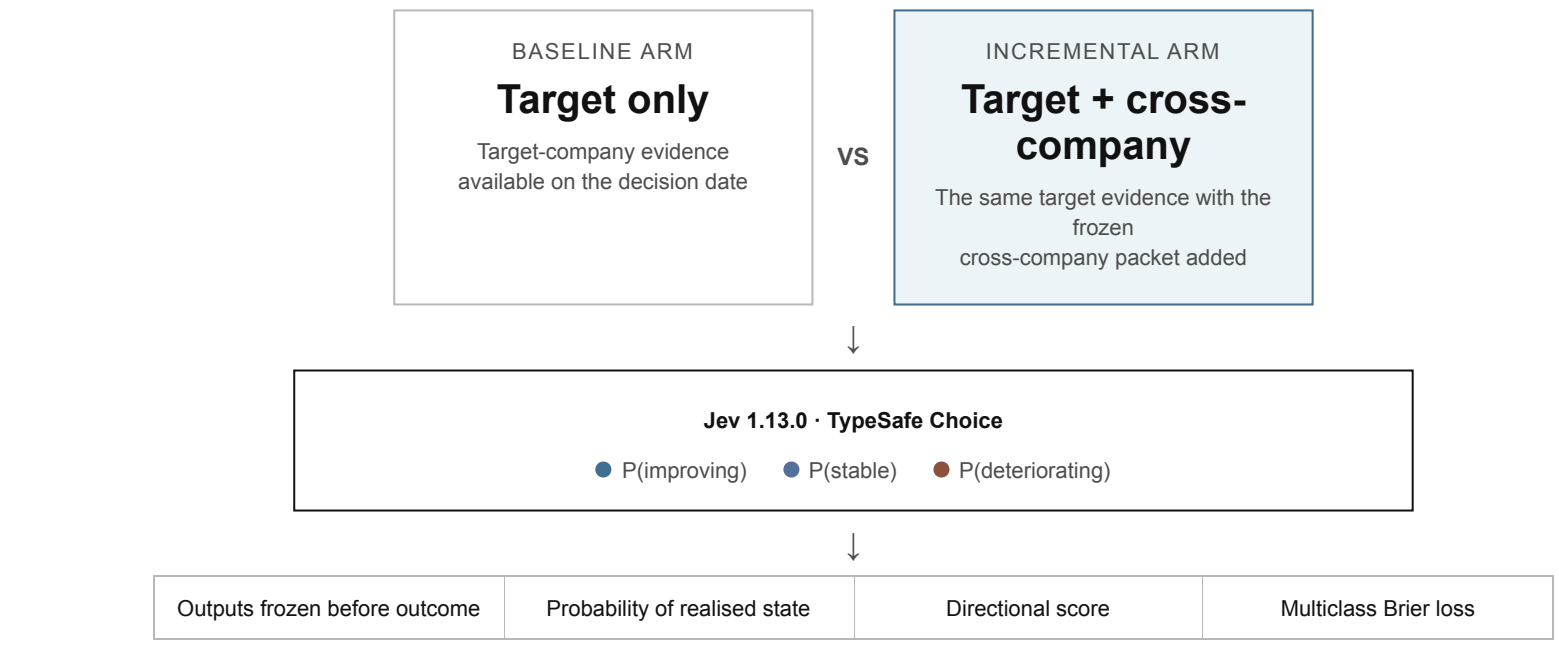


Figure 4. Matched evaluation protocol. The estimand is the change in forecast quality after adding the frozen cross-company packet to an otherwise identical target information set.

7 Example 1: Dover / SWEP heat-exchanger activity

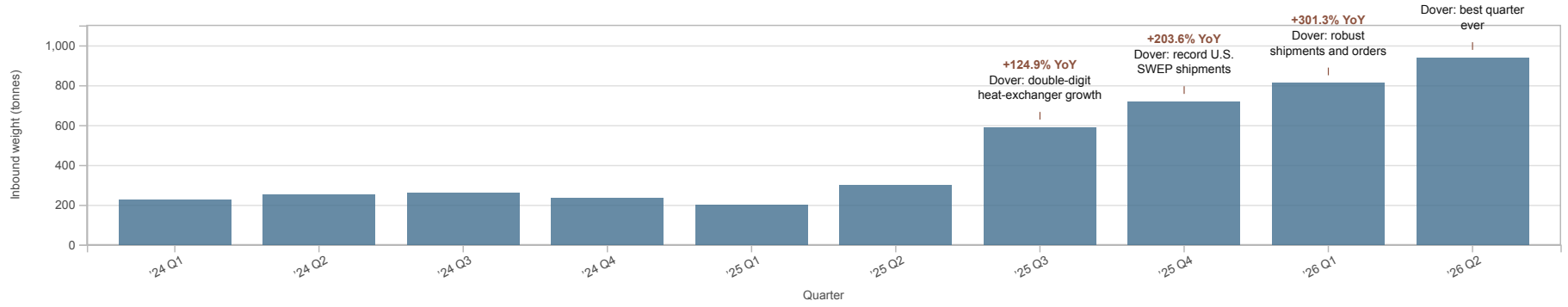
COMPANY	BUSINESS UNIT	PRODUCT
Dover Corporation	SWEP, within Climate & Sustainability Technologies	Brazed-plate heat exchangers, measured through U.S. ocean-inbound weight

INVESTOR INTERPRETATION

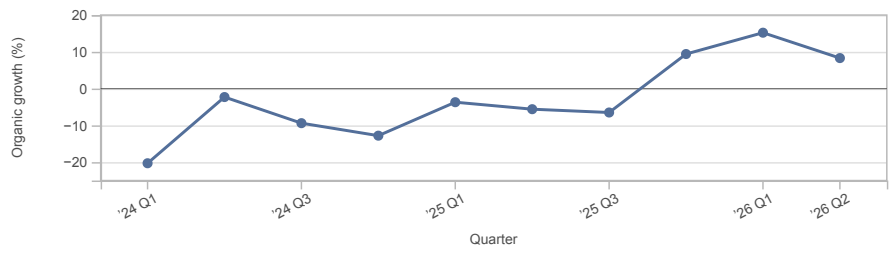
Goods entering SWEP's United States channel accelerated sharply before Dover's broader Climate and Sustainability Technologies segment returned to growth. The series may therefore provide an early read on SWEP itself, but it is not a direct measure of either SWEP revenue or the wider segment.

Public ocean-manifest records provide a partial measure of goods entering SWEP's United States channel. Dover reports SWEP within a much broader segment, which means the manifest series can reveal product-level activity that is not separately disclosed in the segment result.

A · SWEP US inbound weight with contemporaneous Dover commentary



B · DCST reported organic growth



C · SWEP inbound-weight growth

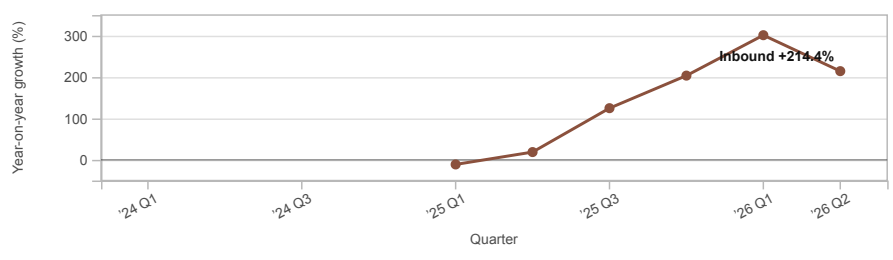


Figure 5. SWEP inbound weight and reported organic growth for Dover's Climate and Sustainability Technologies segment. The two series are not definitionally equivalent. Source: public ocean-manifest records and Dover disclosures; calculations by the author. A final methods appendix should report the manifest provider, shipment-matching rules and revision policy.

What was found. In Q3 2025, SWEP inbound weight increased 124.9% year on year while DCST organic revenue declined 6.5%. Dover was already describing double-digit growth in heat exchangers, and the broader segment returned to growth in Q4. The physical flow data are therefore consistent with an improvement in SWEP becoming visible before the change appeared in the mixed segment result.

Limitation. An inbound shipment is not a customer sale. Inventory policy, freight timing and the gap between an arrival and recognised revenue can all change the apparent lead. Approximately three quarters of the reported segment also falls outside the working SWEP exposure estimate. This is a product-level nowcast candidate that warrants prospective testing, not a mechanical forecast for Dover's segment.

8 Example 2: Danaher / Pall microelectronics filtration

COMPANY	BUSINESS UNIT	PRODUCT
Danaher Corporation	Pall industrial/applied filtration, within Life Sciences	Filtration and purification products used in semiconductor manufacturing

INVESTOR INTERPRETATION

A combination of three semiconductor indicators gave a better historical read on Pall microelectronics than simply carrying forward Danaher's own last disclosed direction. It corrected four target-only calls and made one call worse, but the result remains retrospective and specific to Pall.

What was tested. Before each Danaher report, the rule looked at three observations that describe different stages of semiconductor demand: ASML equipment sales, ASML installed-base and service sales, and the first-published global semiconductor sales figure. A direction was recorded when at least two of the three pointed the same way.

Why it matters. No single indicator has to stand in for the whole market. New equipment reflects fab investment, installed-base activity captures use of equipment already in the field, and semiconductor sales provide a read on end demand and inventory. Agreement across these different clocks supplied information that was not present in Pall's own prior direction.

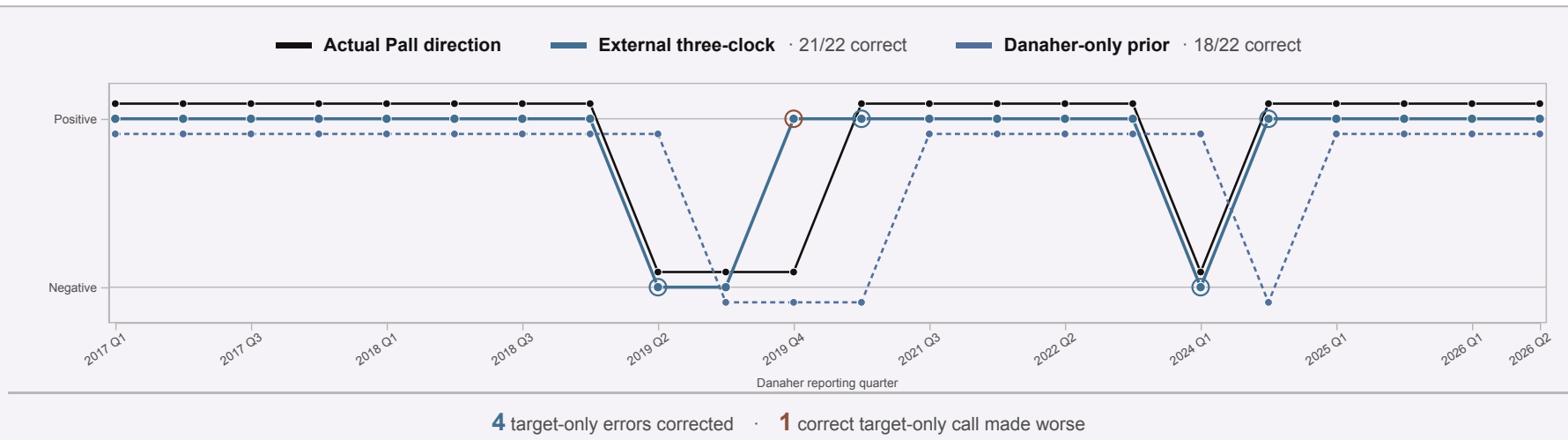


Figure 6. Actual Pall microelectronics direction, the external three-clock call and the Danaher-only prior across the 22 quarters with a strict standalone Pall disclosure. Missing disclosure quarters are not filled. Source: ASML quarterly disclosures, first-published SIA/WSTS global semiconductor sales and Danaher disclosures; calculations by the author.

Limitation. This is model-development evidence rather than a clean prospective test. Sixteen calendar quarters could not be scored because Danaher did not disclose a strict standalone Pall direction, several target labels were qualitative, and the final rule was developed after simpler versions failed. Most importantly, the same rule did not travel to Entegris: it scored 11/13 there, compared with 13/13 for Entegris's own prior. The result supports a Pall-specific monitoring hypothesis, not a universal semiconductor-filtration formula.

9 Example 3: Dover / CPC liquid-cooling connectors

COMPANY	BUSINESS UNIT	PRODUCT
Dover Corporation	CPC, within Pumps & Process Solutions	Quick-connect couplings used in data-centre liquid-cooling systems

INVESTOR INTERPRETATION

Activity at a CPC-linked manufacturer and an independent liquid-cooling peer remained exceptionally strong during Q3. That makes continued positive growth at CPC's thermal-connector business more likely, but it does not reveal CPC's growth rate or predict Dover's wider segment.

CPC makes quick-connect couplings used in data-centre liquid-cooling systems, but Dover reports the business inside its much broader Pumps & Process Solutions segment. The investigation therefore had to establish a product-level route before using activity at another company as evidence about CPC.

Figure 7. The verified product route used to interpret customer and peer activity. NVIDIA also places AVC and CPC in the Blackwell cooling ecosystem, although this does not prove the exact bill of materials in any individual system.

Evidence	Source	Observation	What it adds
Named physical flow	U.S. customs records; AVC Vietnam manufacturing filings	Identified CPC parts shipped into AVC's liquid-cooling production route	Establishes a product-level mechanism
Historical alignment	Customs records; Dover Q1 2025 commentary; AVC thermal-products disclosure	Part shipments rose almost ninefold quarter on quarter; Dover reported thermal-connector growth above 100% and AVC thermal-products revenue grew 75%	Shows the route coincided with strong activity at both ends
Q3 2026 breadth check	AVC and Auras monthly revenue filings	July-August revenue grew 56% at AVC and 91% at independent liquid-cooling supplier Auras; both ran above their Q2 monthly rates	Supports continued positive market activity

Table 2. The CPC mosaic combines an authenticated commercial route, historical operating alignment and an independent market-breadth check. Source: public customs records, AVC manufacturing disclosures, NVIDIA ecosystem disclosures, Dover filings and monthly AVC and Auras revenue reports; calculations by the author.

What was found. The evidence is stronger than a broad comparison between two companies because it begins with identified CPC components moving into a documented liquid-cooling manufacturing process. The historical shipment pulse coincided with strong thermal-connector commentary from Dover, while the later AVC and Auras figures suggest that the surrounding market remained active after Dover's Q2 report.

Limitation. AVC is a manufacturing node rather than a disclosed end customer for every part, and its published revenue covers the whole company rather than purchases from CPC. Customs shipments are batchy and can reflect inventory or contract-manufacturing flows, the Q3 figures exclude September, and Auras reports consolidated revenue without an organic bridge. The evidence modestly raises the probability that CPC growth remained positive; it cannot specify the rate or forecast Dover's overall segment.

10 Limitations

The present evidence is preliminary. The FinRetrieval result concerns exact retrieval and does not establish accuracy in economic interpretation. The three examples also have different evidential status: SWEP is a physical-flow nowcast candidate, Pall is a retrospective backtest, and CPC is an authenticated customer-and-peer mosaic. None is prospective proof or a precise segment forecast.

11 Conclusion

Language models make a larger cross-company search feasible. The proposed architecture places that search inside a controlled research process: source-bound retrieval, explicit economic mapping, versioned method instructions and matched evaluation. The three examples show three distinct outputs from that process: a physical-flow indicator, a historically useful directional rule and an authenticated customer-and-peer mosaic.

The probabilistic model may possess historical knowledge of an outcome even when the supplied packet is clean. Retrospective scores should therefore be treated as development evidence and used to define future tests, not presented as if the model had operated live at the original decision date.

A credible validation programme requires prospective cases with fixed decision dates, preregistered evidence packets and an unchanged scoring rule. Results should include null cases and failed relationships; otherwise, the relational map will accumulate confirmed stories faster than useful negative knowledge.

Every material case should be reviewed by a challenger who checks source identity, measurement, timing, transmission, contrary periods and outcome arithmetic. A relationship may be retained as explanatory without being accepted as predictive, while rejected cases should remain in the register so that the same path is not rediscovered and presented as new evidence.

The map and the skills serve different forms of memory: the map records claims about companies and markets, while the skills record changes to the research procedure. Preserving that boundary makes it possible to improve the method without silently hard-coding past conclusions.

The central empirical question remains open; cross-company evidence has value only when it improves judgement relative to the information already available on the target. The next step is to keep building the relational map by having the system search for, test and record new cross-company relationships. Each investigation adds verified product links, financial definitions, failed hypotheses and unresolved questions that can be reused as context for later work, while matched evaluation tests whether the growing body of evidence improves judgement.

R References

1. CFA Institute. *Standards of Practice Handbook*, 11th ed. Standard II(A): Material Nonpublic Information, p. 62. 2014.
2. Kim, E. Y., and Huang, J. “FinRetrieval: A Benchmark for Financial Data Retrieval by AI Agents.” *arXiv* 2603.04403. 2026.
3. Brier, G. W. “Verification of Forecasts Expressed in Terms of Probability.” *Monthly Weather Review* 78(1): 1-3. 1950. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.