Systematic Alternative Data Research with Language Models

Applying analyst judgement to alternative data in equity research

Alistair Smallwood, Head of Applied AI at Primer

ABSTRACT

Rigorous alternative-data research requires more than comparing the apparent co-movement of two time series. The target must be interrogated as carefully as the external dataset: what each measure captures, which business is exposed, when the effect should appear, and whether reported growth needs adjustment for price, mix, currency, acquisitions or accounting changes. The dataset may also contain revisions, missing observations or shifts in coverage that look like changes in the company. A relationship can therefore look strong on a chart and still be wrong.

This paper presents an autonomous research engine that applies analyst-grade semantic judgement to those problems. Language models direct the investigation through versioned research skills and search for potentially relevant external datasets. Primer's primary-document tools provide low-latency, high-accuracy retrieval from company filings, releases, presentations and transcripts, allowing the engine to verify company evidence rather than rely on model memory. Primer Data Workspace stores the source data and derived series, while deterministic code reconstructs point-in-time datasets, creates comparable quarterly series and scores fixed rules against simple baselines. Each investigation ends in a dated decision to reject, monitor or freeze the relationship for a future test.

Our initial investigation produced three interesting artifacts, each requiring a different judgement before the data could be interpreted. Housing data had to be mapped to the point in construction when IBP installs insulation, while keeping the failed quarters and resisting a better-fitting lag. The RV work required rebuilding original monthly releases, matching them to THOR's non-calendar fiscal year and separating wholesale units from revenue affected by price and mix. In dairy, the system removed milk output when it added no information, treated commodity prices as a supply-and-demand signal and used an FX bridge to compare the series with Saputo's Canadian-dollar reporting. Without these adjustments and judgements, each correlation would be easy to calculate and easy to misread.

1 Introduction

Most alternative-data work starts with a dataset that someone already believes is relevant. An analyst cleans it, charts it against a company result and checks whether the two move together. But a similar-looking chart does not show that the comparison makes sense. An industry shipment series, for example, might track goods leaving manufacturers while the company reports sales to customers. It might cover a different geography, measure units rather than revenue, or be affected by pricing and product mix. Unless those differences are understood, a strong correlation can still tell the wrong story.

High-quality equity analysis involves more than putting two series on a chart and comparing how they move. A thorough process requires deeply understanding the dataset: what it measures, how it is constructed and which company or part of a business it might relate to. It also requires understanding the company's accounting practices and whether price, product mix, currency, acquisitions or revenue-recognition rules might distort the comparison. Just as importantly, analysts need to obsess over the periods when the relationship breaks down. Those gaps may reveal inventory movements, changes in pricing or market share, a shift in business mix, or a problem with the dataset itself. Understanding why the relationship failed can be more informative than studying the periods when it worked.

This paper asks whether a language model can carry out more of that investigative work on its own, starting by searching for potentially relevant data rather than being handed a dataset selected by an analyst. For each candidate it finds, the model examines the material explaining how the dataset was built, alongside company filings and earnings releases, and sets out why the signal should or should not relate to a specific business. Code then performs the calculations and tests the proposed relationship using only information that would have been available at the time. Keeping those roles separate helps prevent a convincing explanation from being mistaken for evidence.

2 Research question and scope

The research question is whether a language-model-led system can search for potentially relevant external data and turn a promising source into a qualified observation about a named company or business unit. Qualification requires a defensible measurement definition, economic mechanism, reporting perimeter, decision-time cutoff and target metric. It also requires a record of rival explanations and failed periods.

The initial target is directional. The system asks whether year-on-year growth is improving or deteriorating rather than attempting to predict the exact reported value. A retrospective match creates a candidate only. Stronger forecasting language requires a rule fixed before the next outcome, with no later change to the source, lag, threshold or target.

3 System architecture and control points

The architecture separates semantic reasoning from numerical execution. GPT-6-Astra conducts the investigation through the Codex harness, which supplies reliable access to retrieval, web research, computer use and executable tools. Versioned research skills define how source identity, comparability, operating mechanisms and quarterly trajectories must be tested.

The company-evidence path depends on Primer Tools for low-latency, high-accuracy retrieval from primary documents. The system repeatedly returns to original filings, releases, presentations and transcripts to verify definitions, reporting boundaries, dates and management explanations rather than relying on what the model remembers. External data enters through a separate path because historical values can be revised, coverage can change and publication may occur after the company result.

Primer Data Workspace is the persistent system of record for numerical data. It stores raw observations, originally reported company figures, first-published vintages, entity mappings, formulas and derived series. Executable Python or SQL performs joins, quarterly aggregation, currency and price bridges, vintage reconstruction, scoring and charting. The model queries and interprets those outputs; it does not hold the tables in its context or calculate the series in prose.

Every completed investigation updates durable research memory. The record includes source definitions, company mappings, formulas, failed variants, structural breaks and the current status of the relationship. Later investigations can reuse what has been verified without silently turning an earlier conclusion into a permanent fact.

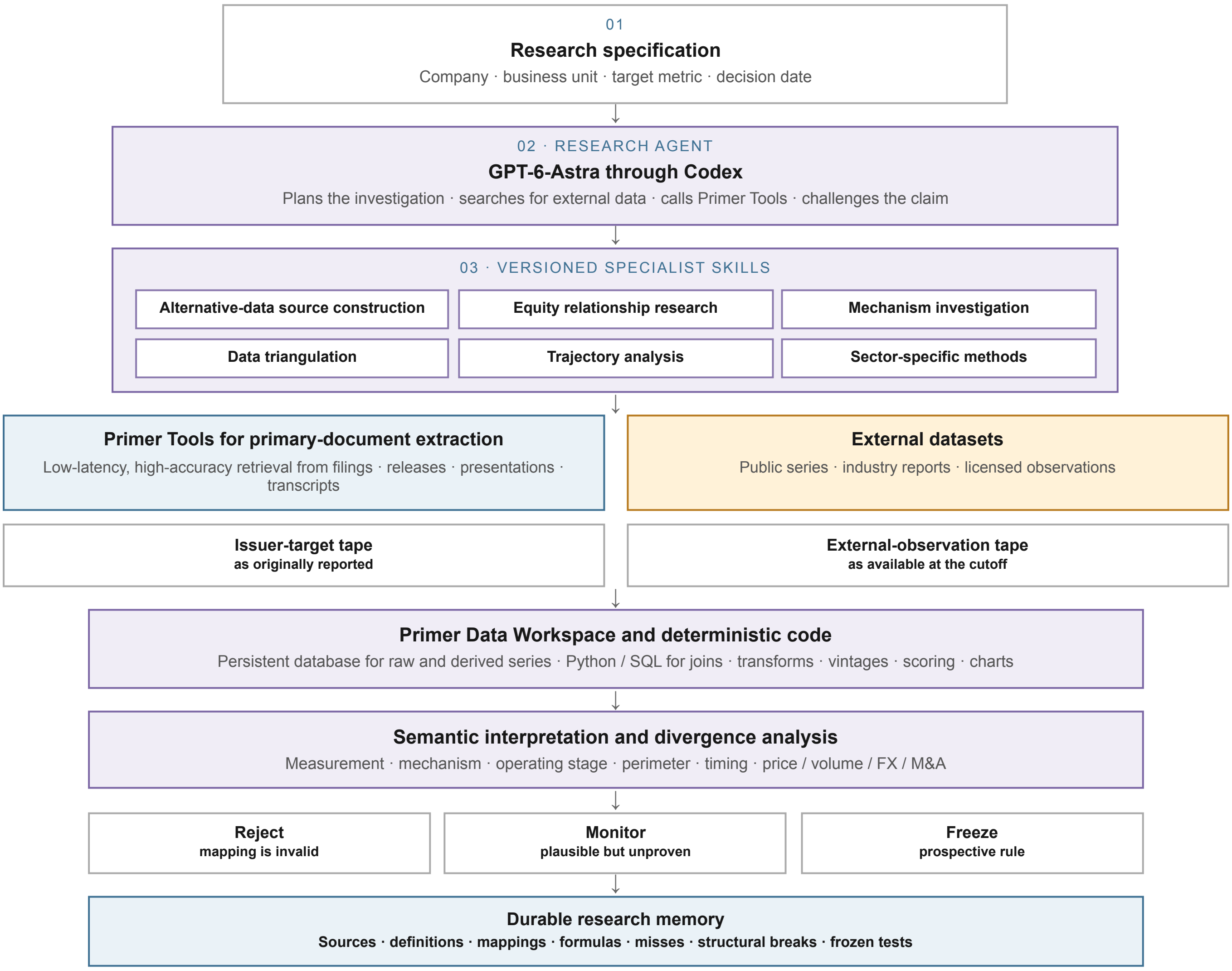

Figure 1. *Proposed architecture for autonomous signal discovery. Primer Tools ground the investigation in accurately retrieved primary company documents; Primer Data Workspace persists the numerical record; and deterministic code performs the transformations and scoring. The language model handles search, interpretation and challenge without becoming the database or calculator.*

4 The semantic task

A relationship is not defined by two column names. It is a chain of claims about identity and economic transmission. The external observation must refer to the right product, geography and operating stage; the company target must refer to the business unit actually exposed to it; and the expected timing must follow from how goods, prices or demand move through the business.

This is where language models add the most value. Source documentation and corporate reporting rarely use the same vocabulary, while the relevant qualification may sit in a footnote, product description or management comment. The engine can resolve those differences and produce a structured mapping that code alone cannot infer from the values.

5 How the engine discovers and tests relationships

The process does not begin with a predefined company question or a dataset selected by an analyst. Versioned specialist skills form the research harness around GPT-6-Astra. Astra supplies the general reasoning capacity, while the skills provide the domain expertise and repeatable instructions that govern how it searches, interprets sources, maps economic mechanisms, challenges failed periods and decides what to investigate next. They are the main drivers of reliable execution. Combined with optimised Primer Tools for low-latency, high-accuracy primary-document retrieval, and Primer Data Workspace plus saved code for persistent data handling and calculation, the harness can move repeatedly between source research, deterministic testing and investigation of failed relationships without relying on model memory or improvised methods.

1. Search for possible relationships

The system searches across external datasets and company disclosures for observations that may describe connected economic activity. A candidate might link an industry shipment series, commodity price or construction-stage measure to a company product or business unit. At this stage it is only a reason to investigate, not a signal.

2. Understand and map each candidate

Astra establishes what the external dataset measures, how it was constructed, its units, geography, operating stage, publication timing and revision history. Primer Tools retrieves the original filings, releases, presentations and transcripts needed to identify the corresponding company activity, reported metric and accounting basis.

3. Require a plausible economic path

Before any correlation is calculated, the system must explain how the observed activity could reach the company result and when the effect should appear. It considers inventory, price, product mix, currency, acquisitions and revenue recognition. If the meaning or transmission path cannot be supported, the candidate is rejected and the search continues.

4. Build and test outside the model

Primer Data Workspace stores the raw observations, company figures, historical vintages, mappings, formulas and derived series. Saved Python or SQL aligns periods, applies documented adjustments and calculates growth rates, lags, correlations, directional scores and simple baselines. The model does not carry the numerical record in its context or perform these calculations itself.

5. Investigate the misses and repeat

Code returns the fixed comparison to Astra, which studies failed periods as closely as successful ones and uses Primer Tools to retrieve further evidence. It may discover a timing mismatch, accounting distortion, source change or weak economic link. The process can return to the source, mapping or test, but every revised hypothesis is versioned and the original result is retained.

6. Decide what survives

The purpose of the loop is not to keep changing the test until it works. A relationship is rejected when the source or mapping fails, monitored when it remains plausible but unresolved, and frozen only when the source, cutoff, transformation, target, rule and abstention conditions can be fixed before the next company result.

6 Research memory and prospective testing

Every investigation updates a durable record containing the source definition, company mapping, transformations, attempted variants, failed periods and final decision. Rejected and unresolved cases remain visible so that the system does not rediscover the same weak relationship, while verified definitions and mappings can inform later searches.

Prospective testing cannot erase the selection that produced a candidate, but it prevents the next quarter from being fitted after the event. The useful asset is therefore not a collection of charts that happened to work. It is the accumulated record of what was searched, what was tested, what failed and which relationships earned a fixed future test.

7 Example 1: the U.S. housing pipeline and Installed Building Products

COMPANY	COMPANY MEASURE	EXTERNAL DATASET
Installed Building Products	**U.S. single-family new-construction installation revenue from branches owned for more than 12 months**	**U.S. Census Bureau and HUD one-unit homes under construction**

INVESTOR INTERPRETATION

This is a useful directional relationship rather than a direct revenue forecast. Changes in the pool of homes approaching installation often coincide with changes in IBP's comparable-branch single-family revenue growth, even though the company result also includes price, product mix, customer exposure and market-share changes.

The U.S. Census Bureau and HUD publish the number of one-unit homes under construction each month. Because insulation and other fitted products are normally installed after a home starts and before it is completed, this stock provides a plausible pipeline for IBP's single-family work. The company series is year-on-year revenue growth from single-family new-construction jobs completed by Installation branches owned for more than 12 months. It is not a measure of homes, jobs or installation volume.

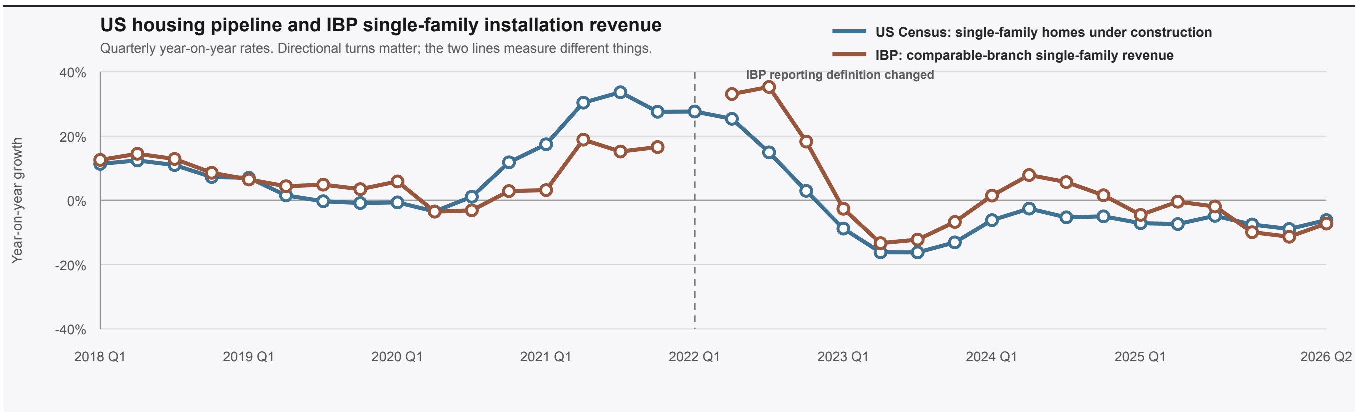

Figure 2A. *U.S. Census Bureau and HUD, New Residential Construction, Table 4a: quarterly average of seasonally adjusted one-unit homes under construction, shown as year-on-year growth. The comparison is IBP's reported year-on-year single-family same-branch Installation revenue growth. The Q1 2022 company observation is omitted because IBP changed its reporting perimeter. The September 2025 Census release arrived after IBP reported, so that quarter is shown for context rather than treated as decision-time evidence.*

What the agent did. The Census series measures housing activity in units, while IBP reports revenue growth. The agent identified that mismatch and used IBP's disclosed Installation price/mix contribution to adjust the company's single-family revenue growth, creating a closer proxy for underlying activity. It also identified IBP's reporting-definition change and began the adjusted comparison in Q2 2023, once the year-on-year figures were entirely within the new definition. IBP does not disclose price separately for single-family work, so the adjustment uses the wider Installation segment.

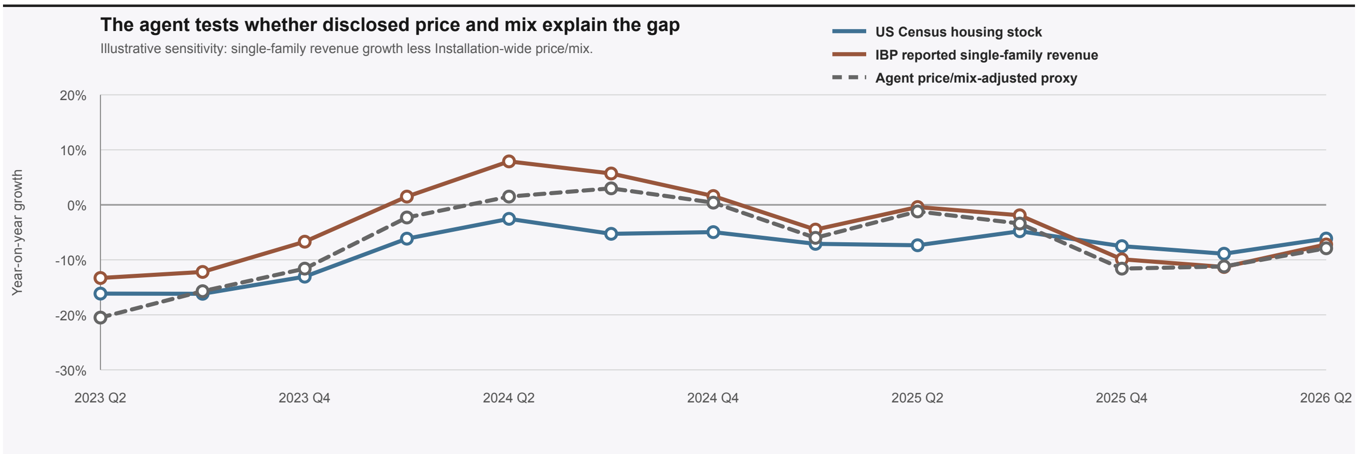

Figure 2B. *Agent-created sensitivity from Q2 2023 to Q2 2026, after all year-on-year comparisons fall within IBP's post-realignment reporting definition. The dashed line is IBP single-family same-branch revenue growth less the disclosed Installation price/mix contribution excluding heavy commercial. IBP does not publish price or job volume for single-family work alone, so this is not a company-reported single-family volume series.*

8 Example 2: RV wholesale shipments and THOR Towable sales

COMPANY	COMPANY MEASURE	EXTERNAL DATASET
THOR Industries	North American Towable Recreational Vehicles reported-sales growth	RVIA first-published monthly wholesale shipments of towable RVs

INVESTOR INTERPRETATION

Across the available history, every change in RVIA towable-shipment growth pointed in the same direction as the corresponding change in THOR Towable sales growth. This is a useful directional relationship, but the industry series measures units shipped to dealers rather than retail demand or THOR revenue, so price and product mix can still create important gaps.

RVIA reports the number of towable recreational vehicles shipped by manufacturers to dealers each month. THOR reports North American Towable sales in fiscal quarters ending in October, January, April and July, so the monthly industry data cannot be compared with the company series without first rebuilding it around THOR's calendar. The test uses the figures printed in the original RVIA reports rather than the revised history now shown online.

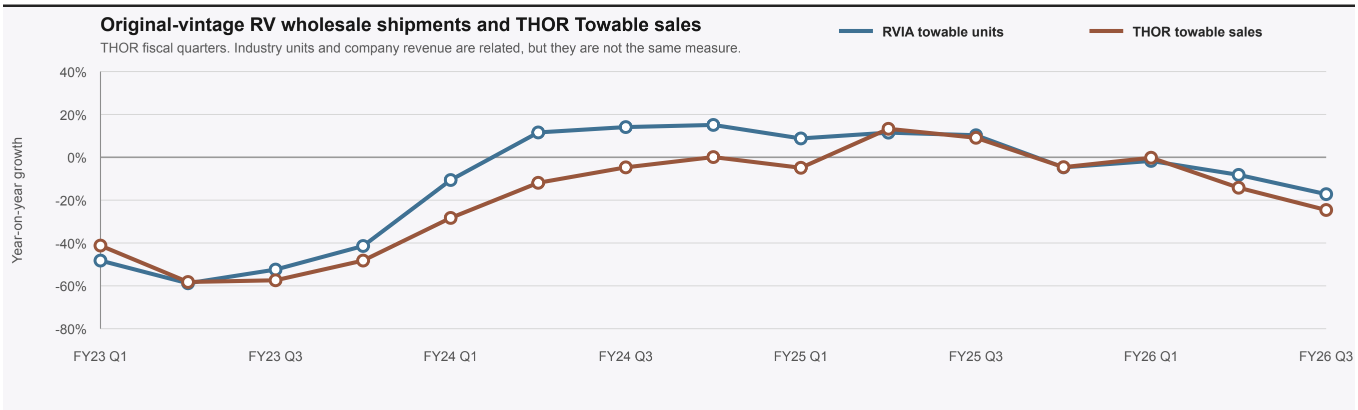

Figure 3. *First-published RVIA towable-unit growth and THOR North American Towable reported-sales growth by THOR fiscal quarter. The scored result concerns the direction of change in the two lines, not equality of their levels.*

What the agent did. The agent identified the calendar mismatch and grouped each set of monthly RVIA reports into the corresponding THOR fiscal quarter. It reconstructed the point-in-time series from 45 original reports so later revisions could not leak into the test, then retained THOR's unit growth as a diagnostic alongside sales. That check exposed FY25 Q1, when THOR units rose but sales fell because net price and product mix per unit declined.

Boundary of the finding. THOR contributes to the RVIA industry total, and company backlog or guidance may already contain the same direction. The result is therefore a qualified trajectory signal, not an exact revenue forecast or evidence of investment alpha.

9 Example 3: dairy prices and Saputo USA reported sales

COMPANY	COMPANY MEASURE	EXTERNAL DATASET
Saputo	USA Sector reported sales, translated from Canadian dollars into U.S. dollars	USDA weekly block cheese, butter and dry whey prices

INVESTOR INTERPRETATION

The longer year-on-year history shows that dairy prices provide a useful directional read across several commodity cycles, although the two lines should not move by the same amount. The basket matched the direction of Saputo USA reported-sales growth in 11 of 12 timely quarters during the original audit window, but each point must be read with its prior-year comparison base in mind.

The USDA publishes weekly prices for block cheese, butter and dry whey. Saputo explains that commodity prices affect USA-sector sales alongside volumes, customer mix, contractual repricing and currency. The comparison therefore asks whether a simple basket of market prices can provide a timely read on the direction of Saputo USA reported sales, rather than treating commodity prices as a direct measure of the company's operating performance.

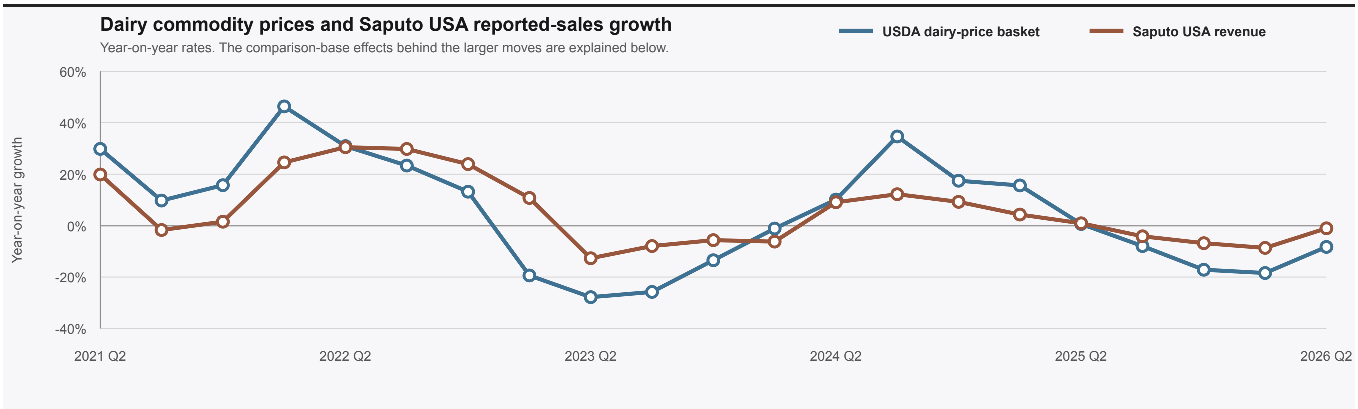

Figure 4. *Year-on-year growth in an equal-weight USDA dairy-price basket and Saputo USA-sector revenue translated into U.S. dollars using the quarter's average Bank of Canada exchange rate. The chart extends to Q2 2021; the quoted score remains the original Q3 2023 to Q2 2026 audit window. Earlier observations are context and include reporting-perimeter effects.*

What the agent did. The agent recognised that every year-on-year point inherits the same quarter's level from a year earlier, so a move in Q4 2021 mechanically affects Q4 2022 and the following annual comparisons. It kept the chart in year-on-year terms because that is the measure tested, but checked the underlying levels and sequential changes before interpreting the larger moves. The apparent Q3 2024 basket spike was 34.7% year on year but only 11.8% from Q2, because Q3 2023 had itself been 25.8% below the prior year. Dry whey rose 78%, butter 19% and block cheese 15%. Saputo said block and butter added approximately C$175 million to revenue, while removing a C$33 million FX benefit reduced growth only from 14.1% to 12.2%. The conclusion is genuine price pass-through, amplified by the YoY base and the basket's equal weights rather than financial hedging.

Boundary of the finding. The equal weights do not represent Saputo's product mix, and the earlier history is contextual rather than additional validation. Saputo also does not disclose a complete quarterly USA price, volume and currency bridge, so the evidence supports a reported-sales and market-conditions signal rather than a precise forecast of company volume, margin or earnings.

10 Research memory

The research memory is a versioned file system built for agents, not a transcript of earlier conversations or one growing summary document. An agent can enter through a short project-state file, retrieve only the records needed for the current question and follow explicit links back to the evidence, code and earlier decisions.

projects/ Current question, target outcome, status, evidence gaps and next test.	**sources/** Dataset definitions, rights, coverage, publication timing and revision behaviour.	**entities/** Raw names mapped to parents, brands and business units with confidence and validity dates.
relationships/ Directional economic edges, products, geographies, mechanisms and expected timing.	**signals/** Exact filters, units, aggregation, transformations, lag hypotheses and known failure modes.	**experiments/** Frozen cutoffs, rules, baselines, outcomes and rejected or inconclusive runs.

The files are deliberately small and split by claim type. Plain-English Markdown records preserve rationale, limitations and unresolved questions; JSON and CSV preserve typed facts, source receipts and numerical series; saved Python or SQL preserves the calculation. Stable paths, headings and links let an agent find the relevant record with simple search and load a narrow evidence packet instead of placing the whole archive in the model's context.

The record is written while the investigation is running. A source call creates a receipt and retains the raw observation. Entity resolution preserves the original name beside the mapped company or business unit. A proposed mechanism becomes a relationship record, each reproducible series gets a versioned signal definition, and every transform, lag, exclusion and material mismatch is added to the experiment ledger before the result is accepted.

Large or licensed source bodies, downloaded data, charts and model outputs remain in local artifact storage rather than Git or the prompt. A compact manifest records provenance, rights, hashes and the command that produced the output. This keeps the agent-facing memory cheap to read while retaining a path back to the underlying evidence.

New evidence creates a new version or experiment run; it does not overwrite the old one. The memory therefore keeps observed evidence, entity and period mappings, derived calculations and economic inference separate. A failed test can weaken an edge or expose a coverage break without erasing the source definition or teaching the next agent that the relationship must work.

11 Limitations

The evidence remains preliminary and retrospective. Many companies, datasets, transformations and lags were considered during the wider research programme, so attractive short histories are subject to selection bias. Adjacent year-on-year quarterly observations are also autocorrelated and do not represent independent market cycles.

All three cases use a broad market or industry measure, and the dairy-price mechanism was already discussed by management. A relationship can be economically real without adding information beyond company guidance, consensus or a specialist's existing dashboard. This paper does not test earnings surprise or stock returns.

Point-in-time reconstruction is incomplete for the housing case. Source publication dates and historical revisions can be difficult to recover perfectly. The absence of a clean vintage is a limitation of the evidence, not an invitation to substitute the current series.

Language models can know historical outcomes even when the supplied evidence packet excludes them. Retrospective results should therefore be treated as method-development evidence. The frozen future rules are the first credible test of whether the system can operate without outcome knowledge.

12 Conclusion

Alternative data research is often described as a data-engineering problem followed by a backtest. The three cases here suggest that the harder task sits between those stages. Someone must determine what the source measures, how it enters the company's economics, which reported number it could affect and whether the information was actually available before the result.

Language models are well suited to that work because the answer is spread across source documentation, company disclosures, product descriptions and historical commentary. A controlled agent can read those materials, construct a company-specific mapping and investigate failures. Deterministic tools then produce the series and scores on which the claim depends.

The resulting engine does not automatically convert external data into forecasts. It creates a disciplined path from an observation to a research status: reject it, monitor it or freeze it for a prospective test. Housing data remained a monitor, while RV shipments and dairy prices earned qualified directional tests, each with a different boundary on what could be claimed.

The longer-term value lies in the accumulated relational context, but that context is held in the linked research files rather than inside the model. A later agent can begin with a company or dataset, traverse verified entity mappings and economic relationships, retrieve the exact signal definition and inspect every prior test, including the failures. Each investigation therefore makes the next search better informed without turning an earlier conclusion into a permanent fact.

R Evidence and references

1. U.S. Census Bureau and HUD. *New Residential Construction*, Table 4a, one-unit homes under construction; Installed Building Products quarterly earnings releases, 2018 Q1 to 2026 Q2. Research records: *IBP x Census construction stage: first gate* and *IBP x Census housing-stage series: longer quarterly audit*, 21 September 2026.
2. RV Industry Association. Original monthly wholesale shipment reports, August 2022 to April 2026; THOR Industries original quarterly earnings releases, FY23 Q1 to FY26 Q3. Research record: *RVIA towable shipments to THOR North American Towable RVs: vintage gate*, 22 September 2026.
3. U.S. Department of Agriculture, Agricultural Marketing Service. First-published weekly block-cheese, butter and dry-whey prices; Saputo quarterly USA-sector results; Bank of Canada USD/CAD rates. Research records dated 22 September 2026.
4. RV Industry Association. Industry wholesale shipment reports and publication archive. rvia.org/news-insights/rv-shipments.
5. USDA My Market News. Dairy Products Mandatory Reporting Program weekly price observations. mymarketnews.ams.usda.gov.

Research status. Prepared 22 September 2026. All case scores are retrospective unless a prospective freeze is explicitly identified. The paper makes no claim of earnings surprise, investment alpha or future stock performance.