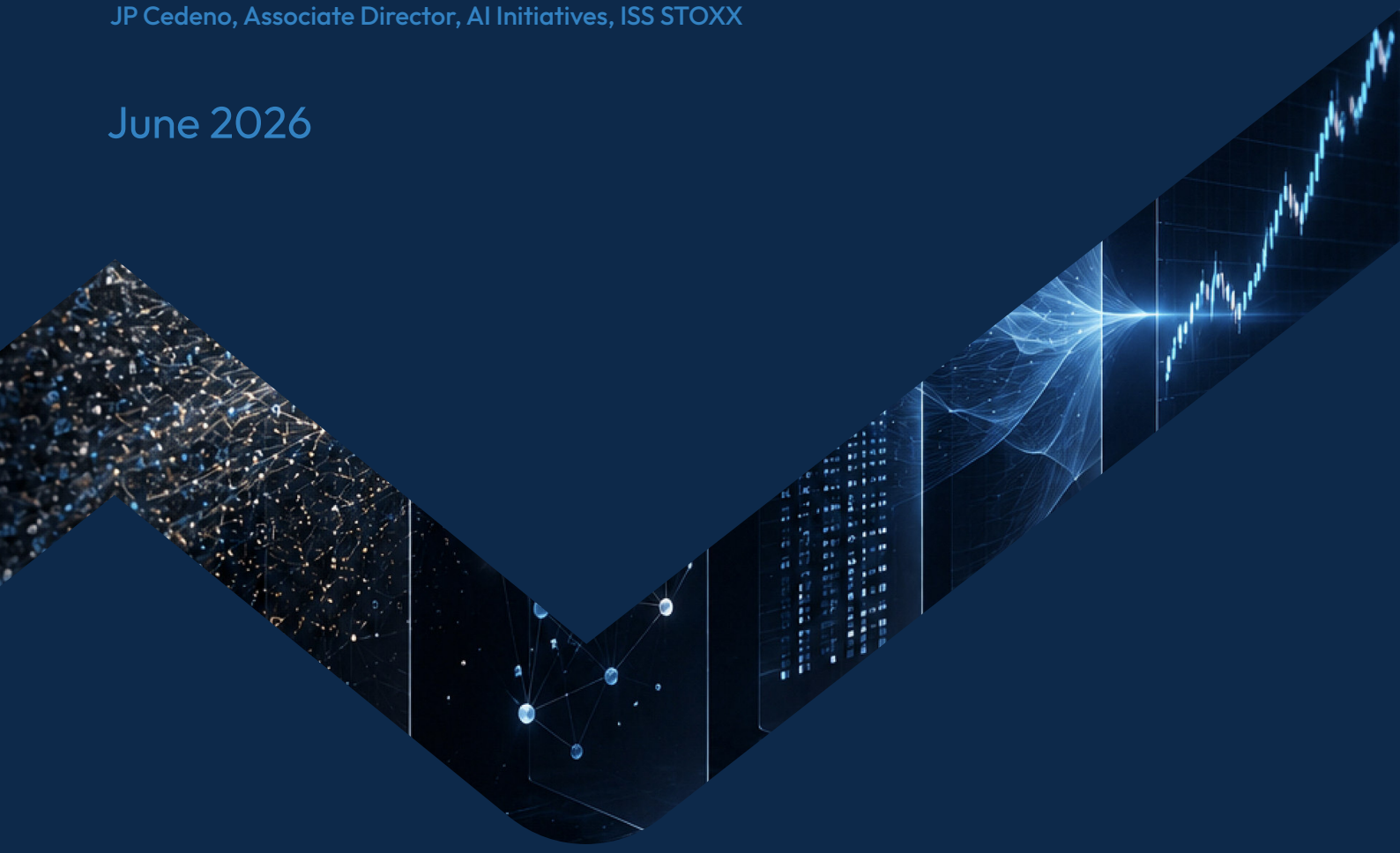
WHITEPAPER

From narrative to signal: Agentic AI for systematic indexing

Gopal Erinjippurath, Managing Director, Global Head of AI Initiatives, ISS STOXX
Hamish Seegopaul, Managing Director, Global Head of Index Product Innovation, STOXX

Acknowledgments to
Taylor Andrews, Associate Vice President, AI Initiatives, ISS STOXX
JP Cedeno, Associate Director, AI Initiatives, ISS STOXX

June 2026



STOXX

Table of contents

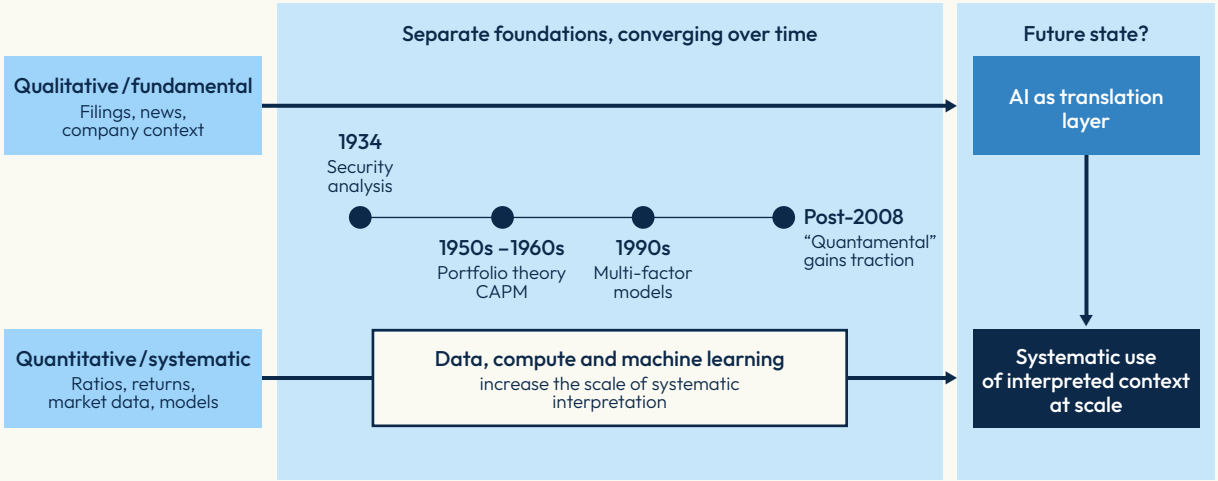
1. Introduction

The research behind this paper was driven by the notion that the future belongs to teams that combine human judgment with machine speed. Artificial intelligence (AI) gives us leverage, but intention is the operator.

The advent of AI makes for one of the most exciting times in quantitative investing that the authors can remember, since ideas can be tested more quickly and easily, new data points can be synthesized to drive investment decisions and human workflows can potentially be mimicked. What makes a good index – clear objectives, understandable rules and intuitive outcomes – doesn't change, but index construction options have multiplied.

In recent years, the world of investing has seen an ongoing convergence of fundamental and quantitative investment styles to produce what is sometimes referred to as a "Quantamental" approach. A key distinction between the two original styles is between qualitative inputs (filings, transcripts, news, etc.) and quantitative inputs (ratios, prices, trading data, etc.). AI offers the potential for a high degree of scale that, if done correctly, can continue bridging this gap.

Figure 1: Convergence in analytical approaches.



Source: STOXX.

This paper presents a structured approach to enhancing the implementation of qualitative information in systematic frameworks. We created a methodology for deriving signals from web-scale text to capture market sentiment, mimicking human judgment on qualitative factors through an agentic workflow. We then used these signals to quantify the degree of positivity or negativity in publicly available coverage of individual issuers.

Our research in this area has revealed promising results in the form of a complementary signal to Price Momentum, with attractive backtested performance and the robustness and auditability required for use cases such as indexing.

2. Factor motivation: Why Narrative Momentum, and why now?

The quantitative world of factor investing, which focuses on systematic drivers of returns, is undergoing constant evolution. Whereas factors themselves are relatively static,[1] the *definition* of factors is where that evolution occurs. New technology, data and research enable novel ways of identifying the phenomena which factors aim to capture.

Momentum is a core pillar of this world and has traditionally been defined as the tendency for prices to increase or decrease. This concept of Price Momentum remains a core approach, but other ways of capturing Momentum, such as Fundamental Momentum (directions of earnings), have also been identified. Modern factor indices, such as the iSTOXX® Ang Research Enhanced index suite, use multiple definitions to define Momentum, which can help support performance.

Like Momentum, sentiment became established in research decades ago and has been a foundational use case for natural language processing (NLP) in finance. This has led to specialized businesses, services and software libraries (e.g., FinBERT, FinRobot), which support the reading of news, filings, transcripts and other documents to detect the tone of the response. These pipelines typically start with textual data, identify the context and the issuers, and result in scores.

This paper does not challenge these approaches per se but highlights that it is not easy to detect context from words alone, or to connect them to the right securities. Consider the following sample text (which was developed by the authors for this research):

> Following the sudden collapse of First Republic's capital position, regulators intensified scrutiny of US regional banks, raising concerns about liquidity and contagion risk across the sector. J.P. Morgan, however, appears largely insulated from the turmoil, having avoided material losses and benefited from deposit inflows as customers seek safety. Despite broader market uncertainty, analysts note that the stress event may ultimately strengthen the competitive position of the largest money-center banks.

This text is negative for First Republic and positive for J.P. Morgan. However, J.P. Morgan's sentiment is more difficult to tease out by finding negative words – although it is used as a contrast to First Republic, it is described using negative terms such as "material losses." A "bag of words" approach may not pick up this nuance, but a modern large language model (LLM) should have no issues detecting the right interpretation "out of the box."

LLMs and Agentic AI – the name given to systems which operate autonomously to achieve goals – offer an alternative approach, since they are able to *start* with an issuer, gather text and understand context without identifying individual words or phrases. Although this is not a "free lunch," since it creates other issues to manage, agentic workflows can enable a best-of-both-worlds approach. In the upcoming sections, we explore the concept of Narrative Momentum – the degree of positivity or negativity in publicly available coverage of an issuer – using an agentic setup.

As with many AI use cases, building prototypes is easy, but robustness requires systems-level engineering around LLM inference[2].

[1] Factors are drivers of returns and risk. Investment factors need to have academic backing, show evidence of performance, have a reason for existence and be differentiated.

[2] LLM inference is the process of converting a prompt into a response. This is distinct from training a model.

3. Design principles for production-grade agentic signal generation

Transitioning from traditional quantitative factors to AI-driven narrative sentiment requires a rigorous, systematic approach. Historically, including text data in index construction has been notoriously difficult because natural language processing (NLP) is inherently unstructured and highly context-dependent. Traditional NLP pipelines often struggled with accurate entity mapping and required constant, costly retraining to interpret financial nuances without introducing noise.

To effectively deploy Agentic AI workflows in a systematic framework, the signal generation process must seek to be as robust as a mathematical model. We established four foundational design principles to ensure our AI signals meet the rigorous demands of systematic strategy consumers. These are described in the subsections below.

3.1 Point-in-time correctness (i.e., no look-ahead bias)

The integrity of any index backtest relies on point-in-time accuracy. The workflow implements strict temporal guardrails to ensure the AI agent only uses information available at a specific historical moment:

- **Enforced date cutoffs:** We enforce hard date cutoffs for all data retrieval, with default quarterly logic being strictly version-controlled to prevent look-ahead bias.
- **Leakage detection:** We use system-level checks to monitor source publication dates and detect leakage of future information when extracting Narrative Momentum for time windows in the past.
- **Automated retries:** If leakage is detected, the system automatically triggers targeted retries to pull the correct point-in-time references.
- **Index alignment:** Signal timestamps are meticulously aligned with standard index conventions, matching universe snapshots and rebalancing schedules.

3.2 Auditability and evidence traceability

Institutional signals cannot operate in a "black box." Our framework requires the AI agent to show its work and rely solely on verifiable evidence:

- **Grounded outputs:** The agent's conclusions must be explicitly grounded in retrieved sources.
- **Hallucination prevention:** We intentionally allow the AI to return sparse sources rather than pressurizing it to invent or fabricate links.
- **Standardized payloads:** The system emits a strict JSON contract containing the top verifiable sources, the per-signal rationale, confidence labels and the exact calculation string used.
- **Governance logging:** We maintain comprehensive run logs tracking token usage, cost, retries and prompt configurations so as to allow for auditability and operational governance.

3.3 Evaluation and repeatability

AI models are not deterministic by nature, whereas a core tenet of systematic indexing is reproducibility. We therefore engineered the workflow to absorb model variance while minimizing the impact the final scoring:

- **Expert-aligned, multisignal coherence evaluation:** Sentiment is checked the way an analyst would read a short memo: We use several simple checks at once, tied to how experts actually grade things, and we automate the process so that it remains consistent rather than merely using "does this label match?" as a basic classifier.
- **Burst consistency testing:** We rerun identical issuer and period configurations 20 to 30 times during testing so as to encourage stable outputs across repeated executions.
- **Ordinal scoring:** By preferring ordinal outputs (e.g., discrete categories from –2 to +2) and fixed thresholds over continuous scoring, we significantly reduce the system's sensitivity to minor wording drifts in the AI's analysis.
- **Strict rerun triggers:** We have defined automatic rerun triggers that activate whenever the system detects missing attribute fields, low-confidence labels or suspect reference links during attribution.

3.4 Scalability and cost control

Operating across a universe such as the STOXX® USA 900 requires massive data processing. Our architecture is designed to handle this scale efficiently and cost-effectively:

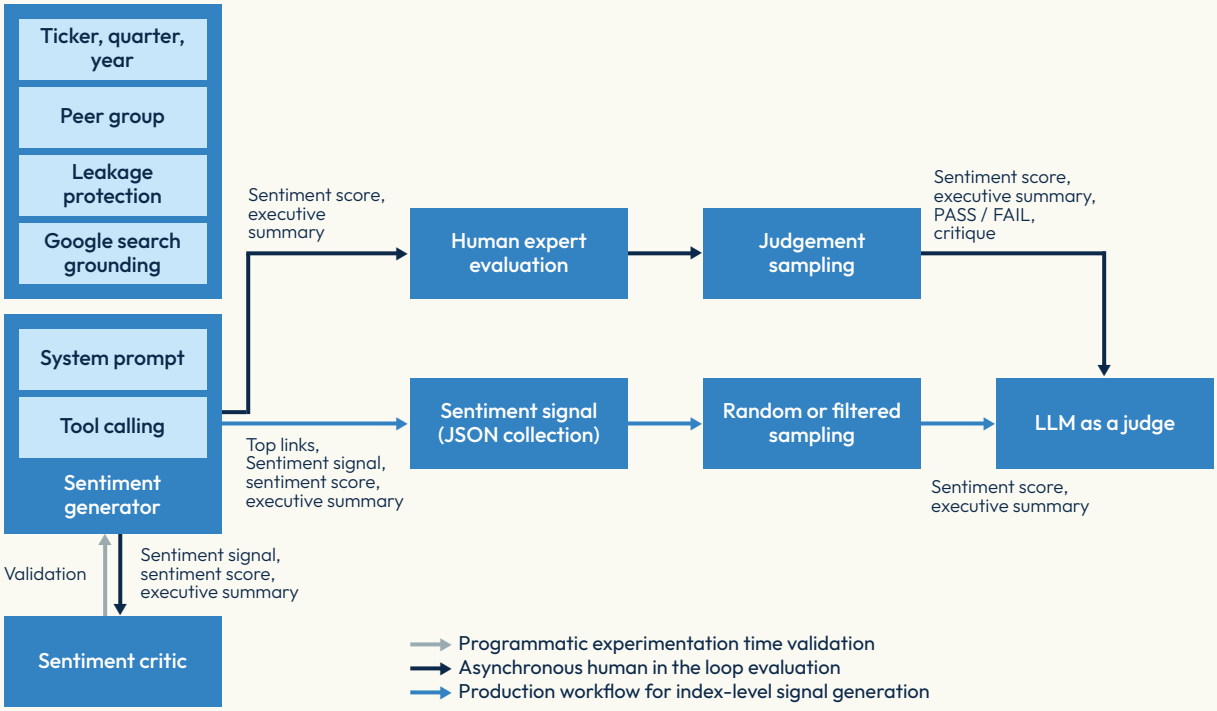
- **Parallel execution:** We used batch execution with parallel jobs to process large swaths of the universe simultaneously, completing quarterly runs in a matter of hours.
- **Model benchmarking:** We benchmark LLM model families to identify the optimal "sweet spot" that balances speed, variance and cost.
- **Predictable operations:** Active token and cost logging is enabled by default, ensuring high visibility for operational planning and run-to-run comparability.

4. Architecting the agentic workflow: Moving beyond statistical signals

Traditional quantitative factors rely on a passive, statistical consumption of structured data, ingesting fixed datasets, applying static weights and outputting a score. By contrast, Agentic AI requires a fundamental paradigm shift toward an "active mindset."

Instead of merely processing the numbers provided, an AI agent autonomously retrieves information, reasons through conflicting evidence, tests counterfactual scenarios and synthesizes a conclusion. This active capability unlocks narrative sentiment that traditional models miss, but also introduces new architectural requirements. Challenges such as hallucination prevention, source grounding and temporal leakage, which are not issues in traditional statistical models, become the primary design constraints of an agentic system.

Figure 2: Agentic signal generation and testing workflow.



Source: STOXX.

4.1 Workflow overview: Generator, critic and signal schema

To ensure institutional-grade reliability, our architecture moves away from single-prompt interactions in favor of a multiagent workflow.

- **Initial setup:** Our system is built on the Google Agent Development Kit (ADK) framework for AI agents. A primary "generator" agent, acting under the persona of a senior equities research analyst, produces a draft sentiment analysis. We strictly enforce a JSON schema for the output.
- **Adversarial validation:** We deploy an optional "critic" or validator agent pattern during testing and iteration. This critic stress tests the generator's outputs, challenging its assumptions with the aim of reducing inconsistency and minimizing positive bias. The critique from the critic agent also provides indicators of how to improve the sentiment generator process.
- **Inputs:** The system is seeded with the target ticker, the specific quarter and year, and, if available, peer group context to anchor the analysis in relative terms.
- **Outputs:** The final output is a strict, structured JSON payload containing ordinal scores, the rationale and verifiable citations. This is systematically stored to permit downstream backtesting and optimizer ingestion.
- **Operational logging:** Comprehensive logging is enabled by default. We track token usage, cost, rate-limit retries and run metadata to allow for operational transparency and auditability.

4.2 Retrieval and grounding approach

An agent is only as good as the content it retrieves. Building high-performance agentic workflows for generating Narrative Momentum requires a search and retrieval protocol that loads the right content into the agent's context window. We implemented a rigorous data retrieval protocol to ensure that the models are grounded in reality:

- **Standardized tooling:** We use Google Search as a tool for retrieval and for grounding the agentic signal generation so as to ensure high-quality, varied results during our iterations. The purpose is to support score generation and auditability, not to redistribute source content.
- **Iterative search and sparse sources:** The agent employs iterative search patterns with varied queries and deduplicates the results. Crucially, we allow the agent to return sparse sources if evidence is thin. By removing the pressure to "fill a quota" of top links, we drastically reduce the model's incentive to invent information.
- **Strict verification:** A hard system requirement is enforced: The AI system must never fabricate links or dates. Unverifiable or broken URLs are strictly rejected.
- **Look-ahead bias mitigation:** To preserve point-in-time correctness, a hard quarter cutoff date is appended to every query, simulating the exact information environment available at the historical moment in question.

Running evaluations: We need to anchor scalable scoring in expert judgments, then judge each output as a coherent bundle of claims, not as an isolated label. We use an LLM that evaluates agent generated responses across three dimensions in an "LLM-as-a-judge" framework:

- **Human rubric first:** The judge is calibrated on human-generated pass/fail decisions and critiques rather than simply on generic LLM preferences. This means that "good" remains tied to how experts actually decide.
- **Multifaceted quality:** Generated sentiment is evaluated beyond "did we pick the right category?" by checking point-in-time correctness, whether tone matches the numeric stance, and whether the summary and score tell the same story.
- **Operational repeatability:** Verdicts are structured (pass/fail + critique) and usage is tracked so that evaluation can be run frequently and runs compared, instead of relying on one-off manual reads.
- **Sampling over issuers and over time:** We run human expert-driven evaluations of narratives on a sampling of tickers over a long period of time so as to capture a wide spread of human judgment. This ensures that the LLM can be used to judge signal generation performance for tickers across the index.

4.3 Infrastructure stack

Operating this architecture at the scale of the STOXX USA 900 requires a robust, scalable technology stack:

- **Gemini Enterprise Agent Platform:** This serves as the core orchestration layer for defining agent personas, behaviors and tool-calling parameters.
- **Gemini model family:** This powers the core reasoning engine. We benchmarked continuously across different variants (for example: Gemini 2.5 vs. Gemini 3 Flash) to select a configuration that is optimized for the "sweet spot" of processing speed, output repeatability and cost efficiency.
- **Data orchestration & quality controls:** Serving as the operational backbone, a robust Python-based orchestration layer handles high-volume batch execution, automated leakage detection, QA scripting, and structured payload post-processing to ensure absolute data pipeline integrity.

- **GitHub and Docker:** We treat prompts and agent configurations as code. GitHub provides strict version control, while Docker ensures reproducible, containerized runtime packaging.
- **Internal data services:** The workflow integrates directly with our internal database via cloud APIs to map stable identifiers, and seamlessly connects with internal portfolio construction tools and universe data to translate raw sentiment into actionable portfolio weights.

5. Data and signal definition: What is Narrative Momentum?

The goal of the Narrative Momentum signal we propose is to quantify the degree of positivity or negativity in publicly available coverage of an issuer. This distinguishes it from Price Momentum (which directly tracks market reactions) and Fundamental Momentum (which tracks the trajectory of realized or expected earnings).

Signal creation thus has certain high-level parameters: the issuer,[3] the period in question and the allowable sources. However, most of the value is derived from how the signal is constructed, which is described below.

5.1 Prompt blueprint

The prompt was engineered with the design goals set out in section 3 in mind, enabling control, stability and scalability. The prompt and setup underwent several iterations to achieve these objectives.

Note that we start with the issuer here and gather information to transform it into quantitative information. This is in contrast to standard NLP pipelines, which start with text and then try to ascertain the context and tie information back to an issuer.

Table 1: Prompt blueprint.

Prompt section	Purpose
Role and tooling	To set the agent's persona and ensure the agent uses its search tool for evidence-based outputs.
Overall goal	To produce a comprehensive, investor-useful sentiment report and score for a single ticker.
Data collection protocol	An iterative search with varied queries and a freshness window. Mandates the use of multiple lanes of evidence and the use of collected results only.
Scoring framework	For each of the five categories, the agent assigns a confidence level and an ordinal score ranging from –2 to +2. These scores quantify the prevailing quarterly sentiment for a specific issuer: –2 (strongly negative), –1 (moderately negative), 0 (neutral), +1 (moderately positive), and +2 (strongly positive).
Overall score calculation	To produce equal weighted signals that are mapped to an ordinal using strict thresholds. Controls for biases with high standards for extreme scores and counterfactual checks.
Output schema (JSON)	Output a JSON with required fields such as scores, confidence levels, rationales, key findings, qualitative analysis and grounded search results.
Examples (pass/fail)	To calibrate behavior using concrete do/don't cases.

Source: STOXX.

[3] We model this at the issuer level, as opposed to the level of individual securities, as the context will typically be related to the company as opposed to its individual equity or debt issuances.

6. Narrative Momentum signal characteristics

We used the workflow and prompts in the sections described above to run the signal generation for Narrative Momentum for each component of the STOXX USA 900 benchmark for each quarter (as of the end of March, June, September and December) between Q4 2014 and Q4 2025. The characteristics of the outputs are shown in the following sections.

6.1 Distributions

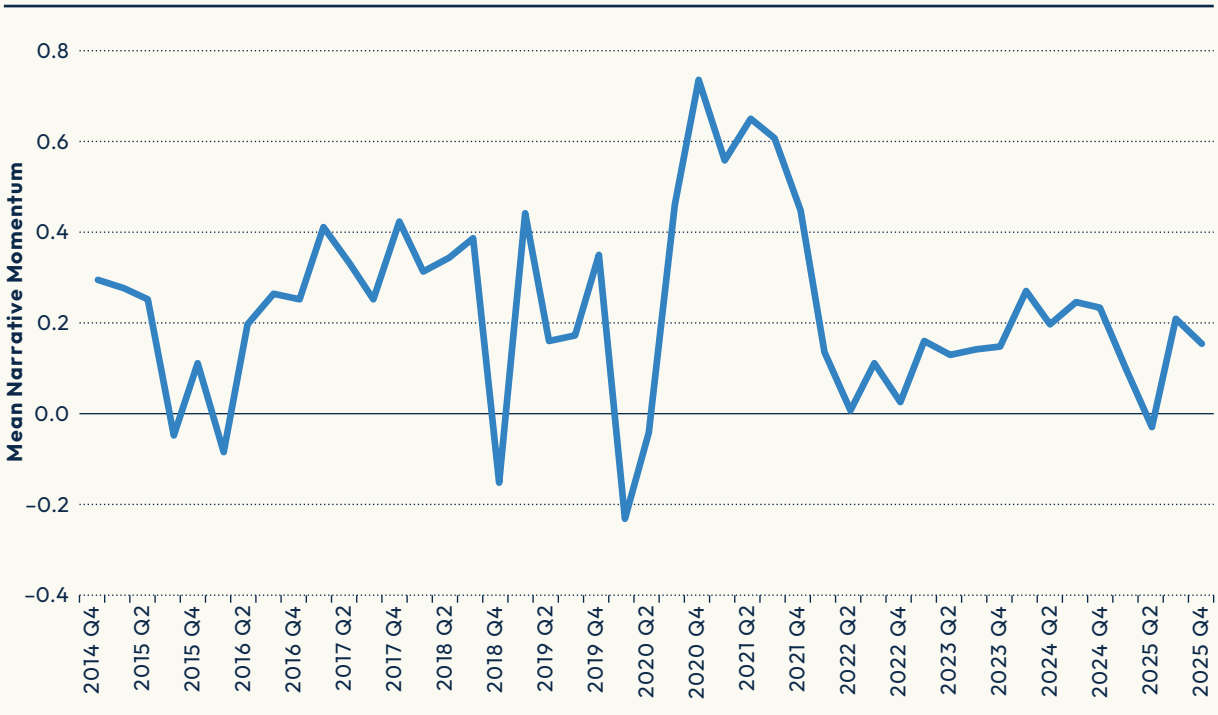
We plotted the average score for each quarter, calculated as a simple average across the 900 securities, the breakdown of scores by categories (–2, –1, 0, +1, +2) for each quarter, and the distribution of scores, segregated by sector. The sectors assigned complied with the top-level Industry Classification Benchmark (ICB) industry allocation as used in the STOXX indices.

Our takeaways from the data for the time window in question are as follows:

- Over time, the major moves in average sentiment coincide with market events. There were dips during crises (notably COVID-19 and the 2025 tariff shock) and rises in subsequent recoveries. This suggests that the signal detects changes appropriately.

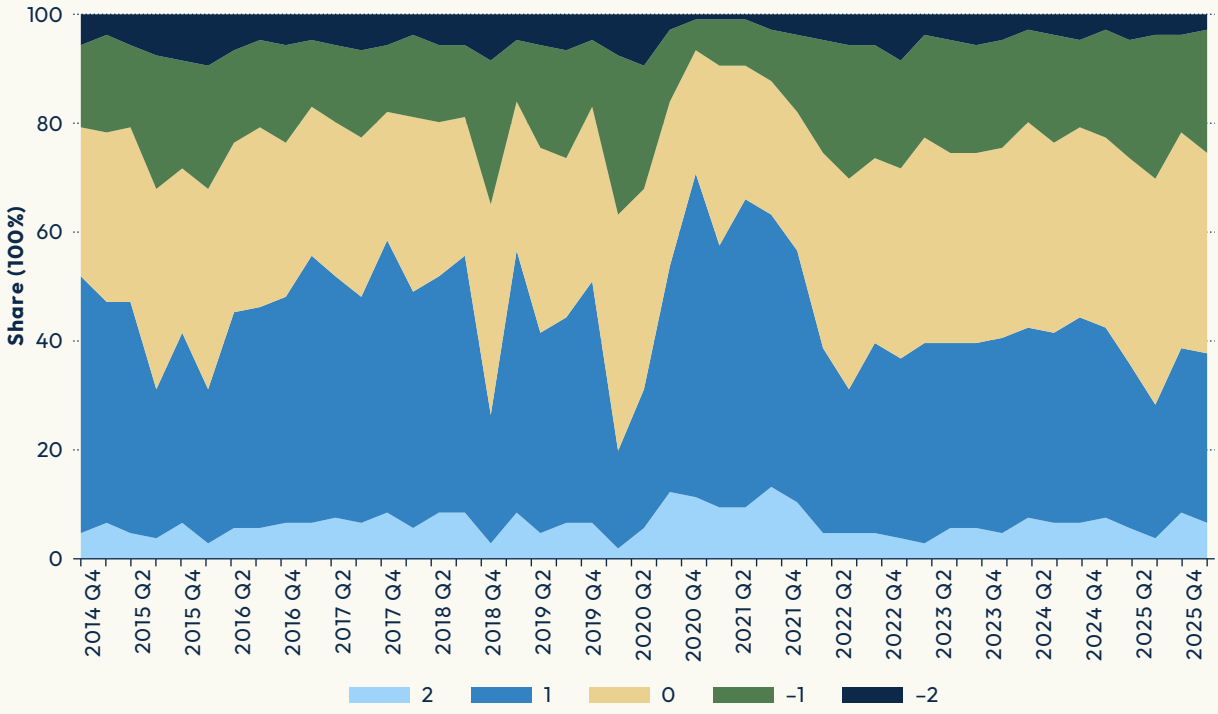
- Looking at the distributions, we still see a positively biased score (with most scores clustering at 1 instead of 0), even accounting for our bias controls. This is not homogenous for the different sectors: Some sectors such as Utilities and Telecommunications have much more normal (0-centered) distributions, while sectors such as Industrials and Technology have more positively biased samples. This also hints towards using the score cross-sectionally in portfolio construction.

Figure 3: Quarterly mean of Narrative Momentum over the STOXX USA 900.



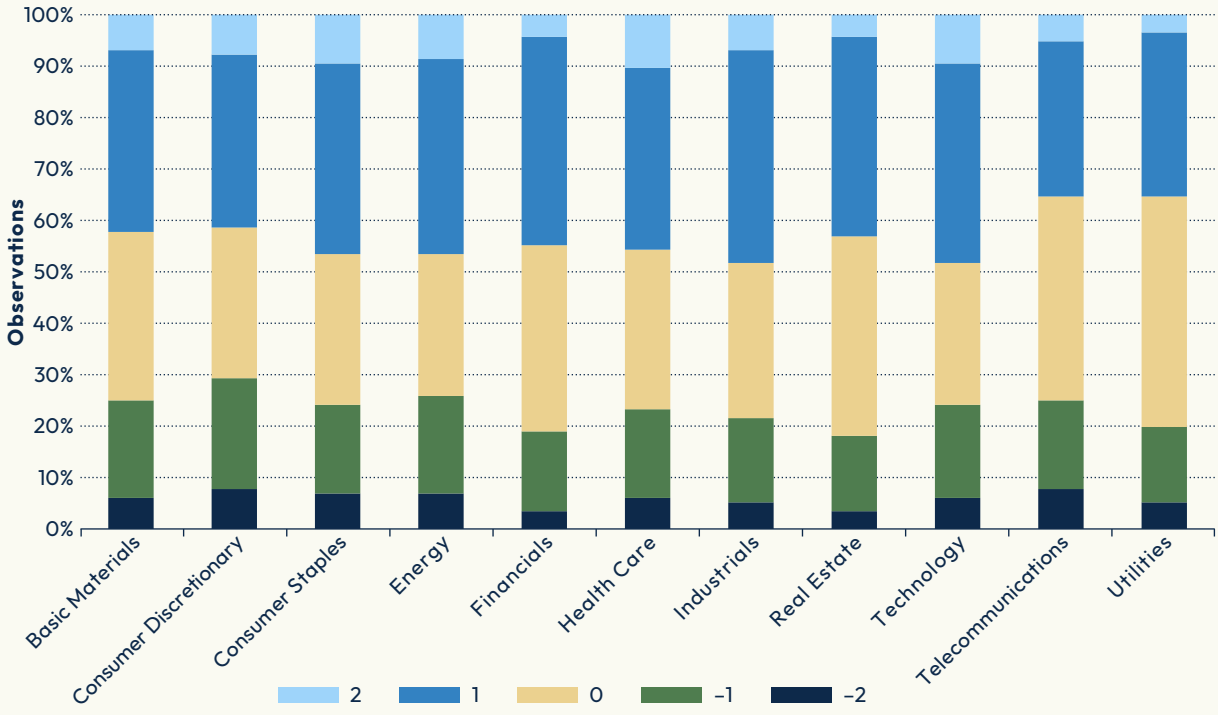
Source: STOXX. Simulated data, December 31, 2014 – December 31, 2025.

Figure 4: Narrative Momentum category mix over time.



Source: STOXX. Simulated data, December 31, 2014 – December 31, 2025.

Figure 5: Sector Narrative Momentum distribution.



Source: STOXX. Simulated data, December 31, 2014 – December 31, 2025.

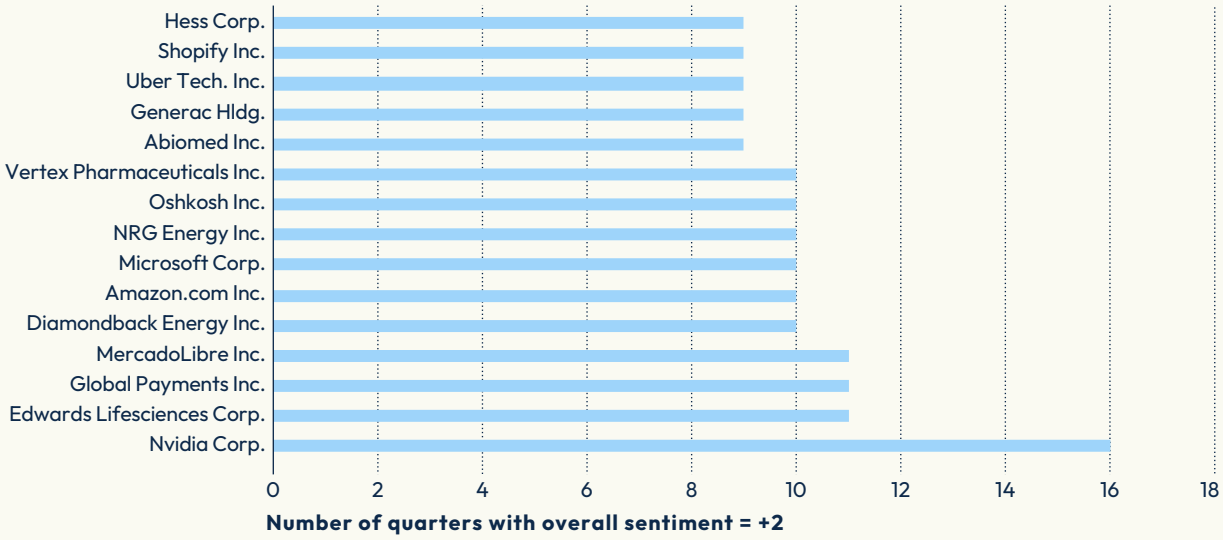
6.2 Individual issuers, crisis reaction and correlations

With nearly 40,500 scores (900 securities × 45 quarters) generated for this research, it would have been impractical for a subject matter expert to review whether each score matched reality. We therefore relied on the evaluation system described in section 4, but could also inspect results for intuitiveness.

6.2.1 Individual issuers

The charts below show the number of extreme scores (2 or –2) across the full sample. While this is not an exhaustive list of issuers associated with bullish/bearish sentiment, some expected issuers (e.g., Nvidia) are to be seen.

Figure 6: Companies with the most +2 Narrative Momentum observations.



Source: STOXX. Simulated data, December 31, 2014 – December 31, 2025.

Figure 7: Companies with the most –2 Narrative Momentum observations.

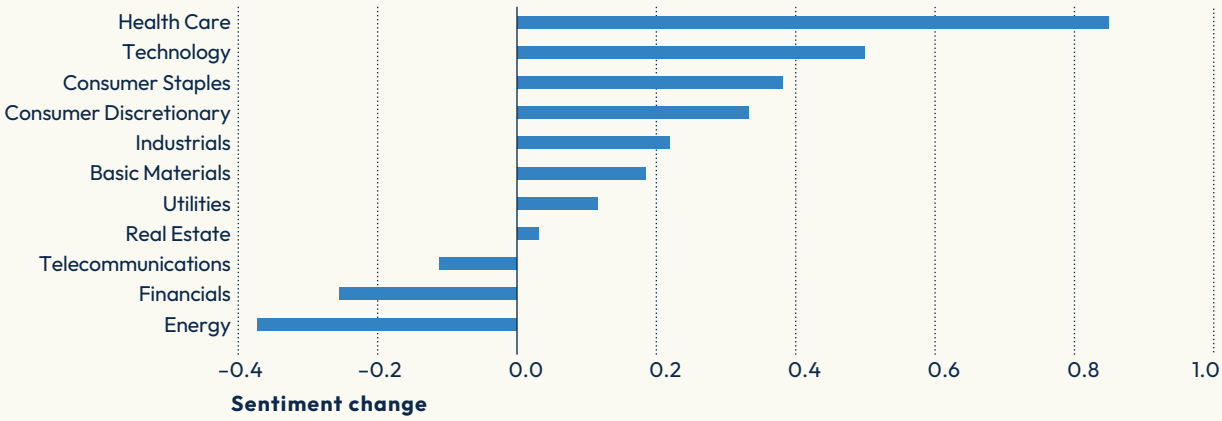
Source: STOXX. Simulated data, December 31, 2014 – December 31, 2025.

6.2.2 COVID-19 changes

We also zoomed into the COVID-19 period to see how the scores evolved before and after lockdown.

The chart below shows the change in average sentiment at the sector level, which corresponds to intuition: Health Care, Technology and Staples were the biggest beneficiaries of changes in spending, while Energy was hit hardest.

Figure 8: Top sector-level sentiment changes Q1 2020 – Q2 2020.



Source: STOXX. Simulated data.

6.2.3 Correlations

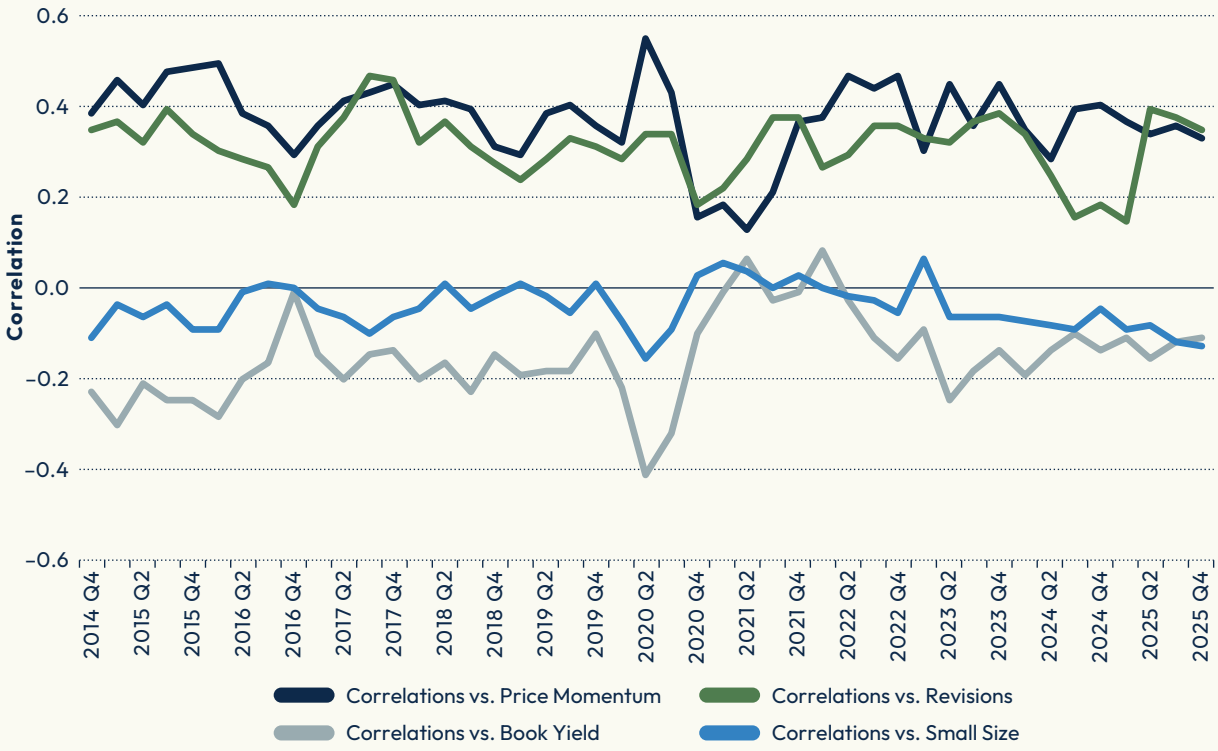
We analyzed the correlations between Narrative Momentum and more traditional factors used in factor investing. Our a priori view was that, if sentiment is indeed a potential Momentum signal, it should have a weakly/moderately positive correlation with other Momentum signals and a negative correlation with Value signals. These correlation levels, if true, would give confidence that the sentiment signal is related to Momentum but is not redundant to other factors.

The results met our expectations for an encouraging result, as can be seen in the exhibit below.
We compared Narrative Momentum to:

- **Price Momentum:** The "classic" definition of Momentum, this is the security return over the previous 12 months, omitting the most recent month.
- **Analyst Revisions:** An updated "fundamental" view of Momentum, this uses analyst ratings to capture the change in analyst sentiment. It represents the ratio of upgrades to downgrades for future earnings per share (EPS) in the trailing quarter.
- **Book Yield:** The "classic" definition of Value, this is the book value-to-price ratio.
- **Size:** The log of the market cap.

Each signal is observed quarterly and standardized for the STOXX USA 900 universe. Correlations are measured cross-sectionally.

Figure 9: Quarterly correlation of Narrative Momentum with external signals.



Source: STOXX. Simulated data, December 31, 2014 – December 31, 2025.

We included the Size factor to examine any large-cap bias in the signal and note that there does not appear to be a correlation between the signal and Size – a good result. We will revisit correlations in the next section.

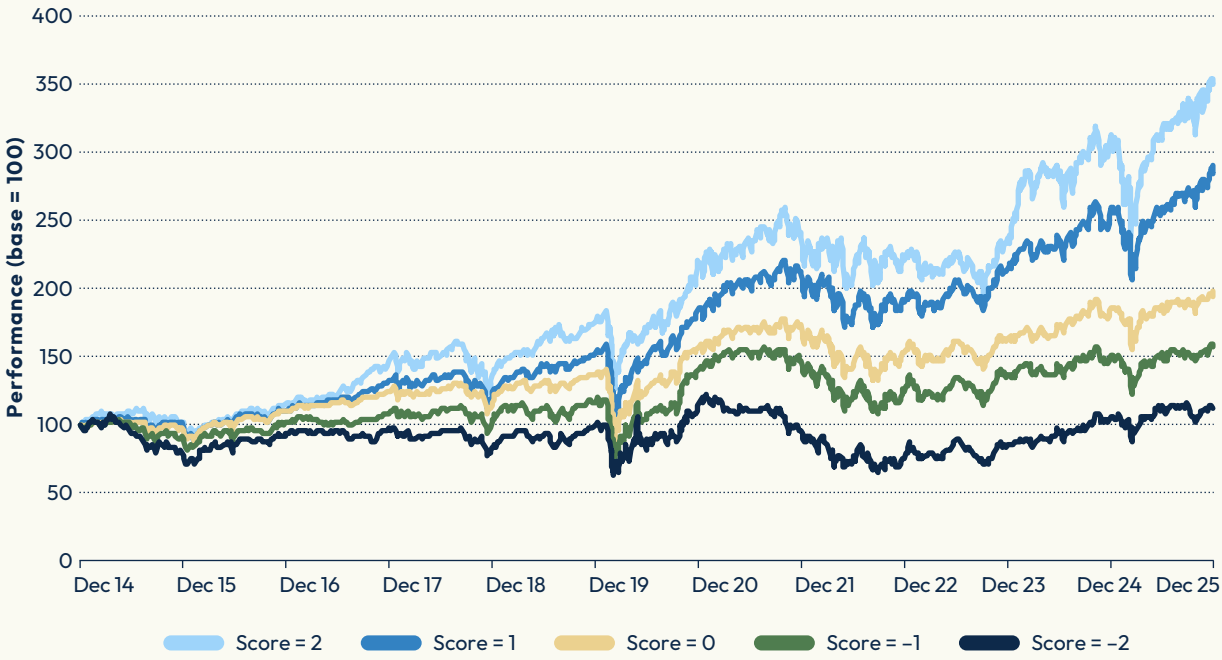
7. Signal performance

7.1 Initial performance view

We started testing the ultimate value of the signal – performance – simply, using portfolios of securities in each scoring category (-2 to +2). Each component was weighted equally and a quarterly review cycle was used. The universe and time period were the same as for signal creation (STOXX USA 900 and Q4 2014 – Q4 2025).

As can be seen in the chart and table below, the results are directionally consistent with expectations for factor performance. There is a monotonic increase in performance from the worst sentiment bucket to the best.

Figure 10: Narrative Momentum performance.



Source: STOXX. Simulated data, December 31, 2014 – December 31, 2025.

Table 2: Performance of score categories.

Category	Annual return (%)	Annual risk (%)	Return/risk
Issuers with score = 2	12.19	19.30	0.63
Issuers with score = 1	10.16	17.32	0.59
Issuers with score = 0	6.38	16.93	0.38
Issuers with score = –1	4.28	19.43	0.22
Issuers with score = –2	1.11	23.32	0.05

Source: STOXX. Simulated data, December 31, 2014 – December 31, 2025.

7.2 Real-world performance view

For a more real-world example of how a signal like this could be implemented, we use an optimized portfolio construction methodology similar to that used with modern STOXX factor indices. The goal of this portfolio construction methodology was to maximize the exposure of the index to the Narrative Momentum signal while closely tracking the benchmark, reducing unintended biases and controlling transaction costs (via turnover). The construction details are as follows:

- **Benchmark** = STOXX USA 900
- **Rebalance** = quarterly
- **Objective:** Maximize alpha exposure (Narrative Momentum signal)
- **Constraints:** Long only, fully invested. | Only securities in the benchmark can be held. | Maximum tracking error to benchmark = 1%. | Maximum one-way, quarterly turnover = 5% | Maximum active sector (ICB Level 1) exposures = +/– 2% | 0.98 < beta < 1.02 | Liquidity: maximum percentile days to trade | Min. asset weight = MAX (0%, ben. wgt – 2%) | Max. asset weight = MIN (ben. wgt + 2%, 20 × ben. wgt)

With this setup, we looked at the risk/return, exposure and correlation metrics for both standard Price Momentum (defined as in the prior section, i.e., the price performance for months 12 – 1) and our Narrative Momentum signal. Again, we found positive results.

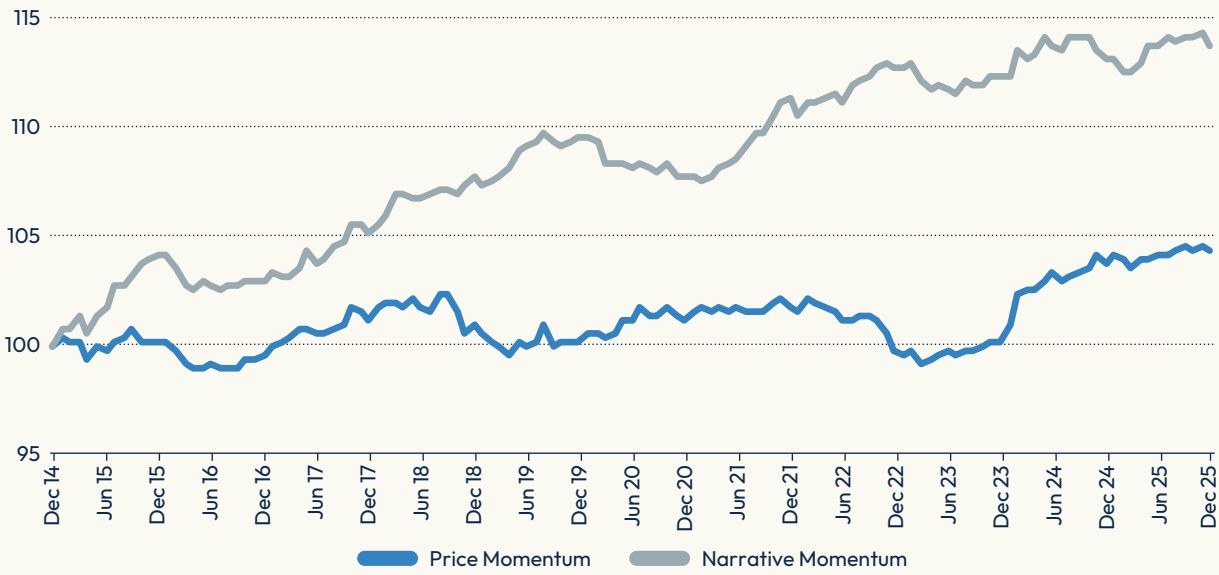
The Narrative Momentum signal, while correlated to standard Price Momentum, outperforms it over the backtest time period mentioned. While the results are encouraging, they should be considered in the light of the limitations discussed later in the paper.

Table 3: Performance of real-world setup.

	STOXX USA 900	Price Momentum	Narrative Momentum
Realized return (% annualized)	13.12%	13.55%	14.43%
Realized risk (% annualized)	15.32%	15.40%	15.45%
Sharpe ratio	0.86	0.88	0.93
Realized active return (% annualized)	-	0.42%	1.30%
Realized active risk (% annualized)	-	1.29%	1.29%
Information ratio	-	0.33	1.01
Average securities held	897.8	364.5	359.1

Source: STOXX. Simulated data, December 31, 2014 – December 31, 2025.

Figure 11: Benchmark outperformance, real-world setup.



Source: STOXX. Simulated data, December 31, 2014 – December 31, 2025. The active return chart series shows the compounded active monthly returns vs. the STOXX USA 900 benchmark.

In addition, the real-world Narrative Momentum strategy has intuitive exposures, with tilts towards Price Momentum and Growth, but no strong overlap. The table below shows active (benchmark-relative) exposures to the Axioma factors. A value of, for example, -8.3% corresponds to an exposure that is 0.083 standard deviations lower than the benchmark. Given the tight tracking error constraint, muted exposures are expected. The tilt away from Dividend Yield warrants further exploration.

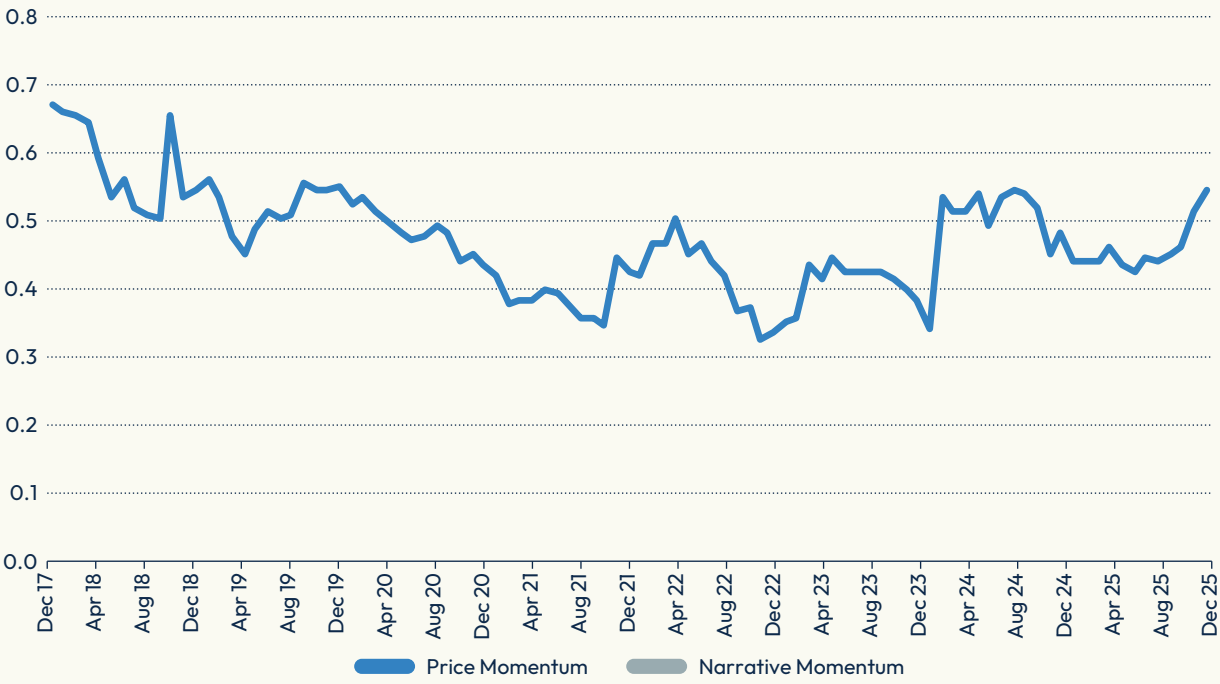
Table 4: Exposures to Axioma style factors.

	Price Momentum (%)	Narrative Momentum (%)
Dividend Yield	–0.3	–8.3
Earnings Yield	0.6	-0.1
Exchange Rate Sensitivity	0.5	0.3
Growth	0.7	3.9
Leverage	4.8	-2.6
Liquidity	3.0	0.8
Market Sensitivity	1.4	0.9
Medium-term Momentum	20.3	8.3
Profitability	1.9	–1.4
Size	–1.1	–3.6
Value	–0.5	0.9
Volatility	2.0	–1.2

Source: STOXX, Axioma. Simulated data, December 31, 2014 – December 31, 2025. Exposures calculated relative to the STOXX USA 900 benchmark and averaged over time at each quarterly rebalance. The Axioma US4-MH risk model was used.

Finally, the correlation of the active returns between the definitions of Standard Momentum and Narrative Momentum is slightly more positive than the correlation levels of the raw signals, although it still points to a positive but not strongly overlapping relationship. This is again an encouraging result, since a likely use case for Narrative Momentum is as part of a multisignal approach to Momentum.

Figure 12: Correlation of Active Price Momentum returns to Active Narrative Momentum returns, 36m rolling.



Source: STOXX. Simulated data, December 31, 2014 – December 31, 2025.

8. Case study: Addressing data leakage

As an example of one of the unique aspects of building a scoring engine with an LLM at its core, we examined the case in which qualified web search results were used to derive the score. While we specified allowable date ranges in the prompt, we also enforced strict controls on the outputs and clearly defined what "good" looks like. While this sounds logical in practice, it could also create a vicious cycle in which the LLM "wants" to conform to our structure and goals and does not fully conform to the rules we provide.

Cases in which an agent returns search results outside the specified window are referred to as "data leakage". To remedy this, we use a redundant security approach. On the one hand, we enforce a strict "before date" with the agent, so that all search results incorporate this date as a criterion. On the other hand, a secondary process detects any residual leakage and reruns the process. This can potentially be done in an infinite loop until the leakage is 0. However, for practical purposes, we found that one retry produced a material improvement in limiting leakage. The improvement in the results is shown below.

Table 5: Leakage improvement statistics.

Protection	Leakage
None	~60% for earlier periods, 40% for more recent periods
Before date filter	~15%
Before date filter + one retry	~7%

Source: STOXX. Percentages based on the number of tickers across the full ~10-year quarterly sample.

Note that this is primarily an issue for backtesting, since (forward-looking) data leakage is not possible when running a live strategy.

9. Limitations, risks and how not to misuse this signal

While we have tried to ensure that our backtested results closely mirror a production setup, there are inherent limitations. Although this is the case with all signal testing, there are unique considerations that must be borne in mind with the LLM-based approach.

Look-ahead bias: We sought to control data leakage by controlling the links used for scoring. For backtesting, there are two potential issues:

- Firstly, due to the nature of LLMs, we cannot know for certain that these were the only inputs considered during score creation.
- Secondly, the LLM used to evaluate sentiment was trained over a period that may already include the evaluation window, i.e., we may be evaluating Q1 2020, but the LLM was trained on data after this period.

Since neither problem can be solved easily, they must be considered alongside the historical results. The real litmus test will be a live track record.

Uncaught hallucinations/violations: Identifying and reducing hallucinations was a key challenge in this project. Some hallucinations, such as suspicious links, were easy to find but other violations, such as using information from outside the specified time window, were more subtle. In addition, we found that the more output structure we enforced, the more hallucinations were encouraged and the more subsequent controls were needed. Our final output produced an acceptable result, but this will need ongoing monitoring.

Human calibration: The sentiment agent mimics what a human analyst might do and follows the guidelines we constructed for this. We also qualitatively judged the agent's output, particularly during the quality assurance (QA) process, in which we made distinctions between good and bad assessments. However, a different set of humans might have generated a different set of guidelines and assessments, leading to different results. This is similar to the quantitative investing world, though, in which researchers may disagree not only on the signals, but also on data sources, normalization techniques and other design choices.

Hidden biases: Thus far, we have tested US large-/mid-cap securities. As the sentiment agent is powered by Google search grounding, there is an intuitive linkage between the data available and the signal quality. While we have not observed a bias towards large-cap securities, we have noted different distributions of sentiment in different sectors/industries. A broader test including small-cap and international securities would be needed to rule out any additional biases associated with this approach.

Model sensitivity: While innovation in LLM capabilities is a hallmark of the space, it can be an enemy of quantitative strategies requiring reproducibility. As mentioned in section 3, we benchmarked different models for scalability; however, we expect the models we chose to be retired at some point. In contrast to what happens during the switch to a different version of portfolio construction software, we expect that sentiment results produced using newer models will differ from our current set of results. While it is possible that they will be qualitatively better, they might equally introduce biases or behavior that were not previously seen. This means that additional effort and setup changes will be needed to control for them.

Version control: Reproducibility is a key tenet of indexing, but it is at odds with what happens with generative AI. This is another reason why we use ordinal scores, since these increase the probability of repeatable performance. Section 4 sets out our approach and the testing we do to avoid drift in results with every run, but this is nevertheless unavoidable in the end. Consequently, strict version control is needed: each run of signals must have a clear version identifier so that the data used for production can never be in doubt.

Taken together, these caveats point to a few areas in which the LLM-driven strategy backtest warrants additional caution compared to traditional quantitative signals:

- Residual look-ahead bias may be more difficult to eliminate.
- Model choice may have a larger impact than other quantitative construction choices (e.g., the risk model).
- "Live" strategy performance becomes even more important, and
- Greater governance is needed on production runs and reruns.

10. Conclusion and next steps

This paper details the potential for new investment signals that might be suitable for index incorporation. The resulting performance profiles are encouraging – both from a growth and a diversification perspective – and warrant further exploration of this concept. Such exploration should include deeper and broader coverage, higher frequency runs and the tracking of out-of-sample performance.

Arriving at these signals required both discovery and engineering. Far from an out-of-the-box solution, achieving this level of performance with agents demanded rigorous systems engineering and intentional context management.

What excites us most, however, is not the signal itself, but the chances it offers to continue merging the qualitative and quantitative worlds. The approach described would also allow us to run agentic signal generation more frequently than once a quarter.

The significant opportunity we see is the ability to translate qualitative information into structured, auditable and scalable inputs for systematic investing. If implemented with appropriate controls, this opens up new possibilities for index design and for investors seeking transparent exposure to information that has historically been difficult to capture systematically.

11. Bibliography

Ang, A., Azimbayev, N., and Kim, A. "The Self-Driving Portfolio: Agentic Architecture for Institutional Asset Management," arXiv:2604.02279, 2026, https://arxiv.org/abs/2604.02279

Ang, A., Brown, M., Hum, B., Renshaw, A., Schwaiger, K., Seegopaul, H., Singhal, A., and Smart, L. "How Do Low Tracking Error, Multifactor ETFs Fit into the Factor Investment Landscape?" *The Journal of Beta Investment Strategies*, 15(1), 2024, pp. 28–38. https://doi.org/10.3905/jbis.2024.1.054

Araci, D. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models." arXiv:1908.10063, 2019, https://arxiv.org/abs/1908.10063

Axioma Worldwide Equity Factor Risk Model www2.simcorp.com/l/48862/2025-04-08/ndsc4r/48862/1744106540wUFaT59d/Axioma_WW_Equity_Factor_Risk_Model.pdf?_gl=1*1o03p5a*_gcl_au*MjA1MzIxNDAzMy4xNzgwMzQ5MzY3

Du, K., Zhao, Y., Mao, R., Xing, F., and Cambria, E. "Natural Language Processing in Finance: A Survey," *Information Fusion*, 115, 102755, 2025, https://doi.org/10.1016/j.inffus.2024.102755

Jegadeesh, N., and Titman, S. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance*, 48(1), 1993, pp. 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

Jegadeesh, N., and Titman, S. "Profitability of Momentum Strategies: An Evaluation of Alternative Explanations," *Journal of Finance*, 56(2), 2001, pp. 699–720. https://doi.org/10.1111/0022-1082.00342

Loughran, T., and McDonald, B. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance*, 66(1), 2011, pp. 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

STOXX Ltd. "STOXX Equity Factor Indices", 2026, https://stoxx.com/solutions/stoxx-equity-factor-indices/

Tetlock, P. C. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market.," *Journal of Finance*, 62(3), 2007, pp. 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. "More Than Words: Quantifying Language to Measure Firms' Fundamentals," *Journal of Finance*, 63(3), 2008, pp. 1437–1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

Yang, H., Zhang, B., Wang, N., Guo, C., Zhang, X., Lin, L., Wang, J., Zhou, T., Guan, M., Zhang, R., Dan Wang, C. "FinRobot: An Open-Source AI Agent Platform for Financial Applications using Large Language Models" https://arxiv.org/abs/2405.14767

12. Offices and contacts

Learn more about STOXX & DAX Indices on STOXX.com

Zug
Theilerstrasse 1A
6300 Zug
Switzerland
P | +41 43 430 71 60

London
1 London Bridge
4th Floor, West Building
London, SE1 9BG
United Kingdom
P | +44 20 7862 7680

Frankfurt
Taunus Tower
Mergenthalerallee 73–75
65760 Eschborn
Germany
P | +49 6196 7719 257

Paris
5 Rue du Renard
75004 Paris
France
P | +33 1 55 27 38 38

Milan
7th Floor
Piazza Velasca 3/5
20122 Milan
Italy

Prague
Futurama Business Park Building E
Sokolovska 662/136e
186 00 Prague 8
Czech Republic
P | +420 226 238 419

ISS Inc. (New York)
1177 Avenue of the Americas
14th Floor
New York, NY 10036
USA
P | +1 646 680 6350

Hong Kong
19th Floor, Nexxus Building
41 Connaught Road Central
Hong Kong
P | +852 3107 8030

ISS K.K. (Tokyo)
Sumitomo Fudosan Kanda, Building 16F
7 Kanda Mitoshiro-cho, Chiyoda-ku
Tokyo 101-0053
Japan
P | +852 3107 8030

ISS Australia Pty. (Sydney)
7F, 55 Clarence Street
Sydney, NSW 2000
Australia
P | +61 2 8048 3999

New Dehli
404-407, Tower-B, 4th Floor
Unitech Cyber Park, Sector-39
Gurugram, Haryana 122003
India

Call a STOXX representative
Customer support
customersupport@stoxx.com
P | +41 43 430 72 72

STOXX

Part of

ISS STOXX

DEUTSCHE BÖRSE GROUP