Unlocking Alpha from Textual Data

PRACTICAL INSIGHTS FOR INSTITUTIONAL INVESTORS USING LARGE LANGUAGE MODELS

Machiel Westerdijk, Olivera Rakic and Ashraf Mansur

INTRODUCTION: BEYOND THE AI HYPE – PRACTICAL ALPHA FROM TEXTUAL DATA

The asset manager's competitive landscape has intensified, and traditional investment factors – such as value, momentum, and quality – are increasingly well-understood, resulting in diminished returns (e.g. McLean & Pontiff, 2015; Calluzzo, Moneta, & Topaloglu S., 2019; Jacobs, Kenneth, & Lee, 2025).

One promising yet still underutilized source of alpha lies in textual data (e.g. see review by Sun et al. 2024). While traditional investment strategies rely on structured financial data, a vast amount of valuable and complementary information remains hidden within unstructured textual sources, such as annual reports, patents, earnings calls transcripts, and employee reviews.

Until recently, systematically leveraging these textual sources for investment signals remained impractical, due to technological and infrastructural limitations.

The rapid advancement of Large Language Models (LLMs) – AI systems capable of deeply understanding and interpreting human language – has changed this. LLMs now offer asset managers a powerful means of systematically extracting and quantifying insights from massive textual datasets. Leaders in the AI space, such as Open AI, Alphabet, and Anthropic, offer the most advanced models through developer-friendly interfaces, making it straight-forward to start experimenting with. While the transformative potential of LLMs is often discussed in broad terms, practical implementation in investment processes requires clear understanding, specialized expertise, and robust infrastructure.

While institutional investors must also navigate critical aspects such as ESG, sustainability, and risk management, this article focuses exclusively on identifying and capturing alpha in publicly traded equity markets using AI. We highlight academic evidence supporting alpha generation from textual data, discuss practical aspects of integrating textual signals into existing investment strategies, and address important AI-related risks and mitigation techniques. A practical case study further demonstrates how advanced peer group detection (AI-TNIC) significantly enhances traditional momentum strategies, delivering measurable alpha.

Machiel Westerdijk, PhD
Managing Director at Entis

Olivera Rakic, PhD
Head of Quant Research at Entis

Ashraf Mansur, MEng
Head of Quant Products at Entis

Crucially, we provide practical guidance for translating academic insights into actionable investment strategies, emphasizing robust infrastructure, disciplined risk management, and strategic considerations around internal versus external development of signals, to help institutional investors effectively position themselves for sustained and differentiated performance.

ALPHA POTENTIAL FROM TEXT DATA – EVIDENCE AND TYPES OF SIGNALS

Traditional quantitative strategies have historically relied primarily on structured financial data – standardized, machine-readable numbers such as financial statement line items, market prices and volumes, and accounting ratios – capturing only a small fraction of all available company information. Information about companies that is less likely to be priced in, however, is qualitative and embedded in textual data – e.g. corporate filings, patents, analyst reports, employee reviews – which had limited usage in systematic investment processes due to the high information processing and data quality management barriers. Recent advancements in LLMs have revolutionized this, enabling automated, large-scale extraction of alpha-generating signals from unstructured text data.

ACADEMIC EVIDENCE OF ALPHA FROM TEXTUAL DATA

Academic research confirms the alpha potential of signals derived from text data across several critical investment domains:

- Innovation and Patents: Patents predict long-term market leadership.
- Peer and Network Effects: A firm's strategic network positioning predicts competitive advantage and stock returns.
- Governance Quality: Consistent governance disclosures predict stock outperformance.

Table 1 summarizes key academic findings.

In addition to being used as idiosyncratic sources of alpha, the above described NLP-derived textual signals can also significantly enhance traditional factor-investing approaches by capturing insights into innovation, governance, and corporate culture. Table 2 highlights some opportunities.

Table 2

Factor	Textual Signal Enhancements
Value	Improved peer benchmarking, intangible asset valuation (patent quality)
Momentum	Narrative-driven sentiment, refined industry momentum via semantic peer identification
Quality	Governance credibility, innovation strength, corporate culture
Low Volatility	Early detection of hidden risks, crisis management clarity
Size	Enhanced small-cap coverage via niche textual sources
Investment	Capital allocation clarity, R&D direction, sustainability alignment
ESG	Real-time ESG risk detection, corporate ethics, and culture benchmarking

In summary, LLM-derived textual signals represent a transformative new source of alpha, complementing traditional factor strategies. Additionally, although annual reports and patents are public, the high information processing barrier limits adoption and slows alpha decay.

INCORPORATION IN INVESTMENT STRATEGIES

Having established the alpha potential of AI-derived textual signals, the next step is to consider their integration into existing investment processes. Practical implementation requires clearly defined selection criteria, adjustments to portfolio construction processes, and awareness of infrastructural implications. Below,

Table 1

Signal Category	Key Insight	Reported Alpha (bps/month)
Peer & Industry Similarities (e.g. Hoberg & Phillips 2016, 2018)	Highlights a company's position within its competitive landscape by analyzing similarities in business models, market focus, and risk profiles.	70–97
Governance Quality (e.g. Fee, Li, & Peng 2022, Martin, Xu, & Zhou 2021)	Assesses impact of management stability, compensation policies, and governance practices on expected returns.	30–55
Strategic Networks (e.g. Eisdorfer et al. 2022)	Identifies strategic partnerships, and competitor network connections that influence competitive advantage and market positioning.	70–90
Qualitative Disclosure Shifts (e.g. Cohen, Malloy, & Nguyen 2020)	Evaluates shifts in company disclosures, their causes, and their impact on expected returns.	48–70
Communication Discrepancies (e.g. Zhang 2024)	Examines gaps between qualitative statements and quantitative financial metrics to detect potential misalignments.	48–55
Innovation & Technology (e.g. Drechsler, Müller, Wagner 2021, Kim 2022)	Measures a company's focus on emerging technologies and its ability to innovate, indicating future growth potential and industry leadership.	20–80
Corporate Culture (e.g. Au 2019)	Analyzes employee reviews to evaluate company's culture, such as teamwork and integrity as key drivers of long-term performance.	30–76

Note: The alpha estimates listed above are based on academic studies and should be interpreted with care. These results are typically derived from idealized backtests using clean historical data, and do not account for real-world frictions such as trading costs, turnover, slippage, or capacity constraints. Furthermore, the signal categories are not mutually exclusive and often exhibit significant correlation, meaning the reported alpha values cannot be assumed to be additive in a portfolio context.

we provide a structured approach for incorporating these signals into both quantitative and fundamental investment strategies.

SIGNAL SELECTION AND VALIDATION

Given the abundance of possible signals derived using LLMs, investment teams must first prioritize signals based on robust statistical evidence, economic rationale, and incremental predictive value beyond traditional metrics. Standard quantitative validation methods such as historical backtesting, forward testing, and out-of-sample validation apply equally to these textual signals. The goal is not just statistical robustness, but also economic intuition: signals must reflect meaningful and understandable business realities.

PRACTICAL INTEGRATION INTO INVESTMENT PROCESSES

Integration into existing investment strategies varies significantly depending on whether the approach is quantitative, fundamental, or a hybrid "quantamental" method.

> NLP-DERIVED TEXTUAL SIGNALS CAN BE USED AS IDIOSYNCRATIC SOURCES OF ALPHA, BUT ALSO TO ENHANCE TRADITIONAL FACTOR-INVESTING APPROACHES

For purely quantitative strategies, the incorporation of LLM-generated textual signals is relatively straightforward once signals have been validated (see Figure 1). These signals typically come as numerical scores – such as governance quality ratings or innovation intensity metrics – similar in structure to traditional quantitative factors like value or momentum. Portfolio managers can directly integrate these scores into factor models or predictive algorithms, enhancing predictive power and portfolio diversification.

A crucial practical step is assessing interactions between new textual signals and traditional factors. Portfolio managers should explicitly test how the introduction of textual signals affects existing factor exposures, ensuring diversification benefits and avoiding unintended biases.

For fundamental or quantamental strategies, textual signals serve primarily as systematic decision-support tools. They provide structured insights into qualitative aspects such as corporate culture, strategic innovation, or peer comparability. Analysts integrate these signals into their investment analyses to sharpen conviction levels, identify hidden risks, or reassess valuations.

A clear example is the textual signal capturing digital innovation (e.g. Drechsler, Müller, Wagner 2021). Using LLMs, investment teams can systematically measure the degree of digital innovation by comparing corporate filings to a carefully constructed textual benchmark describing the concept of "digital innovation." This method yields a quantitative digital innovation score ranging from 0 to 1. Practically, such a score allows quantitative managers to construct straightforward long-short portfolios, for example, by taking long positions in companies ranked in the top 20% and short positions in those ranked in the bottom 20%. Alternatively, fundamental or quantamental analysts integrate this digital innovation score as part of a quality assessment, improving their evaluation of competitive advantages and strategic resilience.

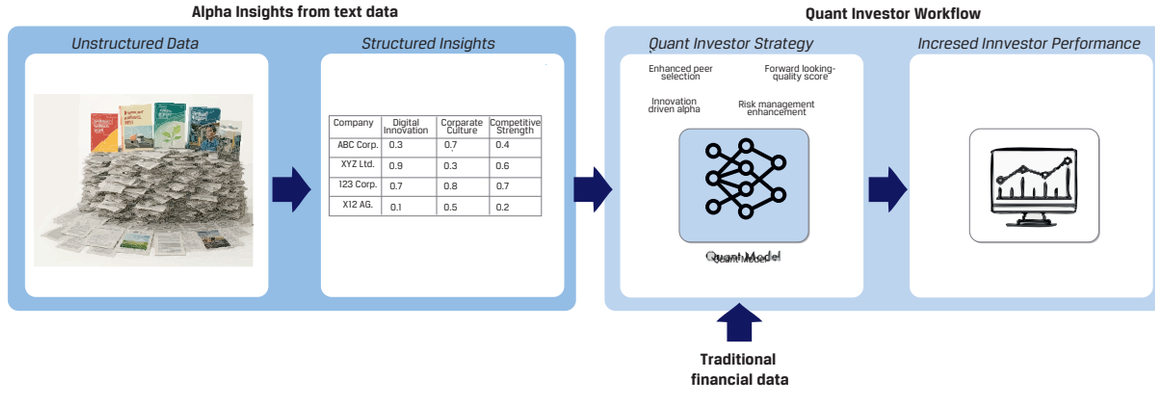
EXTERNAL PROCUREMENT VS. IN-HOUSE DEVELOPMENT OF SIGNALS

The practical considerations of incorporating textual signals depend greatly on whether these signals are externally procured or internally developed.

Procuring AI-derived textual signals from specialized external providers simplifies practical implementation considerably. Platforms like Neudata and Eagle Alpha provide overviews of data providers in this space. NLP-based quant signals can be acquired either directly from data producers such as Entis, or through data aggregators such as Quandl.

Such signals typically arrive as structured numerical data that easily integrate into existing investment workflows without major operational adjustments. Asset managers can choose between standardized 'off-the-shelf' signals or custom-developed signals tailored exclusively to their specific strategy and investment universe. Off-the-shelf signals typically offer cost efficiency, ease of integration, and rapid deployment. Tailored signals, although more expensive, provide strategic differentiation, enhanced

Figure 1
Incorporation of insights from textual information in the quant investor's workflow



alignment with investment objectives, and potentially stronger and more durable alpha due to reduced availability to competitors.

Conversely, internal development requires substantial infrastructural investment and deep technical expertise. Processing tens of thousands of corporate documents globally on a daily basis, including historical records spanning decades, demands specialized infrastructure for text extraction, automatic translations, bias corrections (e.g., correcting country-specific language nuances), and rigorous validation processes. While the technical complexity is considerable, the primary advantage is the potential for stronger alpha due to exclusive signal ownership, enhanced customization capabilities, and deeper strategic alignment.

BUILD OR BUY DECISION DEPENDS ON STRATEGIC OBJECTIVES, ALPHA EXPECTATIONS, AVAILABLE RESOURCES, AND DESIRED COMPETITIVE DIFFERENTIATION

Ultimately, the choice along this spectrum – from standardized external signals through customized outsourced solutions to fully internal development – depends critically on strategic objectives, alpha expectations, available resources, and desired competitive differentiation.

HUMAN-IN-THE-LOOP VALIDATION AND RISK MANAGEMENT

Regardless of the chosen integration route, "human-in-the-loop" validation is critical to the practical success of using AI-generated signals. While automated extraction of information continues to improve with each new generation of NLP and LLM tools, a small fraction of cases will inevitably require manual intervention to ensure the quality needed for trading models. For example, when extracting business descriptions from annual reports (needed for the computation of the AI-TNIC signal described below), in many cases filings are well structured and models can accurately extract the correct pieces of text. But in a non-negligible fraction of cases filings have a very irregular, free-format structure, and business descriptions could be mistaken with e.g. CSR disclosure, making it important to have an analyst manually review high-risk cases. Therefore, infrastructure that facilitates human oversight and allows analysts to manually correct or refine lower-quality outputs remains essential.

MANAGING AI RISKS IN PRACTICE

Generative AI models, particularly LLMs, offer powerful capabilities for interpreting textual data. Yet, institutional investors must approach their use cautiously due to inherent risks. Effective risk management practices are essential, given regulatory demands and the high standards of transparency and accountability expected from institutional investors.

KEY AI RISKS

When employing LLMs, four primary risks need explicit attention:

- **Transparency (the "black box" problem):** LLMs are inherently opaque, making it difficult to precisely trace how inputs lead to specific outputs. Unlike traditional quantitative models with clearly defined formulas, the reasoning within LLMs is challenging to interpret.
- **Biases:** Biases may originate directly from the LLM training data, causing the model to produce systemic errors. Additionally, biases can emerge from the LLM itself, through the selection of input texts provided to the model. For example, when consistently evaluating texts about digital innovation from specific countries or sectors, the resulting scores may reflect unwanted country- or sector-specific textual characteristics rather than genuine innovativeness.
- **Unintended use of prior knowledge and forward looking bias:** LLMs are trained on years of historical data, which poses a challenge for backtesting trading strategies. If an LLM model is to analyze a news item that lies inside its training window, it already "knows" what might have happened next in the real world, so any forecast it gives might be contaminated by hindsight (e.g. Glasserman & Lin 2024).
- **Hallucinations:** Hallucinations occur when LLMs produce plausible yet entirely fabricated responses. This can significantly compromise data integrity and lead to erroneous investment signals.

PRACTICAL RISK MITIGATION STRATEGIES

While a complete overview of all available methods goes beyond this article, practical risk mitigation generally revolves around two core principles: input control and output validation. Effective risk management begins with careful preparation and standardization of the input texts provided to LLMs. Treat the model as a highly intelligent analyst, explicitly instructing it to objectively interpret only the provided texts without using prior knowledge. Practical measures include:

- Removing company, product, and person names to prevent forward-looking biases, ensuring the model evaluates texts based purely on their content at a specific historical moment (time stamping).
- Providing clear, informative prompts and relevant textual sections to minimize hallucinations. Additionally, embedding-based methods (converting texts to numeric vectors) are inherently deterministic and do not lead to hallucinations.

Rigorous validation of the model's outputs is equally essential. Practical validation methods include:

- Statistical anomaly detection to identify unusual or inconsistent signals,
- Cross-validation with the same model, independent models, or simpler baseline methods to confirm the consistency and accuracy of generated scores,

- Checking the output against the ground-truth data, e.g. labeled by analysts,
- Correcting structural biases directly in numeric outputs (embeddings) to eliminate subtle but systematic distortions.

These methods can be partially automated, although human involvement and judgement is still necessary at the moment. However, one area that is gaining traction is Agentic AI, which holds promise of enabling further automation of extracting quant signals from large textual sources, by chaining specialized LLM workers into an auditable and modular full signal pipeline.

PRACTICAL RISK MITIGATION WHEN USING AI GENERALLY REVOLVES AROUND TWO CORE PRINCIPLES: INPUT CONTROL AND OUTPUT VALIDATION

It is important to note that internal biases originating from an LLM's training data typically pose fewer challenges when models are used primarily for semantic interpretation rather than factual knowledge retrieval. Regular model evaluation and periodic fine-tuning remain important, but primary bias mitigation efforts are best directed toward careful input selection, preparation, and embedding-based adjustments.

CASE STUDY – ADVANCED PEER GROUP DETECTION WITH AI (AI-TNIC)

This chapter provides a concrete illustration of how LLMs can practically generate alpha by improving peer group identification, which can be used for any application that relies on industry classifications. Specifically, we focus on enhancing a well-known factor investing strategy: peer momentum.

WHY BETTER PEER IDENTIFICATION MATTERS FOR ALPHA

Peer momentum, a proven alpha-generating factor, captures the tendency of stocks to follow trends observed in related companies. Traditional peer momentum typically relies on standard industry classifications, which are rigid and often fail to identify all genuine company peers. A more accurate method – introduced by Hoberg & Phillips (2016) through their Text-based Network Industry Classification (TNIC) – uses business descriptions from company filings to capture similarities between companies that are often missed by traditional industry classifications. However, their initial method is limited, as it simply matches words without considering their context. For example, their approach might mistakenly identify the fashion company Ralph Lauren and toy manufacturer Hasbro as close peers because they share common terms such as "brand," "design," and "marketing," despite using these terms in entirely different contexts.

LEVERAGING LLMS FOR CONTEXTUAL UNDERSTANDING

To address this limitation, Entis developed AI-TNIC, an advanced approach leveraging LLMs to capture deeper semantic context and meaning within business descriptions. By transforming texts into numerical embeddings, AI-TNIC accurately identifies truly relevant peers across the global company landscape.

However, applying general-purpose LLMs effectively to financial texts requires specialized pre- and post-processing methods. Table 3 briefly summarizes our practical solutions.

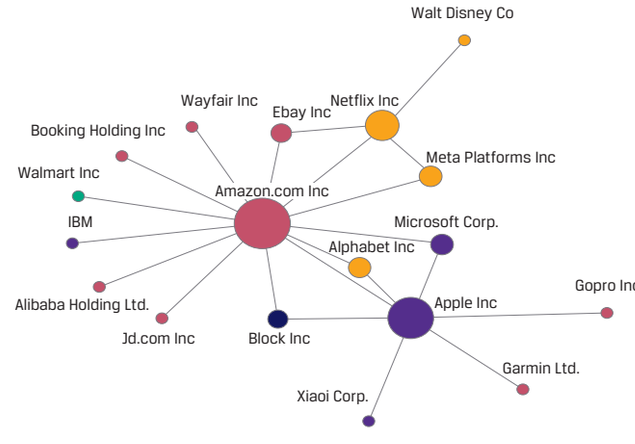
Table 3

Limitation of standard LLMs	Our Solution
Text documents too long	Split text into smaller chunks to preserve meaningful context
Broad general context of embeddings	Emphasize specific business content through embedding adjustments
Geographical bias in embeddings	Remove geographic references and correct country-specific biases
Uninformative or overly formatted texts	Identify and filter out irrelevant text through pre-defined rules

PRACTICAL ALPHA GENERATION: PEER MOMENTUM WITH AI-TNIC

Using AI-TNIC, we constructed an improved peer momentum factor based on the insight that market-relevant information about similar companies moves more slowly across less obvious peer relationships (Hoberg & Phillips 2018). This improved identification of peers results in stronger and more consistent alpha generation compared to traditional industry momentum strategies. For example, traditional industry classifications do not consider Amazon and Netflix as direct peers, despite both competing directly in the video streaming market. Figure 2 illustrates how the AI-TNIC approach correctly identifies this competitive relationship.

Figure 2
According to traditional industry classifications, Amazon and Netflix are not peers, but they compete in the video streaming space – TNIC captures this



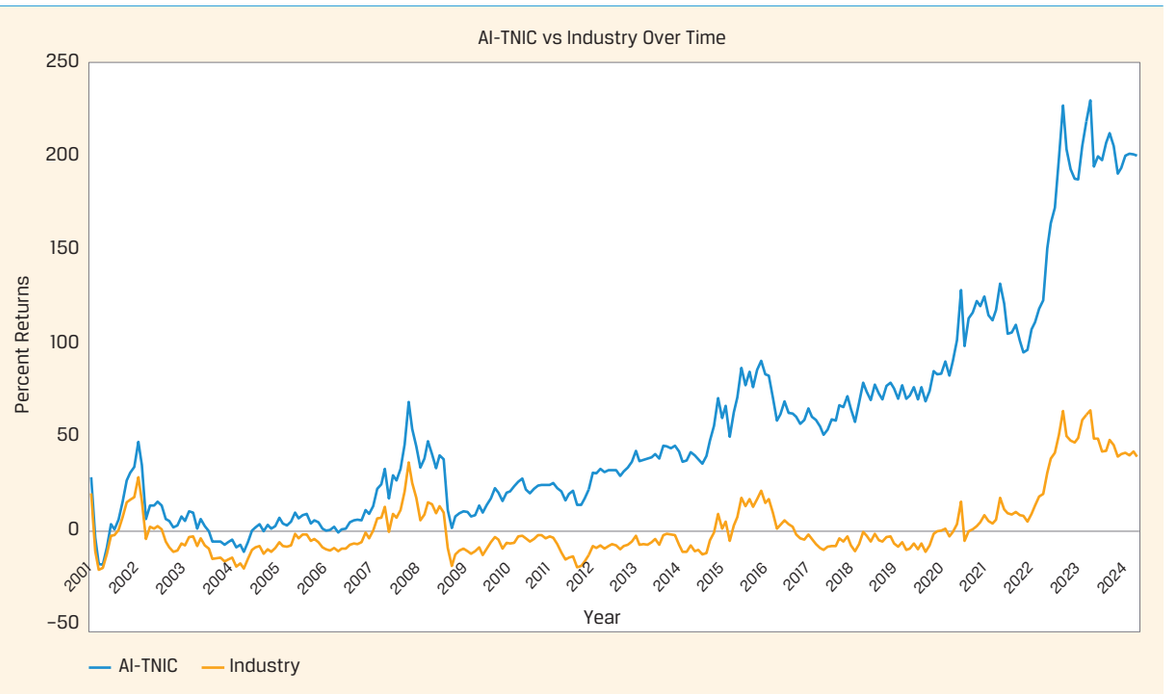


Figure 3
Cumulative Long-Short Performance of AI-TNIC Industry Peer Momentum compared to standard industry classification momentum (2001-2024)

BACKTESTING RESULTS

We applied AI-TNIC peer momentum to a universe of approximately 2,500 U.S. equities (equal-weighted), achieving robust and consistent outperformance. The figure below illustrates the long-term performance of this factor from 2001 to 2024, with cumulative returns exceeding 200% (approximately 5% annualized). Moreover, the AI-TNIC factor experiences similar but less severe drawdowns compared to using standard industry classifications, making it a stable alpha source in volatile markets.

Quantitative analysis in Hoberg & Phillips 2018 using the Fama-French three-factor model indicates a statistically significant monthly alpha of 88 basis points (t-statistic = 2.78) based on a data sample from 1997-2012. The figure above covers a longer time period, through 2024, demonstrating that TNIC-based signals have not yet been arbitraged away. Additionally, TNIC-based signals have high information processing and data quality management barriers so that limits adoption and slows alpha decay. Regarding signal turnover, it is on average 150% annually (two-way), which is comparable to the turnover of traditional momentum strategies. Please note that this analysis does not account for trading costs, slippage, or capacity constraints.

CONCLUSION: AI INSIGHTS BEYOND HUMAN CAPABILITIES

AI-TNIC demonstrates that LLM-driven alpha generation is feasible, and that it provides insights beyond human capability. While analysts traditionally relied on broad industry classifications, LLMs allow precise, dynamic, and global comparisons of companies. This reveals previously hidden patterns of emerging global businesses, capturing subtle similarities and differences far beyond human analytical capacity. Thus, advanced LLM-driven methods like AI-TNIC represent more than a productivity improvement – they offer entirely new ways of understanding market dynamics and generating sustainable alpha.

CONCLUSION: PRACTICAL RECOMMENDATIONS FOR INSTITUTIONAL INVESTORS

This article has explored the potential and challenges of using Large Language Models (LLMs) to systematically extract alpha signals from textual data, for investors looking to uncover new, differentiated sources of alpha.

Achieving robust alpha from text-based signals requires more than simply deploying advanced AI models. It demands disciplined implementation, specialized infrastructure, and rigorous risk management. Based on the insights shared in this article, we offer institutional asset managers the following practical recommendations:

1. **Prioritize Robust Infrastructure with Human Oversight:** Combine automated text processing pipelines with structured human validation processes. This ensures efficiency while maintaining transparency and compliance, and allows for manual data corrections where needed.
2. **Start Small, Then Scale:** Begin with carefully controlled proof-of-concept implementations that allow you to test signal validity, infrastructure robustness, and risk management procedures in a manageable setting before scaling up.
3. **Explicitly Address AI-Specific Risks Early:** Proactively manage key AI risks, including model transparency, biases, forward-looking biases, and hallucinations. Clear input preparation, embedding-based methods, and systematic output validation processes are critical for ensuring reliable, compliant outcomes.

4. **Make Strategic Choices on Outsourcing vs. In-House Development:** Carefully consider your strategic priorities, internal capabilities, and available budgets when deciding whether to develop LLM-driven signals internally or procure externally. Off-the-shelf datasets are quick and cost-effective but may offer limited differentiation. In contrast, fully internal development is resource-intensive but can yield significant alpha advantages. Intermediate approaches – such as collaborative research and co-development with specialized external providers – often offer a balance between cost efficiency and strategic differentiation.
5. **Leverage First-Mover Advantage:** Given that systematic text-based alpha signals remain underexploited, early adopters have a meaningful competitive advantage.

Sustainable alpha generation through AI and textual data is not a one-time effort but an ongoing strategic capability. Successful deployment involves methodical implementation, continuous validation, and systematic improvements to keep pace with technological advancements and changing market dynamics. Ultimately, disciplined, structured AI implementation can empower institutional investors to consistently generate differentiated, high-quality alpha well into the future.

References

- Au S-Y., Dong M., Tremblay A., 2019, Employee Flexibility, Exogenous Risk, and Firm Value, Journal of Financial and Quantitative Analysis, Volume 56, Issue 3: 853-884
- Calluzzo P., Moneta F., Topaloglu S., 2019, When Anomalies Are Publicized Broadly, Do Institutions Trade Accordingly?, Management Science 65(10):4555-4574
- Cohen L., Malloy C, Nguyen Q., 2020, Lazy Prices, The Journal of Finance, Vol. 75, Issue 3: 1371-1415
- Drechsler K., Müller S., Wagner H-T., 2021, The "Digital Premium": Why Does Digitalization Drive Stock Returns?, available at SSRN: https://ssrn.com/abstract=3972173
- Eisdorfer A., Froot K., Ozik G., Sadka R., 2022, The Review of Financial Studies, Vol. 35, Issue 9: 4300-4340
- Fee E., Li Z., & Peng Q., 2023, Journal of Accounting and Economics, Vol 75, Issue 1
- Glasserman P. & Lin C., 2024, Assessing Look-Ahead Bias in Stock Return Predictions Generated by GPT Sentiment Analysis, The Journal of Financial Data Science, Volume 6, Issue 1
- Hoberg G. & Phillips G., 2016, Text-Based Network Industries and Endogenous Product Differentiation, Journal of Political Economy, Vol. 124, Number 5
- Hoberg G. & Phillips G., 2018, Text-Based Industry Momentum, The Journal of Financial and Quantitative Analysis, Vol. 53, No. 6: 2355-2388
- Jacobs I. B., Kenneth N. L., & Lee S., 2025, How Misunderstanding Factor Models Set Unreasonable Expectations for Smart Beta, The Journal of Portfolio Management, Vol 51, Issue 3: 10-21
- Kim J., 2022, New Technologies and Stock Returns, available at SSRN: https://ssrn.com/abstract=4299577
- Martin, X., Xu, J., & Zhou G, 2021, Does Compensation Matter? Evidence from CD&A Disclosures, available at SSRN: https://ssrn.com/abstract=3819394
- McLean D. & Pontiff J., 2015, Does Academic Research Destroy Stock Return Predictability?, The Journal of Finance, Vol. 71, Issue 1: 5-32
- Sun Y., Liu L., Xu Y., Zeng X., Shi Y., Hu H., Jiang J. & Abraham A., 2024, Alternative data in finance and business: emerging applications and theory analysis (review), Financial Innovation, 10, 127
- Zhang, T., 2024, Manager Uncertainty and the Cross-Section of Stock Returns, Available at SSRN: https://ssrn.com/abstract=4854534