Pricing Non-linear Soft Information in Analyst Reports: A Large Language Model Approach to Chinese A-Shares*

Jun Hu[a,*], Peng Zhang[a]

[a] *Southwestern University of Finance and Economics, Chengdu, 611130, PR China*

Abstract

This paper investigates whether unstructured textual narratives in analyst reports contain non-linear expectation structures that explicit quantitative metrics do not fully capture. We leverage a Large Language Model (LLM) to extract semantic embeddings from over 770,000 Chinese A-share analyst reports and apply a Multilayer Perceptron (MLP), which significantly outperforms linear and tree-based alternatives, to map these high-dimensional narratives to future stock returns. We find that the text-to-return mapping is non-linear, with the baseline framework delivering a value-weighted CH4 alpha of 1.62% per month ($t = 4.28$). Furthermore, we disentangle textual soft information from traditional hard forecasts via cross-sectional orthogonalization. The residual soft information score yields a monthly alpha of 1.36% ($t = 2.77$), which is primarily concentrated in the long leg. Furthermore, we attribute the persistence of this predictability to the interaction between cognitive processing costs, limited investor attention, and short-sale constraints. Consequently, while quantitative metrics are quickly incorporated into prices, soft information is absorbed more gradually. Finally, this premium survives transaction costs, and is more pronounced among liquid, large-cap stocks with high analyst coverage.

Keywords: Textual Asset Pricing, Soft Information, Analyst Reports, Large Language Models, Short-sale Constraints

*We gratefully acknowledge the financial support from the National Natural Science Foundation of China (Grant No. 72402182) and the Humanities and Social Science Fund of the Ministry of Education of China (Grant No. 24YJC790061). We remain responsible for all errors and omissions.

*Corresponding author.

Email addresses: hujun@swufe.edu.cn (Jun Hu), zhangpeng_swufe@163.com (Peng Zhang)

1. Introduction

A core proposition in empirical asset pricing is that asset prices efficiently incorporate all available information (Fama, 1970). While the pricing implications of explicit quantitative disclosures—such as earnings surprises (Ball and Brown, 1968; Bernard and Thomas, 1989) and balance sheet fundamentals (Fama and French, 1992, 1993)—are extensively documented, extracting and pricing the latent "soft information" embedded in textual narratives remains a rapidly evolving critical frontier (Gentzkow et al., 2019; Bybee et al., 2024; Zhou et al., 2024). Sell-side analyst reports act as essential information intermediaries, persistently demonstrating substantial price discovery value through both long-term explicit metrics and short-term alternative signals (Womack, 1996; Brav and Lehavy, 2003; Jegadeesh et al., 2004; Altınkılıç et al., 2013; Birru et al., 2022). However, the extant literature evaluating these intermediaries has predominantly relied on explicit quantitative outputs, leaving the rich, unstructured qualitative narratives largely underutilized. Consequently, systematically decoding this textual soft information is critical to fully unpacking the pricing implications of analyst-generated information.

We focus on the Chinese A-share market, which serves as a unique laboratory for studying the price discovery process of textual soft information. Characterized by a high concentration of retail investors (Ng and Wu, 2006; Jones et al., 2021; Chen and Wu, 2022) and binding short-sale constraints (Chang et al., 2014), this market provides an ideal environment to observe the delayed price discovery and asymmetric reactions induced by frictions in information extraction and processing. Just as analysts use short-term trade ideas to bypass the rigidity of standard stock recommendations (Birru et al., 2022), they may similarly exploit unstructured textual disclosures to convey genuine convictions that institutional constraints, compliance pressures, and reputational incentives prevent them from expressing through explicit quantitative metrics. Drawing on recent evidence that analysts use such unstandardized narratives to navigate complex belief formations (Ke, 2024), we hypothesize

that an analyst's true informational edge frequently lies in these nuanced, subjective narratives, where they can embed extreme idiosyncratic optimism or subtle risk warnings with less likelihood of triggering immediate regulatory or internal scrutiny.

Despite the rapid evolution of textual analysis in empirical finance, a critical gap remains in isolating and explaining the precise expectation structure embedded in these qualitative narratives. While foundational work demonstrates the incremental value of textual disclosures, prior approaches—ranging from dictionary-based sentiment (Tetlock, 2007; Loughran and McDonald, 2011) and linear topic models (Hanley and Hoberg, 2010; Dyer et al., 2017; Huang et al., 2018; Cohen et al., 2020) to early machine learning techniques (Huang et al., 2014; Jiang et al., 2019)—generally impose rigid additive assumptions, such as the "bag-of-words" representation. Consequently, they structurally fail to capture the high-order compositional topology and long-range contextual dependencies inherent in complex human language. This methodological limitation creates a fundamental disconnect from the economic reality of human information processing. In practice, market participants do not parse financial reports as disjointed word counts; rather, they evaluate the document holistically, decode deep contextual cues, and jointly synthesize textual nuances with explicit quantitative metrics to engage in high-order fundamental reasoning.

Recently, the application of Large Language Models (LLMs) has revolutionized this space, demonstrating that high-dimensional textual signals offer strong predictive power beyond explicitly stated quantitative forecasts (Chen et al., 2022; Lv, 2024, 2025). However, a deeper empirical question remains largely unexplored: *why* does this semantic premium survive in equilibrium? While existing studies show that text provides incremental R^2, the literature has yet to formally orthogonalize this "soft information" to define its distinct behavioral expectation structure. Furthermore, it remains untested how this isolated semantic signal interacts with structural market frictions—such as short-sale constraints and attention-driven cognitive costs—to dictate the staggered and asymmetric incorporation of

information into asset prices.

To formally test the independent pricing power of textual narratives and their interaction with structural market frictions, we construct an empirical framework pairing the contextual depth of a financial-domain LLM with a Multilayer Perceptron (MLP) architecture. Unlike linear models (e.g., Ridge) or tree-based algorithms (e.g., XGBoost) that rely on discrete threshold-based splits, our continuous MLP architecture is well-suited to capture complex, non-linear interactions in the high-dimensional embedding space. Processing over 770,000 full-text analyst reports spanning 2004 to 2024, we introduce rigorous safeguards to replicate the real-time information environment. To prevent the dilution of subtle, report-specific semantics, individual document embeddings are processed through the network prior to any cross-sectional aggregation. Additionally, we enforce a 60-month rolling training window with strictly out-of-sample dynamic Z-score standardization to mitigate look-ahead bias. Following Liu et al. (2019), we exclude the bottom 30% of stocks by market capitalization across the entire market to purge potential shell-value contamination.

We begin our empirical investigation by evaluating the cross-sectional ranking capability of our models. The empirical results demonstrate the predictive dominance of the LLM-MLP framework in decoding unstructured narratives. Specifically, the MLP model achieves an average cross-sectional Rank IC of 7.11%, substantially outperforming both the tree-based XGBoost (3.90%) and the linear Ridge model (2.93%). This predictive hierarchy provides robust statistical evidence that continuous neural network architectures are more effective than discrete or linear baselines for mapping textual semantics.

Economically, we translate these predictions into a monthly-rebalanced long-short portfolio to assess their practical profitability. Before risk adjustments, the strategy delivers unadjusted raw returns of 0.92% and 1.24% per month for the equal-weighted (EW) and value-weighted (VW) hedge portfolios, respectively. To establish a stringent test for genuine text-driven predictability, we employ the China-tailored CH4 factor model (Liu et al.,

2019) as our primary risk-adjustment baseline, because it explicitly controls for retail-driven sentiment and shell-value premia. Under this strict framework, both weighting schemes generate statistically significant alphas. The EW portfolio yields a robust premium of 1.24% ($t = 3.79$), indicating the signal's breadth across the cross-section. Crucially, the VW long-short portfolio generates an even larger alpha of 1.62% per month ($t = 4.28$), suggesting that the semantic premium is practically implementable and not solely driven by micro-cap illiquidity.

To isolate the unique pricing power of textual narratives, we explicitly disentangle soft information from traditional hard forecasts. We execute a monthly cross-sectional ordinary least squares (OLS) orthogonalization, regressing the LLM-derived signal against explicit consensus expected earnings (FEP), target price implied returns (TPR), and recommendation ratings (REC). Even after statically purging these salient quantitative metrics, the residual textual soft information score continues to deliver a statistically significant monthly alpha of 1.36% ($t = 2.77$). Notably, this profitability exhibits an asymmetric tail effect: the predictive power is heavily concentrated in the extreme positive decile (the long leg), which yields an alpha of 1.46% ($t = 3.94$), whereas the short leg alpha is statistically indistinguishable from zero. This asymmetry corroborates our behavioral hypothesis. Because analysts are reluctant to issue aggressively optimistic target prices due to reputational risks, their strongest idiosyncratic convictions are displaced into textual narratives, creating a "hidden optimism" that is systematically underpriced.

To explain the resilience and survival of this semantic premium, we examine the synergistic interaction between information complexity and market frictions. We propose that this predictability operates through three distinct mechanisms that collectively establish a barrier to instantaneous arbitrage. First, the high-dimensional topology of human language imposes a steep cognitive processing cost. Decoding this continuous semantic manifold requires advanced, non-linear processing, creating a technological entry barrier for the broader

market. Second, this cognitive barrier interacts with limited investor attention (Hirshleifer and Teoh, 2003; Cohen et al., 2020). Under binding cognitive constraints, the marginal investor naturally prioritizes salient, low-cost hard metrics, forcing complex textual nuances to diffuse into asset prices gradually. Third, institutional realities dictate binding short-sale constraints (Miller, 1977). Even when sophisticated arbitrageurs decode the latent pessimism embedded in textual nuances, prohibitive borrowing costs restrict them from aggressively shorting overvalued equities. Consequently, negative semantic beliefs can only enter market prices sluggishly through the staggered unwinding of long positions by existing shareholders.

Synthesizing these mechanisms, our empirical tests reveal a distinct and asymmetric price discovery process. By comparing signal decay across extended lookback windows, we document that while the baseline full-information signal attenuates rapidly, the isolated textual soft information remains remarkably persistent across multi-month horizons. Concurrently, event studies analyzing buy-and-hold abnormal returns (BHAR) on the baseline signal reveal a stark asymmetry: the high-prediction portfolio stabilizes quickly, whereas the low-prediction portfolio exhibits a steady downward drift over 63 trading days. This sluggish empirical digestion of latent pessimism tightly aligns with the theoretical constraints of limited attention, staggered information diffusion, and short-selling frictions (Miller, 1977; Hong and Stein, 1999, 2003).

This paper contributes to three distinct strands of literature. First, we advance the literature on the price discovery value of sell-side analyst reports (Womack, 1996; Brav and Lehavy, 2003; Gleason and Lee, 2003; Ke, 2024). Recent working papers have begun deploying LLMs to process analyst narratives, primarily utilizing Shapley value frameworks to attribute return predictability to specific report topics, such as financial statement analysis or strategic outlooks (Lv, 2024, 2025). We fundamentally diverge from and complement this topic-attribution approach by opening the economic mechanism "black box." Rather than identifying *which* topics matter, we investigate *why* the semantic premium persists.

By formally orthogonalizing narrative expectations from quantitative proxies, we construct a pure soft information score. This allows us to provide robust micro-behavioral evidence on how institutional frictions—namely, short-sale constraints and attention-driven cognitive costs—dictate the asymmetric and staggered incorporation of unstructured text into financial markets.

Second, we contribute to the textual analysis literature in empirical finance. We document the methodological transition from traditional dictionary-based sentiment and linear topic models (Tetlock, 2007; Loughran and McDonald, 2011; Hanley and Hoberg, 2010) to advanced Large Language Models capable of decoding high-dimensional semantics (Gentzkow et al., 2019; Jiang et al., 2019; Wu et al., 2023; Lopez-Lira and Tang, 2023; Chen et al., 2022; Tan et al., 2023; Zhou et al., 2024). In contrast to recent studies that query LLMs directly using zero-shot or few-shot prompts, we instead employ the LLM strictly as an encoder to extract deterministic semantic embeddings. This architectural choice offers several practical advantages for empirical asset pricing. First, it avoids the risk of generative hallucinations, as the model does not produce free-form text. Second, by using a static pre-trained encoder with a fixed training cutoff, we reduce concerns about look-ahead bias that may arise when models are queried on data potentially seen during training. Third, the embedding-based approach is less sensitive to prompt wording, thereby enhancing reproducibility and reducing empirical variance relative to prompt-dependent methods. Overall, we show that this stable, encoding-focused framework enables LLMs to extract a reproducible and economically meaningful signal from textual narratives that is not fully captured by traditional quantitative metrics.

Third, our methodology directly extends the burgeoning literature demonstrating the superiority of non-linear machine learning architectures and textual data in empirical asset pricing (Ke et al., 2019; Gu et al., 2020; Freyberger et al., 2020; Leippold et al., 2022; Bybee et al., 2024; Chen et al., 2024). We formally establish the text-to-return mapping hierarchy

(MLP > XGBoost > Ridge). By demonstrating that continuous neural networks systematically outperform discrete tree-based algorithms and linear models, we provide empirical validation that financial narratives exhibit complex, non-linear interactions that are better captured by continuous neural networks than by simple additive models.

The remainder of the paper is organized as follows. Section 2 describes the sample selection, LLM embedding extraction, and the predictive model architectures. Section 3 presents the baseline predictive performance and rigorous risk adjustments. Section 4 conducts the orthogonalization procedure to disentangle soft from hard information. Section 5 investigates the underlying economic mechanisms, including event studies and limits to arbitrage. Section 6 provides comprehensive robustness checks. Section 7 concludes.

2. Data and Methodology

2.1. Sample Selection and Data Sources

Our sample comprises Chinese A-share analyst reports published between January 2004 and December 2024. The initial textual corpus of approximately 770,000 full-text research reports is sourced from Sina Finance. We restrict the stock universe to firms listed on the Shanghai Main Board (stock codes starting with 60), Shenzhen Main Board (starting with 00), and ChiNext market (starting with 30). Reports covering stocks on the Beijing Stock Exchange, STAR Market (688), or other segments are excluded. This screening yields a final sample of 684,466 report-stock observations.

Figure 1 plots the annual distribution of these reports and the number of unique covered stocks. During the initial five-year training window (2004–2008), both report volume and stock coverage are relatively sparse, reflecting the early development stage of the Chinese sell-side research industry. In the subsequent out-of-sample evaluation period (2009–2024), both series expand substantially. Our rolling-window design utilizes this initial sparse period

to estimate baseline model parameters prior to out-of-sample evaluation in the information-dense period.

Stock returns (measured with cash dividend reinvestment), firm-level characteristics, and quantitative forecasts are obtained from the China Stock Market & Accounting Research (CSMAR) database. Firm characteristics include market capitalization (Size), book-to-market ratio (BM), earnings-to-price ratio (EP), turnover, institutional ownership (IO), and shortability.

To benchmark the incremental predictive power of textual soft information, we use objective proxies for explicit market expectations. Instead of employing Named Entity Recognition (NER) pipelines to extract quantitative targets from unstructured text—a procedure susceptible to idiosyncratic extraction noise—we utilize standardized consensus metrics from CSMAR. For each stock-month observation, we calculate the consensus forward earnings-to-price ratio (FEP), target price implied return (TPR), and recommendation rating (REC). These metrics are defined as the cross-sectional equal-weighted average of all individual forecasts issued for stock i in month t.

To ensure data quality, we exclude financial firms and stocks designated as ST or *ST. During portfolio construction, following Liu et al. (2019), we exclude the bottom 30% of stocks by market capitalization to mitigate the impact of shell-value pollution in the Chinese market, ensuring the economic implementability of the strategy.

2.2. Semantic Feature Extraction: LLM Embeddings

We convert unstructured analyst report narratives into machine-interpretable pricing signals utilizing the Qianwen (Qwen) Large Language Model. Recent natural language processing literature highlights Qwen's architectural advantages and state-of-the-art performance in decoding complex Chinese corpora compared to baseline open-source models (Bai et al., 2023). Built upon the foundational Transformer architecture (Vaswani et al., 2017), the LLM generates a 1024-dimensional dense semantic embedding for each full-text analyst

report. This high-dimensional vector encapsulates complex contextual nuances, implicit professional tone, and latent analytical convictions that traditional dictionary-based sentiment counts or linear bag-of-words models structurally fail to capture.

At the stock-month level, it is common to have multiple analyst reports covering the same firm. Let $e_{i,t}^{(k)} \in \mathbb{R}^{1024}$ denote the document-level embedding for the k-th report covering stock i published in calendar month t. To avoid diluting idiosyncratic semantics through early aggregation, we preserve the report-level resolution. Each individual embedding is directly fed into the neural network, allowing the model to process report-specific information prior to any cross-sectional or temporal aggregation.

2.3. Predictive Architecture and Estimation Strategy

We employ a Multilayer Perceptron (MLP) to map narrative embeddings to future returns. Econometric theory demonstrates that deep neural networks possess superior approximation capabilities for complex, high-dimensional non-linear functions compared to shallow learning methods (Farrell et al., 2021). To ensure out-of-sample validity and prevent look-ahead bias, we implement a monthly rolling-window estimation scheme. For each out-of-sample prediction month τ (from January 2009 to December 2024), the training sample comprises the preceding 60 months, $[\tau - 60, \tau - 1]$. This continuous updating accounts for market regime shifts and the evolution of analyst rhetoric.

Within each window, the model maps individual report embeddings $e_{i,t}^{(k)}$ to future returns. The prediction target, $R_{i,t+1}^{(k)}$, is the cumulative stock return over the 21 trading days following the publication of the k-th report. To reduce the influence of outliers, target returns are cross-sectionally winsorized at the 1% and 99% levels exclusively within the active training set. We apply Z-score standardization to the embeddings; the mean and standard deviation vectors are computed solely from the training window and applied out-of-sample to test month τ, preventing data leakage.

The MLP generates a report-specific return forecast, $\hat{R}_{i,t+1}^{(k)}$, using a feedforward architecture with two hidden layers:

$$h_1^{(k)} = \sigma\big(\mathrm{BN}(W_1 e_{i,t}^{(k)} + b_1)\big)$$

$$h_2^{(k)} = \sigma\big(W_2[\mathrm{Dropout}(h_1^{(k)})] + b_2\big)$$

$$\hat{R}_{i,t+1}^{(k)} = W_3 h_2^{(k)} + b_3$$

where $W_1 \in \mathbb{R}^{256\times1024}$, $W_2 \in \mathbb{R}^{64\times256}$, and $W_3 \in \mathbb{R}^{1\times64}$ are trainable weight matrices, and b_1, b_2, b_3 are bias terms. $\mathrm{BN}(\cdot)$ represents Batch Normalization applied to the first layer. $\sigma(\cdot)$ is the Rectified Linear Unit (ReLU) activation function, defined as $\sigma(x) = \max(0, x)$. Dropout (rate $= 0.3$) is applied after the first hidden layer to regularize the network.

Network parameters are estimated by minimizing the Mean Squared Error (MSE) via the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-3} and a batch size of 1024. We enforce a fixed training duration of 10 epochs. Given the low signal-to-noise ratio of financial data, traditional early stopping based on a validation set is prone to overfitting the validation noise itself (De Prado, 2018). Instead, a fixed number of epochs serves as a structural regularization mechanism to constrain model capacity (Goodfellow, 2016). The 10-epoch hyperparameter is selected based on an empirical evaluation over the 2009–2016 subsample, which covers diverse market regimes (as formally validated in Section 6.1).

To assess the structural advantage of neural networks in processing high-dimensional semantic vectors, we benchmark the MLP against Ridge regression (a linear baseline) and XGBoost (Chen and Guestrin, 2016) (a tree-based non-linear baseline). This specific setup is designed to evaluate the theoretical alignment between model architectures and the topology of LLM embeddings. Ridge regression imposes a strict additive structure, which structurally

precludes the capture of cross-dimensional semantic interactions. XGBoost captures non-linearities but relies on discrete, axis-aligned recursive partitioning. Because dense embeddings distribute semantic meaning continuously across all dimensions rather than isolating them in discrete features, such step-wise, threshold-based splits are theoretically inefficient. In contrast, the MLP architecture, utilizing matrix multiplications and continuous activation functions, enables smooth, high-order geometric transformations. This setup allows us to explicitly test whether continuous neural manifolds provide an empirical advantage over linear constraints and discrete partitioning in decoding entangled semantic spaces.

2.4. Signal Aggregation and Evaluation Metrics

Once the network generates the report-specific out-of-sample forecasts $\hat{R}_{i,t+1}^{(k)}$, these individual predictions must be temporally aggregated to form the final, investable stock-level signal. We achieve this through a two-step aggregation procedure.

First, to prevent months with clustered report issuances (e.g., during earnings seasons) from disproportionately dominating the historical signal, we cross-sectionally average all report-specific forecasts for stock i within a single month t. Let $N_{i,t}$ denote the total number of reports published for stock i in month t. The monthly consensus expected return $\bar{R}_{i,t+1}$ is computed as:

$$\bar{R}_{i,t+1} = \frac{1}{N_{i,t}} \sum_{k=1}^{N_{i,t}} \hat{R}_{i,t+1}^{(k)}$$

Second, to accommodate different signal-smoothing horizons and capture longer-term narrative memory, we temporally aggregate these monthly scores over a specified trailing lookback window (LB). Let $M_{i,t}^{LB}$ denote the set of months within the LB-month window up to month t in which stock i has available report data, and $|M_{i,t}^{LB}|$ be the number of such months. The final aggregated stock-month expected return $\hat{R}_{i,t+1}$ is given by a simple moving average of these monthly scores:

$$\hat{R}_{i,t+1} = \frac{1}{|M_{i,t}^{LB}|} \sum_{m \in M_{i,t}^{LB}} \bar{R}_{i,m+1}$$

This two-step approach intentionally applies equal weighting at the monthly level rather than at the individual report level. This design mitigates the issue where months with clustered report issuances disproportionately dominate the aggregated signal, thereby preserving the relative importance of sparse updates during off-peak periods while accommodating different signal-smoothing horizons.

While the temporally aggregated signal ($\hat{R}_{i,t+1}$) forms the basis for subsequent portfolio construction, evaluating the pure semantic extraction capability of our deep learning architecture requires analyzing the predictions prior to any cross-sectional aggregation. Therefore, to evaluate the predictive efficacy of our model, we compute our statistical metrics strictly at the individual report level. Following the empirical asset pricing literature (Gu et al., 2020; Leippold et al., 2022), we employ the out-of-sample R^2 (R^2_{OOS}). In addition, to address the specific distributional properties of textual semantic signals, we introduce the Spearman Rank Information Coefficient (Rank IC) as our core evaluation metric.

First, we define R^2_{OOS} based on Mean Squared Error (MSE), which quantifies the proportional reduction in MSE relative to a naive zero-return forecast benchmark at the report level:

$$R^2_{OOS} = 1 - \frac{\sum_{k \in T_{test}} (R^{(k)}_{i,t+1} - \hat{R}^{(k)}_{i,t+1})^2}{\sum_{k \in T_{test}} (R^{(k)}_{i,t+1})^2}$$

where T_{test} represents the set of all out-of-sample individual report observations. While widely used, evaluating soft information exclusively through absolute point-prediction poses certain methodological limitations. LLM-extracted semantic signals often capture tail expectations or extreme idiosyncratic optimism. When target returns are forced into normalization via strict winsorization to eliminate microstructure noise, it alters the target

distribution, meaning that MSE-based evaluation tends to penalize directionally accurate tail estimates.

Consequently, absolute point prediction is secondary to cross-sectional ordinal ranking capability. Thus, our core evaluation metric is the report-level Rank IC, defined as the Spearman rank correlation between the model-predicted returns and the subsequent realized returns across all individual analyst reports published in month t:

$$Rank\ IC_t = \rho_{spearman}(\hat{R}^{(k)}_{i,t+1}, R^{(k)}_{i,t+1})$$

The Rank IC is less sensitive to extreme outliers and provides a robust statistical measure of the model's cross-sectional ranking capability, which aligns with our objective to identify relative semantic optimism or pessimism.

3. Empirical Results: Baseline Predictive Power

This section bridges the methodological framework established in Section 2 with cross-sectional asset pricing outcomes. We first evaluate the raw ranking efficacy across competing machine learning architectures to test the non-linear semantic topology hypothesis (Section 2.3). Subsequently, we construct standard decile portfolios to quantify the economic magnitude and time-series consistency of the semantic premium, testing its robustness against a comprehensive suite of classical factor models to assess its independent pricing power.

3.1. Model Comparison: Target Normalization and Semantic Topology

To evaluate out-of-sample predictive performance, we begin by analyzing the absolute point prediction accuracy measured by the out-of-sample R^2 (R^2_{OOS}) formalized in Section 2.4. Panel A of Table 1 reports these results for both unwinsorized and winsorized target returns. When predicting raw, unwinsorized returns, the MLP architecture achieves

an R^2_{OOS} of 7.10%, outperforming both the tree-based XGBoost (3.42%) and the linear Ridge model (1.58%). Diebold-Mariano (DM) tests indicate that the MLP's superiority over these benchmarks is statistically significant ($t = 3.93$ and $t = 4.45$, respectively).

[Insert Table 1 about here]

However, Panel A reveals a methodological trade-off regarding target return normalization when adapting machine learning to textual asset pricing. As discussed in Section 2.4, while forced normalization (winsorization) is conventionally employed to mitigate microstructure noise, it standardizes the target distribution, thereby diminishing the extreme idiosyncratic tail expectations that text embeddings may capture. When target returns are strictly winsorized, the MSE-based R^2_{OOS} for the MLP decreases to -1.69%. This contrast is consistent with our behavioral intuition: the dense 1024-dimensional embeddings extracted by the LLM primarily capture extreme tail expectations and idiosyncratic optimism/pessimism embedded by analysts. Consequently, MSE-based optimization disproportionately penalizes these directionally accurate point estimates.

Given this sensitivity, our evaluation focuses on the cross-sectional Spearman Rank Information Coefficient (Rank IC), which is more robust to outliers. Panel B of Table 1 reports these ordinal ranking metrics, revealing a clear predictive hierarchy. The linear Ridge baseline exhibits a Rank IC of 2.93%. The discrete tree-based XGBoost model improves upon this to 3.90%. The MLP architecture demonstrates the strongest performance, delivering a mean Rank IC of 7.11% with a stable ICIR of 0.72 ($t = 10.00$).

This performance wedge provides empirical support for our core hypothesis: textual semantics encode high-order, compositional interactions that behave as a continuous high-dimensional manifold. Linear models are limited by rigid additive constraints, while tree-based algorithms rely on discrete, threshold-based cuts that may not map smooth semantic transitions effectively. These results suggest that continuous neural network architectures are

better suited for exploiting unstructured financial narratives, aligning with recent findings in the non-linear asset pricing literature.

3.2. Portfolio Sorts and the Anatomy of the Semantic Premium

To translate this ordinal ranking capability into economic significance, we construct investment portfolios following standard empirical asset pricing protocols. At the end of each month t, we sort all eligible A-share stocks into deciles based on the out-of-sample baseline expected returns $\hat{R}_{i,t+1}$. We form value-weighted (VW) portfolios based on lagged market capitalization, rebalancing monthly to preclude look-ahead bias. The primary metric of interest is the zero-cost long-short (H-L) hedge portfolio.

Leveraging the temporal aggregation mechanism formalized in Section 2.4, we first configure our predictive signal utilizing a strict one-month lookback window ($LB = 1$). This configuration preserves the highest frequency of semantic resolution without blending historical data.

As reported in Table 2, we observe a positive relationship between the MLP-predicted expected returns and subsequent realized returns. The VW High-decile portfolio (Group 10) yields a mean monthly return of 1.42%, whereas the VW Low-decile portfolio (Group 1) yields 0.18%. The resulting VW H-L spread generates 1.24% per month with a monthly standard deviation of 5.98%, translating to an annualized Sharpe ratio of 0.72.

[Insert Table 2 about here]

In empirical asset pricing, relying solely on raw cross-sectional dispersion is insufficient; we must evaluate risk-adjusted alpha to address the possibility of passive exposure to known systematic risks. Due to institutional frictions in the Chinese market—namely restrictive listing policies and short-sale constraints—small-cap stocks carry a significant idiosyncratic "shell value" premium. To address this, we adopt the China-specific four-factor (CH4) model

(Liu et al., 2019) as our primary baseline, which explicitly purges the bottom 30% of stocks by market capitalization.

To further assess the robustness of our semantic premium, we subject our baseline portfolio to a comprehensive battery of global and domestic asset pricing models. Table 3 presents the risk-adjusted alphas across the Fama-French three- (Fama and French, 1993), five- (Fama and French, 2015), and six-factor models (Fama and French, 2018) (FF3, FF5, FF6), alongside the CH3 and CH4 frameworks (Liu et al., 2019).

[Insert Table 3 about here]

The empirical results provide strong evidence of independent pricing power. Under the primary CH4 model, the VW long-short portfolio generates a statistically significant monthly alpha of 1.62% ($t = 4.28$). This outperformance remains robust after controlling for the Fama-French six-factor model (FF6), maintaining a monthly alpha of 0.79% ($t = 1.99$). The persistence of this premium across diverse factor environments suggests that the LLM extracts an orthogonal pricing signal rather than a mechanical reflection of traditional metrics like momentum or value.

Recognizing the structural reality of the Chinese A-share market—where short-selling is constrained by institutional limits and borrowing costs—we also evaluate the performance of the purely long-only portfolio. The practical implementability of trading strategies often diminishes when profitability is driven primarily by the short leg. In contrast, the value-weighted top decile portfolio (Group 10) independently yields an economically large and statistically significant CH4 alpha of 1.21% per month ($t = 4.48$).

Furthermore, our results challenge a prevalent critique in the machine learning asset pricing literature, which argues that algorithmic anomalies are often concentrated in highly illiquid, micro-cap stocks. As detailed in Table 3, our semantic strategy consistently exhibits higher alphas in value-weighted (VW) portfolios compared to equal-weighted (EW)

configurations (e.g., VW CH4 alpha of 1.62% vs. EW CH4 alpha of 1.24%). This finding indicates that the LLM-extracted semantic signal is effective among larger, more liquid stocks, mitigating concerns regarding shell-value noise and supporting the strategy's practical implementability.

3.3. Signal Dynamics and Time-Series Robustness

To explore the economic mechanisms driving this predictability, we examine the signal's sensitivity to varying smoothing horizons by extending the lookback window ($LB = 1, 2, 3,$ and 4 months) as reported in Table 2.

The results document a rapid decay in predictive power as the temporal window lengthens. By $LB = 4$, the raw VW H-L spread diminishes to 0.71%, and the CH4 alpha declines from its peak of 1.62% ($t = 4.28$) at $LB = 1$ to 1.24% ($t = 2.33$). This attenuation is consistent with limited investor attention models: aggregate narrative content functions as a transient signal, and averaging over longer horizons incorporates stale information, thereby diluting predictive power. Based on these decay characteristics, we select the $LB = 1$ specification as our testing baseline for the remainder of the study.

Finally, we evaluate the time-series consistency of our benchmark strategy. Figure 2 plots the log cumulative returns over the out-of-sample period. The H-L portfolio exhibits a persistent upward trajectory, remaining relatively resilient during severe market contractions, such as the 2015 crash and the 2022 bear market. Concurrently, the monthly return heatmap (Figure 3) demonstrates that the strategy balances periodic drawdowns with frequent profit realizations to drive long-term cumulative growth.

Having empirically established a robust semantic premium, we now proceed to the next methodological step: examining whether this premium is derived from genuine narrative text or if it simply replicates the explicit numerical forecasts already embedded in the reports.

4. Disentangling Soft from Hard Information

A challenge when applying Large Language Models to financial texts is the presence of quantitative metrics. Analyst reports are populated with hard targets—such as expected earnings, target prices, and discrete rating upgrades. If the predictive efficacy of our baseline signal ($\hat{R}_{i,t+1}$) primarily reflects the neural network's capacity to aggregate these embedded numbers, its incremental asset pricing contribution would be limited (Gleason and Lee, 2003). While foundational accounting literature establishes that textual tone contains incremental value over explicit quantitative forecasts using traditional probabilistic models (Huang et al., 2014), it remains empirically untested whether the high-dimensional, nonlinear semantic structures extracted by LLMs encode a fundamentally distinct belief system. To address this, we empirically isolate the independent pricing power of qualitative narratives from traditional explicit forecasts.

4.1. Methodological Design: Isolating Textual Soft Information

To extract the pure expectation structure embedded within the text, we implement an orthogonalization procedure. Each month t, we run the following cross-sectional ordinary least squares (OLS) regression for all available stock observations:

$$\hat{R}_{i,t+1} = \alpha_t + \beta_{1,t} FEP_{i,t} + \beta_{2,t} TPR_{i,t} + \beta_{3,t} REC_{i,t} + \epsilon_{i,t} \tag{1}$$

where $\hat{R}_{i,t+1}$ is the return prediction generated by our MLP architecture. $FEP_{i,t}$ is the consensus expected forward earnings-to-price ratio, $TPR_{i,t}$ is the consensus target price implied return, and $REC_{i,t}$ is the average recommendation rating.

Market prices typically react to aggregate consensus (FEP, TPR, REC) as the baseline hard information proxy. By regressing the LLM's aggregate textual prediction against the objective, database-derived market consensus established in Section 2.1, we separate the textual soft information from widely available public benchmarks.

The estimated residual, $\hat{\epsilon}_{i,t}$, represents the latent semantic information strictly orthogonal to the analysts' explicit quantitative metrics. We define this residual as the textual soft information score, denoted as $\hat{R}^{soft}_{i,t+1}$. This linear projection acts as a stringent statistical filter, allowing us to build upon the baseline established by Huang et al. (2014) and investigate whether deep learning architectures capture a distinct, high-order belief structure that standard numerical forecasts structurally fail to encode.

4.2. Orthogonalized Portfolio Performance and Asymmetric Tail Alpha

We reconstruct our decile portfolios using the textual soft information score ($\hat{R}^{soft}_{i,t+1}$). Consistent with our baseline strategy, we focus on the one-month lookback window ($LB = 1$) and evaluate risk-adjusted performance using the China-specific four-factor (CH4) model. The portfolio sorting procedure and remaining evaluation metrics are identical to our initial setup. The results are reported in Table 4.

[Insert Table 4 about here]

Even after linearly purging these explicit quantitative signals, the value-weighted long-short hedge portfolio (H-L) delivers a statistically significant CH4 alpha of 1.36% per month ($t = 2.77$, $p = 0.006$). This indicates that the textual soft information extracted by the LLM provides robust independent pricing power.

The performance structure of the orthogonalized deciles reveals a distinct asymmetric tail effect. The profitability is heavily concentrated in the extreme positive decile (Group 10), which generates a significant alpha of 1.46% ($t = 3.94$, $p < 0.001$). In contrast, the alpha of the extreme short leg (Group 1) is 0.11% ($t = 0.26$) and statistically indistinguishable from zero. This asymmetric predictability is consistent with our behavioral hypothesis regarding dual institutional frictions. Foundational literature establishes that sell-side analysts face asymmetric disclosure constraints due to severe career concerns (Hong and Kubik, 2003) and investment banking conflicts of interest (Michaely and Womack, 1999). On the upside,

they hesitate to issue unrealistically high explicit target prices that may damage their long-term forecasting reputation, opting instead to embed their strongest idiosyncratic optimism within qualitative textual nuances. On the downside, analysts are generally reluctant to publicly issue bearish forecasts or explicitly short a stock, which limits the expression of extreme pessimism in any form. Consequently, the LLM systematically captures this latent, "hidden optimism" which traditional quantitative metrics structurally fail to express.

4.3. The Pricing Persistence of Textual Soft Information Across Horizons

Consistent with existing literature (Tetlock, 2007), our baseline empirical results in Section 3.3 demonstrate that the predictive power of aggregate analyst signals decays rapidly as the lookback window lengthens. However, it is necessary to determine whether such attenuation is driven by the immediate pricing of explicit quantitative metrics or the transient nature of textual soft information. To examine this, we replicate our extended lookback window (LB) analysis using our residualized textual soft information signal ($\hat{R}^{soft}_{i,t+1}$).

Table 5 reports the value-weighted decile portfolio performance of this orthogonalized signal across four smoothing horizons. Risk-adjusted alphas are estimated using the China-specific CH4 factor model (Liu et al., 2019). We highlight three key empirical findings.

[Insert Table 5 about here]

First, the textual soft information signal delivers statistically significant excess returns across all evaluated horizons. At the four-month lookback window ($LB = 4$), the long-short hedge portfolio generates a monthly CH4 alpha of 1.14% with a Newey-West adjusted t-statistic of 2.91. This persistence suggests that the residualized text captures longer-term fundamental insights rather than purely transient noise.

Second, the pricing persistence of textual soft information is structurally stronger than that of the full information signal. Table 6 contrasts the attenuation of the full information signal against the strictly orthogonalized soft information signal. As documented in Panel

A, the full information's predictive reliability declines rapidly, with its t-statistic dropping by 45.56% from $LB = 1$ to $LB = 4$. In contrast, Panel B demonstrates that the residualized textual soft information exhibits a relatively stable signal-to-noise ratio. Its statistical significance resists decay across the extended horizons, peaking at $LB = 3$ ($t = 3.43$) before maintaining a robust level at $LB = 4$. This persistence is consistent with models of limited investor attention (Hirshleifer and Teoh, 2003). It suggests that, compared to easily interpretable quantitative metrics, decoding complex textual narratives requires higher cognitive effort, leading to a smoother, multi-month price incorporation process.

[Insert Table 6 about here]

Third, the asymmetric profitability documented in Section 4.2 holds consistently across all lookback horizons. Across every LB specification, the semantic premium is primarily driven by the extreme positive decile (the long leg). This asymmetry reinforces the hypothesis that institutional constraints lead analysts to channel their extreme optimism into textual nuances, which are then absorbed more gradually by the market.

This divergence in pricing cycles helps isolate the economic friction driving the semantic premium. Soft information exhibits persistent pricing power that is shielded from rapid arbitrage, likely due to its intrinsic complexity. This naturally leads to the formal examination of the underlying economic mechanisms and limits to arbitrage driving this predictability, which we execute in the following section before evaluating its exposure to established systematic risk factors in Section 6.

5. Price Discovery and Economic Mechanisms

The decomposition analysis in Section 4 demonstrated that textual soft information possesses significant and asymmetric pricing power that is independent of explicit quantitative forecasts. Having provided empirical evidence for this independent expectation structure, we now transition our focus to the real-world market microstructure.

In financial markets, investors do not observe orthogonalized mathematical residuals; they process the analyst report as a single, holistic signal. Therefore, to capture how market participants digest complex narratives and to observe the true price discovery process, our mechanism analysis in this section returns to the baseline full-information semantic signal ($\hat{R}_{i,t+1}$). This holistic measure accurately reflects the actual information set available to market participants at the time of publication, providing a behaviorally sound foundation for uncovering real-world economic frictions.

We propose that the predictability of semantic information is sustained by an interaction between cognitive processing costs, limited investor attention, and institutional short-sale constraints. We empirically test these mechanisms through event studies of buy-and-hold abnormal returns and cross-sectional double-sorted portfolios.

5.1. Event Study: Market Reaction and Asymmetric Price Discovery

To observe the information-processing channel, we conduct an event study analyzing buy-and-hold abnormal returns (BHAR) around report publication dates. This setup isolates the immediate "information shock" of textual narratives from long-term common risk factors. We sort reports into deciles based on our baseline semantic signal ($\hat{R}_{i,t+1}$), focusing on the extreme optimism (High, H) and extreme pessimism (Low, L) groups. Abnormal returns are calculated relative to the characteristic-adjusted benchmark developed by Daniel et al. (1997) (DGTW):

$$BHAR_{0,T} = \prod_{t=0}^{T}(1 + R_{it}) - \prod_{t=0}^{T}(1 + R_{it}^{DGTW}) \tag{2}$$

where R_{it} is the daily return and R_{it}^{DGTW} is the matched DGTW benchmark return.

Figure 4 plots the 63-day $BHAR$ paths, revealing an asymmetric discovery process that highlights the interplay of behavioral and institutional frictions. For the H-group (long leg), we observe a positive drift that stabilizes relatively quickly after the initial publication.

This is consistent with limited investor attention: the salient positive recommendation triggers an initial surge, while the deeper non-linear semantic value is gradually absorbed over subsequent weeks.

In contrast, the L-group (short leg) demonstrates a continuous downward drift over the entire 63-day horizon, ultimately underperforming the DGTW benchmark by over 3.0% in the value-weighted setting. This prolonged drift provides visual evidence consistent with compounding frictions. Because decoding negative soft narratives requires higher cognitive effort, attention-constrained investors may be slower to identify the underlying pessimism. Furthermore, even when sophisticated agents decode this information, binding short-sale constraints—a core institutional feature of the Chinese A-share market (Miller, 1977)—prevent them from aggressively correcting the overvaluation immediately. Consequently, negative soft information primarily enters prices slowly through the gradual unwinding of positions by existing shareholders. This temporal asymmetry naturally extends to specific cross-sectional constraints, which we examine next.

5.2. The Complexity Premium: Information Environments and Limits to Arbitrage

To provide cross-sectional evidence for our hypothesis, we conduct a series of double-sorting portfolio analyses. This approach allows us to observe the marginal pricing power of textual semantics while sequentially neutralizing firm-level characteristics. The results, presented in Table 7, decompose the cross-section into two theoretical channels: the depth of the information environment and the severity of limits to arbitrage.

[Insert Table 7 about here]

Information Environment and Analyst Coverage

A common critique of machine learning anomalies is their historical concentration in highly illiquid, micro-cap stocks. However, Panel A demonstrates that the semantic premium

is more pronounced among large-cap stocks. The strategy delivers a value-weighted (VW) CH4 alpha of 1.17% ($t = 3.53$) for large firms, compared to 0.78% ($t = 2.53$) for small caps. This outperformance suggests a reliance on the depth of the information environment, as large firms possess higher market visibility and attract a denser ecosystem of sell-side research.

This reliance on information richness is further supported in Panel G: high analyst coverage amplifies the VW alpha to 1.78% ($t = 4.73$), compared to 0.69% ($t = 2.34$) generated in low-coverage environments. While massive textual volume may trigger information overload for attention-constrained human investors, it provides the LLM-MLP framework with a deeper semantic topology to extract latent consensus. Furthermore, Panel B shows the premium is concentrated in growth (low-BM) stocks, yielding a VW alpha of 1.26% ($t = 3.30$). Because growth firms rely heavily on intangible prospects, their valuations often necessitate qualitative narratives that rigid accounting metrics may fail to capture.

Institutional Frictions and Limits to Arbitrage

Panel E directly tests the impact of short-selling restrictions. Among non-shortable stocks, the semantic signal generates a VW alpha of 1.76% ($t = 3.08$). However, once the short-sale constraint is relaxed (shortable stocks), the alpha drops to 0.94% ($t = 2.83$). This divergence suggests that the inability of sophisticated capital to efficiently arbitrage away negative soft information is a primary friction sustaining the premium.

Panel F reinforces this by examining institutional ownership (IO). The strategy yields a VW alpha of 1.30% ($t = 3.16$) in low-IO, retail-dominated stocks, compared to 0.96% ($t = 3.10$) in high-IO stocks. Retail investors are subject to tighter cognitive constraints and typically allocate their limited attention toward highly salient, easily interpretable metrics rather than conducting complex fundamental analysis (Barber and Odean, 2008). Consequently, in retail-dominated environments, the LLM's non-linear advantage in extracting

latent, complex narratives persists longer before being fully impounded into market prices.

Synthesizing the empirical evidence from both the temporal event study (Section 5.1) and the cross-sectional double sorts discussed above, these results collectively support that the semantic premium is sustained by three synergistic mechanisms. First, the high-dimensional topology of narratives imposes a substantial *cognitive processing cost*, granting the LLM a technological advantage over human practitioners. Second, *limited investor attention* leads market participants to prioritize salient hard metrics, resulting in a more gradual incorporation of soft information into prices. Finally, institutional *short-sale constraints* restrict the arbitrage of negative narratives, contributing to the prolonged pricing ceiling observed in our event study. Together, these frictions drive the asymmetric and staggered incorporation of textual nuances into asset prices.

6. Robustness Checks and Economic Implementability

In the preceding sections, we integrated our core empirical analyses with cross-sectional mechanism tests to establish the independent pricing power of textual soft information. To complete our study, we subject the semantic premium to a battery of stress tests. The objective of this section is to verify the strategy's independence from systematic risk factors, its firm-level multivariate robustness, its immunity to industry clustering, the structural stability of our machine learning architecture, and its economic implementability net of transaction frictions.

6.1. Systematic Risk and Factor Exposures

Before examining firm-level characteristics, it is necessary to examine the factor exposures (betas) of both the long-only (top decile) and the zero-cost (H-L) portfolios to ensure the strategy does not mechanically load on conventional systematic risks. The estimations are reported in Table 8.

[Insert Table 8 about here]

We first evaluate standard fundamental exposures. The H-L portfolio exhibits near-zero and statistically insignificant loadings on the Size (SMB) and Value (HML/VMG) factors across all specifications, indicating that the hedge strategy is neutral to static fundamental tilts. Furthermore, we observe two heterogeneous exposures that help characterize the textual signal.

First, we document a significant *negative* loading on the PMO (retail sentiment) factor in the CH4 model for both the purely long portfolio ($\beta = -0.37, t = -3.45$) and the H-L portfolio ($\beta = -0.52, t = -3.05$). This indicates that the LLM-extracted semantic signal is distinct from conventional dictionary-based sentiment measures. Rather than capturing retail exuberance, the model extracts a dimension of professional belief structure that contrasts with aggregate market sentiment.

Second, we observe a significant positive loading on the Momentum (MOM) factor (Carhart, 1997) in the FF6 model for both the long-only ($\beta = 0.31, t = 3.70$) and the hedge portfolios ($\beta = 0.43, t = 3.38$). This implies that the textual signal partially captures a form of narrative momentum embedded in analyst reports that precedes price adjustments. The persistence of a statistically significant alpha after controlling for MOM indicates that the semantic premium provides incremental predictive power not entirely subsumed by historical price continuation.

6.2. Firm-Level Predictability: Fama-MacBeth Regressions

While portfolio sorting provides an intuitive non-parametric evaluation of the semantic premium, it is limited in its ability to simultaneously control for multiple continuous firm characteristics. To test whether the predictive power of the textual signal is subsumed by a linear combination of known anomalies, we estimate firm-level Fama-MacBeth (FMB) regressions (Fama and MacBeth, 1973).

The FMB framework allows us to deploy the baseline semantic signal ($\hat{R}_{i,t+1}$) and explicitly control for analysts' quantitative forecasts (FEP, TPR, REC). This process proceeds in two steps. First, for each month t, we run the following cross-sectional regression:

$$R_{i,t+1} = \gamma_{0,t} + \gamma_{1,t}\hat{R}_{i,t+1} + \sum_{j=1}^{J} \lambda_{j,t} Controls_{i,j,t} + \epsilon_{i,t} \tag{3}$$

Second, to test whether the risk premium associated with the semantic signal is significantly different from zero over time, we compute the time-series average of the estimated coefficients and apply Newey-West (Newey and West, 1987) standard errors to account for potential autocorrelation:

$$\bar{\gamma}_1 = \frac{1}{T}\sum_{t=1}^{T} \gamma_{1,t}, \quad t(\bar{\gamma}_1) = \frac{\bar{\gamma}_1}{\sigma(\hat{\gamma}_1)/\sqrt{T}} \tag{4}$$

where $\bar{\gamma}_1$ represents the average monthly risk premium of our textual signal, and T is the number of months in the sample. To ensure coefficients reflect standardized economic magnitudes and to mitigate outlier influence, all independent variables are winsorized at the 1% and 99% levels and cross-sectionally standardized each month.

As reported in Table 10, we progressively expand the control vector. Model 1 shows the univariate baseline. Models 2 to 5 systematically incorporate standard firm characteristics (Size, BM, EP, Mom, Rev), explicit quantitative forecasts, and industry fixed effects. Model 6 serves as a comparative benchmark relying solely on hard metrics.

[Insert Table 10 about here]

Two empirical findings emerge. First, the coefficient on the semantic signal ($\hat{R}_{i,t+1}$) exhibits stability, ranging between 0.21 and 0.22, and remains statistically significant at the 1% level. Economically, a one-standard-deviation increase in the semantic signal translates to an incremental monthly return of roughly 21 to 22 basis points. Diagnostic tests indicate

a Variance Inflation Factor (VIF) of 1.19, verifying that the semantic topology extracted by the neural network is generally orthogonal to established anomalies.

Second, a direct comparison of the goodness-of-fit (R^2) helps quantify the incremental value of soft information. The hard-information-only specification (Model 6) yields an average R^2 of 0.198. Upon introducing the text-derived signal in Model 5, the average R^2 increases to 0.203. This incremental improvement is consistent with the hypothesis that the LLM captures a dimension of subjective belief formation outside the span of traditional quantitative variables.

6.3. Industry Neutrality and Sector Rotation

An alternative hypothesis is that the baseline signal's predictability reflects industry-level risk premiums or mechanical sector rotation, rather than firm-specific soft information. To address this, we conduct an industry neutralization test. We isolate the firm-specific component of the signal via monthly cross-sectional regressions of the baseline predicted returns ($\hat{R}_{i,t+1}$) on a full set of industry dummy variables constructed using the China Association for Public Companies (CAPC) classification:

$$\hat{R}_{i,t+1} = \alpha_t + \sum_{j=1}^{K-1} \beta_{j,t} \cdot \text{Industry}_{i,j} + \epsilon_{i,t} \tag{5}$$

where $\text{Industry}_{i,j}$ is an indicator variable equal to 1 if stock i belongs to industry j, and 0 otherwise. The estimated residual $\hat{\epsilon}_{i,t}$, orthogonal to industry membership, is subsequently utilized as the neutralized signal for portfolio sorting.

Table 11 reports the decile portfolio performance sorted on this industry-neutralized signal. The results remain economically meaningful and statistically significant. For the value-weighted (VW) specification, the zero-cost portfolio generates a CH4 alpha of 0.90% per month ($t = 2.41$). For the equal-weighted (EW) specification, the corresponding alpha is 0.76% per month ($t = 3.10$). This performance wedge indicates that the extracted infor-

mation relies on intra-industry stock-selection ability, suggesting that sector rotation is not the primary driver of profitability.

[Insert Table 11 about here]

6.4. Hyperparameter Sensitivity and Structural Regularization

A methodological concern when applying deep neural networks to financial data is the risk of overfitting due to low signal-to-noise ratios. Unlike standard machine learning tasks where early stopping via a validation set is common practice, financial validation sets are frequently noisy realizations; optimizing against them can lead the network to overfit the validation sample itself. To regularize the neural network and constrain its effective capacity, we enforced a fixed training duration of 10 epochs in the baseline MLP architecture.

To empirically evaluate this early-stopping threshold and verify that the results are not artifacts of hyperparameter optimization, we assess the cross-sectional ranking performance of the MLP architecture across a grid of training epochs (ranging from 5 to 12). This evaluation is conducted on a subsample spanning January 2009 to December 2016.

[Insert Table 9 about here]

As reported in Table 9, the results illustrate a bias-variance trade-off. At 10 epochs, the model achieves a balance of signal extraction and temporal stability, maximizing both the Information Coefficient Information Ratio (ICIR of 0.86) and the statistical significance (t-statistic of 8.43). Beyond 10 epochs (e.g., at epoch 12), although the raw Rank IC marginally increases to 7.43%, the risk-adjusted stability metrics degrade. This divergence signals the onset of overfitting to idiosyncratic noise, supporting the 10-epoch constraint as a practical regularization mechanism.

6.5. Economic Implementability: Transaction Costs and Turnover

A common critique of machine learning signals is their propensity to generate high turnover, which can render theoretical alphas unattainable in practical deployment (Novy-Marx and Velikov, 2016). To evaluate the net profitability of the strategy, we recalculate the performance of the H-L hedge portfolios after deducting transaction costs.

Let c denote the round-trip transaction cost rate (e.g., 20, 40, or 60 bps), which accounts for trading commissions, stamp duties, and estimated bid-ask spread slippage in the Chinese A-share market. Because both legs of the hedge portfolio incur execution frictions, the transaction cost drag is cumulative. The net return is defined as:

$$R_{H-L}^{Net} = \left(R_H^{Gross} - R_L^{Gross}\right) - c \times (TO_H + TO_L) \tag{6}$$

where TO_H and TO_L represent the average monthly single-sided turnover for the long and short portfolios, respectively.

Table 12 reports the net-of-fee CH4 alphas across varying lookback windows (LB). Optimizing at a one-month lookback window ($LB = 1$) maximizes the raw signal-to-noise ratio but necessitates an average monthly turnover of 83.27% for the long leg and 89.20% for the short leg. While this baseline remains profitable under moderate friction assumptions, such turnover levels pose implementation barriers for institutional capital.

Consequently, we examine the two-month smoothing window ($LB = 2$) as a potential threshold for real-world deployment. This smoothing window suppresses portfolio turnover (dropping to 67.96% and 75.11%) while preserving the underlying semantic insights. Under an institutional friction assumption of 40 bps per round-trip, the $LB = 2$ strategy yields a net value-weighted CH4 alpha of 1.05% ($t = 2.33$). Under a more conservative assumption of 60 bps, the net alpha is 0.76% ($t = 1.69$). This net profitability indicates that the LLM-extracted soft information can generate an actionable alpha net of transaction costs.

[Insert Table 12 about here]

7. Conclusion

This paper investigates the independent pricing power of unstructured soft information in the Chinese A-share market. By integrating a financial-domain Large Language Model (LLM) with a Multilayer Perceptron (MLP) architecture, we process over 770,000 full-text analyst reports to map high-dimensional narrative semantics to future stock returns.

Our empirical results indicate that continuous neural networks can effectively capture complex semantic interactions, significantly outperforming linear and discrete tree-based models in this setting. By explicitly orthogonalizing the neural network's predictions against quantitative consensus metrics, we isolate a textual soft information score that generates a robust and asymmetric risk-adjusted alpha. Furthermore, our evidence suggests that the persistence of this semantic premium is partly attributable to the interaction between information complexity and structural market frictions—specifically cognitive processing costs, limited investor attention, and short-sale constraints.

The core findings and economic implications of our study are threefold:

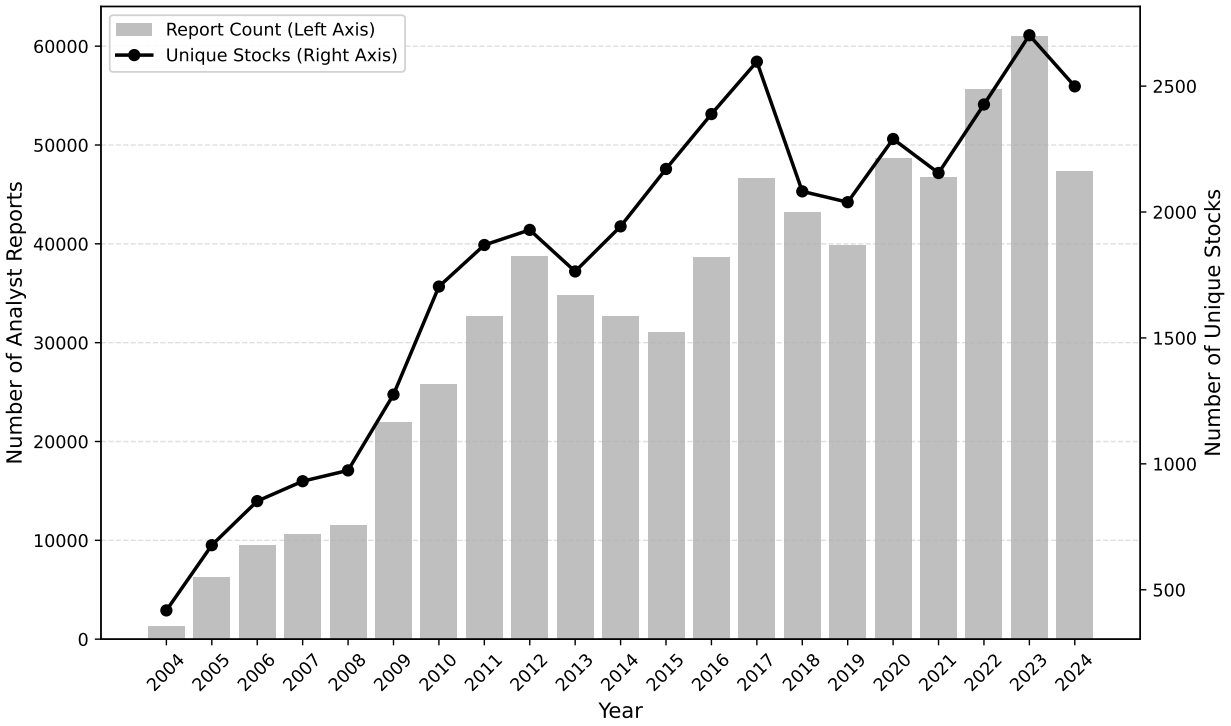
- **Methodological evidence:** We document a clear performance hierarchy (MLP > XGBoost > Ridge) in mapping analyst text to returns, providing empirical evidence that unstructured financial narratives contain non-linear predictive signals that are better exploited by continuous deep learning architectures than by linear or discrete tree-based alternatives.

- **Behavioral microstructure insights:** By disentangling qualitative narratives from quantitative forecasts, we find patterns consistent with analysts' "hidden optimism"—extreme positive views expressed in text but not fully reflected in hard forecasts. Our mechanism analysis offers micro-behavioral evidence on how cognitive constraints and

limits to arbitrage may contribute to the staggered and asymmetric incorporation of soft information into asset prices.

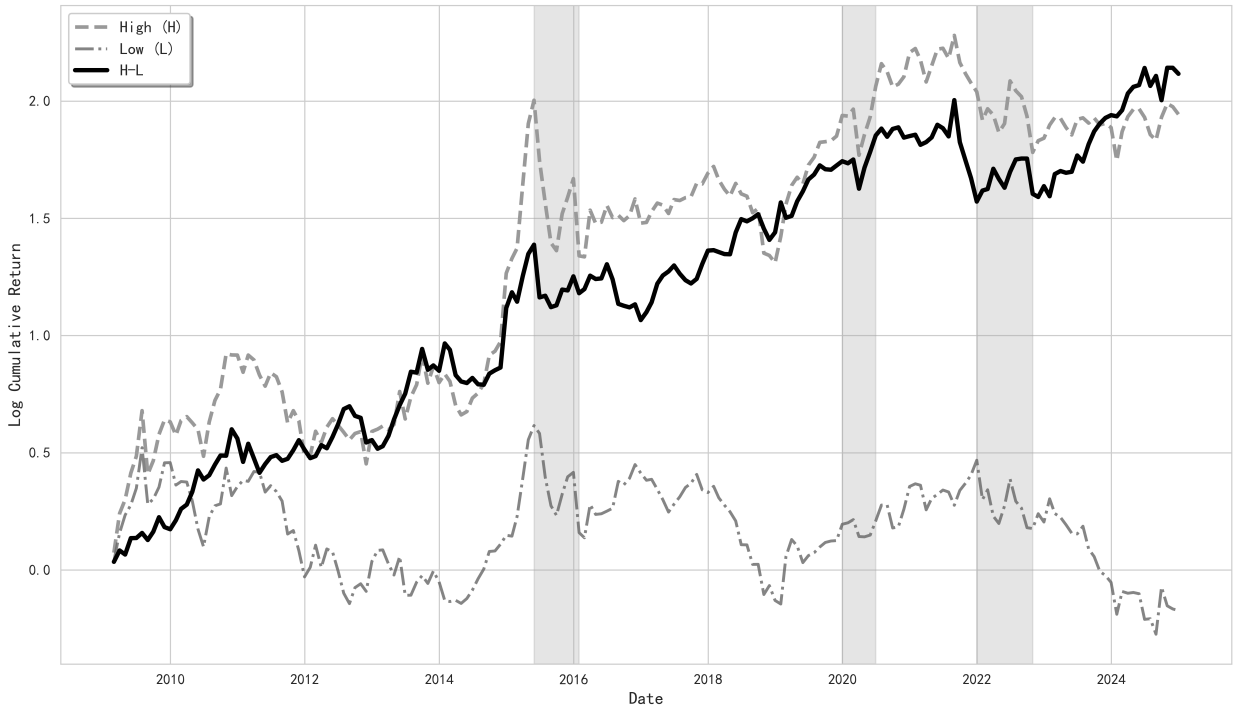
- **Economic implementability:** In contrast to several well-known machine learning anomalies that concentrate in micro-cap stocks, the semantic premium we identify is most pronounced among liquid, large-cap equities with high analyst coverage, and it remains economically viable after accounting for reasonable transaction costs.

Figure 1: Annual Distribution of Analyst Reports and Covered Stocks



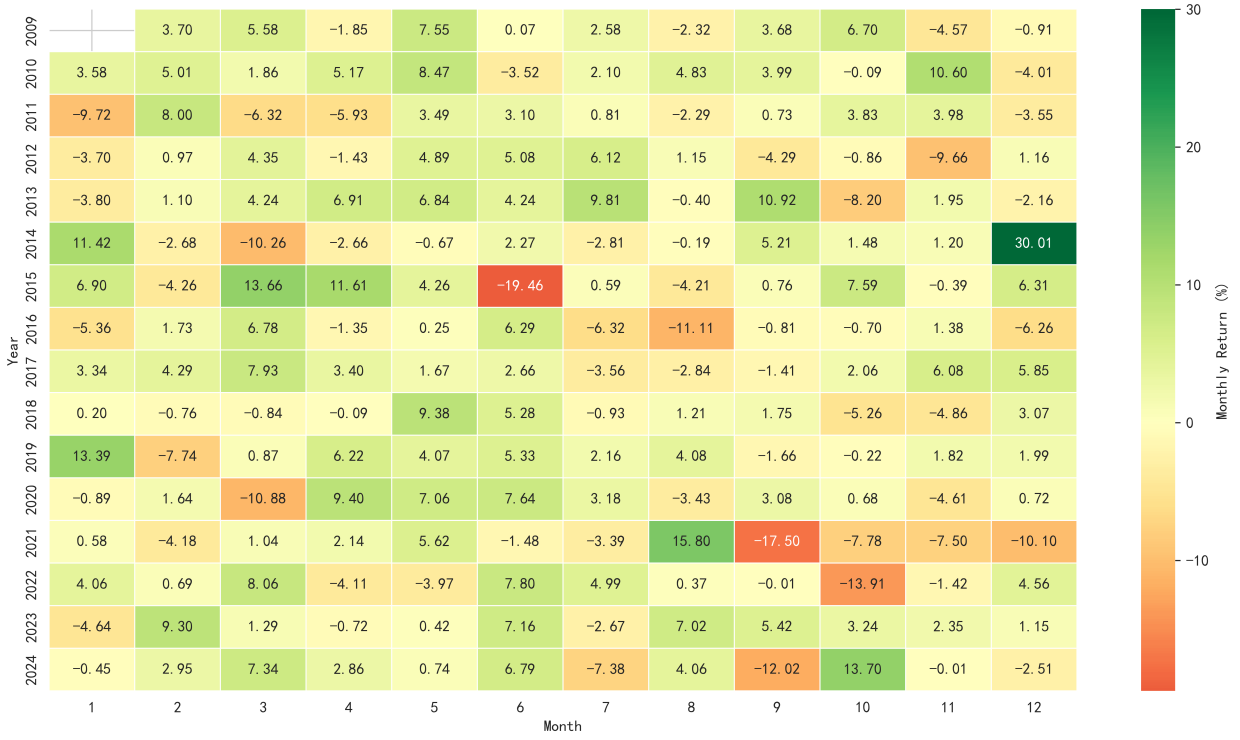
Notes: This figure plots the annual distribution of analyst reports and the number of unique covered stocks over the full sample period 2004–2024. The grey bars (left axis) represent the total number of valid analyst reports published each year. The solid black line (right axis) represents the corresponding number of unique A-share stocks covered by these reports.

Figure 2: Log Cumulative Return



Notes: This figure plots the log cumulative returns of value-weighted decile portfolios sorted by the baseline LLM-MLP signal ($\hat{R}_{i,t+1}$) over the out-of-sample period 2009-2024. H (High) = top decile, L (Low) = bottom decile, H-L = zero-cost long-short portfolio. Shaded areas denote major market downturns: 2015 market crash, 2020 COVID-19 outbreak, and 2022 bear market.

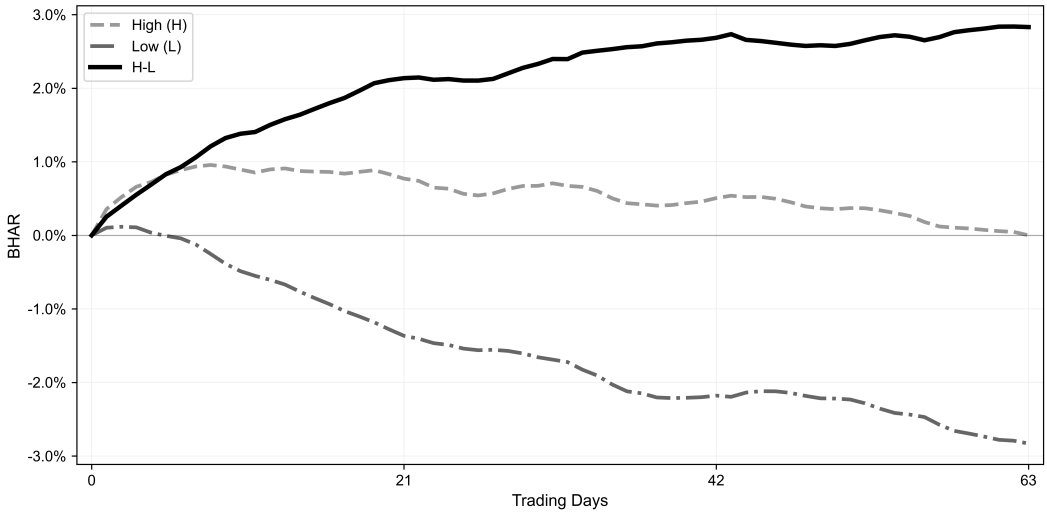
Figure 3: Monthly Return Heatmap



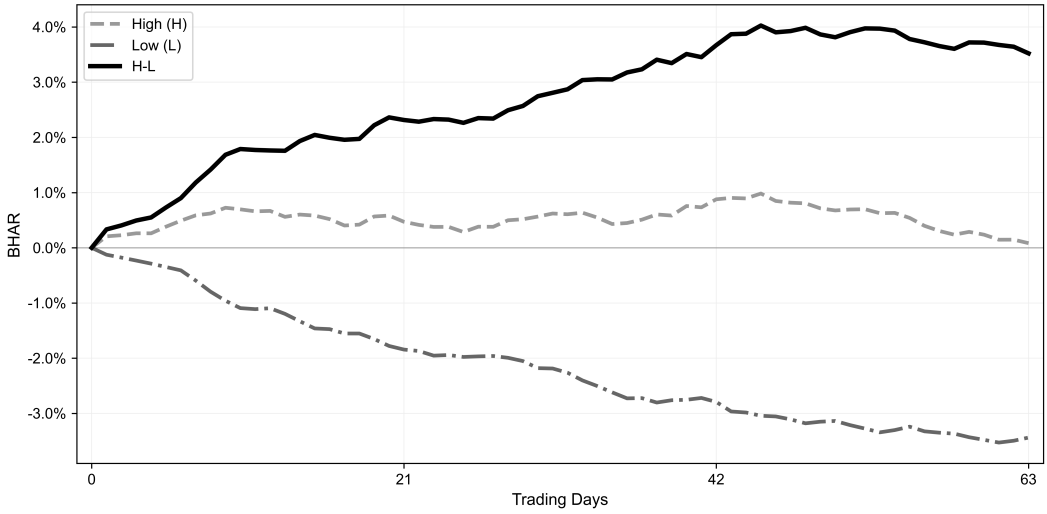
Notes: This heatmap displays monthly returns (in percent) of the value-weighted H-L portfolio over 2009–2024. Green cells = positive returns, red cells = negative returns.

Figure 4: Report-based Buy-and-Hold Abnormal Returns

Panel A: Equal-Weighted BHAR



Panel B: Value-Weighted BHAR



Notes:This figure plots the 63-day cumulative buy-and-hold abnormal returns ($BHAR$) around analyst report publication dates over the 2009–2024 period. Reports are sorted into deciles based on the baseline semantic signal ($\hat{R}_{i,t+1}$). H denotes the top decile (High), L denotes the bottom decile (Low), and $H-L$ represents the zero-cost long-short portfolio. Abnormal returns are benchmark-adjusted following the DGTW (Daniel et al., 1997) characteristic-matching procedure. Panel A reports equal-weighted (EW) returns, and Panel B reports value-weighted (VW) returns based on lagged market capitalization.

Table 1: Model Comparison: Point Prediction vs. Cross-Sectional Ranking

Panel A: Point Prediction Accuracy (OOS R^2)		
Model	Unwinsorized R^2_{OOS} (%)	Winsorized R^2_{OOS} (%)
Ridge (Linear Baseline)	1.58	−1.98
XGBoost (Tree-based)	3.42	0.67
MLP (Neural Network)	**7.10**	−1.69
Diebold-Mariano Test (MLP vs. Ridge): Unwinsorized $t = 4.45^{***}$*, Winsorized* $t = 0.11$		
Diebold-Mariano Test (MLP vs. XGBoost): Unwinsorized $t = 3.93^{***}$*, Winsorized* $t = -1.61$		

Panel B: Cross-Sectional Ranking Performance			
Model	Mean Rank IC (%)	ICIR	t-stat
Ridge (Linear Baseline)	2.93^{***}	0.29	4.08
XGBoost (Tree-based)	3.90^{***}	0.38	5.32
MLP (Neural Network)	$\mathbf{7.11}^{***}$	**0.72**	**10.00**

Notes: This table reports the out-of-sample predictive performance of three competing models from 2009 to 2024. All models use identical 60-month rolling training windows and take dynamic Z-score standardized 1024-dimensional LLM embeddings as inputs. The prediction target is defined as the 21-day cumulative stock return following the report publication date. To strictly constrain model capacity and prevent the neural network from overfitting to idiosyncratic noise, the Multilayer Perceptron (MLP) architecture is structurally regularized with a fixed training duration of exactly 10 epochs. **Panel A** presents the out-of-sample R-squared (R^2_{OOS}) based on Mean Squared Error (MSE). The unwinsorized column uses raw target returns, while the winsorized column uses target returns strictly winsorized at the 1% and 99% levels exclusively within the active training set. Diebold-Mariano (DM) tests evaluate whether the predictive accuracy of the MLP is significantly different from the respective benchmarks. **Panel B** evaluates the cross-sectional ordinal ranking capability. Report-level predictions are aggregated into stock-level signals via simple cross-sectional averages before computing the Spearman Rank Information Coefficient (Rank IC). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 2: Portfolio Performance Across Different Lookback Windows

	$LB = 1$					$LB = 2$				
Portfolio	Mean(%)	SD(%)	SR	Alpha(%)	t-stat	Mean(%)	SD(%)	SR	Alpha(%)	t-stat
1 (Low)	0.18	7.41	0.09	−0.40	−1.35	0.20	7.52	0.09	−0.44*	−1.77
2	0.63	7.00	0.31	−0.23	−1.17	0.51	7.05	0.25	−0.29	−1.49
3	0.64	6.85	0.32	−0.13	−0.83	0.86	6.39	0.47	0.18	0.94
4	0.96	6.93	0.48	0.26	1.06	0.84	6.99	0.42	−0.02	−0.08
5	0.76	6.61	0.40	0.06	0.44	0.44	6.59	0.23	−0.39**	−2.29
6	0.95	6.92	0.48	0.28	1.27	0.89	7.25	0.43	0.39*	1.80
7	0.98	7.33	0.46	0.35	1.54	0.91	7.40	0.42	0.26	1.25
8	0.75	7.74	0.33	0.25	0.93	1.25	7.31	0.59	0.67**	2.26
9	0.86	7.52	0.40	0.50	1.63	0.79	7.57	0.36	0.42	1.27
10 (High)	1.42	9.03	0.55	1.21***	4.48	1.19	9.09	0.45	1.19***	3.60
H-L	**1.24**	**5.98**	**0.72**	**1.62*****	**4.28**	**0.99**	**5.95**	**0.58**	**1.62*****	**3.61**

	$LB = 3$					$LB = 4$				
Portfolio	Mean(%)	SD(%)	SR	Alpha(%)	t-stat	Mean(%)	SD(%)	SR	Alpha(%)	t-stat
1 (Low)	0.46	7.50	0.21	−0.32	−1.15	0.39	7.39	0.18	−0.32	−1.13
2	0.46	7.23	0.22	−0.22	−1.04	0.75	6.81	0.38	0.01	0.04
3	0.74	6.75	0.38	0.04	0.29	0.53	7.17	0.25	−0.21	−1.16
4	0.59	6.88	0.30	−0.17	−0.82	0.76	6.80	0.39	−0.02	−0.13
5	0.68	6.92	0.34	0.11	0.79	0.79	6.89	0.40	0.19	1.02
6	1.00	6.93	0.50	0.23	1.23	0.95	6.64	0.50	0.30	1.30
7	1.12	7.34	0.53	0.36	1.25	0.85	7.39	0.40	0.18	0.77
8	0.96	7.31	0.45	0.56	1.45	0.76	7.16	0.37	0.34	1.12
9	0.79	7.77	0.35	0.44	1.30	1.09	7.90	0.48	0.61*	1.68
10 (High)	1.24	9.20	0.47	1.12***	3.56	1.10	9.02	0.42	0.92**	2.57
H-L	**0.78**	**6.14**	**0.44**	**1.44*****	**3.15**	**0.71**	**6.29**	**0.39**	**1.24****	**2.33**

Notes: This table reports the performance of value-weighted decile portfolios across four lookback window (LB) specifications over the 2009–2024 out-of-sample period. To strictly mitigate shell-value pollution, the bottom 30% of stocks by market capitalization across the entire market are excluded prior to portfolio construction. Eligible stocks are sorted monthly into deciles based on the temporally aggregated baseline LLM-MLP signal ($\hat{R}_{i,t+1}$) and value-weighted using lagged market capitalization. *Mean* denotes the average monthly return (%), *SD* represents the monthly return standard deviation (%), *SR* is the annualized Sharpe ratio, and *Alpha* indicates the risk-adjusted alpha from the China-specific four-factor (CH4) model of Liu et al. (2019) (%), which explicitly accounts for shell-value contamination and retail-driven sentiment (PMO). The t-statistics (reported as t-stat) are calculated using Newey and West (1987) adjusted standard errors with 12 lags. G1 represents the bottom predicted return decile, G10 is the top decile, and H-L designates the zero-cost long-short portfolio (long G10, short G1). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 3: Alpha and t-statistics Across Factor Models and Weighting Schemes

		EW		VW	
Model	Portfolio	Alpha(%)	t-stat	Alpha(%)	t-stat
CH3	1 (Low)	−0.08	−0.52	−0.29	−1.03
	10 (High)	0.96***	3.00	0.95***	3.04
	H-L	**1.03*****	**3.11**	**1.24*****	**3.24**
CH4	1 (Low)	−0.14	−0.80	−0.40	−1.35
	10 (High)	1.10***	3.69	1.21***	4.48
	H-L	**1.24*****	**3.79**	**1.62*****	**4.28**
FF3	1 (Low)	−0.41***	−3.52	−0.53**	−2.36
	10 (High)	0.42*	1.82	0.62**	2.39
	H-L	**0.83*****	**3.38**	**1.15*****	**3.29**
FF5	1 (Low)	−0.43***	−3.72	−0.50**	−2.18
	10 (High)	0.30	1.32	0.57**	2.34
	H-L	**0.73*****	**3.11**	**1.07*****	**3.31**
FF6	1 (Low)	−0.36***	−3.21	−0.42*	−1.73
	10 (High)	0.13	0.55	0.37	1.28
	H-L	**0.49****	**2.09**	**0.79****	**1.99**

Notes: This table reports monthly risk-adjusted alphas (in percent) and corresponding t-statistics for extreme decile and zero-cost long-short (*H-L*) portfolios across five mainstream asset pricing models over the 2009–2024 out-of-sample period. To strictly mitigate shell-value contamination, the bottom 30% of stocks by market capitalization across the entire market are excluded prior to portfolio construction. Eligible stocks are sorted monthly into deciles based on the baseline LLM-MLP semantic signal ($\hat{R}_{i,t+1}$) utilizing a one-month look-back window ($LB = 1$). *1 (Low)* and *10 (High)* denote the bottom and top predicted return deciles, respectively. *H-L* represents the hedge portfolio that longs the top decile and shorts the bottom decile. *EW* and *VW* denote equal-weighted and value-weighted (based on lagged market capitalization) portfolio returns, respectively. The benchmark models include the China-specific three- and four-factor (CH3, CH4) models of Liu et al. (2019), and the standard Fama-French three-, five-, and six-factor (FF3, FF5, FF6) models. The t-statistics are calculated using Newey and West (1987) adjusted standard errors with 12 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 4: Portfolio Performance of Textual soft information (Orthogonalized)

	EW					VW				
Portfolio	Mean(%)	SD(%)	SR	Alpha(%)	t-stat	Mean(%)	SD(%)	SR	Alpha(%)	t-stat
1 (Low)	0.48	8.18	0.20	−0.05	−0.16	0.51	8.08	0.22	0.11	0.26
2	0.75	7.62	0.34	0.07	0.24	0.85	6.91	0.43	0.20	0.73
3	0.85	7.16	0.41	0.07	0.25	0.72	6.28	0.40	0.06	0.19
4	1.00	7.45	0.46	0.22	0.86	1.03	6.54	0.55	0.38*	1.78
5	1.05	7.31	0.50	0.32	1.19	0.71	6.56	0.38	0.15	0.44
6	1.29	7.41	0.61	0.50	1.48	0.91	7.09	0.44	0.09	0.36
7	1.21	7.55	0.55	0.50	1.47	0.66	6.98	0.33	0.06	0.15
8	1.18	7.59	0.54	0.55*	1.63	0.94	7.11	0.46	0.41	1.23
9	1.19	7.83	0.53	0.65**	1.94	0.98	7.16	0.47	0.56	1.12
10 (High)	1.85	8.57	0.75	1.41***	3.45	1.76	8.77	0.70	1.46***	3.94
H-L	**1.37**	**4.33**	**1.10**	**1.45*****	**4.08**	**1.26**	**6.28**	**0.69**	**1.36*****	**2.77**

Notes: This table reports the performance of value-weighted (VW) and equal-weighted (EW) decile portfolios sorted by the textual soft information score ($\hat{R}^{soft}_{i,t+1}$) over the 2009–2024 out-of-sample period. To strictly mitigate shell-value pollution, the bottom 30% of stocks by market capitalization across the entire market are excluded prior to portfolio construction. Following the baseline specification, the signal utilizes a one-month lookback window ($LB = 1$). To isolate genuine narrative expectation, $\hat{R}^{soft}_{i,t+1}$ is strictly orthogonalized against database-derived quantitative consensus metrics—forward earnings-to-price ratio (FEP), target price implied return (TPR), and average recommendation rating (REC)—via monthly cross-sectional OLS regressions. *Mean* denotes the average monthly return (%), *SD* represents the monthly return standard deviation (%), *SR* is the annualized Sharpe ratio, and *Alpha* indicates the risk-adjusted alpha from the China-specific four-factor (CH4) model of Liu et al. (2019) (%). The t-statistics are calculated using Newey and West (1987) adjusted standard errors with 12 lags. G1 represents the bottom predicted return decile, G10 is the top decile, and H-L designates the zero-cost long-short portfolio (long G10, short G1). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 5: Residualized Portfolio Performance Across Different Lookback Windows

	$LB = 1$					$LB = 2$				
Portfolio	Mean(%)	SD(%)	SR	Alpha(%)	t-stat	Mean(%)	SD(%)	SR	Alpha(%)	t-stat
1 (Low)	0.51	8.08	0.22	0.11	0.26	0.23	7.84	0.10	−0.16	−0.48
2	0.85	6.91	0.43	0.20	0.73	0.50	6.94	0.25	−0.22	−0.71
3	0.72	6.28	0.40	0.06	0.19	0.88	6.77	0.45	0.34	1.12
4	1.03	6.54	0.55	0.38*	1.78	0.79	6.38	0.43	0.20	0.74
5	0.71	6.56	0.38	0.15	0.44	0.62	6.35	0.34	0.02	0.09
6	0.91	7.09	0.44	0.09	0.36	1.13	6.93	0.56	0.25	1.10
7	0.66	6.98	0.33	0.06	0.15	0.60	7.12	0.29	−0.05	−0.15
8	0.94	7.11	0.46	0.41	1.23	1.17	6.72	0.61	0.44	1.24
9	0.98	7.16	0.47	0.56	1.12	1.30	7.46	0.60	0.92**	2.01
10 (High)	1.76	8.77	0.70	1.46***	3.94	1.26	8.99	0.49	1.05**	2.26
H-L	**1.26**	**6.28**	**0.69**	**1.36*****	**2.77**	**1.04**	**6.34**	**0.57**	**1.21*****	**2.91**

	$LB = 3$					$LB = 4$				
Portfolio	Mean(%)	SD(%)	SR	Alpha(%)	t-stat	Mean(%)	SD(%)	SR	Alpha(%)	t-stat
1 (Low)	0.10	7.79	0.05	−0.29	−0.98	0.38	7.73	0.17	−0.13	−0.32
2	0.44	7.12	0.21	−0.13	−0.41	0.70	6.69	0.36	0.38	1.54
3	0.86	6.60	0.45	0.22	0.98	0.46	7.14	0.23	−0.21	−0.91
4	1.07	6.11	0.61	0.48**	1.98	1.09	6.17	0.61	0.54**	2.06
5	0.71	6.70	0.37	0.08	0.27	0.84	6.49	0.45	0.22	0.74
6	0.96	6.73	0.50	0.15	0.66	0.90	6.69	0.47	0.06	0.20
7	0.98	6.54	0.52	0.44	1.28	1.02	6.96	0.51	0.32	1.03
8	0.93	6.93	0.47	0.21	0.64	0.87	6.47	0.47	0.18	0.57
9	0.94	7.11	0.46	0.17	0.46	0.89	7.25	0.43	0.22	0.52
10 (High)	1.16	9.11	0.44	1.12**	2.50	1.18	8.91	0.46	1.01**	2.24
H-L	**1.06**	**6.02**	**0.61**	**1.41*****	**3.43**	**0.79**	**5.95**	**0.46**	**1.14*****	**2.91**

Notes: This table reports the performance of value-weighted decile portfolios across four lookback window (LB) specifications for the residualized textual soft information signal ($\hat{R}^{soft}_{i,t+1}$) over the 2009–2024 out-of-sample period. To strictly mitigate shell-value pollution, the bottom 30% of stocks by market capitalization across the entire market are excluded prior to portfolio construction. To isolate genuine narrative expectations, $\hat{R}^{soft}_{i,t+1}$ is extracted by strictly orthogonalizing the baseline semantic signal against database-derived quantitative consensus metrics—forward earnings-to-price ratio (FEP), target price implied return (TPR), and average recommendation rating (REC)—via monthly cross-sectional OLS regressions. *Mean* denotes the average monthly return (%), *SD* represents the monthly return standard deviation (%), *SR* is the annualized Sharpe ratio, and *Alpha* indicates the risk-adjusted alpha from the China-specific four-factor (CH4) model of Liu et al. (2019) (%). The t-statistics (reported as t-stat) are calculated using Newey and West (1987) adjusted standard errors with 12 lags. G1 represents the bottom predicted return decile, G10 is the top decile, and H-L designates the zero-cost long-short portfolio (long G10, short G1). ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 6: Information Decay: Full Information vs. Textual Soft Information

Portfolio Metric	Lookback Window (LB)				Decay (1m to 4m)		
	$LB = 1$	$LB = 2$	$LB = 3$	$LB = 4$	Absolute (Δ)	Rate (%)	Δ t-stat
Panel A: Full Information Signal ($\hat{R}_{i,t+1}$)							
H-L Mean (%)	1.24	0.99	0.78	0.71	-0.53	-42.74%	
CH4 Alpha (%)	1.62	1.62	1.44	1.24	-0.38	-23.46%	
t-stat	(4.28)	(3.61)	(3.15)	(2.33)		-45.56%	**-1.95**
Panel B: Textual Soft Information Signal ($\hat{R}^{soft}_{i,t+1}$)							
H-L Mean (%)	1.26	1.04	1.06	0.79	-0.47	-37.30%	
CH4 Alpha (%)	1.36	1.21	1.41	1.14	-0.22	-16.18%	
t-stat	(2.77)	(2.91)	(3.43)	(2.91)		+5.05%	**+0.14**

Notes: This table compares the predictive power decay between the full information signal ($\hat{R}_{i,t+1}$, **Panel A**) and the strictly orthogonalized textual soft information signal ($\hat{R}^{soft}_{i,t+1}$, **Panel B**) across one- to four-month lookback window (LB) specifications. Consistent with baseline construction, the bottom 30% of stocks by market capitalization are excluded to mitigate shell-value pollution, and the decile portfolios are value-weighted based on lagged market capitalization. H-L represents the zero-cost long-short hedge portfolio. To explicitly quantify the differential pricing speeds, the **Decay** columns report the absolute drop (where Δ denotes the difference between $LB4$ and $LB1$), the percentage decay rate of the CH4 *Alpha* (%), and the absolute change in the t-statistic (Δ t-stat) between the four-month and one-month horizons. *Alphas* are estimated using the China-specific four-factor (CH4) model of Liu et al. (2019). The t-statistics (reported in parentheses) are calculated using Newey and West (1987) adjusted standard errors with 12 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 7: Double Sorts: Information Environment and Limits to Arbitrage

Control Variable	Control Group	Low Pred	Group 2	Group 3	Group 4	High Pred	H-L (VW)	H-L (EW)
Panel A: Size								
Size	Small	0.09	0.24	0.16	0.55***	0.87***	0.78***	0.84***
		(0.44)	(1.40)	(0.60)	(2.75)	(3.40)	(2.53)	(2.67)
	Big	-0.40*	0.10	0.09	0.35	0.77***	1.17***	1.23***
		(-1.77)	(0.66)	(0.47)	(1.57)	(2.61)	(3.53)	(3.63)
Panel B: Book-to-Market (BM)								
BM	Low BM	-0.15	0.65*	0.21	0.36	1.11***	1.26***	1.18***
		(-0.32)	(1.81)	(0.52)	(0.72)	(2.64)	(3.30)	(4.15)
	High BM	-0.53**	-0.39***	-0.04	-0.11	0.25	0.77**	0.54**
		(-2.50)	(-2.61)	(-0.14)	(-0.47)	(0.82)	(2.27)	(2.33)
Panel C: Turnover								
Turnover	Low Turnover	-0.58***	0.11	0.01	0.09	0.57**	1.15***	0.84***
		(-2.73)	(0.57)	(0.06)	(0.37)	(1.94)	(3.91)	(3.38)
	High Turnover	-0.03	0.25	0.60**	0.55	1.29***	1.31***	1.19***
		(-0.09)	(1.01)	(2.65)	(1.39)	(3.44)	(3.33)	(3.42)
Panel D: Momentum								
Momentum	Loser	-0.61***	-0.29	-0.36*	-0.07	0.23	0.84***	0.63**
		(-2.69)	(-1.26)	(-1.92)	(-0.24)	(1.02)	(3.47)	(2.57)
	Winner	0.24	0.50	0.51	0.42	0.90**	0.66*	0.74***
		(0.65)	(1.38)	(1.43)	(0.94)	(2.48)	(1.82)	(3.16)
Panel E: Shortability								
Shortability	Non-Shortable	-0.27	0.18	0.32	0.73**	1.49***	1.76***	1.57***
		(-0.89)	(0.50)	(0.78)	(1.74)	(2.30)	(3.08)	(4.03)
	Shortable	-0.30*	0.23	0.10	0.29	0.64**	0.94**	0.95***
		(-1.35)	(1.25)	(0.50)	(1.10)	(2.51)	(2.83)	(3.05)
Panel F: Institutional Ownership (IO)								
IO	Low Inst.	-0.13	0.05	0.65**	0.80***	1.17***	1.30***	0.92**
		(-0.43)	(0.19)	(2.31)	(2.45)	(3.18)	(3.16)	(2.74)
	High Inst.	-0.32**	-0.05	0.02	0.00	0.64**	0.96***	0.99***
		(-1.51)	(-0.25)	(0.13)	(0.02)	(2.25)	(3.10)	(4.17)
Panel G: Analyst Coverage								
Coverage	Low Coverage	-0.28*	0.02	-0.27*	0.19	0.41	0.69**	0.70**
		(-1.50)	(0.12)	(-1.41)	(0.88)	(1.53)	(2.34)	(2.61)
	High Coverage	-0.61**	0.28	0.46**	0.34	1.18***	1.78***	1.63***
		(-2.25)	(1.09)	(2.02)	(1.22)	(3.73)	(4.73)	(4.76)

Notes: This table reports monthly risk-adjusted alphas (%) for dependent double-sorted portfolios over the 2009–2024 out-of-sample period. Eligible stocks are first partitioned monthly into two groups by a control variable, then sorted into quintiles based on the LLM-MLP signal. Columns 3–7 report value-weighted alphas for intersection portfolios. *H-L (VW)* and *H-L (EW)* denote value-weighted and equal-weighted hedge portfolio alphas, respectively. Alphas are estimated using the China-specific four-factor (CH4) model of Liu et al. (2019). The t-statistics (in parentheses) are calculated using Newey and West (1987) adjusted standard errors with 12 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 8: Factor Exposures of Long and Long-Short Portfolios Across Asset Pricing Models

	CH3		CH4		FF3		FF5		FF6	
	Long	H-L	Long	H-L	Long	H-L	Long	H-L	Long	H-L
Alpha (%)	0.95***	1.24***	1.21***	1.62***	0.61**	1.14***	0.57**	1.07***	0.37	0.79**
	(3.04)	(3.24)	(4.48)	(4.28)	(2.39)	(3.29)	(2.34)	(3.31)	(1.28)	(1.99)
Market	1.18***	0.18	1.13***	0.11	1.16***	0.18	1.19***	0.21	1.23***	0.27**
	(12.49)	(1.32)	(14.40)	(1.04)	(12.37)	(1.36)	(12.79)	(1.55)	(13.31)	(2.13)
SMB	-0.14	-0.10	-0.05	0.02	-0.07	-0.02	0.02	0.12	0.04	0.15
	(-1.31)	(-0.88)	(-0.55)	(0.26)	(-0.68)	(-0.18)	(0.18)	(1.24)	(0.42)	(1.38)
VMG	-0.24*	-0.03	-0.24*	-0.03						
	(-1.77)	(-0.15)	(-1.93)	(-0.19)						
HML					-0.42***	0.00	-0.37**	0.06	-0.30**	0.16
					(-2.87)	(-0.02)	(-2.53)	(0.29)	(-2.40)	(0.87)
RMW							0.22	0.31	0.19	0.27
							(0.91)	(0.88)	(0.81)	(0.84)
CMA							-0.14	-0.05	-0.08	0.04
							(-0.57)	(-0.15)	(-0.42)	(0.17)
PMO			-0.37***	-0.52***						
			(-3.45)	(-3.05)						
MOM									0.31***	0.43***
									(3.70)	(3.38)
Adj. R^2	0.76	0.02	0.78	0.11	0.75	0.02	0.76	0.02	0.79	0.16
N	191	191	191	191	191	191	191	191	191	191

Notes: This table reports the factor loadings and monthly risk-adjusted alphas (in percent) for the long-only (top decile) and zero-cost long-short (*H-L*) value-weighted portfolios across five mainstream asset pricing models over the 2009–2024 out-of-sample period. To strictly mitigate shell-value contamination, the bottom 30% of stocks by market capitalization across the entire market are excluded prior to portfolio construction. Portfolios are formed monthly based on the baseline LLM-MLP semantic signal ($\hat{R}_{i,t+1}$) utilizing a one-month lookback window ($LB = 1$). The benchmark models include the China-specific three- and four-factor (CH3, CH4) models of Liu et al. (2019), and the Fama-French three-, five-, and six-factor (FF3, FF5, FF6) models. The t-statistics (reported in parentheses) are calculated using Newey and West (1987) adjusted standard errors with 12 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 9: Hyperparameter Sensitivity: Model Performance Across Training Epochs

Epochs	Mean Rank IC (%)	ICIR	t-stat
5	6.21	0.66	6.51
6	6.51	0.71	6.91
7	7.00	0.75	7.37
8	6.80	0.76	7.49
9	7.26	0.82	8.02
10	7.31	0.86	8.43
11	6.69	0.75	7.37
12	7.43	0.85	8.29

Notes: This table reports the cross-sectional ranking performance of the MLP architecture across different training epochs (5 to 12) to empirically justify the optimal early-stopping threshold. To ensure structural robustness across different macroeconomic regimes, the evaluation is conducted on a highly representative subsample spanning from January 2009 to December 2016. The metrics include the Mean Spearman Rank Information Coefficient (Rank IC) and the Information Coefficient Information Ratio (ICIR). The results demonstrate a clear bias-variance trade-off. At exactly 10 epochs, the model achieves the optimal balance of signal extraction and stability, maximizing both the ICIR (0.86) and the t-statistic (8.43). Beyond 10 epochs (e.g., at epoch 12), although the raw Rank IC marginally increases, the risk-adjusted stability metrics (ICIR and t-stat) visibly decline. This divergence indicates early signs of overfitting to idiosyncratic noise, thereby confirming 10 epochs as the optimal structural regularization threshold for low signal-to-noise ratio financial data.

Table 10: Fama-MacBeth Regressions

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$\hat{R}_{i,t+1}$	0.22**	0.21***	0.22***	0.21***	0.21***	
	(2.28)	(3.32)	(2.61)	(3.27)	(2.73)	
$Size$		−0.28*		−0.29*	−0.29	−0.29
		(−1.67)		(−1.68)	(−1.61)	(−1.64)
BM		0.18		0.07	0.15	0.13
		(1.03)		(0.48)	(1.07)	(0.93)
EP		−0.07		−0.31***	−0.29***	−0.29***
		(−0.72)		(−2.63)	(−2.95)	(−2.95)
Mom		0.19		0.19	0.16	0.20
		(1.57)		(1.56)	(1.33)	(1.61)
Rev		−0.42**		−0.43**	−0.47***	−0.44***
		(−2.42)		(−2.54)	(−2.78)	(−2.60)
FEP			0.13	0.39**	0.43***	0.43***
			(0.64)	(2.36)	(3.03)	(2.98)
TPR			0.19*	0.08	0.07	0.07
			(1.86)	(0.99)	(0.95)	(0.94)
REC			0.04	0.04	0.07	0.07
			(0.66)	(0.90)	(1.50)	(1.55)
Intercept	1.04*	1.04*	1.04*	1.04*	0.61	0.57
	(1.94)	(1.94)	(1.94)	(1.94)	(1.17)	(1.13)
Ind FE	No	No	No	No	Yes	Yes
Avg. R^2	0.014	0.123	0.068	0.142	0.203	0.198

Notes: This table reports the time-series average coefficients from monthly Fama-MacBeth regressions over the 2009–2024 period. Here, $\hat{R}_{i,t+1}$ denotes the baseline LLM-MLP semantic signal. Models 2–5 progressively control for standard firm characteristics (*Size*, *BM*, *EP*), momentum (*Mom*), and short-term reversal (*Rev*), alongside database-derived quantitative consensus metrics—forward earnings-to-price ratio (*FEP*), target price implied return (*TPR*), and recommendation rating (*REC*)—and industry fixed effects. Model 6 strictly excludes $\hat{R}_{i,t+1}$ to establish a hard-information-only explanatory baseline. Prior to each cross-sectional regression, all independent variables are strictly winsorized at the 1% and 99% levels and cross-sectionally standardized. The t-statistics (reported in parentheses) are calculated using Newey and West (1987) adjusted standard errors with 12 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 11: Industry Neutralized Portfolio Performance

	EW					VW				
Portfolio	Mean(%)	SD(%)	SR	Alpha(%)	t-stat	Mean(%)	SD(%)	SR	Alpha(%)	t-stat
1 (Low)	0.85	8.15	0.36	0.15	0.86	0.47	7.67	0.21	0.02	0.07
2	0.83	7.68	0.38	−0.02	−0.08	0.61	7.03	0.30	−0.23	-1.17
3	0.97	7.48	0.45	0.09	0.50	0.75	6.70	0.39	−0.04	-0.25
4	1.23	7.68	0.55	0.39*	1.86	0.90	6.83	0.45	0.47*	1.65
5	1.09	7.61	0.50	0.23	1.41	0.94	6.91	0.47	0.19	1.16
6	1.07	7.52	0.49	0.25	1.24	0.77	6.55	0.41	0.00	0.01
7	1.29	7.56	0.59	0.46**	2.40	0.83	6.91	0.41	0.08	0.38
8	1.10	7.70	0.50	0.36*	1.79	0.65	7.18	0.32	−0.01	-0.02
9	1.23	7.96	0.54	0.65***	2.70	1.03	7.76	0.46	0.67**	2.43
10 (High)	1.34	8.80	0.53	0.91***	3.23	1.18	8.84	0.46	0.93***	3.59
H-L	**0.50**	**3.55**	**0.48**	**0.76*** **	**3.10**	**0.71**	**5.36**	**0.46**	**0.90****	**2.41**

Notes: This table reports the performance of decile portfolios sorted by the industry-neutralized LLM-MLP semantic signal over the 2009–2024 out-of-sample period. To strictly mitigate shell-value pollution, the bottom 30% of stocks by market capitalization are excluded prior to portfolio construction. Industry neutralization is performed by extracting the residual ($\hat{\epsilon}_{i,t}$) from monthly cross-sectional ordinary least squares (OLS) regressions of the raw predicted scores on China Association for Public Companies (CAPC) industry dummies. *EW* and *VW* denote equal-weighted and value-weighted (based on lagged market capitalization) portfolio returns, respectively. *Mean*, *SD* (standard deviation), and *Alpha* are reported as monthly percentages. Alphas are estimated using the China-specific four-factor (CH4) model of Liu et al. (2019), which explicitly controls for shell-value premia and retail-driven sentiment. The t-statistics are calculated using Newey and West (1987) adjusted standard errors with 12 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 12: Transaction Costs Sensitivity and Limits to Arbitrage

LB Window	Turnover (L/S)	Frictions (bps)	Net H-L Spread (%)	Net CH4 Alpha (%)	*t*-stat
LB = 1	83.27% / 89.20%	20	0.90	1.27***	3.37
		40	0.55	0.93***	2.45
		60	0.21	0.58	1.54
LB = 2	67.96% / 75.11%	20	0.70	1.34***	2.97
		40	0.42	1.05**	2.33
		60	0.13	0.76*	1.69
LB = 3	60.70% / 69.50%	20	0.52	1.18**	2.58
		40	0.26	0.92**	2.00
		60	0.00	0.66	1.43
LB = 4	55.92% / 63.11%	20	0.47	1.01*	1.88
		40	0.23	0.77	1.43
		60	-0.00	0.53	0.99

Notes: This table reports the net-of-fee performance of the value-weighted zero-cost long-short (*H-L*) portfolio across four lookback window (*LB*) specifications and varying round-trip transaction costs ($c = 20, 40,$ and 60 bps) over the 2009–2024 period. To mitigate shell-value pollution, the bottom 30% of stocks by market capitalization are excluded. *Turnover* represents the average monthly single-sided turnover for the long (top decile) and short (bottom decile) portfolios. Net returns are calculated by deducting the total execution friction (round-trip cost multiplied by the sum of single-sided turnovers of both legs) from the gross spread. *Alphas* are estimated using the China-specific four-factor (CH4) model of Liu et al. (2019) to account for shell-value and retail sentiment premia. The *t*-statistics are calculated using Newey and West (1987) adjusted standard errors with 12 lags. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

References

Altınkılıç, O., Balashov, V.S., Hansen, R.S., 2013. Are analysts' forecasts informative to the general public? Management Science 59, 2550–2565.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al., 2023. Qwen technical report. arXiv preprint arXiv:2309.16609 .

Ball, R., Brown, P., 1968. An empirical evaluation of accounting income numbers. Journal of Accounting Research , 159–178.

Barber, B.M., Odean, T., 2008. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. The Review of Financial Studies 21, 785–818.

Bernard, V.L., Thomas, J.K., 1989. Post-earnings-announcement drift: delayed price response or risk premium? Journal of Accounting Research 27, 1–36.

Birru, J., Gokkaya, S., Liu, X., Stulz, R.M., 2022. Are analyst short-term trade ideas valuable? The Journal of Finance 77, 1829–1875.

Brav, A., Lehavy, R., 2003. An empirical analysis of analysts' target prices: Short-term informativeness and long-term dynamics. The Journal of Finance 58, 1933–1967.

Bybee, L., Kelly, B., Manela, A., Xiu, D., 2024. Business news and business cycles. The Journal of Finance 79, 3105–3147.

Carhart, M.M., 1997. On persistence in mutual fund performance. The Journal of Finance 52, 57–82.

Chang, E.C., Luo, Y., Ren, J., 2014. Short-selling, margin-trading, and price efficiency: Evidence from the chinese market. Journal of Banking & Finance 48, 411–424.

Chen, L., Pelger, M., Zhu, J., 2024. Deep learning in asset pricing. Management Science 70, 714–750.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

Chen, X., Wu, C., 2022. Retail investor attention and information asymmetry: Evidence from china. Pacific-Basin Finance Journal 75, 101847.

Chen, Y., Kelly, B.T., Xiu, D., 2022. Expected returns and large language models. Available at SSRN 4416687 .

Cohen, L., Malloy, C., Nguyen, Q., 2020. Lazy prices. The Journal of Finance 75, 1371–1415.

Daniel, K., Grinblatt, M., Titman, S., Wermers, R., 1997. Measuring mutual fund performance with characteristic-based benchmarks. The Journal of Finance 52, 1035–1058.

De Prado, M.L., 2018. Advances in financial machine learning. John Wiley & Sons.

Dyer, T., Lang, M., Stice-Lawrence, L., 2017. The evolution of 10-k textual disclosure: Evidence from latent dirichlet allocation. Journal of Accounting and Economics 64, 221–245.

Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. The Journal of Finance 25, 383–417.

Fama, E.F., French, K.R., 1992. The cross-section of expected stock returns. the Journal of Finance 47, 427–465.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fama, E.F., French, K.R., 2018. Choosing factors. Journal of Financial Economics 128, 234–252.

Fama, E.F., MacBeth, J.D., 1973. Risk, return, and equilibrium: Empirical tests. Journal of Political Economy 81, 607–636.

Farrell, M.H., Liang, T., Misra, S., 2021. Deep neural networks for estimation and inference. Econometrica 89, 181–213.

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. The Review of Financial Studies 33, 2326–2377.

Gentzkow, M., Kelly, B., Taddy, M., 2019. Text as data. Journal of Economic Literature 57, 535–574.

Gleason, C.A., Lee, C.M., 2003. Analyst forecast revisions and market price discovery. The Accounting Review 78, 193–225.

Goodfellow, I., 2016. Deep learning.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

Hanley, K.W., Hoberg, G., 2010. The information content of ipo prospectuses. The Review of Financial Studies 23, 2821–2864.

Hirshleifer, D., Teoh, S.H., 2003. Limited attention, information disclosure, and financial reporting. Journal of Accounting and Economics 36, 337–386.

Hong, H., Kubik, J.D., 2003. Analyzing the analysts: Career concerns and biased earnings forecasts. The Journal of Finance 58, 313–351.

Hong, H., Stein, J.C., 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. The Journal of Finance 54, 2143–2184.

Hong, H., Stein, J.C., 2003. Differences of opinion, short-sales constraints, and market crashes. The Review of Financial Studies 16, 487–525.

Huang, A.H., Lehavy, R., Zang, A.Y., Zheng, R., 2018. Analyst information discovery and interpretation roles: A topic modeling approach. Management science 64, 2833–2855.

Huang, A.H., Zang, A.Y., Zheng, R., 2014. Evidence on the information content of text in analyst reports. The Accounting Review 89, 2151–2180.

Jegadeesh, N., Kim, J., Krische, S.D., Lee, C.M., 2004. Analyzing the analysts: When do recommendations add value? The Journal of Finance 59, 1083–1124.

Jiang, F., Lee, J., Martin, X., Zhou, G., 2019. Manager sentiment and stock returns. Journal of Financial Economics 132, 126–149.

Jones, C.M., Shi, D., Zhang, X., Zhang, X., 2021. Understanding retail investors: evidence from china. Available at SSRN 3628809.

Ke, S., 2024. Analysts' belief formation in their own words. Available at SSRN 5025830 .

Ke, Z.T., Kelly, B.T., Xiu, D., 2019. Predicting returns with text data. Technical Report. National Bureau of Economic Research.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Leippold, M., Wang, Q., Zhou, W., 2022. Machine learning in the chinese stock market. Journal of Financial Economics 145, 64–82.

Liu, J., Stambaugh, R.F., Yuan, Y., 2019. Size and value in china. Journal of Financial Economics 134, 48–69.

Lopez-Lira, A., Tang, Y., 2023. Can chatgpt forecast stock price movements? return predictability and large language models. arXiv preprint arXiv:2304.07619 .

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance 66, 35–65.

Lv, L., 2024. The value of information from sell-side analysts. arXiv preprint arXiv:2411.13813 .

Lv, L., 2025. Do sell-side analyst reports have investment value? arXiv preprint arXiv:2502.20489 .

Michaely, R., Womack, K.L., 1999. Conflict of interest and the credibility of underwriter analyst recommendations. The Review of Financial Studies 12, 653–686.

Miller, E.M., 1977. Risk, uncertainty, and divergence of opinion. The Journal of finance 32, 1151–1168.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica , 703–708.

Ng, L., Wu, F., 2006. Revealed stock preferences of individual investors: Evidence from chinese equity markets. Pacific-Basin Finance Journal 14, 175–192.

Novy-Marx, R., Velikov, M., 2016. A taxonomy of anomalies and their trading costs. The Review of Financial Studies 29, 104–147.

Tan, L., Wu, H., Zhang, X., 2023. Large language models and return prediction in china. Available at SSRN 4712248 .

Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance 62, 1139–1168.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Womack, K.L., 1996. Do brokerage analysts' recommendations have investment value? The Journal of Finance 51, 137–167.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G., 2023. Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564 .

Zhou, Y., Fan, J., Xue, L., 2024. How much can machines learn finance from chinese text data? Management Science 70, 8962–8987.

Appendix A.

Table A.1: Variable Definitions

Variable	Definition
Panel A: Semantic Signals	
$\hat{R}_{i,t+1}$	Baseline expected return, calculated as the cross-sectional average of 21-day cumulative return predictions generated by MLP in month t.
$\hat{R}^{soft}_{i,t+1}$	Textual soft information score, extracted as the residual from orthogonalizing $\hat{R}_{i,t+1}$ against quantitative forecasts (FEP, TPR, REC) via cross-sectional OLS.
Panel B: Explicit Quantitative Forecasts (Hard Information)	
FEP	Consensus forward earnings-to-price ratio in month t.
TPR	Consensus target price implied return relative to current stock price.
REC	Consensus average recommendation rating score.
Panel C: Firm-Level Characteristics and Controls	
Size	Natural logarithm of total market capitalization at end of month t.
BM	Book value of equity divided by total market capitalization.
EP	Trailing 12-month (TTM) net income divided by total market capitalization.
Turnover	Monthly stock turnover rate.
IO	Institutional ownership, measured as the percentage of outstanding shares held by institutions.
Shortability	Dummy variable equal to 1 if the stock is eligible for margin trading and short-selling.
Coverage	Total number of analyst reports covering the stock in month t.
Mom	Cumulative return from month $t-11$ to $t-1$, skipping month t to avoid microstructure noise.
Rev	Short-term reversal factor, measured as the stock return in month t.
SOE	Dummy variable indicating state-owned enterprise status.
Panel D: Asset Pricing Factor Variables	
PMO	Retail-driven sentiment factor in the CH4 model (Liu et al., 2019).
VMG	Value factor in the CH4 model, constructed using the EP ratio.