

From Text to Alpha: Can LLMs Track Evolving Signals in Corporate Disclosures?

Chanyeol Choi^{1*}, Yoon Kim², Yu Yu³, Young Cha⁴, V. Zach Golkhou⁵, Igor Halperin⁶, Georgios Papaioannou⁷, Minkyu Kim⁸, Zhangyang Wang⁹, Jihoon Kwon¹, Minjae Kim¹, Alejandro Lopez-Lira¹⁰, Yongjae Lee^{1, 11}

¹LinqAlpha ²MIT ³BlackRock ⁴Blackstone ⁵J.P. Morgan ⁶Fidelity Investments
⁷Qube Research & Technologies ⁸State Street Investment Management
⁹University of Texas at Austin ¹⁰University of Florida ¹¹UNIST

Abstract

Natural language processing (NLP) has been widely used in quantitative finance, but traditional methods often struggle to capture rich narratives in textual data, leaving potentially informative signals under-explored. Large language models (LLMs) offer a promising alternative as they are able to extract nuanced semantics. In this paper, we ask whether semantic signals extracted by LLMs from corporate disclosures predict abnormal returns (alpha) beyond broad market movements and common risk factors. We introduce a simple framework, *LLM as extractor, embedding as ruler*, which extracts context-aware, metric-focused textual spans and quantifies semantic changes across consecutive disclosure periods using embedding similarity. This allows us to measure the degree of metric shifting – how much firms move away from previously emphasized metrics, referred to as moving targets. In portfolio and cross-sectional regression tests against a recent traditional baseline, our method yields a moving-targets signal that delivers significant risk-adjusted alpha. Qualitative analysis suggests that these gains stem from preserving contextual qualifiers and filtering out non-metric terms that traditional approaches often miss.

1 Introduction

Natural language processing (NLP) has been widely adopted for quantitative investment in finance due to its ability to identify complex patterns and relationships in financial text (Li et al., 2023; Lee et al., 2024; Cao et al., 2025). Traditional approaches based on named entity recognition (NER) and text classification have been shown to extract alpha from financial text (Frankel et al., 1999; Tetlock, 2007; Matsumoto et al., 2011; Loughran and McDonald, 2011, 2016). In this context, alpha refers to abnormal returns that are not explained by

broad market movements or common risk factors. However, these methods often struggle to capture rich semantics and nuanced sentiment in corporate disclosures, leaving potentially informative signals under-explored. Large language models (LLMs) have emerged as a promising alternative because they can extract subtle meaning and contextual qualifiers that traditional pipelines tend to miss (Chen et al., 2022; Lopez-Lira and Tang, 2023; Guo and Hauptmann, 2024).

In this paper, we ask a concrete question: do semantic signals extracted by LLMs from corporate disclosures predict alpha? We address this question by quantifying semantic changes across disclosures over time, focusing on how firms shift the performance metrics they emphasize across consecutive disclosure periods. To this end, we introduce a simple and general framework, *LLM as extractor, embedding as ruler*. The two components address complementary challenges. First, we use an LLM to extract metrics from each disclosure, preserving contextual qualifiers that NER-based approaches often collapse. Second, we compare extracted metrics across periods using embedding-based semantic similarity, which can identify equivalent metrics expressed differently. This similarity serves as a ruler for measuring how the underlying metrics evolve over time.

To validate our framework, we compare against the recent baseline proposed by Cohen and Nguyen (2024). That work introduces an NER-based metric called *moving targets*, which captures how firms strategically shift the performance metrics they highlight across disclosures. Such metric shifting has been shown to contain predictive information about future stock returns. Using an LLM-based analogue of moving targets derived from our framework, we conduct experiments in a realistic setting and show that our approach captures information that traditional NER-based methods fail to quantify,

*Corresponding author.

leading to improved prediction of excess returns.

Our empirical evaluation follows two standard protocols in empirical finance. First, we construct portfolios that buy shares with low metric shifting and sell shares with high metric shifting, motivated by prior evidence that greater metric shifting predicts lower future returns. We then measure portfolio performance after controlling for common market factors (Fama and French, 2015). Our method achieves a significant risk-adjusted alpha, whereas the NER-based baseline does not exhibit significant alpha in any specification. Second, we evaluate return predictability using cross-sectional regressions of next-month stock returns (Fama and MacBeth, 1973). In the regressions, the moving targets score derived from our method significantly predicts next-month returns on the broader 500-firm sample. In addition, qualitative analysis indicates that these improvements stem from capturing semantic equivalence and contextual qualifiers that keyword-based approaches often overlook. Together, these results demonstrate that LLMs provide a robust mechanism for tracking evolving signals in corporate disclosures and for predicting abnormal returns.

2 Related Work

Financial LLMs. Recently, domain-adapted pre-trained language models and LLMs have demonstrated performance gains on standard financial NLP tasks (Kaur et al., 2023; Du et al., 2024). Early work focused on further pre-training or fine-tuning BERT-based models on financial text (Araci, 2019; Liu et al., 2021). Subsequent efforts expanded to large-scale finance-specific LLMs, enabling broader applications to complex financial NLP tasks including question answering (Liu et al., 2023; Wu et al., 2023; Islam et al., 2023). A parallel line of work explores instruction-tuning approaches that evaluate multi-task and zero-shot generalization (Xie et al., 2023; Zhang et al., 2023).

Text Analysis on Financial Text. A large body of research in finance has shown that financial text contains information predictive of future stock returns (Tetlock, 2007; Loughran and McDonald, 2011, 2016). In particular, earnings calls have received attention as a primary channel through which managers communicate firm performance (Frankel et al., 1999; Matsumoto et al., 2011). Studies analyzing individual call transcripts have shown that managerial language – such

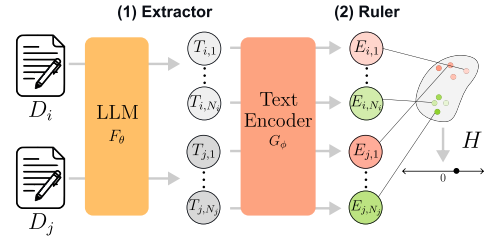


Figure 1: Overview of our framework, *LLM as extractor, embedding as ruler*. Given two documents D_i and D_j from different time periods, an LLM F_θ first acts as an *extractor*, identifying context-aware contents from each document. A pretrained text encoder G_ϕ then serves as a *ruler*, mapping extracted contents into a shared embedding space and measuring their semantic similarity to quantify how extracted contents persistence, disappearance, and emergence.

as tone or readability – can predicts excess returns (Price et al., 2012; Mayew and Venkatachalam, 2012; Fu et al., 2021). More recent work shifts focus from analyzing individual documents to examining changes in content across earnings call transcripts (Cohen and Nguyen, 2024), showing that textual modifications over time carry incremental information for return predictability.

In parallel, a growing literature explores the use of LLMs in finance, applying LLMs to news articles, annual reports, and earnings call transcript to improve return prediction and portfolio performance (Chen et al., 2022; Lopez-Lira and Tang, 2023; Guo and Hauptmann, 2024; Magner et al., 2025). However, existing studies have primarily validated LLMs on single-document tasks; their effectiveness in tracking semantic changes across documents over time remains unexplored.

3 Backgrounds

Earnings Calls and Performance Metrics. Publicly traded companies hold quarterly earnings conference calls to discuss financial results with analysts and investors (Tasker, 1998). These calls typically consist of two parts: a presentation by management summarizing the quarter’s performance, followed by a question-and-answer session with analysts (Frankel et al., 1999). During these calls, managers highlight specific metrics – quantitative measures such as revenue or product-specific figures – to frame the firm’s progress. Because these metrics are chosen by management rather than externally mandated, managers have discretion over which metrics to emphasize and how prominently to feature them (Price et al., 2012).

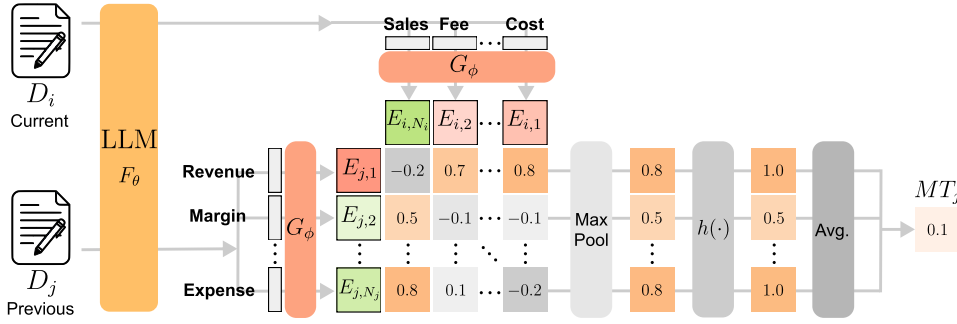


Figure 2: Illustration of the scoring pipeline using embedding similarity. Suppose that two sets of metrics have already been extracted from two transcripts D_i (current) and D_j (previous). First, we encode each metric into a shared embedding space using a pre-trained text encoder G_ϕ , obtaining embeddings E_{i,n_i} and E_{j,n_j} for quarter indices i and j . For each metric E_{j,n_j} , we then compute its maximum cosine similarity $\hat{S}_{n_j}$ to any metric at quarter i , which tests whether a semantically corresponding metric is still present. Next, we apply thresholding $h(\cdot)$ to suppress weak matches and treat only sufficiently similar pairs as retained. Finally, we aggregate these similarities using an averaging operator, quantifying the fraction of previously discussed metrics that are no longer emphasized.

Moving Targets. Cohen and Nguyen (2024) document that managers strategically shift their emphasis on metrics over time. When a previously highlighted metric becomes difficult to sustain – for example, when sales growth stalls – managers often pivot to emphasizing different metrics, such as cost savings or inventories. This phenomenon, termed *moving targets*, is formally measured as the fraction of previously discussed metrics that are absent in the current period. In their approach, metrics are extracted using spaCy (Honnibal, 2017) NER and compared across periods via rule-based matching; we refer to this as the *NER-based* method.

Following their definition, we refer to this measure as the moving targets (MT) score, computed as

$$MT_i = \frac{\sum (\text{Missing Metrics}_i \mid \text{Metrics}_{i-4})}{\sum \text{Metrics}_{i-4}} \quad (1)$$

where i indexes the fiscal quarter, Metrics_i is the set of performance metrics mentioned in quarter i , and $i - 4$ refers to the same fiscal quarter one year prior (e.g., Q1 2024 vs. Q1 2023). Missing Metrics_i denotes metrics that were mentioned in $i - 4$ but are absent in i . Higher values indicate greater departure from previously emphasized metrics.

4 Method

To test whether LLMs can better capture the moving targets, we employ the *LLM as extractor, embedding as ruler* framework. Fig. 1 shows the overall pipeline. The LLM serves as an *extractor* that identifies metrics from transcripts while preserving contextual qualifiers. The embedding model serves as a *ruler* that measures semantic distance between metrics across periods, enabling comparison even when equivalent metrics are expressed differently.

Extractor: Metric Extraction. Let D_i denote the earnings call transcript of a firm at quarter index i . We extract performance metrics using a pre-trained LLM F_θ : given a task-specific prompt and the entire transcript D_i as input, the model outputs a set of mentioned metrics. We denote the obtained set as $\mathcal{T}_i = \{T_{i,1}, \dots, T_{i,N_i}\}$, where N_i denotes the number of metrics extracted from quarter i . Unlike NER-based approaches that often extract only generic terms, the LLM can capture metrics expressed through complex phrases while preserving contextual qualifiers (e.g., “North America cloud revenue” rather than just “revenue”).

Ruler: Semantic Similarity Scoring. The original moving targets score counts the fraction of metrics from $j = i - 4$ that are absent at i via keyword matching. However, this fails when semantically equivalent metrics are expressed differently across quarters (e.g., “sales growth” vs. “revenue increase”). To address this, we use embeddings as a ruler to measure semantic distance. We illustrate the concept of our framework in Fig. 2.

We encode each extracted metric using a pre-trained text encoder G_ϕ , obtaining embeddings $E_{i,n_i} = G_\phi(T_{i,n_i}) \in \mathbb{R}^d$. For each previous metric T_{j,n_j} and its embedding E_{j,n_j} , we compute its maximum cosine similarity to any current metric:

$$\hat{S}_{n_j} = \max_{n_i} \cos(E_{i,n_i}, E_{j,n_j}) \quad (2)$$

Note that raw cosine similarity does not directly correspond to semantic equivalence: pairs below a certain threshold are effectively dissimilar, while pairs above another threshold are highly similar. To reflect this, we apply a piecewise-linear trans-

formation $h(\cdot)$:

$$S_{n_j} = h(\hat{S}_{n_j}) = \begin{cases} 0 & \text{if } \hat{S}_{n_j} \leq \alpha \\ \frac{\hat{S}_{n_j} - \alpha}{\beta - \alpha} & \text{if } \alpha < \hat{S}_{n_j} < \beta \\ 1 & \text{if } \hat{S}_{n_j} \geq \beta \end{cases} \quad (3)$$

This maps low similarities to 0 (dropped) and high similarities to 1 (retained), reducing noise from ambiguous middle-range scores.

Finally, the embedding-based MT score is defined as:

$$MT_i = 1 - \frac{1}{N_j} \sum_{n_j=1}^{N_j} S_{n_j} \quad (4)$$

Since $S_{n_j} \approx 1$ when a metric is retained and $S_{n_j} \approx 0$ when dropped, this formulation yields the fraction of metrics that are absent – matching the original definition. We refer to this approach as the *LLM-based* method. We provide a running example in Appendix A that walks through this pipeline end-to-end on an actual pair of consecutive earnings calls.

5 Experiments

In this section, we empirically evaluate whether LLM-based method of moving targets leads to better prediction of excess returns. Sec. 5.1 documents our experimental setup. Sec. 5.2 describes the results. Codes are available at [this GitHub repository](#).

5.1 Experimental Setup

Data. Our experimental setup follows [Cohen and Nguyen \(2024\)](#). The primary sample covers firms in the S&P 100 index from January 2010 to December 2024, yielding 5,615 firm-quarter observations across 64 quarters; we focus on this universe because LLM-based extraction over full earnings-call transcripts is computationally costly. To verify that our findings are not an artifact of this choice, we replicate the entire pipeline on a broader sample of the firms in the S&P 500.

Evaluation Protocols. To assess whether the moving targets scores carry economically meaningful predictive power, we adopt the evaluation framework of [Cohen and Nguyen \(2024\)](#), which consists of two complementary analyses.

First, we construct a portfolio strategy based on the moving targets score and compare portfolio returns. Following [Cohen and Nguyen \(2024\)](#), we sort firms into quintiles based on their most recent moving targets score and form equal-weighted portfolios. Q1 contains firms with the lowest scores,

Table 1: Statistics of dataset. We use earnings call transcripts from firms listed in S&P 100 index, spanning January 2010 to December 2024.

Statistic	Value
Sample period	Jan. 2010 – Dec. 2024
Number of quarters	64
Number of firms	100
Firm-quarter observations	5,615
Returns observations	16,675

Table 2: Comparison of the top 10 most frequent missing targets extracted by the *NER-based* and *LLM-based* methods.

NER-based Method		LLM-based Method	
Rank	Target	Rank	Target
1	the %	1	market share
2	a % increase	2	share repurchases
3	the range	3	revenue growth
4	% growth	4	pricing
5	q3	5	cash flow
6	q4	6	quarterly revenue growth
7	the % range	7	yearly guidance
8	year	8	free cash flow
9	the quarter	9	dividend
10	share	10	quarterly earnings per share

while Q5 contains firms with the highest scores. We then compute the return spread between Q5 and Q1. A significantly negative Q5–Q1 spread indicates that firms dropping more targets subsequently underperform. To account for other factors that may drive returns, we report risk-adjusted alphas using the Fama–French three-factor and five-factor models ([Fama and French, 1993, 2015](#)).

Second, we test whether the moving targets score significantly predicts next-month stock returns using Fama-MacBeth cross-sectional regressions ([Fama and MacBeth, 1973](#)), a standard protocol in empirical finance for testing whether a variable predicts future returns. Each month, we regress individual stock returns on the moving targets score and control variables. We then average the monthly coefficient estimates across time. A significantly negative coefficient on the moving targets score indicates that more target shifting predicts lower subsequent returns.

5.2 Results

NER and LLM extract qualitatively different metrics at the aggregate level. Tab. 2 compares the top 10 most frequent missing targets extracted by each method. The NER-based method predominantly captures surface-level patterns such as “the %”, “a % increase”, and “the range”, which lack context and do not correspond to meaningful performance metrics. In contrast, the LLM-based method extracts concrete business metrics such as “market share”, “revenue growth”, “cash flow”, and “free cash flow” – terms that align with the definition

NER-based Method	LLM-based Method
And of course, we benefited from the acquisition of Wimm-Bill-Dann, which is a critical addition to our Russia business that further strengthened our already strong position in this key developing market. Importantly, we also continue to generate strong cash flow. Management NER_TARGET operating cash flow exceeded \$6 billion. And as a result, we were able to return \$5.6 billion in cash to our shareholders through share repurchases and dividends NER_TARGET .	And of course, we benefited from the acquisition of Wimm-Bill-Dann, which is a critical addition to our Russia business that further strengthened our already strong position in this key developing market. Importantly, we also continue to generate strong cash flow LLM_TARGET . Management operating cash flow LLM_TARGET exceeded \$6 billion. And as a result, we were able to return \$5.6 billion in cash LLM_TARGET to our shareholders through share repurchases LLM_TARGET and dividends LLM_TARGET .

Figure 3: Comparison of extracted metrics between NER-based and LLM-based methods on a PepsiCo Q4 2011 earnings call transcript. The panel shows the same transcript excerpt with NER extractions (**left**) and LLM extractions (**right**). **Green** highlights indicate NER-extracted terms; **purple** highlights indicate LLM-extracted metrics. Given the same passage discussing cash flow performance, NER extracts only “Management” and “dividends”, missing the key financial metrics mentioned. In contrast, LLM correctly identifies “cash flow”, “operating cash flow”, “cash”, “share repurchases”, and “dividends”.

Table 3: Calendar-time portfolio excess returns and 3-factor / 5-factor alphas based on *Moving Targets*. (A) reports results using the NER-based extraction method, while (B) reports results using the LLM-based extraction method. *t*-statistics are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Q1	Q2	Q3	Q4	Q5	Q5-Q1
(A) <i>NER-based method</i>						
Excess Return	0.0156*** (4.38)	0.0131*** (3.85)	0.0111*** (3.19)	0.0132*** (3.99)	0.0135*** (3.94)	-0.0031 (-1.48)
3-Factor Alpha	0.0047*** (3.43)	0.0027** (2.12)	0.0017 (1.15)	0.0040*** (2.88)	0.0039** (2.47)	-0.0018 (-0.89)
5-Factor Alpha	0.0045*** (3.26)	0.0028** (2.19)	0.0016 (1.07)	0.0040*** (2.79)	0.0041** (2.54)	-0.0014 (-0.70)
(B) <i>LLM-based method</i>						
Excess Return	0.0148*** (4.75)	0.0131*** (3.84)	0.0128*** (3.60)	0.0132*** (3.65)	0.0117*** (3.50)	-0.0041** (-2.08)
3-Factor Alpha	0.0056*** (3.89)	0.0036*** (2.63)	0.0024* (1.89)	0.0029* (1.91)	0.0018 (1.22)	-0.0048** (-2.40)
5-Factor Alpha	0.0055*** (3.71)	0.0038*** (2.78)	0.0026** (2.02)	0.0032** (2.02)	0.0013 (0.87)	-0.0052** (-2.55)

of forward-looking performance indicators. This difference suggests that LLM-based extraction better leverages contextual understanding to identify targets that reflect genuine strategic emphasis.

Individual examples confirm that LLM captures metrics NER misses. Fig. 3 illustrates a representative example from an earnings call transcript. Given the same passage discussing cash flow performance, the NER-based method extracts only “Management” and “dividends”, missing the key financial metrics mentioned. The LLM-based method correctly identifies “cash flow”, “operating cash flow”, “cash”, “share repurchases”, and “dividends” – metrics that also appear among the most frequent targets in Tab. 2. This demonstrates that the LLM’s ability to capture contextually meaningful metrics at the individual level translates to more informative signals at the aggregate level.

LLM-based method yields stronger portfolio returns than the NER-based baseline. Tab. 3 presents calendar-time portfolio excess returns and alphas sorted by the moving targets score. Panel (A) shows that the NER-based method ex-

hibits a negative Q5-Q1 spread: excess return of -0.31% ($t = -1.48$), 3-factor alpha of -0.18% ($t = -0.89$), and 5-factor alpha of -0.14% ($t = -0.70$). In contrast, Panel (B) shows that the LLM-based method yields a significantly negative spread: excess return of -0.41% ($t = -2.08$), 3-factor alpha of -0.48% ($t = -2.40$), and 5-factor alpha of -0.52% ($t = -2.55$)¹, all significant at the 5% level. This indicates that firms with high target shifting reliably underperform those with low target shifting. The stronger predictability suggests that LLM-based extraction captures semantically meaningful changes that the NER-based approach misses. We note that the direct monthly difference between the two Q5-Q1 series (LLM minus NER) is not itself statistically significant ($t = -1.15$); the contribution is thus not that the LLM signal beats the NER signal head-to-head, but that the LLM signal survives standard asset-pricing tests

¹ Alphas are reported before transaction costs. Given the strategy’s one-way turnover of 26–28% per leg per month, about 80% of the gross alpha would be realized at costs realistic for S&P 100 mega-caps (10 bps per dollar traded), and roughly half at 25 bps.

Table 4: Fama–MacBeth regressions of individual firm-level stock returns on *Moving Targets* and standard return predictors. (A) reports results using the NER-based extraction method, while (B) reports results using the LLM-based extraction method. We include standard controls for known return predictors: Log(Size), Log(BM), Ret(-1,0), and Ret(-12,-1). *t*-statistics are shown in parentheses.

	(A) <i>NER-based method</i>	(B) <i>LLM-based method</i>
	Ret	
Mt Score	0.0107 (1.10)	-0.0370 (-0.95)
Log (Size)	-0.0015 (-0.42)	-0.0072 (-1.37)
Log (BM)	-0.0016 (-0.36)	-0.0037 (-0.71)
Ret (-1,0)	0.0404 (0.98)	-0.0501 (-0.81)
Ret (-12,-1)	-0.0019 (-0.22)	0.0006 (0.06)
Constant	0.0129 (0.29)	0.1200 (1.53)
R-squared	0.3897	0.3915
N	5,615	5,615

where the NER signal does not.

LLM-based method shows better predictive power in cross-sectional regressions. Tab. 4 reports Fama-MacBeth regression results testing whether the moving targets score predicts next-month stock returns. Column (A) shows that the NER-based method rather yields a positive coefficient on the moving targets score, suggesting no reliable predictive relationship. In contrast, Column (B) shows that the LLM-based method produces a negative coefficient of -0.0370 , though not statistically significant on this sample ($t = -0.95$); the primary cross-section averages only ~ 34 firms per month, limiting the power of this test. The coefficient becomes significant on the broader 500-firm sample (Tab. 5 below), consistent with the portfolio analysis.

Results generalize beyond specific firms and model choices. Tab. 5 examines whether our findings are tied to the S&P 100 universe or to a particular model configuration. The LLM-based method continues to deliver a significantly negative Q5–Q1 spread, with an excess return of -0.24% ($t = -2.53$), a 3-factor alpha of -0.36% ($t = -2.56$), and a 5-factor alpha of -0.44% ($t = -2.44$), whereas the NER-based baseline remains insignificant in every specification. The

Table 5: Extended results tested on a broader sample of 500 U.S. public firms, using a different LLM (GPT-5.4) and embedding model (text-embedding-3-small). Panel A reports the Fama–MacBeth coefficient on the MT score with full controls; Panel B reports monthly returns and alphas of the equal-weighted Q5–Q1 long-short portfolio. *t*-statistics in parentheses; $**p < 0.05$.

	<i>NER-based method</i>	<i>LLM-based method</i>
<i>Panel A: Fama–MacBeth regression</i>		
MT coefficient	-0.0108^{**} (-2.20)	-0.0326^{**} (-2.35)
<i>Panel B: Calendar-time portfolio (Q5–Q1)</i>		
Excess return	-0.0008 (-0.41)	-0.0024^{**} (-2.53)
3-factor α	-0.0012 (-0.64)	-0.0036^{**} (-2.56)
5-factor α	-0.0014 (-0.72)	-0.0044^{**} (-2.44)

Table 6: Ablation study isolating the source of the gains tested on the S&P 100 firms. Each row varies only the extractor (NER vs. LLM) and the matcher (rule-based vs. embedding), holding the embedding ruler, sample, and evaluation pipeline fixed. Cells report monthly returns and alphas of the equal-weighted Q5–Q1 long-short portfolio. *t*-statistics in parentheses; $**p < 0.05$.

Extractor	Ruler	Excess ret.	3-Factor α	5-Factor α
NER	String	-0.0031 (-1.48)	-0.0018 (-0.89)	-0.0014 (-0.70)
NER	Embedding	-0.0009 (-0.44)	-0.0022 (-1.12)	-0.0024 (-1.20)
LLM	String	-0.0022 (-1.03)	-0.0005 (-0.24)	-0.0001 (-0.04)
LLM	Embedding	-0.0041^{**} (-2.08)	-0.0048^{**} (-2.40)	-0.0052^{**} (-2.55)

Fama-MacBeth regressions show the same pattern: the LLM-based score yields a larger and significant coefficient (-0.0326 , $t = -2.35$) than the NER-based score (-0.0108 , $t = -2.20$), and the larger sample sharpens statistical significance overall. These results indicate that our findings are not an artifact of the firm universe, the choice of LLM, or the embedding model.

The gains originate from the LLM extractor.

Our method differs from the NER-based baseline along two dimensions – the extractor (LLM vs. NER) and the matcher (embedding vs. string matching). Thus, we run an ablation that varies each component in isolation, holding the sample, embedding ruler, and evaluation pipeline fixed (Tab. 6). Pairing the NER extractor with the embedding matcher moves the Q5–Q1 5-factor alpha only partway toward our method (-0.24% , $t = -1.20$) and fails to reach significance, remaining closer to the NER-with-string-matching baseline (-0.14% , $t = -0.70$) than to the LLM-based method (-0.52% , $t = -2.55$). This is consistent

with the qualitative evidence in Tab. 2, as the NER method extracts noisy, context-free spans, and a softer matching criterion cannot recover a persistent signal from targets that are not informative enough to begin with.

6 Conclusion

This study investigated whether LLMs can effectively track evolving signals in corporate disclosures by comparing an LLM-based framework against a NER-based baseline on the moving targets phenomenon. Our results show that LLM-based extraction combined with semantic similarity yields stronger return predictability. Our findings support the view that LLMs can capture meaning beyond surface-level wording, offering a more robust approach to quantifying semantic changes in financial text.

Limitations

We restrict the sample of our primary analysis to S&P 100 firms due to computational costs. Extending the analysis to a broader set of firms, particularly smaller firms with different disclosure practices, would be a valuable direction for future work. In addition, our framework depends on the quality of LLM extraction and embedding models; evaluating across different size of model choices or open-sourced models would help verify robustness. Lastly, we demonstrate predictive gains but do not investigate why LLM-based tracking better captures return-relevant information; understanding the underlying economic mechanisms is a valuable direction for future work. Our extractors are recent LLMs applied to 2010–2024 transcripts, raising a potential memorization concern. We mitigate this by design – the LLM never observes returns or compares periods. We acknowledge that memorization cannot be fully ruled out; re-extraction on anonymized transcripts is left to future work.

References

- Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
- Bokai Cao, Saizhuo Wang, Xinyi Lin, Xiaojun Wu, Hao-han Zhang, Lionel M Ni, and Jian Guo. 2025. From deep learning to llms: a survey of ai in quantitative investment. *arXiv preprint arXiv:2503.21422*.
- Yifei Chen, Bryan T Kelly, and Dacheng Xiu. 2022. Expected returns and large language models. *Available at SSRN 4416687*.
- Lauren Cohen and Quoc Nguyen. 2024. Moving targets. *Available at SSRN 4736129*.
- Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9):1–42.
- Eugene F Fama and Kenneth R French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56.
- Eugene F Fama and Kenneth R French. 2015. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.
- Eugene F Fama and James D MacBeth. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, 81(3):607–636.
- Richard Frankel, Marilyn Johnson, and Douglas J Skinner. 1999. An empirical examination of conference calls as a voluntary disclosure medium. *Journal of Accounting Research*, 37(1):133–150.
- Xi Fu, Xiaoxi Wu, and Zhifang Zhang. 2021. The information role of earnings conference call tone: Evidence from stock price crash risk. *Journal of Business Ethics*, pages 643–660.
- Tian Guo and Emmanuel Hauptmann. 2024. Fine-tuning large language models for stock return prediction using newsflow. *arXiv preprint arXiv:2407.18103*.
- Matthew Honnibal. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. (*No Title*).
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.
- Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. Refind: Relation extraction financial dataset. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 3054–3063.
- Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. 2024. A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315*.
- Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382.

- Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023. Fingpt: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485*.
- Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519.
- Alejandro Lopez-Lira and Yuehua Tang. 2023. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.
- Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.
- Tim Loughran and Bill McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of accounting research*, 54(4):1187–1230.
- Nicolás Magner, Pablo A Henríquez, and Aliro Sanhueza. 2025. Decoding risk sentiment in 10-k filings: Predictability for us stock indices. *Finance Research Letters*, page 107472.
- Dawn Matsumoto, Maarten Pronk, and Erik Roelofsen. 2011. What makes conference calls useful? the information content of managers’ presentations and analysts’ discussion sessions. *The Accounting Review*, 86(4):1383–1414.
- William J Mayew and Mohan Venkatachalam. 2012. The power of voice: Managerial affective states and future firm performance. *The Journal of Finance*, 67(1):1–43.
- S McKay Price, James S Doran, David R Peterson, and Barbara A Bliss. 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4):992–1011.
- Sarah C Tasker. 1998. Bridging the information gap: Quarterly conference calls as a medium for voluntary disclosure. *Review of Accounting Studies*, 3(1):137–167.
- Paul C Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambar, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.
- Boyue Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*.

Table 7: LLM prompt templates used in our method to extraction of targets. Note that the terms *label* and *target* are used in the prompt to refer to what we denote as *metrics* in the main text.

System

Prompt

```
<task>
From the provided inputs, extract all performance metric and target labels
and classify them into two separate lists according to the section
in which they were mentioned:
- presentation: items mentioned in the prepared presentation or opening remarks
- analyst_qa: items mentioned during analyst question and answer sessions
The term "performance metric and target labels" includes every explicitly stated
financial result, non-financial result, and key business indicator discussed or mentioned. Both GAAP and
non-GAAP labels if explicitly distinguished.
## Section classification
- presentation: items mentioned in prepared remarks/opening statements before Q&A begins.
  Typical speakers: executives with "- Executives" in the name, before the first analyst question.
- analyst_qa: items mentioned during the Q&A (analyst questions and executive answers after Q&A starts).
  Q&A typically starts after an Operator prompt like "[Operator Instructions]" or when an analyst first speaks
  (speaker label contains "- Analysts") and continues until the call ends.
- Operator-only lines do not themselves contain items, but they can mark Q&A start.
- Include only labels explicitly stated in content
- Deduplicate near-duplicates to one normalized target per section
</task>
<inputs>earnings-call transcript as indexed JSON dialog</inputs>
<output>
- target:
  - Short noun phrase only
  - **Quarter results:** prefix with "Quarterly" and do NOT include any years or quarter numbers
  - **Fiscal-year results:** prefix with "Yearly" and do NOT include any years
  - **Guidance/targets/other indicators:** no period prefixes unless inherently part of the label; avoid dates,
    years, and quarter numbers
  - No numbers, units, currency symbols, or percent signs anywhere
  - Prefer consistent terminology across products/segments
- index: integer index of the utterance containing the quote
</output>
<format>
{{
  "presentation": List[Dict],
  "analyst_qa": List[Dict]
}}
The dictionary must follow the following structure:
{{
  "target": str # Normalized short label for the metric/target according to the rules
  "index": int # Zero-based index of the utterance in the transcript where the quote appears
}}
</format>
```

A Appendix

Implementation details. For our main analysis, we use Gemini-2.5-Pro for metric extraction and text-embedding-3-large for semantic comparison. For the additional experiments on the broader 500-firm sample, we instead use GPT-5.4 for extraction and text-embedding-3-small for comparison, verifying that our findings do not depend on a specific model choice. We set $\alpha = 0.4$ and $\beta = 0.6$ for thresholding in Eq. 3: α is placed at the top of the similarity range observed for unrelated metric pairs and β at the bottom of the range for paraphrase pairs, so that $h(\cdot)$ zeroes out spurious matches and fully credits genuine ones (see Tab. 8 for illustrative pairs from each regime). The last part of this appendix reports sensitivity of the main results over the full (α, β) grid and a held-out selection protocol. Table 7 shows the full system prompt used to extract metrics in our proposed method.

Running Example We illustrate the full pipeline on the same PepsiCo call as Fig. 3. We take the LLM extractor’s actual outputs for Q4 2010 (104 metrics) and Q4 2011 (121 metrics), compute each Q4 2010 metric’s maximum cosine similarity among the Q4 2011 metrics (Eq. 2), and apply the thresholding $h(\cdot)$ of Eq. 3 with $(\alpha, \beta) = (0.4, 0.6)$. Tab. 8 shows representative rows from each regime of $h(\cdot)$.

Table 8: Running example: PepsiCo Q4 2010 metrics matched against Q4 2011 metrics. $\hat{S}_{n_j}$ is the maximum cosine similarity (defined in Eq. 2) and S_{n_j} its thresholded value (defined in Eq. 3) with $(\alpha, \beta) = (0.4, 0.6)$.

Regime	Q4 2010 metric	Best Q4 2011 match	$\hat{S}_{n_j}$	S_{n_j}
Retained via paraphrase	yearly core constant currency eps growth	core constant-currency eps growth	0.886	1.00
	yearly worldwide snacks volume	yearly snacks volume growth	0.866	1.00
	interest expense	net interest cost	0.647	1.00
Genuine drop	g series fit launch	new product sales	0.319	0.00
	sodium reduction in potato chips	global snacks results	0.364	0.00
Ambiguous middle zone	naked juice sales	liquid refreshment beverage market share	0.467	0.33
	wimm-bill-dann transaction eps impact	eps (usd basis)	0.518	0.59

Table 9: Q5-Q1 five-factor alpha (t -statistic) for each (α, β) pair; setup identical to Tab. 3. [†]Configuration used in the main text.

$\alpha \backslash \beta$	0.4	0.5	0.6	0.7	0.8
0.2	-0.0021 (-0.72)	-0.0053 (-2.64)	-0.0056 (-2.75)	-0.0062 (-3.00)	-0.0051 (-2.40)
0.3	-0.0027 (-0.91)	-0.0053 (-2.61)	-0.0053 (-2.61)	-0.0064 (-3.11)	-0.0051 (-2.42)
0.4	—	-0.0047 (-2.33)	-0.0052 [†] (-2.55)	-0.0065 (-3.16)	-0.0046 (-2.21)
0.5	—	—	-0.0047 (-2.31)	-0.0058 (-2.75)	-0.0051 (-2.53)

Threshold Sensitivity and Held-Out Selection Table 9 examines whether the main result is sensitive to the choice of thresholds. We re-estimate the Q5–Q1 five-factor alpha reported in Table 3 for every admissible (α, β) pair on the grid $\alpha \in \{0.2, 0.3, 0.4, 0.5\}$ and $\beta \in \{0.4, 0.5, 0.6, 0.7, 0.8\}$, subject to $\alpha < \beta$. The result is remarkably stable: all 17 specifications produce negative alphas, and 15 are statistically significant at the 5% level. The only two insignificant estimates occur when $\beta = 0.4$. At this low upper threshold, nearly every metric is classified as retained (the mean score exceeds 0.97), leaving little cross-sectional variation with which to distinguish firms. The baseline choice of $(\alpha, \beta) = (0.4, 0.6)$ is therefore conservative rather than result-maximizing. Raising the upper threshold to 0.7, for example, strengthens the Q5–Q1 alpha to -0.65% per month ($t = -3.16$). Thus, the main finding is not specific to a particular threshold choice but persists across a broad range of economically reasonable parameterizations.

We further address the possibility that the result reflects in-sample threshold selection. Using only portfolio returns through December 2016, we select the (α, β) pair with the most negative alpha t -statistic and then evaluate that choice exclusively over 2017–2024. This procedure selects (0.3, 0.7) and yields a held-out five-factor alpha of -0.66% per month ($t = -2.36$). Importantly, the paper’s baseline specification, (0.4, 0.6), also performs similarly over the same held-out period, producing an alpha of -0.57% per month ($t = -2.04$). These results indicate that the main finding is not an artifact of ex post threshold tuning and remains economically and statistically meaningful out of sample.