Commercial Real Estate AI Text-Evaluation (CREATE): News-Based Signals for Predicting REIT Returns

Ryan G. Chacon
University of Denver
Email: ryan.chacon@du.edu

Pratik Kothari
Oakland University
Email: pkothari@oakland.edu

Thibaut G. Morillon
Elon University
Email:tmorillon@elon.edu

Cayman Seagraves
University of Tulsa
Email: Cayman-Seagraves@utulsa.edu

Draft: November 2025

Abstract: We introduce the Commercial Real Estate AI Text-Evaluation (CREATE) score, a sector-level trading signal extracted from *Wall Street Journal* articles using large language models (LLMs). Unlike traditional natural language processing (NLP) methods based on dictionaries or supervised classifiers, CREATE leverages the contextual reasoning of LLMs to identify relevant information from unstructured text. We apply a filtering technique to mitigate forward-looking bias that afflicts LLMs (Engelberg et al., 2025). A sector-rotation strategy that goes long (short) REIT sectors with high (low) CREATE scores generates an annualized alpha of 6 percent and performs especially well during recessions. Comparing three models, we find that models with more advanced contextual reasoning yield much stronger predictive signals. The results highlight the value of LLM-based text analysis for real estate investment and asset pricing.

Acknowledgements: We sincerely thank David Ling, Stace Sirmans, Thies Lindendthal, Chongyu Wang, Ken McKinnon, Thomas Chan, Yi Zhou, the participants of the 2025 MIT Symposium on Technology, AI, and the Real Estate and the 2025 AI and Real Estate Day at Concordia University. We also acknowledge research support from the Glenn Mueller Real Estate Cycle and Sustainability Fund at the Burns School of Real Estate, University of Denver. Finally, we would like to thank our graduate assistant, Bill Ray. His expertise in LLMs greatly improved the paper.

Previous versions entitled "Artificially Intelligent Sentiment in Commercial Real Estate: A Large Language Model (LLM) Approach"

1. Introduction

The emergence of large language models (LLMs) marks a fundamental shift in how investors and researchers can extract information from unstructured text. Models such as OpenAI's GPT-5 exhibit advanced capabilities in understanding context, nuance, and complex relationships within financial documents and news articles, opening new possibilities for analyzing information embedded in large text corpora that were previously beyond the reach of traditional natural language processing methods.

In public real estate markets, news coverage provides a continuous and detailed record of the economic forces shaping property sector performance. Thousands of articles each year discuss capitalization rates, construction trends, policy shifts, and capital market conditions, as well as narrative and contextual factors not captured by standard fundamentals that directly affect real estate investment trust (REIT) valuations. While this information is theoretically available to all investors, its scale and complexity make systematic analysis difficult without advanced language modeling techniques. The broad availability of LLMs now provides a feasible way to process such information at scale.

A substantial literature documents predictable variation in REIT returns driven by fundamentals, investor expectations, anomalies, and macroeconomic conditions.[1] These studies demonstrate that changes in market perception and information flow can create time varying return patterns exploitable through trading strategies. Yet despite extensive work on valuation ratios, factor exposures, and style based anomalies, relatively little is known about whether the language

[1] See, for example, Cooper et al., 1999; Ling et al., 2000; Chui et al., 2003; Hung & Glascock, 2008; Goebel et al., 2013; Ling et al., 2019; Shen et al., 2021; Jensen & Turner, 2022; Zhu & Lizieri, 2024; Leow & Lindenthal, 2025; Chacon & Evans, 2023

used in financial news contains information that can systematically forecast sector level REIT returns.

Prior studies using text analysis in real estate have mostly relied on dictionary-based methods or early machine learning classifiers applied to news, headlines, filings, and social media. This work shows that text derived measures correlate with commercial real estate pricing and fundamentals (Hausler, 2018; Ruscheinsky et al., 2018; Beracha et al., 2019; Carstens & Freybote, 2021; Paulus et al., 2024), but it typically treats commercial real estate as a single aggregate asset class and stops short of investigating the rich sector level information contained in the text.

Our approach builds on the broader evolution of financial NLP from dictionary based tone measures (Tetlock, 2007; Tetlock et al., 2008; Loughran & McDonald, 2011), through machine learning text representations (Ke et al., 2019; Hausler, 2018), to transformer based sentiment models such as FinBERT (Araci, 2019; Huang et al., 2022), and finally to LLMs. Recent evidence shows that LLMs outperform both dictionaries and FinBERT in extracting sentiment and forecasting returns (Bond et al., 2023; Bybee, 2023; Kirtac & Germano, 2024; Kong et al., 2024; Fairhurst & Greene, 2024). In contrast to earlier tools that require task specific dictionaries or labelled training sets, LLMs can read full length articles, infer which commercial real estate sectors are affected, and evaluate the sign of the news for each sector in a single pass. We exploit these capabilities to construct a novel sector level measure for commercial real estate.

This study introduces the Commercial Real Estate AI Text Evaluation (CREATE) score, a sector level text derived measure constructed from 31,029 Wall Street Journal articles published between 1984 and 2023 that contain commercial real estate related language. Using GPT-5 with medium effort, we generate separate CREATE evaluations for nine property sectors: office, industrial, retail, residential, healthcare, lodging, data centers, self storage, and telecom.

To mitigate forward looking bias, we adopt the entity neutering procedure of Engelberg et al. (2025), which anonymizes company names, dates, numerical identifiers, and other proper nouns before model evaluation. This mitigates the concern that the LLM relies on information unavailable to contemporaneous investors. Validation analyses confirm that CREATE scores capture economically meaningful variation across sectors. Sector scores exhibit expected directional movements during the COVID-19 pandemic, declining sharply for office and lodging while rising for industrial and residential sectors. Cross sector correlations display intuitive economic structure: traditional property sectors such as office, industrial, and retail correlate strongly (0.54 to 0.80), while technology oriented sectors such as data centers and telecom correlate negatively (–0.34 to –0.48) with traditional property types, consistent with their complementary roles during structural market shifts.

We implement a monthly sector rotation trading strategy based on CREATE scores. At the start of each month, sectors are sorted by their prior month CREATE scores. The strategy takes long positions in the NAREIT index of the highest evaluated sector and short positions in the lowest evaluated sector. This produces economically significant returns. The large sectors portfolio, comprising the six most liquid property types, generates mean monthly returns of 0.49 percent with a Sharpe ratio of 0.34. The all sectors portfolio, which includes specialized sectors, yields mean monthly returns of 0.64 percent with a Sharpe ratio of 0.40. After controlling for standard REIT risk factors including market, size, value, and momentum following Letdin et al. (2025), annualized alphas remain substantial at 6.0 percent for large sectors and 9.6 percent for all sectors.

We further compare performance across three models with varying contextual reasoning capabilities: an embedding based classifier, GPT-4.1 nano, and GPT-5. Performance improves monotonically with model sophistication. The embedding model produces no alpha, GPT-4.1 nano generates an annualized alpha of 4.8 percent but exhibits more concentrated exposures and higher volatility, while GPT-5 achieves more balanced sector diversification and the highest risk adjusted returns. This monotonic relationship provides direct evidence that enhanced contextual reasoning translates into superior information extraction and return predictability.

This study contributes to several strands of literature. First, it extends research applying LLMs to financial text analysis (Bybee, 2023; Bond et al., 2023; Engelberg et al., 2025) by demonstrating the value of sector specific evaluation within a single asset class. Second, it advances the emerging literature on text-based analysis in real estate (Hausler, 2018; Ruscheinsky et al., 2018; Beracha et al., 2019; Carstens & Freybote, 2021; Paulus et al., 2024) by showing that LLM approaches can produce richer and more predictive sector level information than dictionary and supervised learning methods. Third, it contributes to the REIT return predictability and trading strategy literature (Cooper et al., 1999; Ling et al., 2000; Chui et al., 2003; Hung & Glascock, 2008; Goebel et al., 2013; Ling et al., 2019; Shen et al., 2021; Jensen & Turner, 2022; Zhu & Lizieri, 2024; Leow & Lindenthal, 2025; Chacon & Evans, 2023) by documenting that news derived information contains robust predictive power for sector level returns. Finally, the cross-model comparison offers broader insight into how advances in contextual reasoning translate into measurable improvements in financial information extraction, with implications extending beyond commercial real estate.

The remainder of the paper proceeds as follows. Section 2 reviews the development of NLP for financial text. Section 3 describes the data sources. Section 4 details the CREATE methodology. Section 5 evaluates the sector rotation trading strategy. Section 6 compares performance across models with varying contextual reasoning capabilities. Section 7 concludes.

2. Natural Language Processing (NLP) of Financial Text

Investors and researchers have leveraged the technological advances of the early 2000s to create algorithmic methods for categorizing and synthesizing unstructured financial text. These methods have been applied to company disclosures (Huang et al., 2022; Cao et al., 2023; Brockman et al., 2015), conference calls (Ewertz et al., 2025; Huang et al., 2022; Alan et al., 2023; Bushee and Huang, 2024; Rennekamp et al., 2022; Price et al., 2012), and financial press (Tetlock, 2007; Tetlock et al., 2008; Bybee et al., 2024; Kong et al., 2024; Fairhurst & Greene, 2024; Lo & Singh, 2023). The approaches used to achieve this task with accuracy and efficiency have progressed exponentially alongside technological innovation.[2]

The earliest generation of financial NLP relied on dictionary-based bag of words techniques. Tetlock (2007) develops one of the first widely used empirical measures of news sentiment by counting negative words in the Wall Street Journal using the Harvard General Inquirer dictionary and shows that daily pessimism predicts short term market reversals. Tetlock et al. (2008) extend this approach to firm specific news and demonstrate that negative language helps forecast both returns and earnings surprises. Loughran and McDonald (2011) provide a major refinement by constructing finance specific word lists for positive, negative, uncertainty, litigious, and modal language. Their dictionaries correct domain specific misclassification problems that arise when general lexicons are used in financial text. However, dictionary methods provide only a simple count of predefined positive or negative words and therefore ignore context, word meaning, and sentence structure.

[2] Lo et al., 2023 provide an excellent review the evolution of NLP for financial applications.

Machine learning methods mark the next stage of development. Instead of relying on predefined dictionaries, researchers begin transforming documents into high dimensional numeric representations using term frequency inverse document frequency scores, word embeddings, and topic models. These representations serve as inputs to predictive models that identify text-based factors or topics associated with returns. Ke et al. (2019) show that text-based features from filings and news improve forecasting performance in both time series and cross sectional tests. Hoberg and Phillips (2016) use cosine similarity across 10 K filings to estimate product market similarity and show that these text-based proximity measures explain variation in competition, mergers, and equity valuation. These methods represent a shift toward data driven identification of textual patterns, although they still treat documents largely as unordered word collections and have limited ability to capture context or linguistic nuance.

Transformer based models further advance how text is processed. BERT introduces an attention mechanism that models context and word meaning far more effectively than earlier approaches. FinBERT adapts this architecture to finance through domain specific pre training and fine tuning for sentiment classification (Araci, 2019; Huang et al., 2022). FinBERT produces sentence level sentiment probabilities and captures negation, conditional phrasing, and subtle contextual signals that dictionary and bag of words methods systematically miss. As a result, FinBERT became widely adopted in studies of conference call tone, managerial communication, and financial news. However, FinBERT is only a classifier trained for a single task and it is not a general reasoning model. It cannot summarize, explain, extract entities, detect forward-looking statements, interpret numerical information, or answer questions about the text.[3]

[3] Although FinBERT can be incorporated into pipelines for tasks such as sector level analysis, these capabilities depend on external rules and models rather than FinBERT's own training.

General purpose LLMs represent the most recent and most advanced stage of NLP. LLMs go much further because they are trained on extremely large and diverse corpora. This allows the model to learn broad world knowledge, economic reasoning patterns, and deep semantic relationships between ideas. The model can therefore perform classification, summarization, explanation, and reasoning in a zero-shot setting. Zero shot refers to the ability of a model to interpret financial text without any task-specific labelled data or fine tuning (Bybee, 2023). Instead, the model follows natural language instructions and applies contextual reasoning to extract sentiment or other economic signals.

Recent evidence shows that these models outperform both dictionary-based and transformer-based approaches. Bond et al. (2023) find that GPT-based sentiment from Reuters end of day market summaries predicts S&P 500 reversals more accurately than Loughran and McDonald (2011) sentiment and more accurately than FinBERT. Kirtac and Germano (2024), using nearly one million financial news articles, show that GPT family models generate substantially higher predictive accuracy and significantly more profitable trading strategies than FinBERT or other transformer based baselines. Bybee et al. (2023) use GPT 3.5 to infer expectations directly from Wall Street Journal articles and show that the resulting time series closely track professional survey expectations and exhibit meaningful forecasting properties. Kong et al. (2024) and Fairhurst and Greene (2024) demonstrate that large language models capture multidimensional market narratives, including uncertainty, risk, and macroeconomic interpretation, at a level that earlier methods cannot match.

Although the real estate literature has begun to incorporate textual analysis, it remains largely situated in earlier generations of NLP. Existing studies rely primarily on dictionary-based sentiment measures applied to news or filings, as in Beracha et al. (2019), Ruscheinsky et al.

(2018), and Carstens and Freybote (2021), or on early machine learning classifiers such as support vector machines, as in Hausler (2018). More recent work, such as Paulus et al. (2024), compares dictionaries with more sophisticated classifiers in the context of REIT related social-media sentiment and documents the incremental benefits of moving beyond lexicons. Our approach advances this literature by using an LLM to generate sector specific sentiment scores directly from full length news articles. The model not only classifies tone but also identifies the relevant commercial real estate sector and provides a justification for its assessment, producing a richer and more economically meaningful representation of real estate news than earlier NLP methods allow.

3. Data Sources

We utilize three primary data sources. First, we obtain Wall Street Journal articles from ProQuest One Businesscovering 1984-2023. Second, we collect sector-level REIT returns from NAREIT for nine property sectors spanning 1994-2023: Office, Industrial, Retail, Residential, Healthcare, Lodging, Data Centers, Self-Storage, and Telecom. The returns for all sectors begin in 1994 except for Telecom and Data Centers which begin in January 2012 and December 2015, respectively. Third, we employ REIT factor returns from Letdin et al. (2025). These factors represent the equivalent of Fama and French (1993) and Carhart (1997) factors carefully applied to the REIT industry. This requires the CRSP Ziman REIT index constituents, CRSP for market data, and COMPUSTAT for accounting data. In our base specification, we use the Market, Size, Value, and Momentum factors and our results hold after including Quality and Low Volatility.[4]

[4] These factors are constructed following the methodology detailed at https://www.reitfactors.ai/methodology and in Letdin et al. (2025), which provides comprehensive documentation of data sources and construction procedures.

4. CREATE scores

4.1 Preventing Look-Ahead Bias with Entity Neutering

A key challenge in using LLMs for forecasting is forward-looking bias. Because models are trained on text that may postdate the study period, they could infer patterns from future events or identifiable entities unavailable to real-time investors. To mitigate this risk, we adopt the entity-neutering procedure of Engelberg et al. (2025), which anonymizes proper nouns, dates, industries, product types, and numerical identifiers within each article. Following the protocol in Table A2, GPT-4o-mini transforms every article into a redacted version where firm names and temporal markers are replaced with generic placeholders (e.g., *Company_1*, *time_x*, *number_a*), forcing the model to evaluate language in abstract, economic terms rather than through specific historical cues.

[Insert Figure 1 Here]

Figure 1 illustrates this transformation: the original *Wall Street Journal* article on the left includes company names, monetary figures, and event dates, while the neutered version on the right substitutes them with generalized tags. This process preserves linguistic structure but removes contextual identifiers that might reveal outcomes known only with hindsight.

By masking these details, the neutering process mitigates the potential for the model to exploit future information while preserving its ability to interpret economic language and tone. Although no approach can fully eliminate forward-looking bias, the procedure meaningfully limits its influence and aligns model evaluation with the contemporaneous information available to investors. The resulting CREATE scores thus provide a more realistic, time-consistent signal of commercial real estate conditions.[5]

4.2 Generating Sector Level CREATE Scores

We begin with 31,029 WSJ articles spanning 1984-2023 containing CRE-relevant keywords. These articles are sourced from ProQuest One Business using a keyword search.[6] An initial review revealed many articles were unrelated to CRE pricing, such as pieces on foreign luxury residential homes.

To enhance relevance and cost efficiency, we screen articles using GPT-5-mini, prompting the model to adopt the perspective of a REIT portfolio manager for each sector and identify sector-specific relevance. This process assigns a binary score (0 or 1) for each sector-article pair. Of the 31,029 articles, 8,183 received zeros across all sectors and were excluded as irrelevant (26.4% of articles). The complete prompt appears in Appendix Table A4.

Once we have the filtered set of WSJ articles, we next use the prompt listed in Table A1 to elicit article-level sector CREATE scores from GPT-5.0, focused on the expected impact of each article on U.S. commercial real estate prices.[7] The prompt returns two outputs for each property sector:

- increase_decrease ∈ {“increase”, “decrease”, “uncertain”}, which we map to a CREATE score of +1, –1, or 0,

[5] If forward-looking bias were a dominant concern, one would expect the CREATE scores to exhibit near-perfect predictive accuracy. In practice, the predictive power of the scores is modest and statistically plausible, suggesting that residual look-ahead effects are limited rather than driving the observed results.

[6] Keywords include "Commercial Real Estate," "Cap Rates," "CMBS," "REITs," and "NOI." and so on.

[7] In later analysis, we use a similar approach to generate CREATE scores from GPT-4.1-nano and a simple embedding model to compare model contextual reasoning and performance.

- explanation, a concise justification from the model.

This format follows the method developed in Bybee (2023); however, our prompt is unique in its direct alignment with CRE valuation and sector level evaluation rather than general macroeconomic or equity expectations. The nine property sectors we obtain scores for are Office, Industrial, Retail, Residential, Healthcare, Lodging, Telecom, Data Centers, and Self-Storage. Given some sectors are much larger than others and we are focused on a trading strategy, we conduct all the analysis using *Large Sectors* (Office, Industrial, Retail, Residential, Healthcare, Lodging) and *All Sectors* (Large Sectors + Data Centers, Self-Storage, and Telecom).[8]

Figure 2 illustrates LLM-generated CREATE scores and justifications for a March 31, 2021 article discussing office subleases flooding the market post-COVID. The article explicitly states that remote work is "hurting demand" in office markets and later elaborates on work-from-home trends. This example demonstrates two key capabilities of our approach.

[Insert Figure 2 Here]

First, it highlights the importance of sector-level evaluation. The LLM assigns CREATE scores of -1 for Office, Retail, and Lodging; 0 for Industrial and Healthcare; and +1 for Residential, Telecom, Data Centers, and Self Storage. The justifications align with economic intuition and capture the differential sectoral impacts. For Office, the model notes "Surging sublease supply and discounts signal structurally lower demand and falling rents from persistent remote work." For Residential, it observes "Remote work raises demand for living space and home offices, modestly supporting housing values."

[8] While some of the largest individual REITs are Telecom stocks, the sector itself is modest in size and relatively new.

Second, the justifications demonstrate contextual reasoning capabilities exceeding traditional NLP methods. Near the article's end, the author compares the current office environment to e-commerce's impact on retail in the late 2000s, stating "demand for office space could be permanently lower at some companies, much like the rise of e-commerce has been driving down demand and rents for street-level retail." However, the LLM's justification for the -1 Retail score references contemporary mechanisms: "Fewer office workers downtown reduces weekday foot traffic, pressuring street-level retail rents." The model correctly identifies the e-commerce reference as analogical rather than contemporaneous, providing compelling evidence of sophisticated contextual understanding.

4.3 CREATE Score Summary Statistics

Table 1 presents descriptive statistics for GPT-5 CREATE scores at both the article and monthly aggregation levels. Panel A reports article-level statistics across 17,754 articles for most sectors. Data Centers and Telecom show fewer articles (3,930 and 6,663, respectively) because NAREIT sector return data for these sectors begins later in our sample period.

[Insert Table 1 Here]

At the article level, mean CREATE scores reveal modest sectoral variation. Data Centers exhibits the highest average score (0.26), followed by Self Storage (0.13) and Residential (0.09), suggesting these sectors received relatively favorable coverage during our sample period. Office shows the most negative average score (-0.06), consistent with concerns about remote work and structural demand shifts. Industrial (0.05), Healthcare (0.02), Lodging (0.00), and Retail (-0.01) cluster near neutral. Standard deviations range from 0.40 (Telecom) to 0.72 (Retail), indicating substantial article-to-article variation in scores.

Panel B presents monthly aggregated scores, averaging article-level CREATE scores within each sector-month. The sample includes 360 months for most sectors and 97-144 months for Data Centers and Telecom, reflecting their later NAREIT index inception dates. Monthly mean scores largely preserve the cross-sectoral patterns from Panel A. Office maintains the most negative average (-0.07), while Data Centers (0.23) and Self Storage (0.13) remain most positive. The monthly distributions show tighter ranges, with Office spanning -0.68 to 0.33 and Data Centers ranging from -0.14 to 0.63.

4.4 CREATE Score Validation

A critical innovation in our analysis is the importance of disaggregating news information to the sector level. To test whether our method captures cross-sector variation, we calculate Pearson correlation coefficients for each sector at the monthly level. The results appear in Table 2.

[Insert Table 2 Here]

The correlation matrix reveals substantial heterogeneity in CREATE score co-movement across sectors. Traditional property types (Office, Industrial, Retail) exhibit strong positive correlations, ranging from 0.54 to 0.80. This pattern is intuitive: macroeconomic news affecting business activity typically impacts these sectors similarly. Office and Retail show the highest correlation (0.80), likely reflecting their shared sensitivity to employment trends and urban economic activity.

In contrast, newer or specialized sectors display markedly different correlation patterns. Data Centers, Self-Storage, and Telecom show negative correlations with Office (-0.34, -0.04, and -0.29, respectively) and Retail (-0.48, -0.11, and -0.42, respectively). These negative correlations suggest that news benefiting traditional property types often has offsetting implications for technology-enabled sectors. For example, articles discussing remote work trends may signal declining office demand while simultaneously indicating rising data infrastructure needs.

Healthcare exhibits low or negative correlations with most sectors, ranging from -0.22 (Lodging) to 0.33 (Industrial). This weak co-movement reflects Healthcare's defensive characteristics and its sensitivity to healthcare policy rather than general business cycle news. The strongest correlations emerge among technology-oriented sectors. Data Centers and Telecom correlate at 0.89, suggesting that news about digital infrastructure, cloud computing, and connectivity affects both sectors similarly.

Overall, these patterns provide compelling evidence that sector-level disaggregation captures economically meaningful variation. The correlations align with intuitive economic linkages while remaining sufficiently distinct to justify separate CREATE measures for each sector.

To further validate our measure, we examine CREATE scores during the COVID-19 pandemic to test whether sectors exhibit expected patterns around this period.[9] Figure 3 displays scores across four property types (Office, Lodging, Residential, Industrial) from 2019-2023. Panel A shows 3-month rolling averages over time, while Panel B presents average scores by COVID period. We define four periods: Pre-COVID (January 2019-February 2020), COVID Peak (March 2020-May 2021), Recovery (June 2021-May 2022), and Post-COVID (June 2022-December 2023). Positive values indicate favorable CREATE scores and negative values indicate unfavorable

[9] In Appendix A5 – we also prompt the LLM to provide an "overall" score that is not sector specific and graph that over time. The overall CREATE scores accurate reflect broader economic cycles.

CREATE scores. Shaded regions in Panel A highlight the COVID Peak (red) and Recovery (orange) periods.

[Insert Figure 3 Here]

The results align strongly with known sectoral impacts during the pandemic. Panel A reveals the temporal dynamics of score shifts. Lodging and Office scores plummeted in early 2020, with Office reaching a trough below -0.60 in mid-2020. Lodging recovered more quickly, approaching neutral scores by late 2021, while Office scores remained persistently negative through 2023. This divergence reflects concerns about structural changes in office demand from remote work. Residential and Industrial scores remained elevated through the Recovery period before moderating in the Post-COVID period, capturing the normalization of pandemic-driven demand surges.

Panel B shows Lodging experienced the most severe negative CREATE scores during COVID Peak (average score of -0.30), followed by Office (-0.20). These sectors faced direct demand shocks from travel restrictions and remote work mandates. In contrast, Residential and Industrial maintained positive CREATE scores during COVID Peak (0.05 and 0.23, respectively), consistent with housing market strength and elevated e-commerce activity driving warehouse demand. The sector-specific patterns during this well-documented event provide strong validation that CREATE scores capture meaningful economic variation across property types.

5. Sector Rotation Analysis

5.1 Strategy Construction and Sector Selection Patterns

We construct a simple long-short portfolio strategy based on monthly CREATE scores. At the beginning of each month, we sort sectors by their CREATE scores from the previous month and take a long position in the sector with the highest score and a short position in the sector with the lowest score. When ties occur, we include all tied sectors in the respective portfolio with equal weight.[10] This approach tests whether CREATE scores contain predictive information about future sector returns. We implement this strategy using two sector universes: "Large Sectors" (Office, Industrial, Retail, Residential, Healthcare, and Lodging) and "All Sectors" (includes Data Centers, Telecom, and Self-Storage).

Table 3 presents the frequency with which each sector appears in the long and short portfolios over the full sample period. Because ties result in multiple sectors being selected in some months, the total percentages exceed 100%. The selection patterns reveal economically intuitive relationships between news coverage and sector characteristics. For the Large Sectors portfolio (Panel A), Residential appears most frequently in the long portfolio (37.5% of months), followed by Industrial (30.3%). Office dominates the short portfolio (46.4%), consistent with persistent negative CREATE scores documented in Table 1. Retail appears in both long and short positions (12.8% and 18.3%, respectively), reflecting its cyclical sensitivity to economic conditions. Healthcare shows relatively balanced representation (15.6% long, 25.3% short), while Lodging appears similarly in both portfolios (13.6% long, 15.0% short).

[Insert Table 3]

The All Sectors portfolio (Panel B) shows more dispersed selection patterns. Self-Storage appears most frequently in the long portfolio (35.8%), followed by Residential (24.7%) and Data Centers (15.3%). Notably, both Self-Storage and Data Centers rarely appear in short positions (0.6% and 0.3%, respectively), aligning with their consistently positive average CREATE scores

[10] For example, if 2 sectors tie for the highest score, the long portfolio has 50% in each. If 3 tie, the long portfolio has 33% in each, and so on.

in Table 1. Office again appears most frequently in the short portfolio (46.4%), identical to its frequency in the Large Sectors universe. Lodging shows balanced representation across both portfolios (10.6% long, 15.0% short), capturing its volatile exposure to travel and economic cycles. Healthcare appears predominantly in short positions (23.3%) compared to long positions (1.9%), despite its defensive characteristics. Telecom shows limited presence in long positions (1.4%) but appears more frequently in short positions (5.0%). Importantly, across both panels, no sector is in the long or short portfolio greater than 50% of the time. This sector selection diversity reduces reliance on any given sector to drive results.

Figure 4 illustrates the temporal evolution of sector selection for both the Large Sectors (Panel A) and All Sectors (Panel B) portfolios, revealing how portfolio allocations respond to changing economic conditions. Each year, we take the percentage of months a given sector is in the long (short) portfolio. The stronger coloring in the heat map suggests higher percentages for a sector and blank white squares indicate the sector was not selected in the long (short) portfolio that year. Several patterns emerge. First, sector selection exhibits meaningful time variation rather than persistent concentration in specific sectors. This dynamic rebalancing suggests CREATE scores capture evolving economic narratives rather than static sector characteristics.

[Insert Figure 4]

Second, certain periods show clustering of specific sectors in long or short positions. During the 2008 financial crisis, Retail and Office predominantly occupied short positions in both portfolios while Healthcare dominated the long position in Panel A. The All Sectors portfolio (Panel B) shows Self-Storage and Data Centers emerging as frequent long positions during the post-crisis recovery period, reflecting positive news coverage about these sectors' defensive characteristics and growth potential.

Third, the COVID-19 period shows dramatic shifts across both portfolios. Office persistently occupies short positions from 2020 through 2023 in both panels with some rotation into retail and lodging, reflecting concerns about remote work and travel restrictions. Industrial and Residential occupied long positions during the pandemic peak in Panel A. Panel B reveals additional nuance: Self-Storage has some presence post-COVID but it is largely concentrated in Data Centers, capturing the surge in digital infrastructure demand. These patterns demonstrate that CREATE scores identify time-varying sector selection consistent with economic intuition.

5.2 Strategy Performance

We evaluate the performance of the CREATE-based sector rotation strategy using both raw returns and risk metrics. Figure 5 displays the cumulative growth of $1 invested in the long portfolio, short portfolio, and the NAREIT index from 1994-2023. Panel A (Large Sectors) shows the long portfolio growing to approximately $35 by 2023, substantially outperforming both the short portfolio (approximately $4) and the NAREIT index (approximately $12). Panel B (All Sectors) exhibits even stronger performance, with the long portfolio reaching approximately $70 by 2023, while the short portfolio grows to approximately $5 and the NAREIT index reaches approximately $12.

[Insert Figure 5 Here]

Both panels reveal several notable features. First, the long portfolios consistently outperform the NAREIT index across nearly the entire sample period, suggesting CREATE scores successfully identify sectors with favorable future returns. Second, the short portfolios exhibit persistent underperformance relative to the market, indicating CREATE scores also identify sectors with unfavorable prospects. Third, the performance gap between long and short portfolios widens substantially after the financial crisis. Both panels display outperformance during the COVID pandemic, suggesting CREATE scores perform well during extreme market distress. The divergence is particularly pronounced in Panel B, where specialized sectors like Data Centers and Self-Storage contribute additional returns beyond what is achievable using only the more liquid, traditional property types in Panel A.

Figure 6 examines the consistency of strategy performance through win rate analysis for the Large Sectors portfolio. The All Sectors analysis is suppressed for brevity. We define a "win" as any month where the long-short spread (long portfolio return minus short portfolio return) exceeds zero. The top left panel shows win rates by decade, ranging from 43.8% in the 1990s to 59.3% in the 2000s. A value less than 50% indicates an unsuccessful strategy. With the exception of the first 5 years of the trading strategy start date (1994), win rates are consistently high and profitable. The top right panel disaggregates performance across five metrics: long versus NAREIT benchmark (56.3% win rate), short versus NAREIT benchmark (56.7%), overall long-short spread (59.6%), long-short spread during recessions (88.9%), and long-short spread during expansions (57.5%). For the short versus NAREIT, a win is identified when the short leg return is less than NAREIT. Therefore, the short leg has returns lower than NAREIT 56.7% of the time. The dashed horizontal line at 50% represents random selection performance.

[Insert Figure 6 here]

The win rate analysis reveals that strategy performance exceeds random selection across most metrics, with particularly strong performance during recessions (88.9% win rate albert with only 18 months of observations). This suggests CREATE scores may be especially valuable during periods of economic stress when sectoral divergence is most pronounced. The bottom panel presents a 24-month rolling win rate over time, with gray vertical bands indicating NBER-dated recessions. For each date, the win rate is calculated using the previous 2 years of data only. The rolling win rate fluctuates between approximately 0.3 and 0.9, showing substantial time variation in strategy effectiveness. Notable peaks occur during the early 2000s recession, the 2008 financial crisis, and the COVID-19 pandemic, consistent with the hypothesis that news-based signals become more informative during periods of heightened uncertainty. Win rates only dip below the 50% threshold at the very beginning of the sample, very briefly in 2003, and between 2016 and 2018.

We next move to more traditional portfolio statistics presented in Table 4. The Large Sectors long-short strategy generates a mean monthly return of 0.49% (5.9% annualized) with a standard deviation of 5.10%, resulting in an annualized Sharpe ratio of 0.34. The All Sectors long-short strategy generates higher returns of 0.64% monthly (7.7% annualized) with slightly higher volatility (5.55%), producing an annualized Sharpe ratio of 0.40. The Sortino ratio, which measures return relative to downside volatility only, follows a similar pattern (0.51 for All Sectors versus 0.42 for Large Sectors), indicating that All Sectors exhibits better risk-adjusted performance using both total and downside risk metrics.

[Insert Table 4]

Tail risk measures reveal substantial downside exposure in both strategies. Value at Risk at the 5% confidence level (VaR 5%) indicates the maximum expected monthly loss in the worst 5% of months is 6.59% for Large Sectors and 6.67% for All Sectors. Maximum drawdowns are severe, reaching 51.98% for Large Sectors and 52.91% for All Sectors, though these peak-to-trough declines are smaller than many individual REIT sectors experienced during the financial crisis. Return distributions show positive skewness (7.04% for Large Sectors, 16.62% for All Sectors), indicating occasional large positive returns. Downside volatility (4.13% for Large Sectors, 4.38%

for All Sectors) is lower than total volatility, consistent with positive skewness in the return distribution.

The performance metrics demonstrate that CREATE-based sector rotation generates economically significant returns with moderate risk-adjusted performance. The All Sectors strategy outperforms the Large Sectors baseline across most metrics, suggesting that the inclusion of specialized sectors (Data Centers, Self-Storage, Telecom) enhances portfolio performance. However, the Large Sectors strategy provides a more conservative and implementable baseline given the greater liquidity and trading volume in traditional property types.

5.3 Risk-Adjusted Performance

To assess whether the CREATE-based sector rotation strategy generates abnormal returns beyond compensation for systematic risk exposure, we estimate a four-factor model following the REIT factor framework of Letdin et al. (2025). We regress monthly long-short portfolio returns on four factors: Market (the excess return on the REIT market portfolio), Size (the return differential between small-cap and large-cap REITs), Value (the return differential between high book-to-market and low book-to-market REITs), and Momentum (the return differential between past winners and past losers). We estimate separate regressions for the long portfolio, short portfolio, and long-short portfolio for both Large Sectors and All Sectors strategies. Standard errors are calculated using the Newey-West (1987) procedure with a lag length of one to account for potential heteroskedasticity and serial correlation in monthly returns.

Table 5 presents the factor model estimates. Panel A shows results for the Large Sectors portfolio. The long portfolio generates a statistically significant alpha of 0.002 per month (t-statistic = 1.68), equivalent to 0.2% monthly or 2.4% annualized. As expected in a long only strategy, the long portfolio exhibits strong positive exposure to the Market factor (β = 1.015, t = 22.03) and significant negative exposure to Value (β = -0.171, t = -3.80). The negative Value loading indicates the strategy tends to select growth-oriented sectors (low book-to-market) rather than value sectors. Size and Momentum exposures are statistically insignificant. The model explains 77.8% of return variation (R-squared = 0.7775).

[Insert Table 5]

The short portfolio in Panel A shows a significant negative alpha of -0.003 per month (t = -1.98), equivalent to -0.3% monthly or -3.6% annualized, indicating the short positions underperform even after controlling for factor exposures. Like the long portfolio, the short portfolio has strong Market exposure (β = 1.082, t = 26.73) and positive Value exposure (β = 0.382, t = 7.69). The positive Value loading suggests the strategy shorts value-oriented sectors, consistent with the long portfolio's growth tilt.

The long-short portfolio (Column 3) combines these effects, generating an alpha of 0.005 per month (t = 2.51), equivalent to 0.5% monthly or 6.0% annualized. This alpha is statistically significant at the 5% level, indicating the strategy produces abnormal returns beyond what can be explained by systematic factor exposures. The long-short portfolio exhibits minimal Market exposure (β = -0.066, t = -0.90), confirming the strategy is approximately market-neutral. The Value factor loading is strongly negative (β = -0.553, t = -7.50), reflecting the divergent value exposures in the long and short portfolios.

Panel B presents results for the All Sectors portfolio. The long portfolio generates a larger and more significant alpha of 0.006 per month (t = 3.40), equivalent to 0.6% monthly or 7.2% annualized. This represents a substantial improvement over the Large Sectors alpha, suggesting the inclusion of specialized sectors enhances risk-adjusted performance. The factor exposures are similar to Panel A, with strong Market exposure (β = 0.900, t = 20.53) and negative Value exposure (β = -0.206, t = -2.91). The model explains 70.1% of return variation.

The short portfolio in Panel B shows a statistically insignificant alpha of -0.002 (t = -1.44), equivalent to -0.2% monthly or -2.4% annualized, indicating neutral risk-adjusted performance after controlling for factors. The short portfolio maintains strong Market exposure (β = 1.056, t = 25.33) and positive Value exposure (β = 0.388, t = 7.82).

The All Sectors long-short portfolio (Column 3) generates an alpha of 0.008 per month (t = 3.28), equivalent to 0.8% monthly or 9.6% annualized. This alpha is statistically significant at the 1% level and represents more than 50% higher annualized alpha than the Large Sectors strategy (9.6% versus 6.0%). The long-short portfolio is slightly negatively exposed to the Market (β = -0.156, t = -2.28) and strongly negatively exposed to Value (β = -0.594, t = -6.58). The model explains 38.2% of long-short return variation.

Several patterns emerge from the factor analysis. First, both strategies generate significant positive alphas, indicating CREATE scores contain information not captured by standard REIT risk factors. Second, the All Sectors strategy produces substantially higher risk-adjusted returns than the Large Sectors baseline, suggesting specialized sectors contribute meaningfully to performance. Third, both strategies exhibit strong negative Value factor exposure in the long-short portfolio, indicating a systematic tilt toward growth-oriented sectors. This growth bias may reflect that positive news coverage disproportionately benefits sectors with stronger growth prospects.

The economic magnitude of the alphas is substantial. The All Sectors long-short alpha of 9.6% annualized represents a sizable abnormal return for a relatively simple sector rotation strategy based solely on news coverage. These results provide strong evidence that CREATE scores capture pricing-relevant information that translates into exploitable return predictability across REIT sectors.

6. Evaluation of Models Across Varying Levels of Contextual Reasoning

6.1 Alternative Model Selection

Our baseline analysis employs GPT-5 to generate CREATE scores through explicit reasoning and justification for each sector-article pair. To assess whether GPT-5's advanced contextual reasoning capabilities translate into superior predictive performance, we compare three model architectures: GPT-5, GPT-4.1-nano, and an embedding-based approach.

Embedding models represent text as high-dimensional numerical vectors that capture semantic relationships. Unlike generative language models that produce text outputs, embedding models compress documents into fixed-length representations where semantically similar texts occupy nearby positions in vector space. These embeddings can then be used as inputs to traditional machine learning classifiers. We implement a sector-level embedding approach using OpenAI's text-embedding-3-small model combined with sector-specific multinomial logistic regression classifiers. Each classifier is trained on synthetic labeled data generated by prompting an LLM to create positive, negative, and neutral examples for each sector-topic combination. The embedding model transforms article text into vectors, and the trained classifiers produce CREATE scores for each sector. Complete methodological details are provided in Appendix A3 for brevity.

GPT-4.1-nano represents an intermediate capability level between embeddings and GPT-5. Like GPT-5, GPT-4.1-nano generates explicit CREATE scores and justifications for each article-sector pair using the same prompting structure described in Section 3. However, GPT-4.1-nano has less advanced reasoning capabilities than GPT-5, providing a natural benchmark to assess whether incremental improvements in language model sophistication translate into measurably better CREATE score generation and return prediction.

Comparing these three approaches allows us to test a central hypothesis: do models with more advanced contextual reasoning yield substantially stronger predictive signals? If the embedding approach performs comparably to GPT-5, this suggests that sophisticated reasoning is unnecessary and that simpler, more cost-effective methods suffice for extracting predictive information from commercial real estate news. Conversely, if GPT-5 substantially outperforms both alternatives, this would provide evidence that advanced contextual understanding is critical for capturing the nuanced, sector-specific implications embedded in financial news coverage.

6.2 Model Performance Comparisons

We focus our model comparison on the Large Sectors portfolio for brevity, as this represents the more liquid and tradeable baseline strategy. Table 6 examines sector selection patterns across the three models, revealing substantial differences in how each approach identifies attractive and unattractive sectors. The frequency distributions show the percentage of months each sector appears in long or short positions from 1994-2023.

[Insert Table 6]

GPT-5 sector selection is identical to that displayed in Table 3 but is added for comparison. GPT-4.1-nano shows substantially different selection patterns, with Lodging appearing in long positions 0.28% of the time and short positions 72.00% of the time. This extreme concentration on a single sector raises concerns about portfolio stability and exposure to idiosyncratic Lodging-specific risks. Healthcare also shows high frequency in long positions (47.50%), while Office appears predominantly in short positions (20.00%). The concentrated Lodging exposure suggests GPT-4.1-nano may lack the contextual sophistication to differentiate sector-specific implications across the full range of commercial real estate news.

The embedding approach produces quite uninformative sector selection patterns. Contrary to the other models, Office dominates long positions (80.00%), while Retail is in the short position 100.00% of the time. This extreme concentration, particularly the invariant Retail short position, indicates the embedding model lacks the dynamic contextual understanding necessary to adapt sector selection to changing economic conditions. The static classifier approach appears to identify persistent cross-sectional patterns rather than time-varying sector attractiveness.

Figure 7 illustrates the performance implications of these selection differences. The top panel displays cumulative growth of $1 invested in each model's long-short strategy alongside the NAREIT index. This figure differs from the growth of $1 presented in Figure 5 because it is the long-short strategy growth instead of disaggregated long and short portfolios. GPT-5 (orange line) shows outperformance, growing to approximately $5 by 2023. GPT-4.1-nano (burgundy line) exhibits more volatile performance, reaching approximately $1.50 with substantial run ups and drawdowns during 2008-2009 and again in 2020-2021. The embedding approach (blue line) shows persistent underperformance, declining to approximately $0.50 by 2023.

[Insert Figure 7]

The bottom panel of Figure 7 presents win rate comparisons across five performance metrics. GPT-5 consistently achieves the highest win rates: 56.3% for long versus NAREIT, 56.7% for short versus NAREIT, 59.6% for overall long-short spread, 88.9% during recessions, and 57.5% during expansions. All metrics substantially exceed the 50% random selection benchmark (dashed line). GPT-4.1-nano shows more modest win rates, ranging from 50.0% to 74.4%, with particularly strong performance during recessions (74.4%) but weaker performance in other metrics. The embedding approach performs at or below random selection across all metrics besides recession periods.

Table 7 provides comprehensive performance statistics and risk-adjusted alphas. Panel A shows that GPT-5 generates mean monthly returns of 0.49% with a standard deviation of 5.10%, producing a Sharpe ratio of 0.34 and Sortino ratio of 0.41. Maximum drawdown is 51.98%, and the return distribution exhibits modest positive skewness (0.07). GPT-4.1-nano produces lower returns (0.29% monthly) with substantially higher volatility (6.45%), resulting in inferior risk-adjusted performance (Sharpe ratio of 0.16, Sortino ratio of 0.19). Critically, GPT-4.1-nano exhibits extreme negative skewness (-1.02), indicating a pronounced left tail with occasional large losses. The maximum drawdown of 71.06% is substantially larger than GPT-5, consistent with the concentrated exposure and lack of dynamic sector diversification.

[Insert Table 7]

The embedding approach produces negative mean returns (-0.19% monthly) with moderate volatility (4.04%), resulting in negative Sharpe (-0.167) and Sortino (-0.227) ratios. Maximum drawdown reaches 81.58%, the worst among all three models. The negative skewness (-0.25) and poor overall performance confirm that static embedding-based classification lacks the contextual sophistication necessary for effective sector rotation.

Panel B presents risk-adjusted alphas from the four-factor model. GPT-5 generates a long-short alpha of 0.005 per month (t = 2.51), equivalent to 0.5% monthly or 6.0% annualized. This alpha is statistically significant at the 5% level. GPT-4.1-nano produces a smaller alpha of 0.004 (t = 1.87), equivalent to 0.48% monthly or 5.76% annualized, which is marginally significant at the 10% level. The embedding approach generates a negative alpha of -0.002 (t = -1.45), indicating no ability to produce risk-adjusted returns.

The component portfolio alphas reveal additional insights. For GPT-5, both long (0.002, t = 1.68) and short (-0.003, t = -1.98) portfolios contribute to long-short performance. For GPT-4.1-nano, the long portfolio generates a significant alpha of 0.003 (t = 2.75), but the short portfolio shows no significant abnormal performance (-0.002, t = -0.89). This asymmetry, combined with the concentrated Lodging exposure, suggests GPT-4.1-nano identifies favorable sectors more effectively than unfavorable ones. The embedding model shows no significant alphas in any portfolio.

These results provide compelling evidence that advanced contextual reasoning capabilities matter substantially for CREATE score generation and return prediction, with performance increasing monotonically across model sophistication. GPT-5's superior performance stems from three key advantages: balanced sector diversification reflecting dynamic economic interpretation, consistent performance across different market regimes, and significant risk-adjusted alphas in both long and short portfolios.

GPT-4.1-nano occupies an intermediate position, demonstrating that even moderate contextual reasoning capabilities generate economically meaningful results. Despite concentrated sector exposure and higher volatility, GPT-4.1-nano produces positive returns (0.29% monthly), positive Sharpe (0.16) and Sortino (0.19) ratios, and a marginally significant long-short alpha (4.8% annualized). This substantially outperforms the embedding approach while falling short of GPT-5. The concentrated Lodging exposure, extreme negative skewness, and larger drawdowns reduce practical viability relative to GPT-5, but the model still captures predictive information beyond what static classification methods achieve.

The embedding approach, while computationally efficient, lacks the contextual sophistication necessary to capture time-varying sector-specific implications in financial news, resulting in persistent underperformance with negative returns, negative risk-adjusted ratios, and no statistically significant alphas. The stark contrast between embedding and GPT-4.1-nano performance confirms that some degree of contextual reasoning is essential, while the gap between GPT-4.1-nano and GPT-5 demonstrates that incremental improvements in reasoning capabilities translate directly into better risk-adjusted returns. This monotonic performance relationship across model sophistication validates our hypothesis that nuanced interpretation of commercial real estate news requires advanced language understanding beyond what traditional NLP methods or basic language models can provide.

7. Conclusion

This study demonstrates that LLMs can extract economically meaningful, sector-specific signals from commercial real estate news coverage. Using GPT-5 to generate CREATE scores from *Wall Street Journal* articles, we show that news-based signals contain substantial predictive power for REIT sector returns. A simple monthly sector rotation strategy produces annualized risk-adjusted alphas of 6.0 percent for large sectors and 9.6 percent for all sectors, with particularly strong performance during recessions. Comparing three models with varying contextual reasoning capabilities, we find that performance increases monotonically with model sophistication, providing direct evidence that advanced language understanding translates into superior information extraction. These results establish that sector-level disaggregation is critical for extracting predictive information from commercial real estate news, as identical articles carry divergent implications across property types.

Our findings have practical implications for real estate investors and portfolio managers. The CREATE methodology offers a scalable approach to processing large volumes of unstructured text that would be infeasible to analyze manually. The sector rotation strategy's strong recession performance suggests particular value during periods of economic stress when sectoral divergence is most pronounced. More broadly, the monotonic relationship between model sophistication and predictive performance indicates that continued advances in language models may yield further improvements in financial information extraction.

Several directions for future research emerge from this work. First, extending the analysis to additional news sources beyond the Wall Street Journal, including trade publications, earnings calls, and social media, could reveal whether CREATE scores capture unique information or reflect broader market narratives. Second, examining the persistence of alphas over time and their sensitivity to transaction costs would inform practical implementation decisions. Finally, investigating the economic mechanisms through which news coverage predicts returns, whether through information revelation, investor attention, or sentiment channels, would deepen our understanding of how textual information moves markets.

References

Alan, N. S., Engle, R. F., & Karagozoglu, A. K. (2023). Impact of Language Complexity on Volatility in Financial Markets: Evidence from Textual Analysis of Earnings Calls. *Journal of Portfolio Management*, *50*(2).

Beracha, E., Lang, M., & Hausler, J. (2019). On the relationship between market sentiment and commercial real estate performance—a textual analysis examination. *Journal of Real Estate Research*, 41(4), 605-638.

Bond, S. A., Klok, H., & Zhu, M. (2023). Large Language Models and Financial Market Sentiment. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4584928

Brockman, P., Li, X., & Price, S. M. (2015). Differences in conference call tones: Managers vs. analysts. *Financial Analysts Journal*, *71*(4), 24-42.

Bushee, B. J., & Huang, Y. (2024). Do analysts and investors efficiently respond to managerial linguistic complexity during conference calls?. *The Accounting Review*, *99*(4), 143-168.

Bybee, L. (2023). Surveying Generative AI's Economic Expectations. *SSRN Electronic Journal*. https://doi.og/10.2139/ssrn.4430515

Bybee, L., Kelly, B., Manela, A., & Xiu, D. (2024). Business news and business cycles. *The Journal of Finance*, *79*(5), 3105-3147.

Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2023). How to talk when a machine is listening: Corporate disclosure in the age of AI. *The Review of Financial Studies*, *36*(9), 3603-3642.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, *52*(1), 57-82.

Carstens, R., & Freybote, J. (2021). Can the textual tone in REIT financial statements improve the information environment for commercial real estate investors? An investigation. *Journal of Real Estate Research*, 43(3), 335-354.

Chacon, R. G., & Evans, J. D. (2023). Investor inattention to earnings surprises: Evidence from REIT and tenant information transmission. *Journal of Real Estate Portfolio Management*, 29(1), 61-77.

Chui, A. C., Titman, S., & Wei, K. J. (2003). Intra-industry momentum: the case of REITs. *Journal of Financial Markets*, 6(3), 363-387.

Cooper, M., Downs, D., & Patterson, G. (1999). Real estate securities and a filter-based, short-term trading strategy. *Journal of Real Estate Research*, 18(2), 313-333.

Engelberg, Joseph and Manela, Asaf and Mullins, William and Vulicevic, Luka, Entity Neutering (March 17, 2025). Available at SSRN: https://ssrn.com/abstract=5182756 or http://dx.doi.org/10.2139/ssrn.5182756

Ewertz, J., Knickrehm, C., Nienhaus, M., & Reichmann, D. (2025). Listen Closely: Measuring Vocal Tone in Corporate Disclosures. *Journal of Accounting Research*.

Fairhurst, D., & Greene, D. (2025). How Much Does ChatGPT Know about Finance?. *Financial Analysts Journal*, *81*(1), 12-32.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3-56.

Goebel, P. R., Harrison, D. M., Mercer, J. M., & Whitby, R. J. (2013). REIT momentum and characteristic-related REIT returns. *The Journal of Real Estate Finance and Economics*, 47(3), 564-581.

Hausler, J., Ruscheinsky, J., & Lang, M. (2018). News-based sentiment analysis in real estate: A machine learning approach. *Journal of Property Research*, *35*(4), 344–371. https://doi.org/10.1080/09599916.2018.1551923

Hung, S. Y. K., & Glascock, J. L. (2008). Momentum profitability and market trend: evidence from REITs. *The Journal of Real Estate Finance and Economics*, 37(1), 51-69.

Jensen, T. K., & Turner, T. M. (2022). Monetary policy shifts, dividends and REIT momentum. *Journal of Real Estate Research*, 44(3), 311-330.

Leow, K., & Lindenthal, T. (2025). Enhancing real estate investment trust return forecasts using machine learning. *Real Estate Economics*, 53(3), 574-606.

Letdin, M., Seagraves, C., & Sirmans, S. (2025). REIT Factors. *Available at SSRN*.

Ling, D. C., Naranjo, A., & Ryngaert, M. D. (2000). The predictability of equity REIT returns: Time variation and economic significance. *The journal of real estate finance and economics*, 20(2), 117-136.

Ling, D. C., Ooi, J. T., & Xu, R. (2019). Asset growth and stock performance: Evidence from REITs. *Real Estate Economics*, 47(3), 884-927.

Lo, A. W., & Singh, M. (2023). From ELIZA to ChatGPT: The Evolution of Natural Language Processing and Financial Applications. *Journal of Portfolio Management*, *49*(7).

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35-65.

Newey, W. K., & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777-787.

Paulus, N. M., Lautenschlaeger, L., & Schaefers, W. (2025). Social media and real estate: Do twitter users predict reit performance?. *Journal of Real Estate Research*, *47*(3), 322-355.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, *36*(4), 992-1011.

Rennekamp, K. M., Sethuraman, M., & Steenhoven, B. A. (2022). Engagement in earnings conference calls. *Journal of Accounting and Economics*, *74*(1), 101498.

Ruscheinsky, J. R., Lang, M., & Schäfers, W. (2018). Real estate media sentiment through textual analysis. *Journal of Property Investment & Finance*, 36(5), 410-428.

Shen, J., Hui, E. C., & Fan, K. (2021). The beta anomaly in the REIT market. *The Journal of Real Estate Finance and Economics*, 63(3), 414-436.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), 1139-1168.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, 63(3), 1437-1467.

Zhu, B., & Lizieri, C. (2024). Local beta: Has local real estate market risk been priced in REIT returns?. *The Journal of Real Estate Finance and Economics*, 69(4), 682-718.

Figure 1: Neutering Articles with LLMs

Original Article (abbreviated)	Neutered Article (abbreviated)
Fairness Becomes the Issue for High-End Funds --- At a Big Goldman Real-Estate Vehicle, A Capital Call Adds Insult to Big Losses Now, the Whitehall Street Global Real Estate Limited Partnership 2007 has told its investors to give it $1 billion in additional capital they had committed to the fund, according to fund documents from late April that were reviewed by The Wall Street Journal. The money is due on Monday and will be used in part to pay back $677 million on a credit line. The money also will be used to finance a business plan that Whitehall says will recover 71% of investors' total equity over the fund's holding period, the documents say.	Fairness Becomes the Issue for High-End Funds --- At a Big Company_1 Real-Estate Vehicle, A Capital Call Adds Insult to Big Losses Now, the Company_2 Global Real Estate Limited Partnership time_3 has told its investors to give it number c in additional capital they had committed to the fund, according to fund documents from late time_4 that were reviewed by link_x. The money is due on time_5 and will be used in part to pay back number d on a credit line. The money also will be used to finance a business plan that Company_2 says will recover number e% of investors' total equity over the fund's holding period, the documents say.

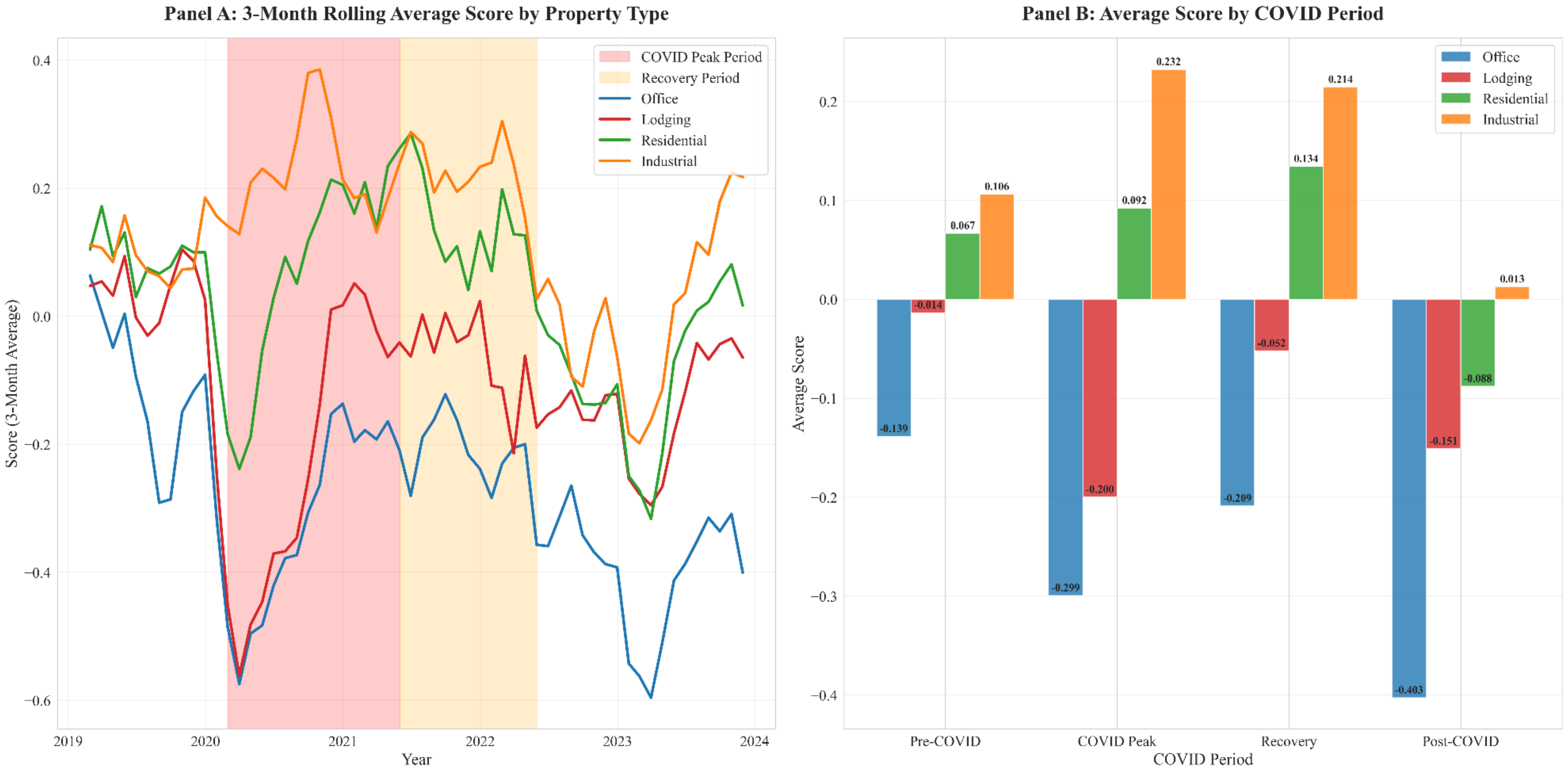
This figure shows an example WSJ article in our dataset in its raw form and its neutered form. Neutered is performed following the method outlined in Engelberg et al. (2025).

Figure 2: Sector-Level CREATE Score Example

Article Dated 3/31/2021	CREATE Score	LLM Justification
Office-Space Subleases Flood Market --- Companies cut rental needs in expectation that working from home will continue JPMorgan Chase &Co., Salesforce.com Inc. and PricewaterhouseCoopers are among the major companies looking to unload big blocks of office space, the latest sign that remote work is **hurting demand for this pillar of commercial real estate.** Large companies typically sign office leases for a decade or longer, giving them few options for reducing their footprint beyond trying to sublease floors to other tenants. At the end of 2020, 137 million square feet of office space was available for sublease across the U.S., according to CBRE Group Inc. That is up 40% from a year earlier and the highest figure since 2003. While sublet space increases during every recession as struggling businesses look to cut costs, firms typically add office space when the economy picks up again. But this time many of the companies ditching real estate are doing well financially; they say they need less space because they plan for more employees to **work at least part time from home even after the pandemic is over.** That raises the prospect that demand for office space could be permanently lower at some companies, much **like the rise of e-commerce has been driving down demand and rents for street-level retail**. *(article truncated for brevity)*	Office (-1)	Surging sublease supply and discounts signal structurally lower demand and falling rents from persistent remote work.
	Industrial (0)	WFH trend largely unrelated to warehouse demand; minimal direct impact.
	Retail (-1)	Fewer office workers downtown reduces weekday foot traffic, pressuring street-level retail rents.
	Residential (+1)	Remote work raises demand for living space and home offices, modestly supporting housing values.
	Healthcare (0)	Healthcare real estate weakly linked to office usage; impact unclear.
	Lodging (-1)	Reduced office-based business travel could depress urban hotel demand.
	Telecom (+1)	Remote work drives mobile/data demand, marginally supporting tower cash flows.
	Data Centers (+1)	Remote work accelerates cloud adoption, supporting data center demand and pricing.
	Self Storage (+1)	Households create home offices and declutter, increasing storage demand.

This figure shows an example WSJ article in our dataset, the resulting CREATE score for each property sector, and the justification for the score provided by the LLM (GPT-5).

Figure 3: Scores During COVID-19 for Select Property Sectors

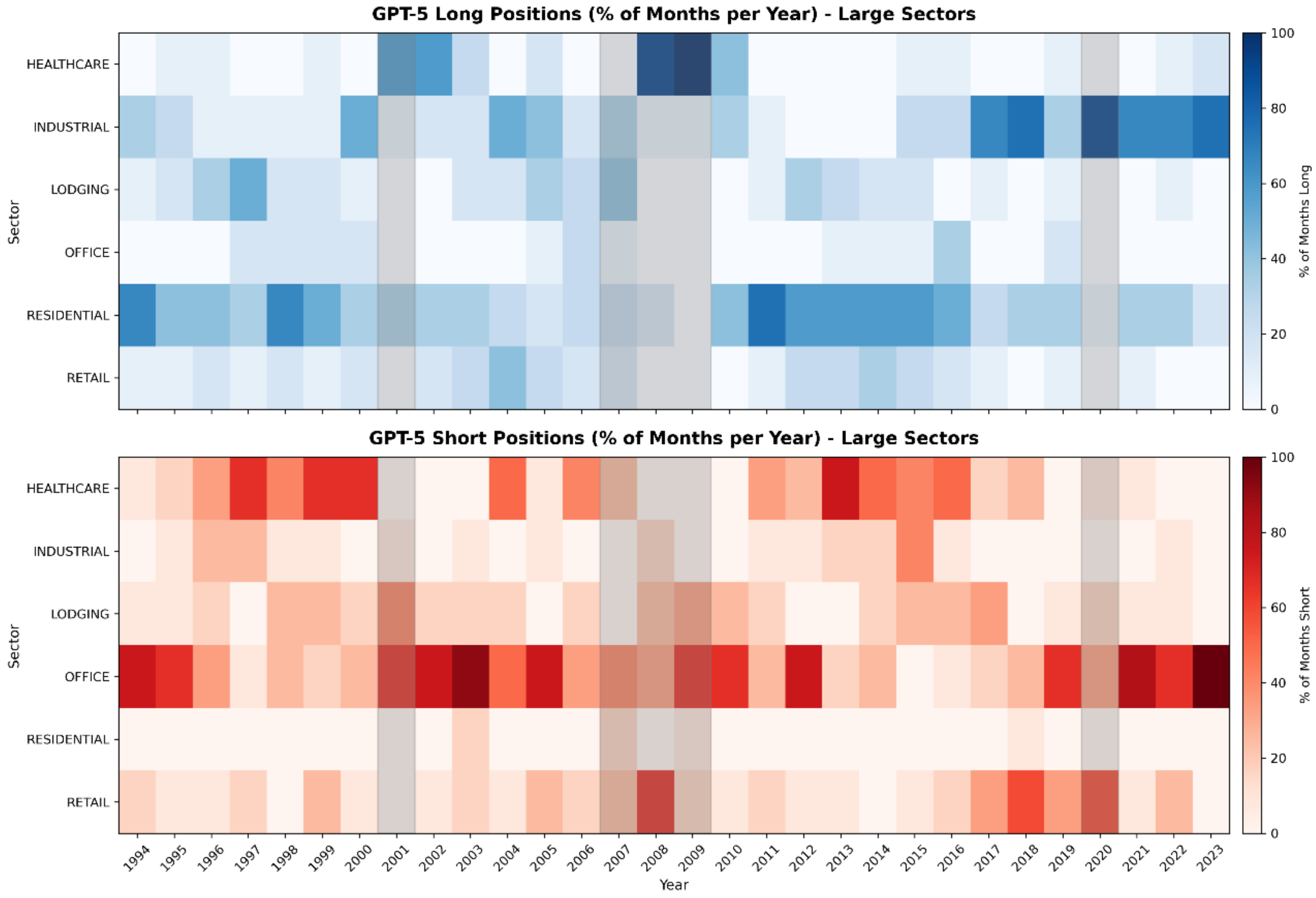


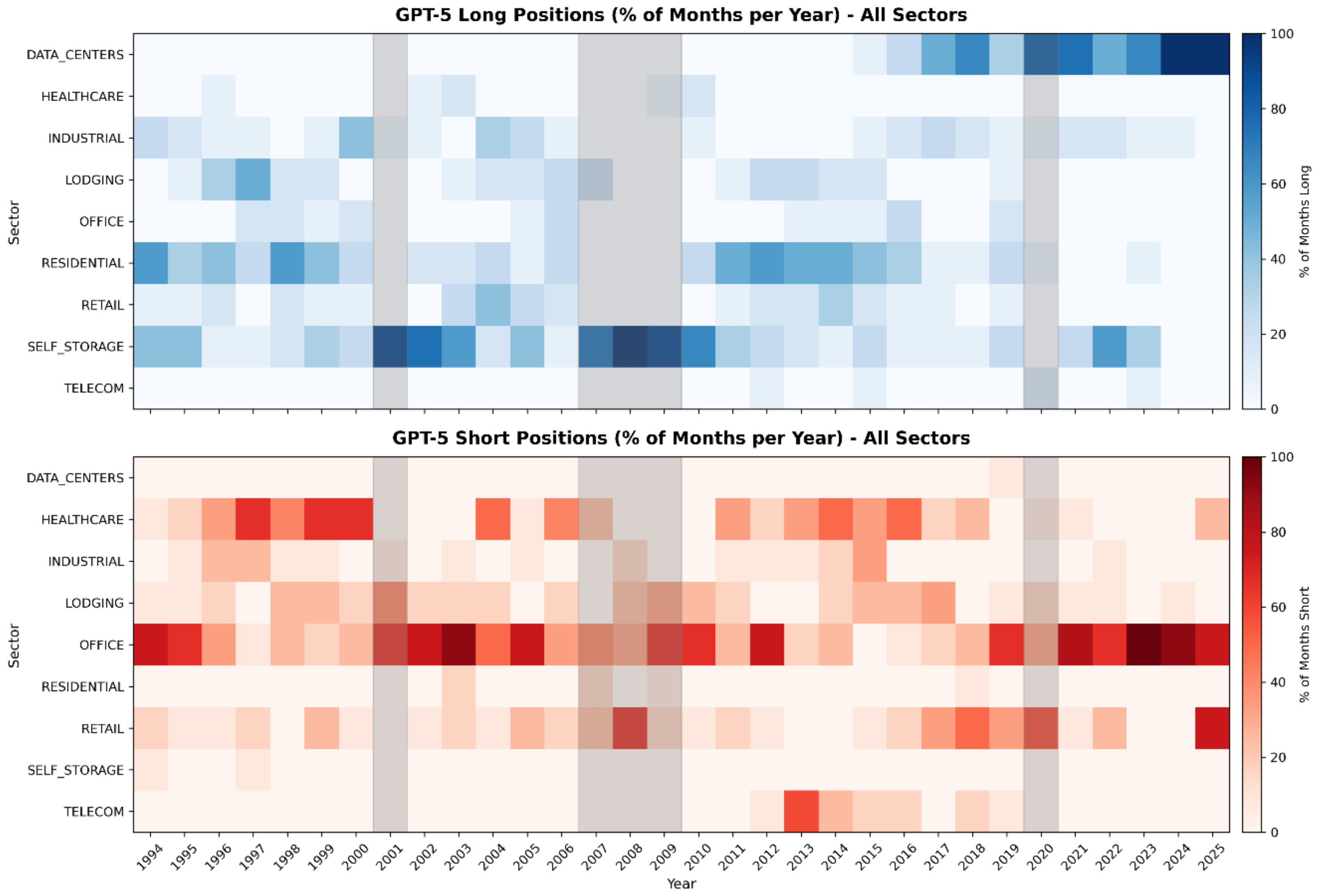
Panel A: 3-Month Rolling Average Score by Property Type

Panel B: Average Score by COVID Period

Figure 3 displays CREATE scores across four property types (Office, Lodging, Residential, Industrial) from 2019-2023. The left panel shows 3-month rolling averages over time and the right panel shows average scores by COVID period (right panel). COVID periods are defined as: Pre-COVID (Jan 2019-Feb 2020), COVID Peak (Mar 2020-May 2021), Recovery (Jun 2021-May 2022), and Post-COVID (Jun 2022-Dec 2023). CREATE scores are derived from LLM analysis, with positive values indicating favorable scores and negative values indicating unfavorable scores. Shaded regions in the left panel highlight the COVID Peak (red) and Recovery (orange) periods.

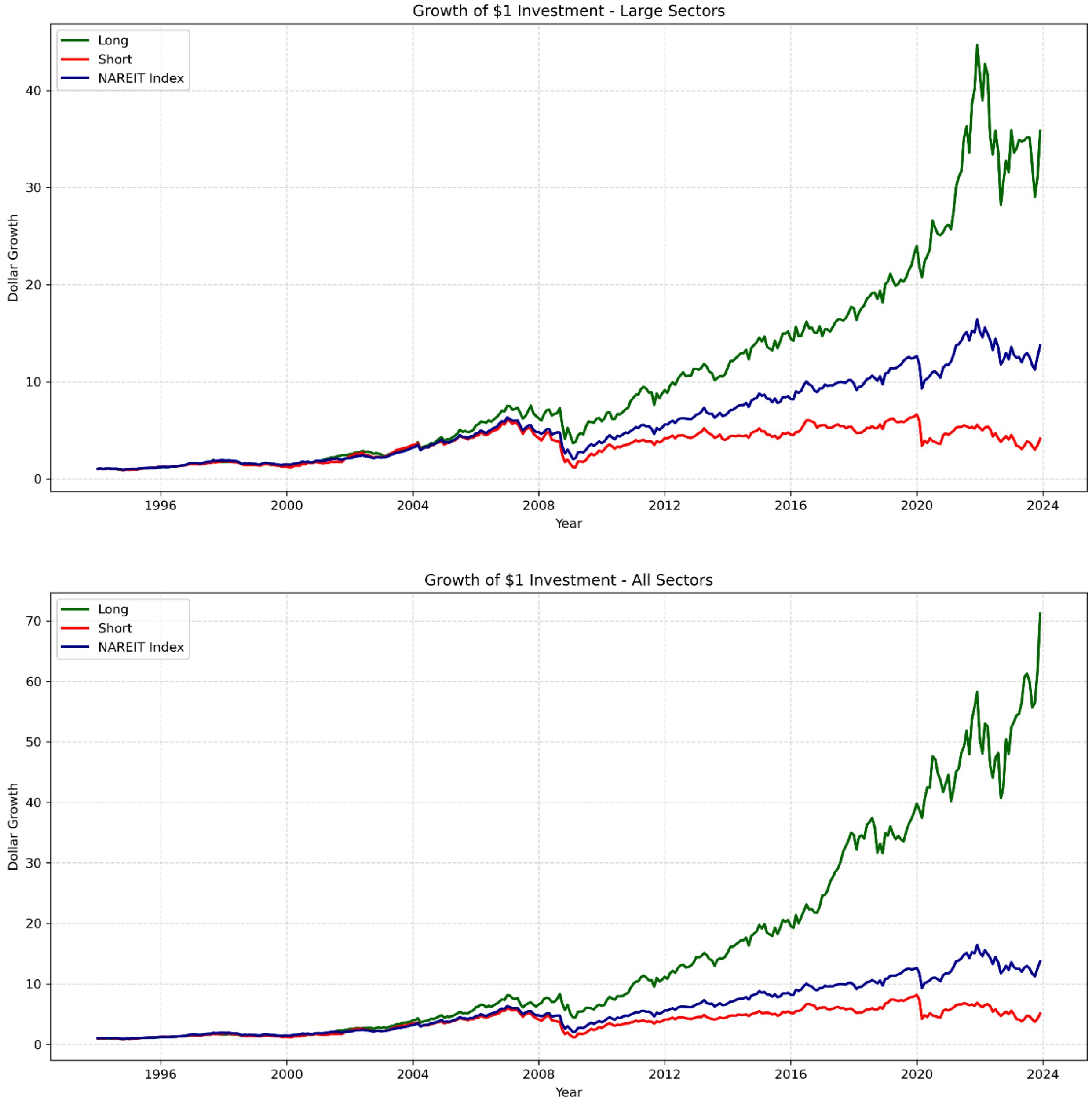
Figure 4: Sector Selection Over Time





This figure displays the annual frequency of GPT-5 long and short sector positions for large sectors (first panel) and all sectors (second panel). The heatmaps display the percentage of months within each year that each sector was selected for a long position or short position. Gray vertical bands indicate NBER-dated recessions.

Figure 5: Growth of $1 Investment in Sector Rotation Strategy

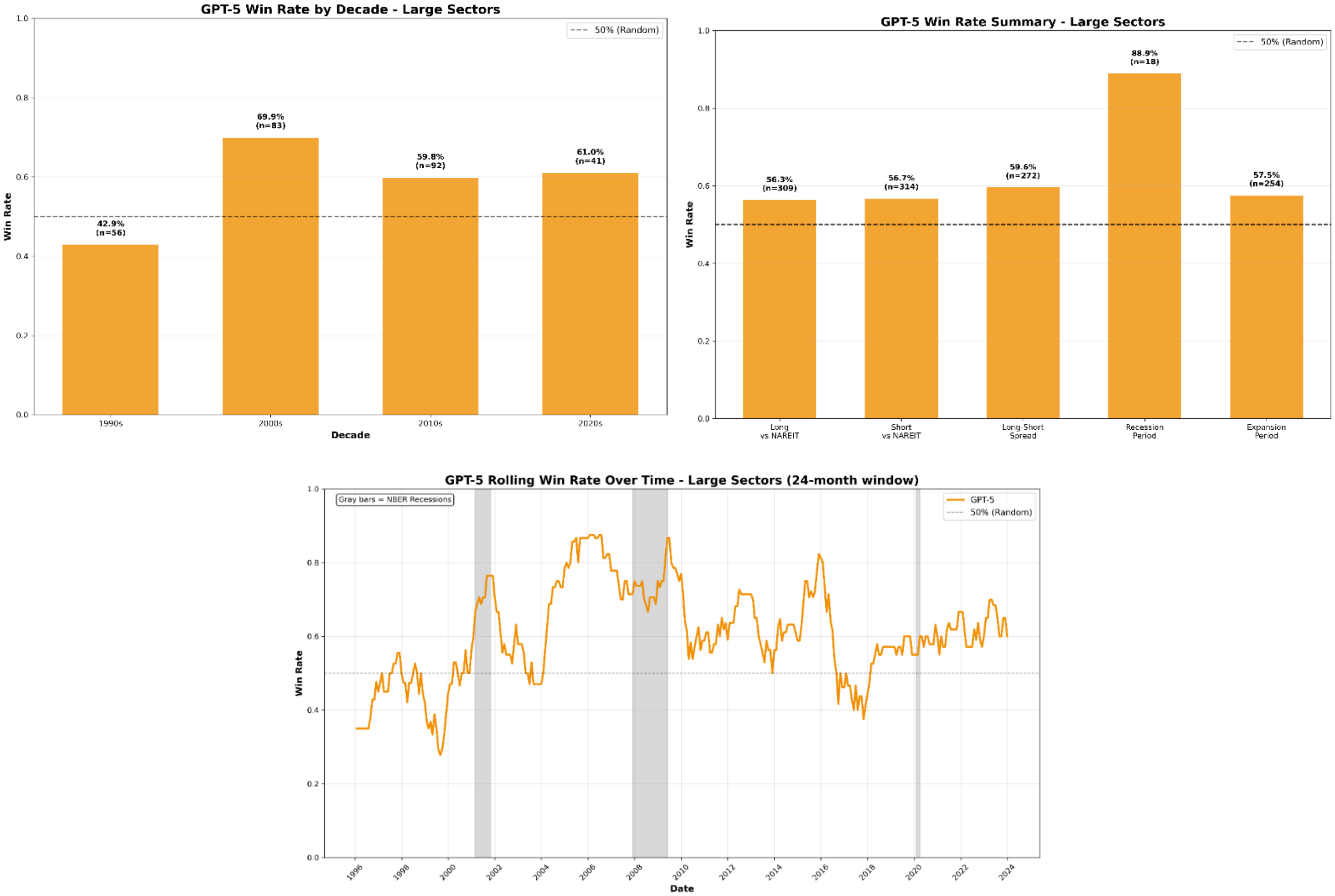


Growth of $1 Investment - Large Sectors

Growth of $1 Investment - All Sectors

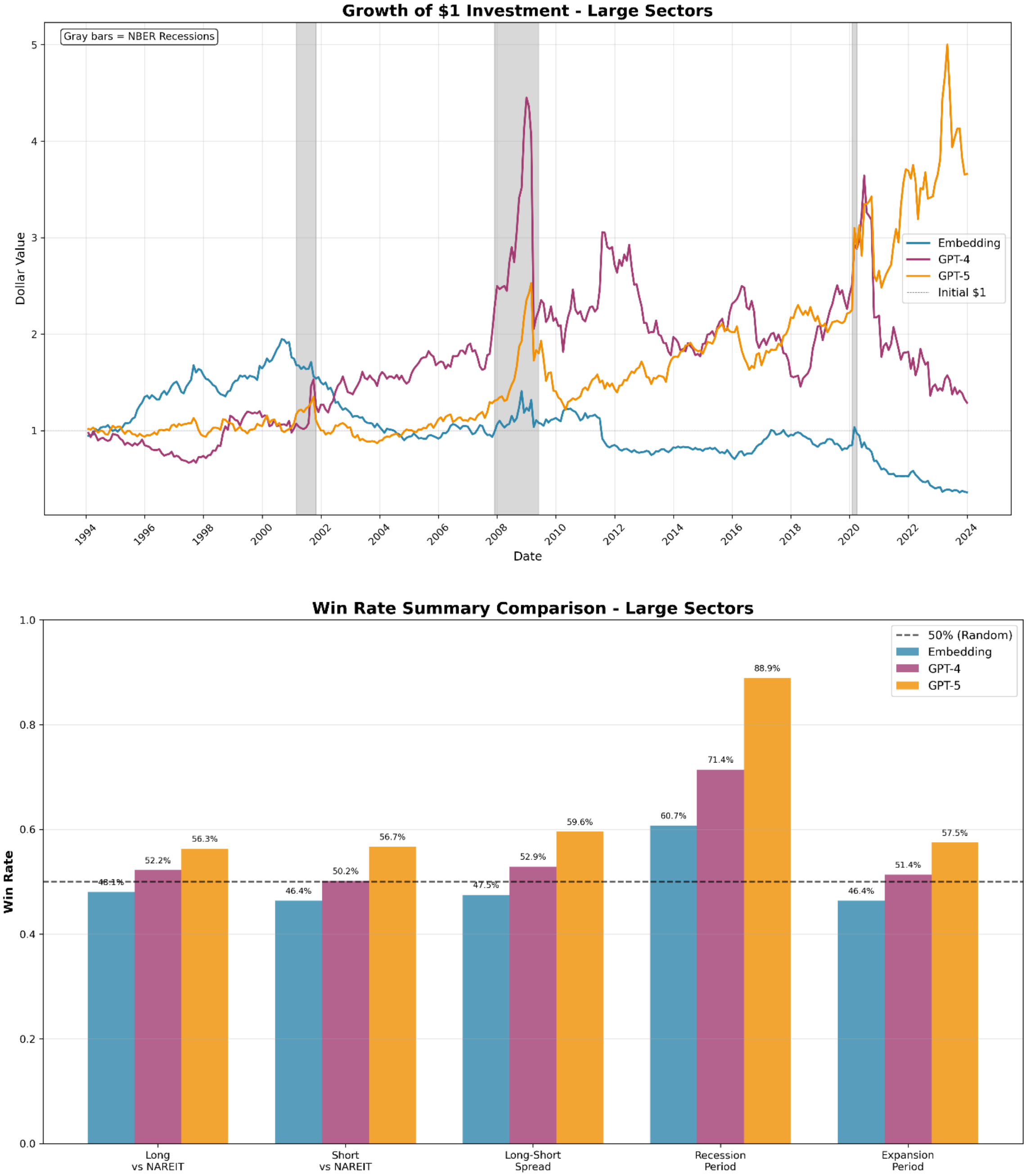
This figure shows the growth of $1 invested in the sector rotation strategy using the GPT-5 model. The green line is the growth of $1 in the long portfolio, the red line is the growth of the short portfolio, and the blue line is the growth of the NAREIT index. The first panel uses only large property sectors and the second panel uses all property sectors.

Figure 6: Win Rates



This figure displays GPT-5 long-short spread win rates for large sectors. Top left panel shows long-short spread win rates by decade (1990s-2020s). Top right panel displays win rates across five performance metrics: long positions versus NAREIT benchmark, short positions versus NAREIT benchmark, overall long-short spread, long-short spread during recession periods, and long-short spread during expansion periods. The bottom panel presents the 24-month rolling win rate for the long-short spread over time, with gray vertical bands indicating NBER-dated recessions. The dashed horizontal line at 50% represents random selection performance. Sample sizes are shown in parentheses for each metric.

Figure 7: Model Comparisons



This figure shows a performance comparison of three models (Embedding, GPT-4.1-nano, and GPT-5) for large sectors. Top panel shows the growth of $1 invested using each model's long-short strategy. Bottom panel compares win rates across five performance metrics: long positions versus NAREIT benchmark, short positions versus NAREIT benchmark, overall long-short spread, long-short spread during recession periods, and long-short spread during expansion periods. The dashed horizontal line at 50% represents random selection performance.

Table 1: Article and Monthly CREATE Score Statistics

Panel A: GPT-5 Scores	Articles	Mean	Std. Dev.	Min	P25	Median	P75	Max
Office	17754	-0.06	0.70	-1.00	-1.00	0.00	0.00	1.00
Industrial	17754	0.05	0.59	-1.00	0.00	0.00	0.00	1.00
Retail	17754	-0.01	0.72	-1.00	-1.00	0.00	1.00	1.00
Residential	17754	0.09	0.60	-1.00	0.00	0.00	0.00	1.00
Healthcare	17754	0.02	0.43	-1.00	0.00	0.00	0.00	1.00
Lodging	17754	0.00	0.65	-1.00	0.00	0.00	0.00	1.00
Data Centers	3930	0.26	0.54	-1.00	0.00	0.00	1.00	1.00
Self Storage	17754	0.13	0.49	-1.00	0.00	0.00	0.00	1.00
Telecom	6663	0.10	0.40	-1.00	0.00	0.00	0.00	1.00

Panel A: GPT-5 Monthly Scores	Months	Mean	Std. Dev.	Min	P25	Median	P75	Max
Office	360	-0.07	0.21	-0.68	-0.21	-0.04	0.09	0.33
Industrial	360	0.06	0.16	-0.53	-0.02	0.08	0.15	0.47
Retail	360	-0.02	0.21	-0.73	-0.14	0.03	0.13	0.42
Residential	360	0.08	0.16	-0.47	0.00	0.10	0.19	0.43
Healthcare	360	0.02	0.09	-0.33	-0.03	0.03	0.07	0.25
Lodging	360	0.00	0.19	-0.70	-0.11	0.04	0.13	0.38
Data Centers	97	0.23	0.15	-0.14	0.13	0.23	0.33	0.63
Self Storage	360	0.13	0.10	-0.25	0.08	0.14	0.19	0.53
Telecom	144	0.10	0.11	-0.14	0.04	0.08	0.14	0.53

This table presents descriptive statistics for GPT-5 CREATE scores across nine real estate sectors at two levels of aggregation. Panel A displays article-level statistics, showing the distribution of scores across individual news articles for each sector. Panel B presents monthly-level statistics, showing the distribution of aggregated monthly average scores for each sector. Scores range from -1 (most negative) to +1 (most positive).

Table 2: Sector CREATE Scores Correlation

Sector	Office	Industrial	Retail	Residential	Healthcare	Lodging	Data Centers	Self-Storage	Telecom
Office	1.00								
Industrial	0.54	1.00							
Retail	0.80	0.57	1.00						
Residential	0.54	0.60	0.54	1.00					
Healthcare	-0.08	0.33	-0.14	0.27	1.00				
Lodging	0.77	0.56	0.87	0.48	-0.22	1.00			
Data Centers	-0.34	0.05	-0.48	0.05	0.64	-0.55	1.00		
Self-Storage	-0.04	0.40	-0.11	0.44	0.68	-0.20	0.69	1.00	
Telecom	-0.29	0.12	-0.42	0.15	0.70	-0.50	0.89	0.72	1.00

This table displays the Pearson correlation matrix of average monthly GPT-5 CREATE scores across nine real estate sectors from 1994-2023. Each cell shows the correlation coefficient between the monthly score series for the corresponding pair of sectors.

Table 3: Long-Short Sector Frequency

Panel A: Largest Sectors	Long	Short	Long %	Short %
Industrial	109	27	30.3%	7.5%
Healthcare	56	91	15.6%	25.3%
Office	22	167	6.1%	46.4%
Residential	135	6	37.5%	1.7%
Retail	46	66	12.8%	18.3%
Lodging	49	54	13.6%	15.0%

Panel B: All Sectors	Long	Short	Long %	Short %
Industrial	39	24	10.8%	6.7%
Healthcare	7	84	1.9%	23.3%
Office	19	167	5.3%	46.4%
Residential	89	6	24.7%	1.7%
Retail	35	64	9.7%	17.8%
Lodging	38	54	10.6%	15.0%
Data Centers	55	1	15.3%	0.3%
Self Storage	129	2	35.8%	0.6%
Telecom	5	18	1.4%	5.0%

This table presents the number and percentage of months each sector was selected for long or short positions by GPT-5. Panel A covers large sectors; Panel B covers all sectors. When ties occurred, multiple sectors were counted for that month, causing totals to exceed 360 months.

Table 4: Sector Rotation Performance

Metric	Large Sectors	All Sectors
Mean	0.49%	0.64%
Std. Dev.	5.10%	5.55%
VaR 5%	-6.59%	-6.67%
Max Drawdown	-51.98%	-52.91%
Skewness	7.04%	16.62%
Downside Vol.	4.13%	4.38%
Sharpe Ratio	0.34	0.40
Sortino Ratio	0.41	0.51

This table presents portfolio performance statistics for the GPT-5 long-short sector rotation strategy from 1994-2023. Results are shown for both large sectors (six sectors) and all sectors (nine sectors) samples. Mean is the average monthly return. Std. Dev. is the standard deviation of monthly returns. VaR 5% is the Value at Risk at the 5% confidence level, representing the maximum expected loss in the worst 5% of months. Max Drawdown is the largest peak-to-trough decline in cumulative returns. Skewness measures the asymmetry of the return distribution (positive values indicate right skew). Downside Vol. is the standard deviation of negative returns only. Sharpe Ratio is the mean return divided by standard deviation annualized. Sortino Ratio is the mean return divided by downside volatility annualized.

Table 5: Sector Rotation Alpha

Panel A: Largest Sectors	(1)	(2)	(3)
	Long	*Short*	*Long - Short*
Alpha	0.002*	-0.003**	0.005**
	(1.68)	(-1.98)	(2.51)
Market	1.015***	1.082***	-0.066
	(22.03)	(26.73)	(-0.90)
Size	0.010	-0.050	0.060
	(0.14)	(-0.59)	(0.47)
Value	-0.171***	0.382***	-0.553***
	(-3.80)	(7.69)	(-7.50)
Momentum	0.069	0.049	0.019
	(0.75)	(0.54)	(0.13)
R-Squared	0.7775	0.8569	0.3560
Number of Observations	360	360	360

Panel B: All Sectors	(1)	(2)	(3)
	Long	*Short*	*Long - Short*
Alpha	0.006***	-0.002	0.008***
	(3.40)	(-1.44)	(3.28)
Market	0.900***	1.056***	-0.156**
	(20.53)	(25.33)	(-2.28)
Size	-0.095	-0.070	-0.025
	(-1.09)	(-0.80)	(-0.19)
Value	-0.206***	0.388***	-0.594***
	(-2.91)	(7.82)	(-6.58)
Momentum	0.012	0.046	-0.034
	(0.11)	(0.49)	(-0.22)
R-Squared	0.7013	0.8513	0.3821
Number of Observations	360	360	360

This table presents the long-short sector rotation performance against a REIT-based 4-factor model. Each of the factors are constructed using the REIT universe as documented by Letdin et al. (2025). Each variable is defined in the appendix. The values in parentheses are Newey and West (1987) corrected t-statistics with a lag length of one. Significance at the 10%, 5%, and 1% level are denoted by *, **, and ***, respectively.

Table 6: Sector Frequency by Model – Large Sectors

Model	EMD	GPT4	GPT5	EMD	GPT4	GPT5
		Long %			Short %	
Industrial	0.00%	13.61%	30.28%	0.00%	4.44%	7.50%
Healthcare	0.00%	47.50%	15.56%	0.00%	2.50%	25.28%
Office	80.00%	11.11%	6.11%	0.00%	20.00%	46.39%
Residential	19.44%	36.11%	37.50%	0.00%	0.28%	1.67%
Retail	0.00%	4.44%	12.78%	100.00%	15.83%	18.33%
Lodging	0.56%	0.28%	13.61%	0.00%	72.50%	15.00%

This table compares the frequency with which each large sector was selected for long and short positions across three models (Embedding, GPT-4.1-nano, and GPT-5) from 1994-2023. Values represent the percentage of months (out of 360 total) that each sector appeared in the long portfolio (left columns) or short portfolio (right columns) for each model. When multiple sectors tied for the highest (lowest) score in a given month, both sectors were included in the long (short) position, resulting in percentages that may sum to more than 100%.

Table 7: Sector Performance by Model – Large Sectors

Panel A: Portfolio Statistics by Model

Metric	EMB	GPT-4	GPT-5
Mean	-0.19%	0.29%	0.49%
Std. Dev.	4.04%	6.45%	5.10%
VaR 5%	-5.38%	-7.67%	-6.59%
Max Drawdown	-81.58%	-71.06%	-51.98%
Skewness	-0.25	-1.02	0.07
Downside Vol.	2.97%	5.29%	4.13%
Sharpe Ratio	-0.167	0.1552	0.335
Sortino Ratio	-0.227	0.1893	0.4136

Panel B: Alphas by Model

Metric	EMB	GPT-4	GPT-5
Long Alpha	-0.001	0.003**	0.002*
Short Alpha	0.001	-0.002	-0.003**
L-S Alpha	-0.002	0.004	0.005**

This table compares the performance of sector rotation strategies across three models (Embedding, GPT-4.1-nano, and GPT-5) for large sectors from 1994-2023. Panel A presents portfolio statistics for the long-short strategy, including mean monthly return, standard deviation, Value at Risk at 5%, maximum drawdown, skewness, downside volatility, Sharpe ratio, and Sortino ratio. Panel B reports risk-adjusted performance (alpha) from factor model regressions for long, short, and long-short (L-S) portfolios. Alphas are estimated using market, size, value, and momentum factors following Letdin et al. (2025). Significance at the 10%, 5%, and 1% level are denoted by *, **, and ***, respectively.

Table A1: LLM Sector Prompt

Given a news article, identify the likely impact of the news on U.S. real-estate values for every NAREIT subsector listed below.

Respond with a single JSON object in which **each key** is one subsector name and **each value** is an object with the following keys:

- explanation: a brief explanation of your reasoning (less than 25 words)

- increase_decrease: "increase", "decrease", or "uncertain"

Subsectors to score (use exactly these names as keys):

Office, Industrial, Retail, Residential, Health Care, Lodging, Self Storage, Telecommunications, Data Centers.

Example format:

```
{
 "Office": {
   "explanation": "",
   "increase_decrease": "increase/decrease/uncertain",
 },
```

Table A2: LLM Neutering Prompt

```
system_message = '''

Your role is to ANONYMIZE all text that is provided by the user. After you have anonymized a
text, NOBODY, not even an expert financial analyst should be able to read the text and know the
identity of the company nor the industry the company operates in. For example, if the text is: The
country's largest phone producer Apple had great phone related earnings but Google did not in
2024 likely because of Apple's slogan Think Different, then you should ANONYMIZE it to: The
country's largest product_type_1 producer Company_1 had great product_type_1 related
earnings but Company_2 did not in time_1 likely because of Company_1's slogan slogan_1. You
should also ANONYMIZE any other information which one could use to identify the company
or make an educated guess at its identity. Stock tickers are identifiers and are usually four
capitalized letters or less (consider TIK as a stand-in for an arbitrary ticker) and are sometimes
referenced in the text in the following formats; SYMBOL:TIK, $TIK, >TIK, $ TIK, SYMBOL
TIK, SYMBOL: TIK, > TIK. Make sure you censor TIK to ticker_x, and any other identifiers
related to companies. This includes the names of individuals, locations, industries, sectors,
product names and types and generic product lines, services, times, years, dates and all numbers
and percentages in the text including units. These should be replaced with name_x, location_x,
industry_x, sector_x, product_x, product_type_x, product_line_x, service_x, time_x, year_x,
date_x and number a, b, c, respectively. Also replace any website or internet links with link_x.
You should never just delete an identifier; instead, always replace it with an anonymous analog.
After you read and ANONYMIZE the text, you should output the anonymized text and nothing
else.

'''.replace('\n', ' ').strip()

MODEL = ‘gpt-4o-mini-2024-07-18’

response = client.chat.completions.create(

            model=MODEL,

            messages=[

                {"role": "system", "content": system_message},

                {"role": "user", "content": doc},

            ],

            temperature=0,

            max_completion_tokens=max(128, len(doc)//3),
```

Table A3: Sector Level Embeddings Process

This appendix provides detailed documentation of our embedding-based approach to generating sector-level CREATE scores. The methodology combines synthetic training data generation, text embeddings, and supervised classification to produce scores comparable to those generated by GPT-5 and GPT-4.1-nano.

A.4.1 Overview

The embedding approach differs fundamentally from generative language models in that it does not produce explicit reasoning or justifications. Instead, it generates synthetic training data with known labels using an LLM, embeds both training snippets and news articles into high-dimensional vector space, trains sector-specific classifiers to predict CREATE scores from embeddings, and applies the trained classifiers to score actual news articles.

A.4.2 Synthetic Training Data Generation

We begin by defining a comprehensive sector-topic prompt space. For each sector, we identify 8 domain-relevant topics that capture key aspects of commercial real estate: Market fundamentals (occupancy, absorption, rent growth, operating metrics), Capital markets and financing (debt/equity conditions, cap rates, spreads), Development and construction (starts, deliveries, costs, delays), Leasing and tenant news (new leases, renewals, spreads, expansions/closures), Regulatory and policy (tax, zoning, incentives, environmental rules), Macroeconomic and external factors (rates, inflation, employment, GDP), ESG and sustainability (green financing, certifications, emissions), and Exit and returns (dispositions, pricing, exit cap rates, realized returns).

For each sector-topic combination, we use GPT-5 with low reasoning effort to generate text snippets with known properties. We request 5-10 snippets in each of four classes: Positive (directionally bullish for the sector, such as improving fundamentals, better pricing power, or easier financing), Negative (directionally bearish, including weakening demand, rising costs, or tighter financing), Neutral (informational or mixed updates with no clear directional signal), and Irrelevant (real estate content not pertaining to the specific sector-topic combination). The output format is structured JSON via OpenAI's Pydantic response format to ensure schema compliance.

A.4.3 Label Mapping and Class Structure

We collapse the four generated classes into a 3-class problem for training. Class 0 (Negative) contains negative snippets, Class 1 (Neutral/Irrelevant) merges neutral and irrelevant snippets, and Class 2 (Positive) contains positive snippets. The rationale for merging neutral and irrelevant is that both represent non-directional information for sector valuation. Creating a stable "middle" class improves classifier robustness and downstream scoring stability.

A.4.4 Embedding Generation

We use OpenAI's text-embedding-3-small model, which produces 1536-dimensional vectors. This model was selected for computational efficiency while maintaining semantic quality. This approach enables deterministic reproduction given fixed inputs.

A.4.5 Classifier Training

We employ multinomial logistic regression from scikit-learn as our model architecture. This choice offers several advantages: fast training, interpretable linear decision boundaries, and suitability for high-dimensional embeddings. The limitation is that the model cannot capture non-linear semantic patterns that more sophisticated models might learn. Our hyperparameters include solver='lbfgs' (a multinomial-compatible optimizer), multi_class='multinomial' (for direct 3-class probability estimation), class_weight='balanced' (which automatically adjusts for class imbalance across sectors), C=1.0 (standard regularization strength), max_iter=10,000 (sufficient for convergence on embedded features), and random_state=42 (a fixed seed for reproducibility).

Our training strategy involves building one classifier per sector, each specialized to its sector. We use a stratified 80/20 train/test split within each sector to preserve class distributions and evaluate performance using macro-F1 score on the held-out test set, which gives equal weight to all classes. Each trained classifier is saved as a Joblib package containing the fitted estimator, class names and label mappings, and embedding model identifier.

A.4.6 Article Scoring (Inference)

The scoring process requires an article corpus with precomputed embeddings using text-embedding-3-small, alignment between article CSV and embeddings array via a key column, and use of the same embedding model for both training and inference. For each article-sector pair, we load the trained sector-specific classifier and extract class probabilities [P(negative), P(neutral/irrelevant), P(positive)], and compute a scalar CREATE score as S = P(positive) minus P(negative). This score ranges from negative one to positive one, where positive one indicates maximally positive, negative one indicates maximally negative, and zero indicates neutral. The output format is a wide CSV with one column per sector, where each cell contains the CREATE score for that article-sector pair. The process is resumable and can merge with previous results to populate only missing scores.

A.4.7 Reproducibility and Quality Controls

We ensure determinism through several mechanisms. We use fixed random_state=42 for all train/test splits and model initialization. The embedding cache ensures identical vectors for identical text inputs. Structured JSON outputs reduce parsing ambiguity. Error handling includes retry logic with exponential backoff for LLM generation, non-fatal warnings for minor deviations (such as snippet count slightly off target), and resumable workflows at both generation and scoring stages. Validation procedures include macro-F1 reporting on held-out test sets per sector and optional correlation analysis across sector scores to assess co-movement patterns.

A.4.8 Limitations

The embedding approach faces several important limitations. First, classifiers are trained on LLM-generated data rather than human-labeled examples. This introduces distributional bias, where synthetic snippets may not fully capture the linguistic diversity and edge cases present in actual financial news, and instruction-following bias, where training data reflects the LLM's interpretation of "positive," "negative," and "neutral" rather than human expert judgment. Second, prompts constrain content to U.S. commercial real estate contexts, and generalization to out-of-domain text (such as international markets or non-CRE assets) depends entirely on the embedding model's semantic coverage. Third, multinomial logistic regression assumes linear decision boundaries in embedding space. Non-linear classifiers (such as neural networks or gradient boosting) might capture more nuanced patterns but risk overfitting to artifacts in the synthetic training data. Fourth, unlike GPT-5 which can reason about context dynamically, the embedding approach applies fixed classifier weights learned from historical patterns. This limits adaptability to novel economic situations or unprecedented sectoral dynamics.

A.4.9 Key Technical Specifications

The class mapping assigns Negative to 0, Neutral/Irrelevant to 1, and Positive to 2. The CREATE score formula is S = P(positive) minus P(negative), yielding values between negative one and positive one. The training configuration uses one classifier per sector, with features derived from 1536-dimensional text-embedding-3-small vectors. Evaluation uses a stratified 80/20 split with macro-F1 reporting. Output consists of per-article, per-sector scalar scores in wide CSV format suitable for merging with other model results and downstream analysis.

Table A4: Article Filtering Prompt

You are a news triage agent for a multi-sector REIT investment firm.

Task: For a single news article, decide which Nareit sectors' portfolio managers should receive the article (if any).

Output: A single JSON object whose keys are the EXACT sector names and whose values are booleans:

true = forward the article to that sector team; false = do not forward.

Forward only when the article likely moves the needle for that sector:

- clear positive or negative implications for fundamentals, valuations, credit, regulation, tax, supply/demand, or cap rates

- material company- or market-level events (M&A, guidance, large developments, policy shocks, disasters)

- credible data with sector-specific consequences

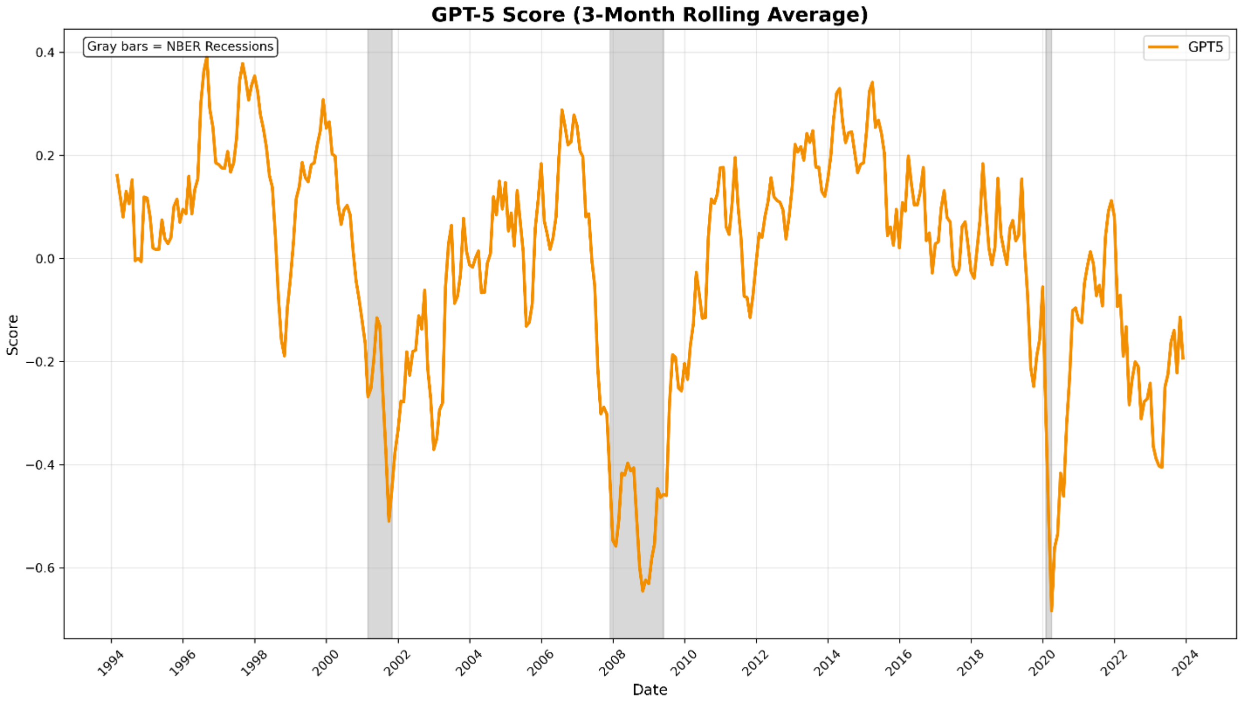
Be conservative. If relevance is weak, return false for that sector. Some articles may not be relevant to any sector. Some news will be relevant for all sectors.

Sectors:

Office, Industrial, Retail, Residential, Health Care, Lodging, Self Storage, Telecommunications, Data Centers

Table A5: Overall CREATE Score

We use a similar prompt to our sector level CREATE score prompt to elicit article-level scores from GPT-5, focused on the expected impact of each article on U.S. commercial real estate prices. Rather than prompting about sector-level information, we simply ask it for an overall CREATE score. The figure below displays the 3-month rolling average score over time alongside NBER recession bars. This figure provides additional validation our measure is highly time varying and consistent with overall economic conditions. The dip in 2022 is also consistent with the CRE focused economic challenges.


GPT-5 Score (3-Month Rolling Average)
Gray bars = NBER Recessions
GPT5
Score
Date