Journal of High School Science

Quantifying legal risk with Large Language Models: A text-based investment signal

Offutt J[1], Xie Y[2]

Submitted: July 13, 2025, Revised: version 1, August 25, 2025
Accepted: September 8, 2025

Abstract
This study developed a novel, text-based investment signal by applying Large Language Model (LLM) technology to quantify changes in legal risk disclosed in twin Securities and Exchange Commission (SEC) 10-Q and 10-K filings. Using the OpenAI GPT-4o LLM, the method compared the *Legal Proceedings* sections and associated footnotes from period-adjacent filings to score directional changes in legal risk. All firms with non-zero scores were incorporated into monthly long-short sub-portfolios, where negative scores were 'longed' and positive scores were shorted. Positions were dynamically weighted using a hyperbolic tangent transformation based on each firm's historical volatility and magnitude of legal risk change score. This approach offers a novel method for extracting firm-specific legal risk signals that have been historically underutilized in asset pricing models. It has potential applications in alpha generation, risk management, litigation monitoring, ESG compliance, and research automation. Ordinary Least Squares (OLS) regression was used to test the portfolio's returns against the standard Fama-French five-factor model plus momentum. The long-short strategy produced a statistically significant mean monthly alpha of 0.471% (annualized to 5.80%) for each sub-portfolio, indicating excess return generation beyond traditional risk factors, though with exposure to a large-to-moderate drawdown. All results were unlevered. Given the strategy's measured volatility and delta exposure, returns could be further scaled according to risk tolerance via leverage. Controlling for standard risk factors, the strategy's returns were partially explained by exposure to both growth and conservative-investment firms, but a meaningful portion of alpha remained unexplained by market risk, size, profitability, or momentum, highlighting the distinct contribution of the legal risk signal. An LLM-only, equal-weight mirror test—portfolios traded solely on the LLM legal-risk score (no position sizing or additional parameters)—was also executed: the normal equal-weight portfolio (long risk decreases / short risk increases) decisively outperformed its mirror (same trades but with reversed longs/shorts) over both control and winsorized iterations. These findings suggest that LLMs can uncover priced legal risk shifts and generate predictive signals beyond traditional textual analysis techniques.

Keywords
Large language model, Artificial intelligence, SEC filings, 10-K, 10-Q, Textual analysis, Alpha generation, Litigation disclosure, Twin SEC filings, GPT-4o, API

[1]James Offutt, The Hotchkiss School, 11 Interlaken Rd, Lakeville, CT 06039, USA. JamesOffutt3@icloud.com
[2]Yale Xie, Liberal Arts and Science Academy, 1012 Arthur Stiles Rd, Austin, TX, USA. YaleXie@gmail.com

Introduction

The classification and interpretation of SEC filing disclosures have proven increasingly valuable in understanding financial markets and academic research. While sections like *Risk Factors* and *Management's Discussion & Analysis (MD&A)* have been thoroughly researched using a multitude of approaches, the *Legal Proceedings* section and its related footnotes remain relatively unexplored in this capacity. This study investigates whether a Large Language Model (LLM)-driven methodology provides superior predictive insights into firm-specific risks and future market performance through textual analysis of SEC 10-Q and 10-K filings. In particular, the research sought to utilize OpenAI's flagship GPT-4o model, which has demonstrated competency in capturing nuanced contextual shifts and changes in tone escalation.

As stated, the analysis focused on the *Legal Proceedings* section (as well as its associated footnotes) of SEC 10-Q and 10-K filings, expanding upon this under-researched portion of the related literature. Existing methodologies have predominantly used Natural Language Processing (NLP) techniques or basic sentiment categorization, overlooking the nuanced semantic contexts inherent in the filings. In contrast, this study used GPT-4o to capture semantic shifts and sentiment variations in these disclosures, which can yield more accurate assessments of firms' legal and economic risks.

Previous textual analysis methods contain limitations that underscore the necessity for methodological improvements. For example, cosine similarity—used in much of the prior literature, such as Cohen, Malloy, and Nguyen (1)—is based solely on word frequency and is limited in accurately interpreting subtle but economically meaningful textual shifts. Meanwhile, FinBERT, while capable of extracting word meaning and surrounding context, is purely a classifier (positive, neutral, negative) and constrained by shallow context understanding due to its 512-token limit. Consequently, these approaches provide only limited insight into how textual disclosures relate to future financial performance.

This study extends prior literature, such as the foundational insights in Cohen, Malloy, and Nguyen (1) and the embedding-based analyses in Adosoglou et al. (2,3) and Yilmaz and Reichmann (4). It further interacts with generative NLP advancements showcased by Gupta (5), Daimi and Iqbal (6), and Bürgler et al. (7), refining these methods through targeted specialization, methodological changes, and recency. The research addresses limitations in these prior methods by adopting a domain-specific LLM and intentional LLM prompt engineering. Although this study's main model (which incorporated both the signal form the LLM as well as risk parameters for position sizing) performed the best, the pure attribution of the LLM scores were tested using equal-weight, mirror portfolios that isolate the raw signal independent of position sizing.

This study parsed the SEC 10-Q and 10-K filings to isolate the *Legal Proceedings* section, along with any additional footnotes containing legal information. GPT-4o was then used to detect the magnitude and direction of legal risk change between "twin sections" (sections from period-adjacent filings detailing the same company). The initial dataset was composed of 1,200 randomly selected firms from the

NASDAQ and NYSE, which were then automatically narrowed to 734 firms that contained "high-quality" SEC extractions. A "high-quality" extraction was defined as extractions with at least 200 characters, an arbitrary count that ensured that immaterial/noise text was not used in our analysis.

A demonstration of the methodology can be seen in the case of the NASDAQ-listed firm AudioEye, Inc. (AEYE). For the SEC filing periods of 2023-09 and 2023-12, a legal risk change score of 1.0—indicating a negative change in the legal situation—was assigned by the LLM. Manually examining these twin filings' *Legal Proceedings* and their associated footnotes, two subtle, yet significant, changes in legal risk could be identified: [1] the introduction of regulatory agency inquiries and [2] a newly acknowledged potential for material adverse effects on future operating results. Specifically, the Q4 filing explicitly referenced exposure to "regulatory agency inquiries," a notable escalation absent from the prior disclosure. Additionally, whereas the earlier filing downplayed potential legal impacts, stating that disputes were "not likely to have a material adverse effect," the subsequent filing explicitly introduced language indicating these matters "could materially affect operating results when resolved in future periods." These nuanced but significant textual shifts supported the LLM's scoring and validated the rationale behind taking a short position in AEYE based on anticipated negative market reactions to heightened legal risk.

In contrast, the methodology identified meaningful legal de-escalation between AudioEye's 2020-12 and 2021-03 SEC filings (assigning a score of -1.0). In the earlier filing, AEYE noted ongoing exposure to "proceedings, lawsuits, regulatory agency inquiries, and other claims," with potential outcomes that "could materially affect operating results," albeit softened by language downplaying overall materiality. However, in the subsequent filing, this language was replaced entirely with the straightforward statement: "Currently, there are no pending material legal proceedings to which the Company is a party to or to which any of its property is subject." This complete removal of qualifying legal risk statements and replacement with an explicit denial of any material litigation represented a clear resolution or dismissal of prior uncertainties. The LLM's score of -1.0 appropriately captured this shift as a strong positive legal development, justifying a long position based on anticipated market relief from diminished legal overhang.

Table 1 and Figure 1 summarize the characteristics of the sample, which comprised over 15,437 firm-quarter observations drawn from the 734 randomly sampled firms from the NASDAQ and NYSE exchanges. The sample contained a diverse set of firms in terms of book-to-market ratio (B/M), market value of equity (MVE), and the natural logarithm of market value of equity, as seen in Table 1, and a robust set of quarterly data for firms, shown in Figure 1. Sample summary statistics were also compared to the S&P 500 and CRSP Total Market indexes to demonstrate how representative our sample was, of the complete U.S. stock market. Table 2 presents the signal's performance and risk characteristics while Table 3 does the same but for the LLM-only test. Following trade statistics, as seen in

Tables 4 and 5, the legal-risk-based signal successfully captured economically meaningful information not reflected in standard asset pricing factors. The long-short strategy, sized according to legal risk change scores and expected position-level volatility, demonstrated a statistically significant average monthly alpha of 0.471% (annualized to 5.80%) when regressed against the Fama-French and momentum factors. The strategy exhibited a mean Sharpe ratio of 0.673 per portfolio, which was slightly less than the S&P 500's ~ Sharpe ratio of 0.81 during the same five-year period. The abnormal returns were not explained by exposures to the market, size, profitability, and momentum factors, but rather reflected in the value and investment factors. The remaining returns may have stemmed from the unique contribution of legal risk disclosures. Complementing these regression findings, the portfolio carried positive returns, reinforcing the practical value of the signal for alpha generation and litigation-aware investing.

The paper begins with an introductory section and a literature review, outlining the existing body of literature and the study's contributions. The data utilized in this study is then discussed, followed by an explanation of the employed methodology. The results and discussion section presents the summary statistics of the chosen sample as well as the specifics of the paper's findings. This section concludes with a comprehensive summary of the key findings, their applications and implications, and the limitations of the paper. Lastly, the paper concludes with a recapitulation of the study and potential further research.

Literature review

Relation to previous literature

The goal of this research was to refine and enhance methods for analyzing textual changes in SEC filings, specifically targeting segments carrying legal significance, such as the *Legal Proceedings* section of annual 10-K and quarterly 10-Q filings. This focus aligned with earlier work by Loughran and McDonald (8), which showed that language in specific sections of SEC filings—such as *Risk Factors* and *MD&A*—contained forward-looking information that is only partially incorporated into prices at the time of release. Notably, while many studies have focused on *MD&A* or *Risk Factors*, relatively little attention has been paid to the *Legal Proceedings* section, despite its potential to signal litigation risk, regulatory scrutiny, and contingent liabilities. This underexplored section—addressed in this study—presents unique modeling challenges, including short length, boilerplate phrasing, and high legal complexity. Meanwhile, Zhu et al. (9) applied sentence-level classification to detect obfuscation in disclosures, and Liu, Zhang, and Wu (10) analyzed the predictive value of forward-looking statements. These studies underscore the importance of localized, context-aware models, but again overlook litigation sections—an area this study specifically addressed.

The foundation of this inquiry is Cohen, Malloy, and Nguyen (1), who demonstrated through their "Lazy Prices" study that measured textual changes in SEC filings using normalized Levenshtein distance and Term Frequency-Inverse Document Frequency (TF-IDF) cosine similarity possesses predictive power regarding future firm performance. Their research also provided the basis for our long-short portfolio building methodology.

Specifically, they created monthly sub-portfolios, which they called "vintages," that longed the top quintile of stocks (non-changers) and shorted the bottom quintile of stocks (changers). They then calculated the aggregate portfolio's monthly change by averaging the three sub-portfolio returns from a given month. Subsequent research has expanded this fundamental insight by exploring advanced natural language processing (NLP) techniques, such as neural embeddings and sentiment analysis, to more effectively quantify semantic shifts and sentiment dynamics within corporate disclosures. Methodologically, this study followed a delta-based approach similar to Yilmaz and Reichmann (4), emphasizing changes in tone and disclosure content across time rather than cross-sectional sentiment levels, consistent with their use of neural embeddings to capture year-over-year shifts in 10-K filings. This enabled a cleaner identification of disclosure evolution while mitigating firm-style or industry-fixed effects. In closely related domains, LLMs have proven especially effective at interpreting nuanced language in small text snippets. For instance, Bunt et al. (11) validated GPT-4o's ability to decode concise psychological text reliably, illustrating that sophisticated inference capabilities can be applied successfully at a micro-text scale.

Early approaches, notably Cohen, Malloy, and Nguyen (1), primarily relied on word-frequency methodologies to detect textual changes, revealing potential economic significance but limited by semantic depth. These early approaches were constrained by their inability to capture context, negation, or semantic subtlety—limitations well-documented by Loughran and McDonald (8) and reinforced by more recent critiques of keyword-counting approaches, which often generated noisy signals in legal contexts due to ambiguous or technical language (e.g., "dismissed," "contingent," or "pending appeal"). Adosoglou, Lombardo, and Pardalos (2) and Yilmaz and Reichmann (4) have since documented the improvement of embedding-based methods. Specifically, neural embeddings more robustly capture semantic nuances compared to traditional word-frequency techniques, significantly improving predictive capabilities for firm-specific returns and risks, such as stock price crash events.

Expanding on these advances, the work by Adosoglou et al. (3) introduced the "Lazy Network" methodology, incorporating neural embeddings into network-based frameworks to identify economically significant textual changes across firms within the same industry or market segment. Their method substantially increased predictive precision and offered a novel dimension to detect firm stability and vulnerability to economic shocks.

Additionally, emerging research has highlighted the role of generative LLMs in systematically analyzing financial texts. Gupta (5), Daimi and Iqbal (6), and Bürgler et al. (7) demonstrated that employing advanced LLMs, such as GPT models fine-tuned for financial data, significantly enhanced the granularity and interpretability of textual analysis in SEC filings and went beyond traditional NLP measures. Related tuned models such as FinBERT (12), SentiBERT, and LegalBERT have shown strong performance in financial and legal sentiment classification. Research by Fatemi & Hu (13) and Rodríguez Inserte et al. (14) confirmed that smaller, fine-tuned LLMs

trained on domain-specific corpora performed on a similar level to large general models in financial sentiment tasks. However, these were often tuned for sentence-level classification, limiting their effectiveness in capturing intertemporal tone drift in long-form regulatory documents. Conversely, a large body of literature, such as Atkinson et al. (15), has shown that large, general-purpose LLMs such as GPT-4o could effectively interpret unstructured free text without task-specific fine-tuning. This corroborates this paper's reliance on GPT-4o for broader semantic understanding across dense, nuanced passages found in SEC filings. These studies collectively underline that advanced LLM approaches, much like this study's methodology, enable a deeper semantic understanding of disclosures, significantly enhancing investment signal extraction and predictive accuracy of market outcomes. The value of analyzing textual disclosures for their informational content has been extensively documented by Gentzkow, Kelly, and Taddy (16), who showed that text-based data could reveal economic signals that were otherwise overlooked by traditional structured data. Similarly, Amel-Zadeh & Faasse (17) demonstrated that investors frequently underreacted to qualitative disclosures—especially those buried in complex legal sections—highlighting the potential alpha embedded in overlooked narrative text.

Contribution to prior literature

This study contributed to existing literature by providing a nuanced and focused analysis of specific text portions of SEC filings that are especially pertinent to assessing corporate legal and financial risks. Specifically, this research emphasized the *Legal Proceedings* section, a comparatively under-researched area within SEC filings, offering potential novel insights into the predictive relevance of litigation-related disclosures. Furthermore, the focus on litigation-related footnotes and section notes is less explored by academia, adding another dimension of contribution to the methodology.

Methodologically, the study leveraged recent advancements in NLP by combining semantic shifts and sentiment analysis, specifically capturing the two by prompting the chosen LLM (GPT-4o) to assess the directional change in legal risk between the examined 10-Ks. This approach is supported by Atkinson et al. (15), who found that GPT-4o was proficient in nuanced, domain-specific textual analysis despite not being fine-tuned. Our study utilized LLM technology to compute the implications of these changes in SEC filings via legal risk change score differentials in modified textual elements. Thus, this methodology differs from previous methods that rely solely on embedding distances or simple sentiment metrics.

Finally, by evaluating the resulting legal-risk change metric through a structured long-short portfolio—long on firms with notably improved legal semantics or tone, and short on firms displaying significantly negative shifts—this research contributed practical insights into actionable investment strategies. Furthermore, the use of a hyperbolic tangent-weighted average for calculating both sub-portfolio and portfolio returns as opposed to a nonweighted average further distinguishes this study from previous research. Rebalancing, backtesting, and incorporating standard factor benchmarks increased both the rigor and practical relevance of the findings. In summary, this research

meaningfully extended prior literature through its specialized, rigorous, and economically focused textual analysis of SEC filings through the lens of LLMs, demonstrating significant potential to refine market prediction and enhance investor decision-making.

Materials and Methods

Data

Using the SEC EDGAR API, this study pulled the 10-Q and 10-K SEC filings from February 2020 to May 2025 for 1,200 randomly selected firms from the NASDAQ and NYSE. The *Legal Proceedings* sections (Item 1. for 10-Qs, Item 3. for 10-Ks) and any accompanying footnotes disclosing material litigation updates were then extracted. These sections were located by a Python program which identified specific headings and delimiters (text boundaries or markers) commonly used in filings. After parsing the full text of each filing, the extraction logic excluded general, standardized, or administrative content and selectively retained disclosures containing legal terminology and substantive litigation detail. To ensure data quality, extracted sections were cleaned to remove formatting noise and non-substantive content, and filings with insufficient or missing legal disclosures were excluded from the analysis. Due to this specificity, only 734 tickers with "high-quality extracts" (extracts with at least 200 characters) of the original 1,200 firm sample made it past the legal text extraction step and were included in the study's final LLM analysis. This meant that the strategy had a total of 734 unique tickers to trade in any given month. While the SEC sample period was narrower than the return and firm characteristic datasets, it was sufficient for capturing meaningful changes in disclosure language. Although future versions of this research should look to expand the sample, extending its time frame might risk picking up signals that no longer exist due to technological advancements and market efficiency improvements.

Adjusted close price data—used to compute daily returns and 90-day historical volatility for all firms in the sample—was sourced from the open-source Yahoo Finance Python library. Firm-level summary statistics, such as market value of equity and book-to-market Ratio, were pulled using the Zacks API, specifically Zacks' Fundamental Collection B. Both datasets served distinct purposes: Zacks was used solely for describing the sample, while Yahoo Finance return data was used in the empirical analysis. As such, any mismatch in their periods was inconsequential.

Scoring methodology

This research employed a novel methodology by utilizing GPT-4o to interpret legal text found in the Legal *Proceedings* section (and relevant footnotes) extracted from quarterly (10-Q) and annual (10-K) SEC filings. After parsing and isolating the legally significant textual content from consecutive filings, each pair of sequential quarters for a given firm was input into GPT-4o. The prompt presented to the model was split into three main segments: context, instruction, and scoring criteria. The context given to the model was intended to frame its position as a legal professional competent in tasks related to the methodology. Specifically, the context was as follows:

"You are a professional legal risk analyst specializing in SEC 10-K/10-Q 'Legal Proceedings' disclosures. Your role is to

evaluate changes in litigation tone and implied legal risk by comparing 10-K/10-Q excerpts from the same company published one quarter apart." For instruction, the model was told to evaluate the given twin filings and instructed on how to apply its role as a "legal risk analyst" in the context of this study's methodology:

"Your task is to assess the directional change in tone between the two disclosures, reflecting shifts in severity, specificity, and materiality of legal risk. You should ignore stylistic differences or general disclaimers and instead focus on whether new language suggests: Escalation (e.g., new government investigations, class actions, regulatory action, trial deadlines, large settlements, or fines), De-escalation (e.g., settlements reached, dismissals, cases closed), No meaningful change, Respond with a single numerical score between -1.0 and +1.0 in intervals of 0.1"

A portion of the instruction was also presented at the end of the prompt to remind the LLM to remain as consistent as possible, measuring only the relative shift in risk between twin filings instead of the absolute risk of a filing, and ensuring that it would only output numbers. After the main set of instructions, the scoring criteria were; +1.0 = Very strong increase in legal risk, +0.5 = Moderate increase in legal risk, 0.0 = No material change, -0.5 = Moderate decrease in legal risk, -1.0 = Very strong decrease in legal risk.

This score characterized the firm's relative change in legal risk exposure: a negative numeric score was indicative of an improvement (a reduction in perceived legal risk), whereas a positive score denoted a deterioration (an increase in perceived legal risk).

Unlike conventional text analysis, which is typically reliant on word frequency-based methodologies such as cosine similarity and often overlooks nuanced shifts in language, GPT-4o leverages its neural architecture to interpret textual subtleties contextually. Internally, the model generates neural embeddings that encapsulate comprehensive semantic understanding, allowing nuanced detection and quantification of subtle legal disclosure changes. This method captures both explicit litigation events and subtle yet material changes in tone or phrasing indicative of changing legal circumstances. While qualitative assessment inherently introduces some subjectivity relative to strictly numeric metrics, consistency was prioritized by setting the model's temperature parameter to 0.1, which limited the variability of the model and ensured highly deterministic behavior. A higher temperature would cause the LLM to output more randomly.

Portfolio construction and backtesting

Once the LLM outputs were generated, we created 64 monthly long-short sub-portfolios (each sub-portfolio representing a specific month of trading, not a division of firms), where all firms with non-zero scores were included in its period-associated sub-portfolio. Firms with a legal risk change score of 0 were excluded from the trading process entirely to ensure materially informative signals. Stocks with negative legal risk change scores—indicating improving legal conditions—were assigned long positions, while stocks with positive scores—indicating deteriorating legal conditions—were assigned short positions.

This methodology contrasts with many previous studies, which used threshold filters (i.e. quantile, quintile, decile, etc.) to assign trading positions. We used all non-zero scores as a result of the LLM's selectivity of these scores, since 81.29% of outputs were zero.

Positions on each sub-portfolio were established at the beginning of each month and held for a fixed three-month period. Consequently, in any given month other than the first two, there would exist three overlapping sub-portfolios. In the scenario where a firm had a positive or negative legal risk change in two consecutive quarter-over-quarter observations (e.g. an increase in legal risk between Q1-Q2 and then another increase in legal-risk from Q2-Q3), the aggregate portfolio would accumulate that stock as the overlapping sub-portfolios would have both bought or shorted more as opposed to just holding and maintaining the original position size. However, it is important to consider that the sub-portfolios themselves are independent and their holdings are never truly related: they are only compiled on the final month of the sample to analyze the strategy across the aggregate portfolio (taking into account all of the sub-portfolios). Position sizes within each sub-portfolio were determined using its legal risk change score and its 90-day historical volatility. All portfolios and results were recorded on an unlevered basis. Because single-name caps, volatility-based scaling, and side risk limits were enforced, realized gross exposure could fall below the nominal target—meaning any residual portion of that sub-portfolio was carried as cash until the next rebalancing event).

As an illustrative example; two-stock sub-portfolio example (composed of only Firm A and Firm B), in month *t*, suppose Firm A receives $S_A = +1.0$ (legal risk increased) and Firm B receives $S_B = -1.0$ (legal risk decreased). A is shorted and B is longed in the *t* month sub-portfolio with the weighting of position A and B depending on the weighting system explained below. Regardless of the number of positions, positions never make up more than 100% of a sub-portfolio's allocation and entries with $S = 0$ are omitted for that sub-portfolio. Once opened, positions for stock A and B are held for three months and then closed without any influence of other sub-portfolios or other factors (positions are never closed early). Since sub-portfolios are independent of each other (only put together at the end of the analysis to examine the total performance of the strategy), whatever scores the subsequent sub-portfolio after month *t* receive are irrelevant to month *t*'s holdings (no prediction of the next score is used or required). Short exposure is implemented via ordinary stock borrowing practices. The backtest is unlevered, and when caps or risk limits prevent full deployment, the residual notional remains in cash until the next rebalancing event.

Specifically, initial position weights ($w_i^{initial}$) were calculated as the product of the portfolio's target volatility (σ_{target}) and the absolute value of the legal risk change score (S_i) and scaled inversely by each stock's historical ninety-day volatility (σ_i). Target volatility was set to a conservative 5% target portfolio-wide exposure, anticipating elevated volatility exposure from trading around legal disclosures and aiming for minimal portfolio drawdowns. Ninety-day historical volatility was calculated as the stock's price volatility over the

preceding ninety trading days. As such, a greater legal risk in either direction would increase the weighting, while a greater volatility would decrease the weighting. The raw weighting equation is defined as,

$$w_i^{initial} = \frac{\sigma_{target}}{\sigma_i} \cdot |S_i| \quad \text{(eq1)}.$$

Note that the initial weightings are all positive due to the absolute value but the sum of these initial weightings do not add up to 1 (addressed later). To manage extreme exposures and reflect diminishing marginal conviction of lower scores, these raw weights were passed through a nonlinear S-curve transformation using the hyperbolic tangent (tanh) function. This approach moderated position sizes, reducing the influence of extremely high or low scores and sensitivity to outliers. This transformation was scaled by a coefficient (κ), itself a function of the portfolio's target volatility and the sub-portfolio-wide average ninety-day historical volatility, ($\bar{\sigma}$) for the month, was defined as, $\kappa = \kappa_0 \cdot \frac{\sigma_{target}}{\bar{\sigma}}$ (eq2).

The kappa coefficient adjusts the steepness of the S-curve based on perceived market volatility at the time, dynamically scaling position weights to maintain consistent risk exposure across different market environments. κ_0 is the initial S-curve scaling factor and was set to 1.5 to provide managed steepness to the transform. Once calibrated, the final kappa coefficient and the previously computed raw position weight were used to establish each stock's raw weight defined by $w_i^{raw} = \tanh\left(\kappa \cdot w_i^{initial}\right)$ (eq3).

These transformed raw weights were then capped at 0.2 per position (to balance exposure across a broad set of firms while mitigating concentration risk) before being proportionally normalized to create final weights within each sub-portfolio, defined by,

$$w_i^{norm} = \frac{w_i^{raw}}{\sum_{j=1}^{m} w_j^{raw}} \quad \text{(eq4)}.$$ Here, $\sum_{j=1}^{m} w_j^{raw}$ represents the sum of all the raw weights in a single sub-portfolio. Note that $\sum_{i=1}^{m} w_i^{norm}$ now equals 1 by design. The return on a sub-portfolio was calculated by taking the weighted average of the long positions' returns minus the weighted average of the short positions' returns as,

$$R_{a,i} = \sum_{j=1}^{k} \left(w_{j,i}^{norm} \cdot R_{j,i} \right) - \sum_{j=k+1}^{m} \left(w_{j,i}^{norm} \cdot R_{j,i} \right) \quad \text{(eq5)}.$$

$R_{a,i}$ denotes the return on the *a*-th sub-portfolio in period *i* (adjusted for realistic trading conditions via a 0.10% transaction cost at both position entry and exit), $w_{j,i}^{norm}$ is the normalized weighting for stock *j* in period *i*, and $R_{j,i}$ is the return for stock *j* in period *i*. Stocks with index 1 to *k* were longed (negative legal risk change), while stocks with index $k + 1$ to *m* were shorted (positive legal risk change) in this formula. Hence, there are *m* total stocks in the *a*-th sub-portfolio in period *i*.

The aggregate portfolio's return for a given month was expressed as the weighted average of the three sub-portfolios active in that month. To calculate the weights of each sub-portfolio, the normalization process was invoked again.

Let $\omega_{i,i}^{raw}$ denote the raw weighting of the *i*-th sub-portfolio in period *i* (in that order for the subscript notation). $\omega_{i,i}^{raw}$ was calculated as,

$\omega_{i,i}^{raw} = \sum_{j=1}^{m} w_{j,i,i}^{raw}$ (eq6). Here, each $w_{j,i,i}^{raw}$ is the raw weighting of one stock in the *i*-th sub-portfolio during period *i* (in that order for the subscript notation). Then, similar to the intraportfolio normalization, the weights were normalized as, $\omega_{i,i}^{norm} = \frac{\omega_{i,i}^{raw}}{\omega_{i,i}^{raw} + \omega_{i-1,i}^{raw} + \omega_{i-2,i}^{raw}}$ (eq7). $\omega_{i,i}^{norm}$ denotes the normalized weighting of the *i*-th sub-portfolio in period *i* (in that order for the subscript notation). The denominator represents the sum of the weightings from the *i*-th, *i* - 1-th, and *i* - 2-th sub-portfolios in period *i*. Note that $\omega_{i,i}^{norm} + \omega_{i-1,i}^{norm} + \omega_{i-2,i}^{norm}$ now equals 1 by design. Finally, the aggregate portfolio's return in the period *i* is given by the following equation, $R_{agg,i} = \omega_{i,i}^{norm} \cdot R_{,i,i} + \omega_{i-1,i}^{norm} \cdot R_{,i-1,i} + \omega_{i-2,i}^{norm} \cdot R_{,i-2,i}$ (eq8).

The aggregate portfolio's return in the first and last months was not calculated because there were only one and two sub-portfolios active during those months, respectively, skewing the precision of the study. Hence, there were a total of 62 observations for regression.

An example is presented for methodological clarity around the weighting. Imagine the three sub-portfolios active in July of 2020 – one was just constructed, one was constructed in June, and one was constructed in May. For simplicity, these sub-portfolios are named A, B, and C, respectively, and each sub-portfolio takes four positions. Sub-portfolio A has raw position weightings: 0.02, 0.03, 0.03, 0.05 (total of 0.13). Sub-portfolio B has raw position weightings: 0.01, 0.03, 0.05, 0.06 (total of 0.15). Sub-portfolio C has raw position weightings: 0.01, 0.03, 0.05, 0.05 (total of 0.14)

To calculate each sub-portfolio's returns for July 2020, each sub-portfolio's raw weightings would first be normalized. For A, the normalized weightings would be 0.02/0.13, 0.03/0.13, 0.03/0.13, and 0.05/0.13. For B, the normalized weightings would be 0.01/0.15, 0.03/0.15, 0.05/0.15, and 0.06/0.15. Then, each of those weightings are multiplied by the return on the associated stock that month and summed together.

To calculate the aggregate portfolio's returns for July 2020, a similar process was followed. Instead of normalizing stock weightings, each sub-portfolio's overall weighting was normalized. To calculate each sub-portfolio's normalized weighting, first, all the total weights were summed (0.13 + 0.15 + 0.14 = 0.42) then each sub-portfolio's total raw weightings was divided by this total. A's normalized weighting would be 0.13/0.42, B's would be 0.15/0.42, and C's would be 0.14/0.42. Finally, each of these weightings would be multiplied by the associated sub-portfolio's return that month, as calculated in the previous step, and finally summed.

LLM-Only Test

We also tested the validity of the LLM legal-risk signal by itself. To demonstrate the contribution of the LLM scores apart from the main model's portfolio-engineering aspects (risk parameters and non-linear transform), we retained the original portfolio and backtesting framework—monthly formation of sub-portfolios, three-month overlapping holds (three sub-portfolios active after the first two months), and symmetric transaction costs of

0.10% at both entry and exit. However, in this LLM-only specification, all volatility-based position sizing and nonlinear transforms were removed such that position weights were equal within each side and then normalized by side. Furthermore, in this test, two portfolios were formed and compared: a normal portfolio (henceforth called "Equal-Weight Portfolio 1" and a "mirrored" or "inversed" portfolio (henceforth called "Equal-Weight Portfolio 2").

As in the main model, each month we formed long-short sub-portfolios from all firms with non-zero LLM scores for that month's filing. In the core methodology, negative scores (legal risk decreased) indicate long candidates and positive scores (legal risk increased) indicate short candidates. Equal-Weight Portfolio 1 took these signs as-is (long negatives, short positives) while Equal-Weight Portfolio 2 took the exact opposite sides (short negatives, long positives) on the same names, magnitudes, and time frames. For a given sub-portfolio in period i, if there are L_n longs and S_n shorts, each long received weight $\frac{+1}{L_n}$ and each short received weight $\frac{-1}{S_n}$. Side totals were normalized so that the long side and short side summed to +1 and -1 respectively. After portfolio construction and equal weighting, this LLM-only signal test followed the exact same backtesting and returns methodology as the main method.

Overall, by construction, the mirror portfolio (Equal-Weight Portfolio 2) flipped the signs of the normal portfolio's positions within each sub-portfolio (same names, equal magnitudes, same periods, but opposite sides). Consequently, the two series were mechanically highly negatively correlated and comparing the two helped to determine the strength of the pure legal risk change signal and ensure that the core methodology's returns were not entirely random. Additionally, it should be considered that in this test we deliberately removed position sizing: every non-zero score was traded at the same weight, hence a score of +1.0 (or -1.0) carried no more capital than +0.1 (or -0.1). This choice was conservative by design and put both equal-weight portfolios at a disadvantage relative to (1) the main model, which scaled positions by signal strength and volatility, or (2) thresholded designs (e.g., deciles/quintiles) that traded only the strongest scores.

These portfolios traded all non-zero scores to avoid selection/tuning, keeping all else aligned with the main methodology, to determine whether the sign of the LLM legal-risk score alone influenced returns. Accordingly, these equal-weight portfolios were expected to carry lower cumulative returns, more volatility and drawdowns, and higher skew as they were put at the inherent disadvantage of being subjected to large exposure to even the LLM's least confident trades. As a sensitivity check, winsorized variants of both equal-weight portfolios were also recorded. Winsorization reduces the influence of single-name outliers (which the equal-weight portfolios significantly exhibited) without introducing selection rules or additional parameters. Specifically, symmetric clipping at the position-level monthly return at 2.5% and 5% tails before aggregating to sub-portfolio and portfolio returns was performed. Since the position-level monthly return contained roughly 7,450 observations, the 2.5% winsorized variant clipped ~ 372 values while the 5% variant

clipped ~ 746. Portfolio formation, side assignment, holding period, costs, and rebalancing remained unchanged and this step was only meant to be overlaid with original results to demonstrate performance after accounting for outliers. Results are shown alongside the originals in Figure 3 and Table 3 and do not alter the directional conclusions.

Statistical evaluation

The financial significance and robustness of the portfolio strategy was evaluated using an asset pricing regression framework, specifically the Fama-French five-factor model, augmented by the momentum factor—specifically, Carhart's momentum factor, which was denoted as "UMD". Monthly returns from the aggregate long-short portfolio were regressed against standard risk factors to assess whether the observed returns could be explained by established asset pricing factors or whether they represented novel, risk-adjusted performance attributable to changes in legal risk as identified by this research's methodology. The factor returns used in the regression were simulated using empirically grounded parameters, with each factor's monthly mean and volatility calibrated to typical historical values documented in asset pricing literature. This included, for instance, an average monthly market risk premium of approximately 0.80% per month (or 10.40% per year) and a risk-free-rate of 0.35% per month (derived from the ~ 4.36% annual 10-year treasury yield). The regression equation was defined as,

$$R_t - R_{ft} = \alpha_t + \beta_{MKT} \cdot \left(R_{MKTt} - R_{ft}\right) + \beta_{SMB} \cdot SMB_t + \beta_{HML} \cdot HML_t + \beta_{RMW} \cdot RMW_t + \beta_{CMA} \cdot CMA_t + \beta_{UMD} \cdot UMD_t + \varepsilon_t$$

(eq9).

The *Notations* section can be referenced for variable definitions. The primary focus was the intercept (α), representing the portfolio's average monthly risk-adjusted return. A significantly positive α would indicate that the legal risk-based investment signal provided predictive power and economic value beyond traditional factors. Factor loadings (β coefficients) were also examined to interpret portfolio exposures to market risk (MKT), size (SMB), value (HML), profitability (RMW), investment activity (CMA), and momentum (UMD).

The significance of regression coefficients was assessed using t-statistics and associated p-values. Further robustness checks included analyses of subperiod stability, lagged predictability to test for look-ahead bias, and size-sorted portfolio regressions to examine heterogeneity (variation across groups or units) in signal effectiveness. Finally, various portfolio metrics were computed to quantify the overall performance relative to residual volatility, providing additional context on the investment signal's consistency and reliability. In particular, the alpha (the excess of a portfolio's returns after stripping out expected returns from its market and factor exposures) and the Sharpe ratio (returns minus the risk-free rate per unit of total risk, measured by volatility) were particularly of interest. The full statistical evaluation was not performed on the equal-weight portfolios, since their role was attribution rather than investability.

Results and Discussion

Summary statistics

Table 1 and Figure 1 present the summary statistics and data completeness for various financial metrics and firm characteristics in the analyzed sample. The sample's statistics are then compared to two main indexes: the S&P 500 and the CRSP Total Market Index. In Table 1, the summary statistics for key financial variables, namely the book-to-market ratio, market value of equity, and the natural logarithm of market value of equity, are shown for the observed firms, totaling approximately 15,437 observations. The discrepancy between the potential maximum number of observations and the actual figure arises due to inevitable gaps in the Zacks Fundamentals database.

Table 1. Various summary statistic metrics for the sample, including their minimum, 10th percentile, 25th percentile, median (50th percentile), 90th percentile, maximum, and mean value, alongside the number of observations for each metric.

	Minimum	10th	25th	50th	75th	90th	Maximum	Mean	Obs
Book-to-Market	-131.8	0.043	0.185	0.454	0.877	1.371	71.066	0.598	15,437
Market Value of Equity (in millions)	0.06	48.9	210.9	1,113.4	6,144.3	21,990.6	3,761,827.5	15,192	15,553
Natural Log of Market Value of Equity	-2.81	3.89	5.35	7.02	8.72	10.00	15.14	7.00	15,553

Table 1 highlights considerable diversity across the sampled firms, with significant variations in size and market valuation. For instance, B/M exhibits a wide range from a minimum of -131.77 to a maximum of 71.066, reflecting substantial differences in financial valuation among the included firms. Similarly, MVE ranges from 0.06 million to 3,761,827.5 million, emphasizing the significant variation in firm size within the dataset. This range indicates a positively skewed distribution, underscored by the mean MVE at 15,191.975 million, which notably exceeds the median value of 1,113.41 million. However, the mean natural log of MVE (6.999) appeared roughly equivalent to its median (7.015), suggesting a distribution not heavily influenced by larger firms. For context, the same natural log of MVE statistics computed on the CRSP Total Market Index yielded a mean of 7.082 and a median of 7.128, while the S&P 500 showed a mean of 10.428 and a median of 10.276. Our sample's central tendency thus closely matched the broad U.S. market and was well below the large-cap S&P benchmark. Consistent with this observation, our sample's firms fell near the CRSP quartiles (25th = 5.539, 75th = 8.556) and well below the S&P quartiles (25th = 9.678, 75th = 11.007), indicating representation across small- and mid-cap issuers rather than the dense concentration of mega-caps as seen in the S&P.

As illustrated in Figure 1, the number of quarterly observations available per firm varied across the 734 company sample included in the final analysis. The majority of the firms (541 to be precise) demonstrated consistent reporting,

detailing 25 quarters. Fewer firms contributed fewer observations, with 19 firms reporting 24 quarters, 13 firms covering 23 quarters, 32 firms detailing 22 quarters, and 12 and 17 firms reporting 21 and 20 quarters each, respectively. A significant decrease was observed for the numbers of firms that reported 19 quarters or less, summing to 38 companies. Variations and discrepancies in the number of available quarters per firm could primarily be attributed to unavoidable limitations within the Zacks Fundamentals database. This distribution suggested a robust consistency in quarterly data availability among the firms analyzed, facilitating uniform and reliable analysis.

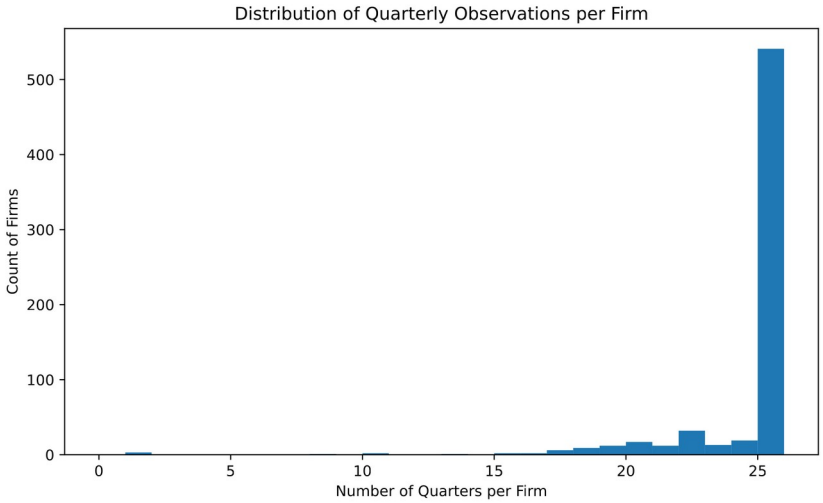

Figure 1. Distribution of the number of quarters per firm for the sample's firms.

Portfolio results

To evaluate the performance of the legal-risk-based signal after adjusting for market frictions and transaction costs (estimated at a total of 0.20% per trade), summary statistics and cumulative returns of the long-short portfolios were analyzed. Figure 2 shows the cumulative returns of the entire strategy across the sample period, while Table 2 summarizes the performance metrics of the portfolios. The cumulative return plot shows the total compounded performance of the long-short portfolio over time, with each point representing the net return from inception through that date. As seen in Figure 2, the signal exhibited relatively consistent and steady cumulative growth, with minimal spikes in both volatility and in sharp drawdowns. The cumulative return across all of the portfolios totaled to 51.43% whereas the S&P 500 returned roughly 83% during the sample period. Although cumulative returns did not outperform the S&P, the generation of alpha (and potentially uncorrelated returns to S&P) presents its own benefits. Realized returns occasionally reflect modest cash balances due to position caps and risk-based scaling, which dampen raw returns but help to contain drawdowns.

The steady upward trajectory of cumulative returns suggested that the legal-risk-based

signal effectively captured economically significant information from litigation disclosures. Although the signal did experience occasional minor episodic drawdowns (e.g., the drawdown in 2024), the signal's reliability was particularly evident through the period coinciding with the COVID-19 pandemic—which overlapped with the first years of the sample and had widespread effects on corporate legal environments (18). During this period, companies across various sectors faced unprecedented litigation risks tied to supply chain disruptions, labor force adjustments, bankruptcy filings, and regulatory scrutiny—many of which triggered expanded or urgent legal disclosures in 10-Q and 10-K filings. External research by Loughran and McDonald (8) identified that firms were generally late in disclosing pandemic-related risks, suggesting an eventual surge of legally material disclosures. The heightened volatility and abnormal profitability observed circa 2024 may thus reflect a temporary concentration of material legal events on the back-end of the pandemic, making text-based signals particularly contributory. However, the subsequent sharp drawdown suggested that part of the spike was transitory—driven by market overreaction or the model's response to atypical disclosure patterns. This underscored the signal's responsiveness to major legal developments, yet also highlighted how exogenous shocks like COVID-19 could amplify or distort disclosure-based signals, requiring caution in real-world deployment and strategy design. The portfolios' consistent performance suggested its effectiveness at withstanding exogenous shocks such as the pandemic without substantial distortion (and even potential improved performance). Thus, the signal demonstrated practical robustness and stability for investment applications, even during times of increased legal disclosure complexity and market stress.

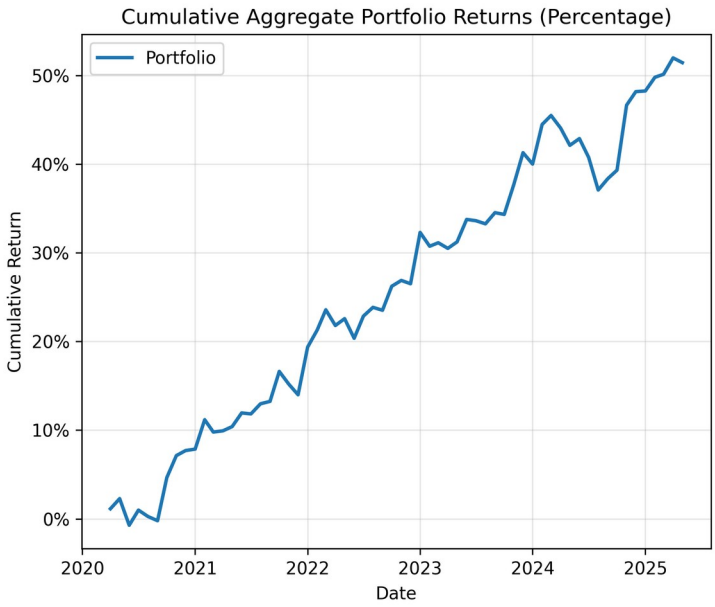

Figure 2. Cumulative compounded return of the strategy, 2020-02 to 2025-05, net of trading costs.

Reflected in Table 2, on average, the aggregate portfolio yielded a return of 0.686% per month with a standard deviation of 1.73%. While seemingly modest, given that unique portfolios were formed each month (based off of the most recent SEC filings), the cumulative and overlapping effect of the portfolios generated economically substantial return magnitude in the broader scheme of the strategy. This is evident from the 51.43% cumulative return across all portfolios of the methodology, seen in Figure 2. Furthermore, these statistically significant returns (according to the return metadata) suggested that changes in legal risk disclosures—as quantified through GPT-4o—may contain valuable, market-relevant information.

Table 2. Investment signal and portfolio-level performance and risk metrics.

Metric	Mean portfolio return (monthly %)	Sharpe (all sub-portfolios annualized)	Win rate (%)	Win-to-Loss ratio	Annualized volatility (%)	Maximum drawdown (%)	Skewness	Kurtosis
Value	0.69	0.67	64.53	1.64	6.00	5.78	0.63	0.58

The strategy exhibited an annualized Sharpe ratio of 0.673, suggesting weak efficiency on a risk-adjusted basis. While the signal's figure did not reach the high benchmarks typical of superb strategies, it remained promising given the portfolio's absolute returns, the novelty of the information source, and the strategy's new, unorthodox signal. It may be helpful to consider this investment strategy's Sharpe ratio in comparison to the S&P 500, which serves as a robust benchmark and demonstrated a Sharpe ratio of approximately 0.81 over the same five-year period.

The portfolio's volatility (5.992%) was low-to-moderate, and it experienced a maximum drawdown of 5.776%, highlighting occasional volatility and exposure to downside risk. The win rate of 64.526%—the percentage of months with positive returns—suggested a moderate but reliable edge, further evidenced by a favorable average win-to-loss ratio of approximately 1.642. Though modest, this directional consistency, when paired with positive returns that not only occurred more frequently but were also larger in magnitude on average, suggested a non-random pattern of success. Additionally, the positive skewness (0.625) and elevated kurtosis (0.575)—statistical measurements that describe the tailing propensity and asymmetry of a probability distribution—reflected the portfolios' minor tendency toward occasional high positive returns relative to a typical Gaussian distribution. Although the skewness indicated that much of the strategy's returns may have come from less frequent, more extreme profits, this behavior is consistent with a signal targeting high-impact, but typically more sporadic, legal-disclosures events. Such skewed and leptokurtic or platykurtic characteristics are common in financial return distributions, which rarely conform to perfect normality. However, it should be noted that despite not being in the typical range of -0.5 to 0.5 of symmetrical skews, the exhibited

skewness is still considered low and mostly symmetrical. The same applies to the kurtosis which is classified as platykurtic due to its sub-three value, indicating lighter tails and fewer outliers.

All results were reported on an unlevered basis. However, the strategy's lower delta exposure, low market beta, and moderate volatility with comparatively shallow drawdowns suggested that calibrated leverage could be used to scale returns to a desired risk tolerance without materially changing the strategy's risk characteristics under ordinary conditions. In practice, leverage simply scales the long and short books proportionally while maintaining net exposure. Expected return and volatility rise approximately in proportion to the leverage multiple, while risk-adjusted performance (e.g., Sharpe) is unchanged in the absence of frictions. Although, standard caveats still apply: financing and borrowing costs, margin requirements, liquidity during stress, and tail risk around legal events must be managed.

Overall, while the strategy's raw return profile was compelling, its occasional exposure to elevated volatility and deep drawdowns solicit wariness. This dichotomy emphasizes the importance of pairing such signals with robust risk management protocols in applied settings. Nevertheless, the results offered strong empirical support for the predictive relevance of textual legal disclosures, especially when measured using advanced LLM methods.

Comparison of Equal-Weight portfolios

As in the core methodology, raw performance of the legal-risk-based signal after adjusting for market frictions and transaction costs (estimated at a total of 0.20% per trade) were analyzed. However, these results solely pertained to the trades taken by the equal-weight portfolios, outlined in the *LLM-Only Test* section of this study. Results were demonstrated by way of summary statistics and cumulative returns of the two mirrored long-short, equal-weight portfolios (Equal-Weight Portfolio 1 and 2). Figure 3 shows a comparison of the cumulative, compounded performance of the Equal-Weight Portfolio 1 (normal) and Equal-Weight Portfolio 2 (mirror) over the same period and market frictions of the main methodology. Similarly, Table 3 compares the two equal-weight portfolios' summary, return, and risk statistics throughout the sample.

As seen in Figure 3, over the sample, the two equal-weight portfolios demonstrated a clear directional split: Equal-Weight Portfolio 1 delivered a cumulative return of +33.76%, whereas Equal-Weight Portfolio 2 returned -86.24%. For context, over the same window the S&P 500 Total Return rose ~83%. Hence, Equal-Weight Portfolio 1's +33.76% was modest in absolute terms but still directionally informative relative to its mirror. This gap was expected—the LLM-only portfolios carried equal-weight positions and low-correlation to the market and the test's purpose was attribution, not benchmark outperformance. Figure 3 also plots 2.5% and 5% winsorized versions of both curves. Clipping extreme monthly position-level returns smoothed the paths and reduced drawdowns and volatility for both portfolios, resolving the mid-2021 spike/crash visible in the originals—which had potential to bias and skew interpretation of the LLM-only test. The normal equal-weight portfolio reached a similar terminal level under winsorization with a steadier trajectory, while

the mirror portfolio showed substantially shallower losses (less negative terminal value). Crucially, the directional ordering remained unchanged—Equal-Weight portfolio 1 beat Equal-Weight Portfolio 2 across all iterations—indicating that the LLM legal-risk direction drove performance beyond isolated outliers.

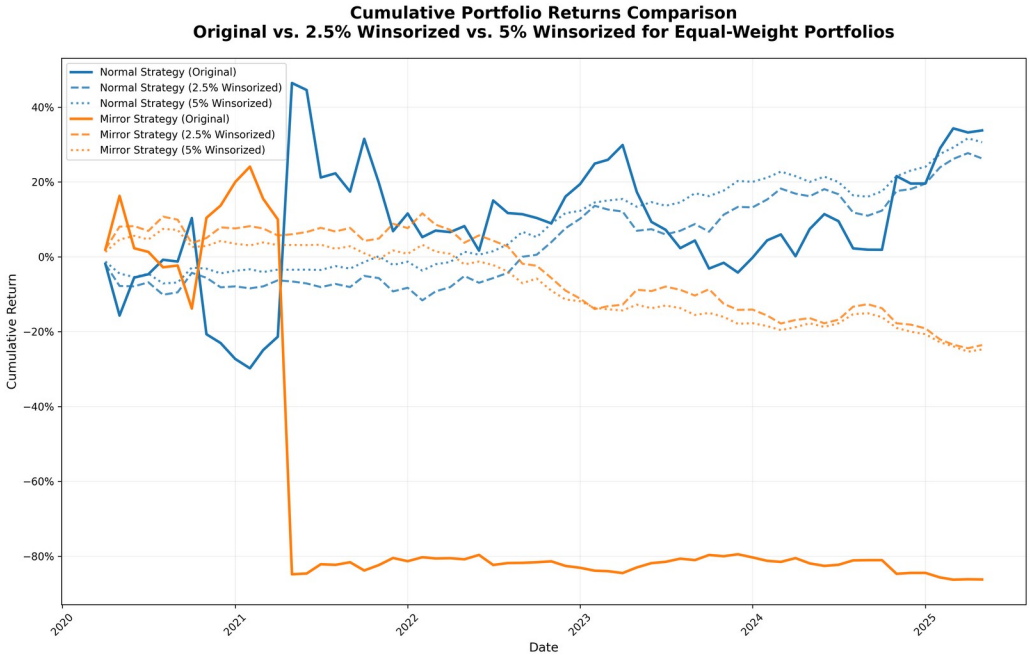

Figure 3. Cumulative returns of the LLM-only strategy, comparing Equal-Weight Portfolio 1 against Equal-Weight Portfolio 2, with original, 2.5% winsorized, and 5% winsorized series, 2020-02 to 2025-05, net of trading costs.

After some early volatility—including the pandemic period noted in Figure 2's analysis—the two curves diverged sharply by mid-2021 and remained far apart through the sample. Because the mirror construction flipped the signs of the same trades at equal magnitudes each month, the series were mechanically highly negatively correlated (~ -1). The resulting directional spread—defined as Equal-Weight Portfolio 1 minus Equal-Weight Portfolio 2—was positive and totaled ~ +120% on a cumulative, compounded basis. This was precisely what the attribution test was designed to detect: when trades were determined solely by the sign of the LLM legal-risk score, the long risk-decreases / short risk-increases logic systematically outperformed the reverse. The *Regression results* section provides complementary, factor-adjusted evidence (the model's alpha) under our research's final model (with sizing and risk controls), reinforcing that the performance gap in Figure 3 reflected the information content of the LLM legal-risk signal rather than coincidence.

As shown in Table 3, the Equal-Weight Portfolio 1 earned an average monthly return of 1.135%, whereas its mirror (Equal-Weight Portfolio 2) recorded -1.135% over the same window. Since the two portfolios held the same names with opposite sides of the trade and equal magnitudes, their risk was essentially

identical and their returns were mechanically opposed. This appears in the metrics: both portfolios exhibited the same annualized volatility (45.713%) and the same (very high) kurtosis (28.792), while their skewness flipped sign (+4.331 for the normal, -4.331 for the mirror). The mirror's negative drift also explained its substantially deeper maximum drawdown (88.957%) relative to the normal portfolio (36.379%). These elevated volatility exposures, drawdowns, and skewed profiles were to be expected given that the equally weighted positions in these portfolios subjected them to large swings in Profit-and-Loss, dramatic losses, and immense skew from outlier returns. Winsorizing position-level monthly returns at 2.5% and 5% materially compressed these tails—volatility and maximum drawdowns decreased to single-digit levels for both portfolios, and skewness/kurtosis moved toward 0—while the normal-over-mirror performance gap remained similar.

Table 3. LLM-only equal-weight portfolio performance and risk metrics, including original, 2.5% winsorized (abbreviated as *Wins.*), and 5% winsorized results for Portfolio 1 (normal) and Portfolio 2 (mirror).

Metric	Portfolio 1 - Original	Portfolio 1 - 2.5% Wins.	Portfolio 1 - 5% Wins.	Portfolio 2 - Original	Portfolio 2 - 2.5% Wins.	Portfolio 2 - 5% Wins.
Mean Portfolio Return (monthly %)	1.135	0.405	0.445	-1.135	-0.405	-0.445
Sharpe (all sub-portfolios annualized)	0.206	0.080	0.200	-0.390	-1.092	-1.674
Win Rate (%)	48.387	56.452	61.290	51.613	43.548	38.710
Win-to-Loss Ratio	1.522	1.181	1.227	0.657	0.846	0.815
Annualized Volatility (%)	45.713	8.299	5.698	45.713	8.299	5.698
Maximum Drawdown (%)	36.379	9.867	5.922	88.957	32.299	30.543
Skewness	4.331	-0.272	-0.089	-4.331	0.272	0.089
Kurtosis	28.792	0.116	-0.191	28.792	0.116	-0.191

On a risk-adjusted basis, the LLM-only specification was structurally undermined. The normal equal-weight portfolio's annualized Sharpe was 0.206 (mirror -0.390). Its win rate was slightly below 50% (at 48.387%), yet the win-to-loss ratio was 1.522, indicating that gains, when they occurred, tended to be materially larger than losses. The mirror showed the opposite profile (win rate 51.613% with a win-to-loss ratio 0.657), consistent with the two series being mirrors: small, frequent moves could favor either side month-to-month, but the directional drift favored the normal portfolio over time. Under winsorization, the normal portfolio's win rate increased into the mid-50s/low-60s, whereas the mirror's win rate decreased into the low-40s/high-30s. Since the mirror's mean return stayed negative while its volatility was sharply reduced, its Sharpe became more negative in magnitude, which is an expected arithmetic consequence rather than a change in attribution. The lower win rates and win-to-loss ratios for the equal-weight portfolios were expected when compared to the

main parameterized model, since portfolios that do not control and position themselves according to risk are expected to demonstrate less months with positive returns and more dramatic losses.

Compared with our study's flagship model summarized in Table 2 (mean monthly return 0.686%, Sharpe 0.673, annualized volatility 5.992%, max drawdown 5.776%), the normal equal-weight portfolio showed a higher average monthly return (1.135%) but markedly poorer risk characteristics—signficantly higher volatility and drawdowns and a significantly lower Sharpe and win rate. As discussed, this pattern was exactly what our design was expected to show. By stripping out volatility scaling, score-magnitude weighting, and the nonlinear transform, the equal-weight portfolios allocated the same capital to weak and strong LLM signals alike. That structure inflated tail exposure and month-to-month noise (hence the extreme kurtosis and large absolute skew), while allowing a few high-impact legal-event months to lift the mean return. In contrast, the production model's risk parameters and sizing reallocated capital toward stronger, lower-volatility signals, smoothing returns, increasing the win rate, and compressing drawdowns—at the cost of a lower raw mean. Winsorization narrowed the gap on tail risk relative to the production model—smoothing paths and compressing drawdowns—yet it could not substitute for targeted sizing. It primarily reduced sensitivity to episodic single-name shocks while preserving the same directional evidence.

In sum, Table 3 provided the intended attribution: when trading was performed on the LLM legal-risk signal at equal weight, the normal equal-weight portfolio outperformed its mirror, while Table 2 showed that the same signal became far more investable once standard risk controls and sizing were applied. The combination of [1] normal > mirror in Table 3 and [2] improved Sharpe/volatility/drawdowns in Table 2 was precisely what would be expected if the LLM-identified direction of legal risk was informative and portfolio engineering primarily improved delivery rather than creating the effect. These conclusions held for the winsorized variants as well, reinforcing that the attribution result conformed with expectation even after accounting for extreme observations.

Together, Figure 3 and Table 3 show that when trades were determined solely by the LLM legal-risk score, the normal equal-weight portfolio outperformed its mirror, providing direct attribution that the signal's direction carries return information independent of portfolio engineering. The LLM-only test carried structurally conservative returns: with no position sizing, every non-zero score received the same capital (±0.1 was treated like ±1.0), which lowered Sharpe and raised volatility/drawdowns and skew; thus these results should be read as a lower bound on what the same signal could deliver when sensibly sized. By contrast, the main specification (Figure 2/Table 2) improved investability by allocating more capital to stronger, lower-volatility names while preserving the same underlying legal-risk signal that drove the normal-over-mirror result.

Despite its hindrance to returns, equal weighting was a necessary adoption for this test as it avoided selection/tuning, preserved the full non-zero universe used by the main

methodology, and tested whether the sign alone of the LLM legal-risk score moved returns. Accordingly, absolute returns here should be read as a conservative lower bound on what the same signal can deliver when sensibly sized. In other words, the primary benchmark was directional ordering—the normal equal-weight portfolio outperforming the mirror—rather than the magnitude of raw performance (especially when compared to properly diversified or sized models/benchmarks such as our study's flagship model or the S&P 500). Also, the winsorized curves confirmed that the attribution result was not a consequence of a handful of extreme observations. Clipping tails narrowed risk (lower volatility and drawdowns) but left the normal-over-mirror ranking intact, consistent with the view that the sign of the LLM legal-risk score contained predictive content, while the main specification (Table 2/Figure 2) improved investability by allocating less capital to weak or high-volatility stocks.

Regression results

To assess the abnormal return attributable to changes in legal risk disclosures, the portfolios' monthly excess returns were regressed on the Fama-French five-factor model plus momentum (FF5 + UMD). Table 4 presents the regression's core findings and summary statistics, and Table 5 details the individual factor exposures and their statistical significance. Likewise, via Panels A, B, and C, Figure 4 presents diagnostic plots evaluating the factor regression modeling the portfolio's excess returns against the Fama-French five-factor model augmented by momentum (FF5+UMD). As seen in Table 4, the signal produced a mean monthly alpha of 0.471%, which annualized to 5.80%, across the various long-short portfolios formed in the trading period. This alpha was statistically significant at the 5% level ($p = 0.023$), suggesting that the legal risk-based signal generated returns not captured by standard risk factors.

Table 4. Regression summary statistics for portfolios' excess returns. **Significant at the 5% level

Monthly α (%)	Annualized α (%)	p-value	t-stat	Adj. R^2	F-statistic (p-value)	N (Months)
0.471**	5.80	0.023	2.34	0.318	5.76 (0.0001)	62

The model's moderate R^2 value of 0.318 suggests that approximately 31.8% of the variance in the portfolio's excess returns was explained by the regression model. While this indicates a modest explanatory power, it further highlights that a substantial portion of the strategy's performance remained unexplained by conventional factor models, pointing to the unique informational content that the legal risk signal may contribute. However, the unexplained variance also suggests the possibility that some of the returns may have stemmed from noise, non modeled risks, or other omitted variables, warranting caution in attributing the entirety of the excess returns solely to methodology.

Additionally, the F-statistic of 5.76 with an associated p-value of 0.0001 strongly confirmed the overall joint significance of the

model at the 5% level. Rejection of the null hypothesis (that all coefficients are jointly zero) indicated that the included factors, along with the abnormal return (α), collectively have explanatory relevance for the observed returns and suggested that the regression specification is a good fit (i.e. not a result of random noise). Together, these results imply that the legal-risk-based strategy produced economically meaningful, unique, and statistically significant excess returns, supporting the notion that the LLM-derived legal risk signal contains distinct and priced information relevant to market participants.

As seen in Table 5, the portfolio exhibited statistically significant exposure to the market risk factor (MKT - RF) the profitability factor (RMW), and value factor (HML) with loadings of 0.119 (p = 0.006), -0.345 (p = 0.000), and 0.166 (p = 0.019) respectively. The negative loading on the RMW factor indicates that the portfolio disproportionately captured returns from firms with weak operating profitability. A potential reason for this may be that less profitable firms are typically associated with more volatile legal controversies, hence whenever a particular change occurs (whether positive or negative), they will tend to be represented in a sub-portfolio. Meanwhile, the positive MKT - RF and CMA loadings reflect a tendency for the strategy to benefit from the overall market and firms with more conservative investment practices.

Table 5. OLS regression outputs for the legal-risk-based long-short portfolio returns against the Fama-French FF5 + UMD model. *Significant at the 10% level, **Significant at the 5% level, ***Significant at the 1% level

Statistic	MKT - RF	SMB	HML	RMW	CMA	UMD
β	0.119***	-0.130	0.166**	-0.345***	-0.068	-0.103*
Std Err	0.042	0.078	0.069	0.086	0.096	0.056
t-stat	2.87	-1.67	2.41	-4.01	-0.71	-1.83
p-value	0.006	0.100	0.019	0.000	0.480	0.072

In contrast, factor loadings for size (SMB), investment (CMA), and momentum (UMD) were statistically insignificant, with p-values of 0.100, 0.480, and 0.072 respectively. These results suggest that the strategy's returns were not primarily driven by exposures to these standard dimensions of risk. The regression framework therefore isolates the legal risk signal's contribution by accounting for traditional pricing factors, reinforcing the conclusion that the observed performance is not entirely explained away by conventional sources of return. However, some of the observed factor exposures may also reflect characteristics inherent to the sampled firms, such as their size, industry composition, or prevailing economic conditions during the sample period.

In Figure 4, Panel A displays a scatter plot of actual versus fitted excess returns, where the fitted values represent the predicted returns from the FF5+UMD regression model. The degree of clustering around the reference line suggests a modest positive relationship between these values and indicates that the

model explained a portion of return variation, though with considerable unexplained returns. However, residuals—defined as the differences between actual and predicted returns—appeared notably dispersed around the regression line, implying the presence of idiosyncratic variation, which may reflect factors—such as legal risk—not captured by traditional risk models.

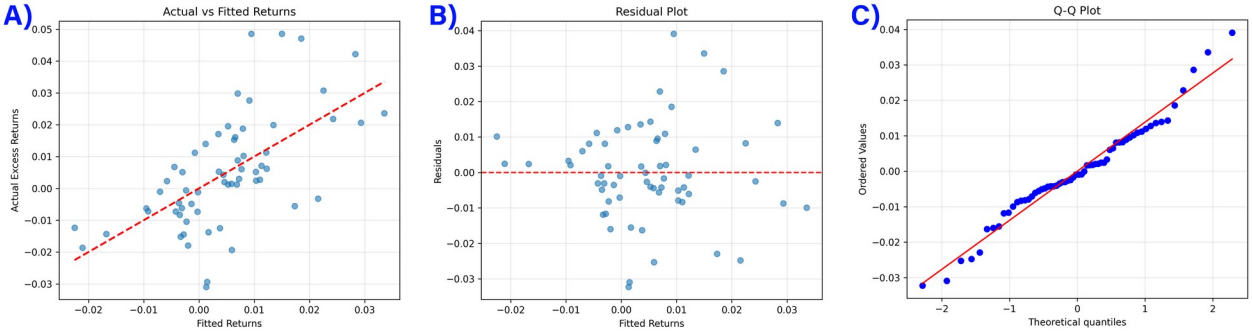

Figure 4. Regression Diagnostic Scatter Plots. **A.** The relationship between fitted and actual excess returns. **B.** Residuals against fitted values, used to assess homoscedasticity. **C.** Distribution of standardized residuals against a normal distribution to evaluate normality assumptions.

Panel B shows the residuals plotted against the fitted values. Residuals appeared randomly dispersed around the zero line with no clear pattern. This random distribution supports homoskedasticity (constant error variance). This pattern suggests that the variability of prediction errors increased with higher predicted returns. Panel C features a Q-Q (quantile-quantile) plot comparing the distribution of standardized residuals to a theoretical normal distribution. Though some deviation is observed in the tails, suggesting mild departure from perfect normality, most points lie close to the 45-degree line, indicating that residuals are approximately normally distributed. Residual normality supports the use of t-tests for factor significance, especially in smaller samples, and supports the integrity of statistical inference under OLS assumptions. However, it should be noted that financial markets do not uniformly follow a normal distribution and typically exhibit a higher frequency of fat tails. Also, regressions were not applied to the equal-weight portfolios, since they served only to demonstrate attribution of the LLM signal and were not intended as investable strategies.

Strategic and Financial Applications

Answering this research question has many real-world financial applications. The most straightforward use of this research is creating trading signals for alpha generation. The long-short portfolio strategy can be adapted into an event-driven trading strategy where algorithms can rapidly detect legal risk changes and respond with high-speed trading decisions.

Another application of the study is in asset management, specifically risk management. Portfolio managers and institutional investors can use LLM-based legal risk scores as an indicator of potential legal trouble and shift their portfolio weightings accordingly. This

research goes beyond traditional volatility measures like CoVaR and MES by incorporating idiosyncratic legal risk in the signal. Similarly, managers may also use the signal as a fundamental workflow input: large negative changes (risk decreased) could flag candidates for increased long conviction while large positive changes (risk increased) could flag candidates for hedging, trimming, or fundamental shorts. Existing positions can be risk-managed when disclosed legal risk worsens (e.g., temporary hedges, tighter limits) and scaled when it improves. This usage integrates cleanly with prevailing investment processes without committing to a higher-turnover standalone strategy and keeps fees/loads primarily a function of ordinary trading activity rather than the signal itself.

Furthermore, beyond a standalone higher-frequency long-short implementation, the legal-risk signal is suitable for lower-turnover deployment or being a smaller portion of a broader multi-strategy portfolio (e.g., ~5% of NAV or risk budget). A more selective version of the model could trade only the strongest signals—names exceeding a threshold in the absolute LLM legal-risk change ($|LLMscore| \geq$ score criteria) or top/bottom N by absolute score in a quintile or decile based system—with monthly/quarterly rebalances and 3 or 6-month holds (or until the score's sign reverses). Positions would continue to use the study's volatility-aware sizing and standard risk limits with single-name/sector caps, and thereby materially reduce turnover and trading costs while preserving the directional content demonstrated by the methodology.

Besides trading, the methodology can assist in corporate decision-making and strategy monitoring. In particular, in investment banking, quantifying changes in legal risk could be used to inform potential merger or acquisition deals. Furthermore, for ESG-focused investment strategies and corporate compliance monitoring, quantifying firm-level changes in legal risk enables automated detection of ongoing litigation related to environmental violations, labor disputes, or governance failures. These granular risk signals can be integrated into broader ESG scoring models, improving the precision of governance assessments and reducing reliance on generalized third-party ESG ratings. This allows compliance officers and responsible investors to flag material legal issues in real-time and respond with greater agility.

Similarly, equity research analysts and sell-side teams can adopt this legal risk scoring framework to systematically track the evolution of firm-level litigation exposure. Rather than manually parsing dense and often unclear legal sections, analysts can reveal meaningful tone and risk shifts across large samples of firms more efficiently. This enhances the robustness of coverage, accelerates risk identification, and may uncover under-reported legal issues earlier than traditional methods allow.

Limitations

One limitation of this research stems from its focus on publicly traded firms listed exclusively on major U.S. exchanges (NYSE and NASDAQ). This focus inherently introduces survivorship bias, as firms that underwent delisting due to bankruptcy (e.g, JCPenney and Bed Bath & Beyond), acquisition, or other reasons within the analyzed period were excluded, potentially

biasing results towards more stable firms. Additionally, while the sample size was substantial, expanding the dataset beyond the selected number of firms and time frame could have enhanced statistical robustness and generalizability. Given that differences in legal risk disclosures are subtle, nuanced, and inherently challenging to quantify precisely, restricting the sample size and duration might limit the ability to detect more subtle yet economically meaningful signals.

Another inherent limitation lies within the use of an LLM for textual interpretation. Although GPT-4o is a well-established model and demonstrates advanced contextual understanding, the numerical scoring of legal risk changes remains inherently subjective. Slight variations in prompt phrasing or instruction clarity could influence the consistency of generated scores. Unlike structured financial metrics, these qualitative evaluations lack an objective benchmark for verification, thus potentially introducing measurement noise.

A core limitation of any filing-based signal is that firms are obligated to disclose material legal proceedings. Matters deemed "immaterial" may be omitted or summarized at management's and counsel's discretion. Consequently, our LLM measures the change in disclosed legal risk, not the full latent legal-risk for a firm. This introduces several implications: (1) censoring of smaller disputes and early-stage matters, (2) potential systematic under-representation of firms or industries with lower litigation prominence, (3) timing discretion around when developments cross the materiality threshold and appear in 10-Q/10-K *Legal proceedings* (or related sections/footnotes). These forces can attenuate the signal (true changes that remain below disclosure thresholds might go unobserved) and may bias coverage toward larger issuers and high-profile cases. Our attribution tests and portfolio results should therefore be interpreted as evidence that disclosed legal-risk changes contain information about subsequent returns: they do not rule out additional, undisclosed dynamics. In future work, overlaying the core SEC legal-disclosure data with higher-frequency sources (e.g., docket updates, regulatory actions, or curated news) into the same LLM change-scoring framework could reduce disclosure-timing lags while capturing more litigation-related data that was originally deemed "immaterial" and preserving the model's comparative, "change-focused" design.

Similarly, there is a natural lag between actual legal events and their public disclosure in 10-K and 10-Q filings. Due of this lag, investors may already have priced in some of the legal risk changes before the filing date, limiting the alpha potential. Moreover, the analysis did not explicitly control for industry-specific litigation dynamics or systematic macroeconomic shocks, both of which could influence changes in legal disclosure language and portfolio performance independently of firm-specific legal risk factors. In terms of study design, transaction costs and market frictions were approximated in the backtesting procedure. Real-world implementation of the strategy could result in differing transaction costs, liquidity constraints, or execution slippage, which could reduce actual portfolio performance relative to simulated results.

Conclusion

This research aimed to establish a more precise and insightful method of analyzing textual disclosures within the legal portions of SEC filings by leveraging LLMs. Specifically, this paper employed OpenAI's GPT-4o model to measure changes in legal risk between consecutive filings. The core objectives of this study were: [1] to evaluate the effectiveness of this novel methodology in capturing economically meaningful changes in firm-specific legal risk compared to traditional text-analysis methods, and [2] to assess whether these LLM-generated scores could produce actionable investment strategies capable of yielding significant alpha.

The study primarily distinguishes itself from previous literature by using GPT-4o for legal analysis of the under-utilized *Legal Proceedings* section and its corresponding footnotes. Traditional methodologies, such as TF-IDF cosine similarity and simpler NLP classifiers like FinBERT, typically face substantial limitations, including insufficient semantic depth, token-length constraints, and a lack of nuanced contextual interpretation. These limitations are notably addressed by the research's methodological approach, which not only captures the contextual subtleties of legal disclosures but also quantitatively translates these subtle textual shifts into actionable numeric scores of legal risk change. Beyond methodological innovation, this study offers practical value across finance and strategy. It can inform alpha generation, portfolio risk management, ESG monitoring, and legal due diligence—providing a scalable alternative to manual analysis of firm-level litigation disclosures.

Data for the analysis was sourced from SEC 10-Q and 10-K filings of 1,200 randomly sampled firms from the NASDAQ and NYSE, coupled with daily returns from Yahoo Finance and firm-level summary statistics from Zacks Fundamental Collection B. This comprehensive combination provided a robust basis for empirical testing and ensured the application of results across a diverse range of U.S. equities.

The empirical strategy involved constructing monthly long-short portfolios incorporating all non-zero signals, with position sizes scaled by signal strength and anticipated risk. Backtesting and portfolio statistical evaluation revealed the signal's ability to produce modest returns, coupled with moderate risk exposure and a mean Sharpe ratio of portfolios slightly below that of the S&P 500 benchmark over the same period. Subsequent performance evaluation through OLS regression analysis against standard Fama-French and momentum factors confirmed the methodology's ability to generate statistically significant alpha. Specifically, on average, an individual long-short portfolio yielded a 0.471% monthly alpha (annualized to 5.80%). Mirror tests stripped of the position sizing and using only the LLM legal-risk scores show the theory-consistent side outperformed the reverse, reinforcing that legal-risk direction contributes to returns independent of risk parameters rather than being largely confounded by randomness. These results demonstrate that LLMs can systematically extract price-predictive information from legal text with 10-K and 10-Q SEC filings. As LLM capabilities advance, these results provide a foundation for more refined, scalable applications of disclosure-based alpha strategies.

Looking ahead, this research opens multiple avenues for future examination. Future researchers could apply this methodology to a broader dataset encompassing additions such as international firms, mitigating survivorship bias and enhancing the implications of results. Additionally, refining the utilized model through fine-tuning on extended legal data or integrating additional quantitative variables—such as litigation costs or revenue impact—could significantly enhance predictive capabilities. Another promising area of future research involves analysis that tracks evolving legal risk profiles over extended periods, potentially forecasting economic resilience or identifying emergent risks well before they become impactful. Creating a similar-style model that can receive real-time news feeds or litigation database information and output some variation of our legal risk change score could also be a promising area of future study. Ultimately, this research represents part of the growing body of literature that harnesses sophisticated LLM technologies for nuanced and significant financial analysis. As LLM capabilities continue to advance, their integration into financial markets research and practical portfolio strategies promises substantial improvements in market efficiency, risk assessment, and strategic investment decision-making.

Abbreviations

Ann. - Annualized, Vol. - Volatility, DD – Drawdown, N – Number, SEC - Securities and Exchange Commission, CRSP – Center for Research in Security Prices, LLM - Large-Language Model, OLS - Ordinary Least Squares, ESG - Environmental, Social, and Governance, BERT - Bidirectional Encoder Representations from Transformers, NLP - Natural Language Processing, B/M – Book-to-Market, MVE - Market Value of Equity, MD&A - Management's Discussion and Analysis, API - Application Programming Interface, S&P 500 - Standard & Poor's 500 (Index Fund), NASDAQ - National Association of Securities Dealers Automated Quotations, NYSE - New York Securities Exchange, GPT - Generative Pretrained Transformer, TF-IDF – Term Frequency Inverse Document Frequency, CoVaR- Conditional Value-at-Risk, MES- Micro E-mini S&P 500 futures contract

Notations

α - The portfolio's intercept or alpha

σ_{target} - Target portfolio volatility

S_i - OpenAI legal risk score of stock *i*

$w_i^{initial}$ - Initial position weight for stock *i*

$\omega_{i,i}^{raw}$ - Raw weighting of the *i*-th sub-portfolio in period i

σ_i - 90-day historical volatility (HV90) of stock i

κ - S-curve scaling factor

$\bar{\sigma}$ - Cross-sectional average HV90 for the month

β - The coefficient of various factor sensitivities (denoted in its subscript)

R_t - The monthly return of the long-short portfolio

RF_t - The monthly risk-free rate
MKT_t - RF_t - The excess market return
SMB_t - The size premium (small minus big)
HML_t - The value premium (high minus low book-to-market)
RMW_t - The profitability premium (robust minus weak)
CMA_t - The investment premium (conservative minus aggressive)
UMD_t - The momentum factor (winners minus losers)
Std Err - Standard error
ϵ_t - The error term

References

1. Cohen, L., Malloy, C., Nguyen, Q. (2020). Lazy prices. The Journal of Finance, 75(3), 1371-1415. https://doi.org/10.1111/jofi.12885

2. Adosoglou, G., Lombardo, G., Pardalos, P. M. (2021). Neural network embeddings on corporate annual filings for portfolio selection. Expert Systems With Applications, 164, 114053. https://doi.org/10.1016/j.eswa.2020.114053

3. Adosoglou, G., Park, S., Lombardo, G., Cagnoni, S., Pardalos, P. M. (2022). Lazy network: A word embedding-based temporal financial network to avoid economic shocks in asset pricing models. Complexity, 2022(1). https://doi.org/10.1155/2022/9430919

4. Yilmaz, Y., Reichmann, M. (2024). Textual changes in 10-Ks and stock price crash risk: Evidence from neural network embeddings. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4643560

5. Gupta, U. (2023). GPT-InvestAR: Enhancing Stock Investment Strategies through Annual Report Analysis with Large Language Models [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2309.03079

6. Daimi, S. A., Iqbal, A. (2024). A scalable data-driven framework for systematic analysis of SEC 10-K filings using large language models (arXiv:2409.17581) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.17581

7. Bürgler, Vivien and Davidson, Brittany and Bürgler, Manuel, Democratizing Financial Information using GPT-4o (June 30, 2024). http://dx.doi.org/10.2139/ssrn.5130935

8. Loughran, T., McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35-65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

9. Zhu, Y., Falahati, A., Yang, D. H., Mohammadi Amiri, M. (2025). SentenceKV: Efficient LLM inference via sentence-level semantic KV caching. arXiv preprint arXiv:2504.00970. https://arxiv.org/abs/2504.00970

10. Liu, X., Zhang, Y., Wu, W. (2023). Forward-looking statements in annual reports and firms' financing constraints. Telfer College of Business, University of Ottawa – University of Illinois Gies College of Business. https://giesbusiness.illinois.edu/docs/default-source/default-document-library/departments/accountancy/tija-23_liu-zhang-wu.pdf

11. Bunt, H. L., Goddard, A., Reader, T. W., Gillespie, A. (2025). Validating the use of large language models for psychological text classification. Frontiers in Social Psychology, 3. https://doi.org/10.3389/frsps.2025.1460277

12. Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models (arXiv:1908.10063). arXiv. https://arxiv.org/abs/1908.10063

13. Fatemi, S., Hu, Y. (2023). A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis. https://doi.org/10.48550/arXiv.2312.08725

14. Rodriguez Inserte, P., Nakhlé, M., Qader, R., Caillaut, G., Liu, J. (2024). Large language model adaptation for financial sentiment analysis (arXiv:2401.14777). arXiv. https://doi.org/10.48550/arXiv.2401.14777

15. Atkinson, T. M., Petrov, A., Lynch, K. A., George, L. S., Cracchiolo, J. R., Daly, B., et al. (2025). Using GPT-4o to interpret patient-reported outcomes without training. JNCI: Journal of the National Cancer Institute, 117(4), 809-811. https://doi.org/10.1093/jnci/djaf016

16. Gentzkow, M., Kelly, B., Taddy, M. (2019). Text as Data. Journal of Economic Literature, 57(3), 535-574. https://www.aeaweb.org/articles?id=10.1257/jel.20181020

17. Amel-Zadeh, A., Faasse, J. (2016). The information content of 10-K narratives: Comparing md&a and footnotes disclosures. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2807546

18. McIntosh, J., Starykh, S. (2021). Recent trends in securities class action litigation: 2020 full-year review. NERA Economic Consulting. https://www.nera.com/content/dam/nera/publications/2021/PUB_2020_Full-Year_Trends_012221.pdf

19. Loughran, T., McDonald, B. (2023). Management disclosure of risk factors and COVID-19. Financial Innovation, 9(1). https://doi.org/10.1186/s40854-023-00459-5