ARTIFICIAL INTELLIGENCE RISK

EVIDENCE FROM LLM-ASSESSED 10-K FILINGS, 2020–2025

ZAKARIAS SJÖBERG DAHLÉN

DANIEL SALAMON

Bachelor Thesis

Stockholm School of Economics

2026

Artificial Intelligence Risk: Evidence from LLM-Assessed 10-K Filings, 2020–2025

Abstract:

We construct a firm-level AI risk exposure index from mandatory 10-K Item 1A disclosures for S&P 500 firms over fiscal years 2020–2025. Three large language models score each firm's risk factor text across four dimensions: displacement, operational, regulatory, and cybersecurity risk. We test whether AI risk exposure predicts differential equity returns using quartile portfolio sorts, Fama–MacBeth cross-sectional regressions, bivariate sorts, and panel regressions on high AI attention days identified from Google Search Volume Index data. In the post-2023 period, equal-weighted long-short portfolios yield monthly Fama–French five-factor alphas of 0.59–0.77%, with the premium concentrating among value and low-profitability firms. A decomposition shows that 54–66% of the spread reflects between-sector variation. Firms with higher AI exposure earn significantly lower market-adjusted abnormal returns on days of elevated public attention to AI regulation and safety. Taken together, these results suggest that AI risk is reflected in equity prices primarily through sector composition rather than through firm-level disclosure intensity, with within-sector pricing concentrated in selected industries.

Keywords:

Artificial intelligence risk, Risk disclosure, Large language models, Asset pricing

Authors:

Zakarias Sjöberg Dahlén (26137)
Daniel Salamon (26121)

Tutor:

Alvin Chen, Assistant Professor, Department of Finance

Examiner:

Ramin Baghai, Professor, Department of Finance

Bachelor Thesis
Bachelor Program in Business & Economics
Stockholm School of Economics

1 Introduction

Mandatory annual filings require U.S. public companies to disclose the risks they face in material detail, generating a body of regulatory text that researchers have increasingly treated as a direct signal of firm level exposure. Whether the language contained in these disclosures is economically meaningful and specifically whether it predicts how risk is priced across firms has emerged as a central question in the literature on information and asset prices. Cohen, Malloy, and Nguyen (2020) examined the relationship between changes to the language and structure of financial reports and a given firm's future returns and operations. They found that changes to 10-K reports predict subsequent earnings outcomes, corporate news, and bankruptcy events. Florackis, Louca, Michaely, and Weber (2023) address this question in the domain of cybersecurity by constructing a firm level risk index derived from the Item 1A sections of mandatory 10-K filings. Scoring each firm's disclosure language and validating the resulting index against realized cyberattack incidents, they demonstrate that the measure predicts both the cross section of stock returns and cumulative abnormal returns around cybersecurity related events.

Artificial intelligence represents an emerging category of firm level risk, operating simultaneously across multiple channels. This risk domain is increasingly being disclosed in 10-K filings, growing from appearing in 4% of all 10-K filings in 2020 to 43% in 2024 (Uberti-Bona Marin, Rijsbosch, Spanakis, and Kollnig, 2025). First, firms face displacement risk. Competition, new entrants, or AI enabled substitutes may render existing offerings or business models obsolete. Second, firms face operational risks arising from the integration of artificial intelligence into their business operations, spanning reliability failures, algorithmic bias, and reputational damage from AI related incidents. Elements of that category are already visible in mandatory 10-K filings, which are often cited AI risk categories (Uberti-Bona Marin et al., 2025). Third, firms face regulatory and legal risk. For example, the EU AI Act (Regulation (EU) 2024/1689) imposes compliance obligations, with failure to comply resulting in fines of up to 7% of global annual turnover. Fourth, firms face cybersecurity and data privacy risks as a result of AI enabled threats. The U.S. National Institute of Standards and Technology's Generative AI Risk Management Framework (2024) establishes that AI simultaneously reduces the barriers against cyberattacks and expands firms' own attack surfaces via increased vulnerabilities such as prompt injection and data poisoning. Cao, Liu, Qiu, and Zhao (2024) used a combination of keyword search and LLM based analysis to classify AI related risk disclosed in Item 1A of 10-K filings into the following six categories: regulatory, operational, competitive, cybersecurity, ethical, and third-party. They find that firms disclosing AI risk in any given category face higher tail risk, measured by stock idiosyncratic volatility and option implied risk metrics. However, their risk measure is binary, and their analysis is centred around linking the binary variables to volatility and options market outcomes. The question of whether a continuous risk disclosure index predicts abnormal equity returns, either persistently or around specific AI-related events, remains unanswered.

This motivates the following question:

> *Does a firm's disclosed AI risk intensity, as measured by its AI Risk Exposure Index score constructed from Item 1A risk factor disclosures, predict differential equity returns for S&P 500 companies over the period 2020–2025, both persistently in portfolio sorted return tests and in cumulative abnormal returns around days with high attention toward AI risk?*

We address this question in four stages. We start by constructing AI risk disclosure scores for each firm-year pair using a standardized prompt with three frontier LLMs. Our empirical analysis begins by sorting S&P 500 firms into quartile portfolios based on their AI risk score, and equal weighted and value weighted long short returns are evaluated against CAPM, Carhart four factor, and Fama-French five factor benchmarks. In addition to the three individual LLM scores, we construct two composite measures: a simple average of the three model scores, and the first principal component of the standardised scores. We further control for five firm characteristics (market capitalisation, book to market, ROA, R&D intensity, and leverage) and use Fama MacBeth cross sectional regressions and bivariate sorts to test whether the observed premium persists after controlling for known return predictors. To address the concern that the observed effects reflect cross sector variation rather than firm-level AI risk, we construct within sector long short sorts and an exact between sector versus within sector decomposition. Finally, firm-day panel regressions are estimated to test whether firms with higher AI risk exposure earn lower abnormal returns on days of elevated public attention to AI related risks, identified through the Google Search Volume Index in a manner similar to that of Florackis et al. (2023).

The sample covers S&P 500 firms across fiscal years 2020 to 2025, with stock return data running from May 2021 through December 2025. AI risk disclosure scores are generated independently by three large language models: OpenAI's ChatGPT, Anthropic's Claude, and Google's Gemini. Each model received the same prompt and a firm level dataset of extracted Item 1A sections retrieved from the SEC's EDGAR database. Each LLM scored firms across four risk dimensions: displacement, operational, regulatory, and cybersecurity.

In the late subsample (May 2024 through December 2025), quartile sorts on all five measures produce positive equal weighted long short spreads. The monthly Q4$-$Q1 excess return ranges from 0.52% for Claude to 0.94% for Gemini, with the composite measures in between at 0.83% (Simple Average) and 0.86% (PC1), corresponding to annualised premiums of approximately 6.2% to 11.2%. Fama-French five factor alphas are statistically significant for ChatGPT (0.59%, $t = 2.46$), Gemini (0.77%, $t = 3.09$), and both composite measures, indicating that the return spread is not captured by standard size, value, profitability, or investment factors. Q1 is the lowest-returning portfolio and Q4 the highest-returning for every measure. Perfect monotonicity across the full quartile range obtains only for Claude and ChatGPT, while Gemini and the composites show Q3 dipping slightly below Q2.

However, several limitations temper these findings. In the full 56 month sample, only Gemini generated scores produce statistically significant risk adjusted results, reflecting the limited cross sectional variation in AI risk disclosures during 2020–2022. An exact decomposition of the long short spread shows that the between sector component accounts for 54–66% of the total premium for the measures with significant spreads. The within sector component, while positive, is marginally significant only for Gemini. Within sector sorts reveal that the premium concentrates in Financials and Consumer Discretionary, while Information Technology shows no within sector variation. A daily frequency robustness check confirms that the inference is not sensitive to the HAC lag parameter and that the point estimates are consistent across frequencies, though the daily t-statistics are lower due to the weaker signal to noise ratio at daily frequency. Together, these results suggest that the majority of the observed premium reflects cross sector variation in both AI risk and returns, though a within sector component probably exists to, in some industries more than others.

Disclosure intensity in mandatory 10-K filings already carries return relevant information for AI risk: long short portfolios deliver economically meaningful alphas in the post-2023 period, and high exposure firms bear greater losses on days of elevated public attention to AI related risks. The majority of this premium reflects between-sector variation rather than

firm level disclosure intensity, and the precision of firm level inference is constrained by the short sample period and the voluntary nature of current AI risk reporting. Both limitations are temporary. Three frontier large language models given identical prompts recover a common signal from the underlying text with strong cross-model agreement, suggesting that ensemble LLM rating is a viable and generalisable methodology for scoring novel risk categories where a labelled training sample is unavailable.

1.1 Literature Review

Textual analysis of mandatory disclosures has become an established tool in finance research since Loughran and McDonald (2011) demonstrated that domain-specific dictionaries extract economically meaningful signals from 10-K filings. Florackis et al. (2023) build on this tradition and establish the methodological framework for this thesis. They construct a risk disclosure index for cybersecurity risk and test whether cybersecurity risk is priced in the cross section of stock returns. They begin by extracting cybersecurity relevant disclosures from mandatory 10-K reports, specifically from Item 1A, using a keyword and context algorithm. They then score each firm's disclosures using cosine similarity to extracted risk disclosures of a training sample consisting of firms that experienced verified cyberattacks. The motivation behind using cosine similarity to the training sample is that hacked firms had higher exposure prior to the attack, and that similarly exposed companies would articulate their risk exposure in a similar fashion. Key validation methods include verifying the intuitiveness of the scores by demonstrating that high scores concentrate in industries that rely heavily on IT, and conducting an event study that returns statistically significant negative CARs for high score companies surrounding an attack disclosure. Finally, they show that equal weighted portfolios long high exposure and short low exposure firms earn an 8.3% annual premium that is robust to Fama-MacBeth controls, bivariate sorts, and the event study mentioned earlier. They further test whether the cybersecurity risk factor earns lower returns on days of elevated public attention to cybersecurity risk, identifying such days through abnormal Google Search Volume Index activity for cybersecurity related search terms and showing that the factor performs significantly worse on high attention days.

We apply much of this methodology to the present thesis. We extract risk disclosures from 10-K Item 1A and construct a similar firm level risk disclosure index. Since AI is a relatively new field and parallel examples of the Florackis et al. (2023) training sample do not exist to the same extent, we use large language models for the scoring. A growing literature applies large language models to financial text, including return prediction from news headlines (Lopez-Lira and Tang, 2023) and the generation of disclosure summaries that retain value-relevant information (Kim, Muhn, and Nikolaev, 2024a), supporting their use where labelled training samples are unavailable. This approach is preferable as LLMs natively understand context and thus constitute a more flexible option for scoring novel risk categories. We adopt the same validation methods, tests, and robustness checks as described previously. One structural difference worth noting is that cybersecurity risk disclosure in Item 1A has been subject to increasing SEC guidance since 2011, whereas AI risk disclosure currently remains voluntary and lacks standardised reporting requirements. This difference may affect the consistency of disclosure language across firms and, consequentially, the results.

Cao et al. (2024) establish a clear narrative regarding the relevance of AI risk disclosure in 10-K Item 1A. They analyse a sample of U.S. publicly listed firms' 10-K filings across 2010 to 2023. The paper decomposes the disclosed risks into six dimensions and shows that firms disclosing one or more of the risk categories exhibit significantly higher tail risk, measured by idiosyncratic equity volatility and option implied volatility. This supports our approach of

applying Item 1A textual analysis to AI related risk disclosures, while leaving room to apply the Florackis et al. (2023) methodology for a more in depth analysis. The Cao et al. (2024) risk measure captures the presence of disclosures across categories, not intensity. We extend their work by applying the Florackis et al. (2023) methodology to construct a continuous risk disclosure index that measures the intensity of total risk as well as the decomposed risks. Cao et al. (2024) further show that certain decomposed risk categories impact tail risk and realised volatility more than others, further motivating our risk disclosure measure decomposition.

2 Data and Methods

2.1 Data

The sample consists of S&P 500 constituents over fiscal years 2020 through 2025, with historical index compositions obtained from Barthwal (2024). After adjusting for index recomposition, delistings, M&A activity, and multiple share classes, between 492 and 497 unique firm-year observations remain per year. Firm-level fundamentals are sourced from the Compustat North America database (S&P Global, 2026), including total assets, total debt, book equity, revenue, research and development expenditure, and market capitalisation. All financial data correspond to the fiscal year-end immediately prior to the relevant 10-K filing. Monthly and daily stock returns come from the Center for Research in Security Prices (CRSP, 2026), and the and daily Fama-French five factors together with the momentum factor are taken from French (2026). Daily Google Search Volume Index data for AI-related search terms are obtained from Google Trends (Google, 2025) and used to identify high-attention days for the SVI regressions described in Section 2.4.1.

Summary statistics for the key firm-level variables are reported in Table 1. The median firm has total assets of $21.8 billion, a market capitalisation of $27.7 billion, and revenue of $10.6 billion. At 0.28, the median book-to-market ratio is low, which is expected given the growth orientation of the S&P 500. Leverage averages 0.34 with considerable cross-sectional dispersion. Return on assets averages 6.4%, and R&D intensity has a median of zero because many S&P 500 firms do not report R&D expenditure. Missing R&D values are imputed as zero, following common practice.

Table 1: Summary Statistics

Variable	Mean	Std. Dev.	P25	Median	P75
Total assets ($M)	79,058	263,216	9,710	21,780	56,287
Revenue ($M)	27,556	56,836	4,799	10,586	22,724
Market capitalization ($M)	72,245	206,439	14,534	27,684	57,766
Book equity ($M)	18,086	41,133	3,006	7,098	17,375
Book-to-market	0.234	8.205	0.117	0.277	0.552
Leverage	0.335	0.242	0.190	0.319	0.440
Return on assets	0.064	0.092	0.022	0.054	0.103
R&D intensity	0.068	1.159	0.000	0.000	0.052

Summary statistics for the pooled sample of S&P 500 firm-year observations, fiscal years 2020–2025. Total assets, revenue, market capitalisation, and book equity are in millions of dollars. Book-to-market is book equity over market capitalisation. Leverage is total debt over total assets. Return on assets is operating income over total assets. R&D intensity is R&D expenditure over revenue; missing values are imputed as zero.

2.2 AI Risk Measure Construction

The AI risk exposure measure is constructed for each firm-year observation from the text of Item 1A ("Risk Factors") of annual 10-K filings, retrieved from the SEC's EDGAR system (SEC, 2026). Item 1A is the mandatory section in which registrants describe the most significant risks to their business, making it a natural source for studying how firms perceive and communicate AI-related exposure.

Each extracted Item 1A section is submitted to three separate large language models (Anthropic's Claude Sonnet 4.6, OpenAI's GPT-5.4, and Google's Gemini 2.5 Pro) using an identical prompt, reproduced in full in Appendix A. As Cao et al. (2024) demonstrate, keyword-based methods can locate candidate AI passages in 10-Ks but tend to miss context and cannot reliably distinguish between different types of AI activity or risk. The prompt is therefore designed to first identify all passages discussing artificial intelligence, machine learning, automation, algorithmic systems, or related technologies, and only then classify those passages into risk categories.

The choice of four risk sub-categories is motivated by the recent disclosure literature. Cao et al. (2024) classify AI risk disclosures in Item 1A into categories including regulatory, operational, competitive, and cybersecurity risk. Uberti-Bona Marin et al. (2025) find a similar pattern, with legal and competitive risks among the most common AI risk themes in 10-Ks, followed by cyberattacks, fraud, privacy and security issues, and technical limitations. Based on this evidence, each passage is classified into one or more of the following groups: displacement risk, operational risk, regulatory and legal risk, and cybersecurity and data privacy risk.

A further concern raised by Uberti-Bona Marin et al. (2025) is that many AI risk disclosures are vague, hypothetical, or boilerplate, and often lack detail on mitigation. The prompt therefore rewards passages that name specific AI technologies, identify affected business lines, quantify likely effects, or describe concrete management responses, while penalising generic language. Each sub-category is scored on an integer scale from 0 to 10, where 0 indicates no relevant disclosure and higher values reflect increasingly detailed and specific discussion.

Scoring is based solely on the text of the filing. The prompt instructs the model not to draw on outside knowledge about the company, its industry, or the broader economy, and to treat silence as absence. This addresses the concern about model priors raised by Kim, Muhn, and Nikolaev (2024b), who anonymise and standardise their inputs to prevent GPT from drawing on memorised company information. We instruct the model to disregard such priors while keeping firm identifiers in the input for dataset construction.

For each model, the firm-level AI risk score is defined as the simple arithmetic average of the four sub-category scores.

Composite measures

Because the three LLMs differ in their score levels but agree closely on relative rankings (Section 2.3), two composite measures are constructed to pool information across models. The first is simply the arithmetic average of the three model scores for each firm-year. The second is the first principal component (PC1) of the three standardised model scores. The PCA is estimated once on the pooled sample of all firm-year observations, with each model's score standardised to zero mean and unit variance prior to extraction, so as to keep the loadings stable and interpretable. Table 2 reports the results. The first component explains 86.6% of the total variance, with an eigenvalue of 2.60, and the loadings are nearly equal across models (0.58, 0.56, and 0.59 for Claude, ChatGPT, and Gemini, respectively). The second and third components together account for 13.4% of the variance and primarily capture pairwise disagreements: PC2

loads positively on ChatGPT and negatively on the other two, while PC3 contrasts Claude and Gemini. The PC1 score is oriented so that higher values correspond to higher AI risk. Given that the loadings are approximately equal, the PC1 and simple average measures are nearly perfectly correlated ($\rho = 0.999$), but both are reported throughout to demonstrate that results do not depend on the aggregation method.

Table 2: Principal Component Analysis of Standardised AI Risk Scores

	PC1	PC2	PC3
Panel A: Eigenvalues			
Eigenvalue	2.599	0.262	0.138
Variance explained (%)	86.6	8.7	4.6
Cumulative (%)	86.6	95.4	100.0
Panel B: Loadings (Eigenvectors)			
Claude	0.582	−0.467	−0.665
ChatGPT	0.562	0.823	−0.086
Gemini	0.587	−0.324	0.742
Panel C: Correlation with PC1			
Claude	0.939		
ChatGPT	0.906		
Gemini	0.947		
Simple Avg	0.999		

PCA estimated on the pooled sample of firm-year observations ($N = 2{,}994$), with each model's score standardised to zero mean and unit variance. Panel A reports eigenvalues and variance shares. Panel B reports eigenvector loadings. Panel C reports Pearson correlations between PC1 and each individual model score, as well as the simple average.

2.3 Validation of AI Risk Measures

Development over time

The evolution of AI risk ratings across the sample period is documented in Table 3 and Figure 1. In fiscal year 2020, between 71% and 83% of firms had no AI-related risk language in Item 1A at all, and the mean scores ranged from only 0.21 to 0.28. A clear break occurs with the fiscal year 2023 filings (filed in early 2024), coinciding with the release of ChatGPT in November 2022 and the surge in public attention to generative AI in the year that followed. Mean scores roughly tripled between 2022 and 2023, and continued to rise to 2.96 (Claude), 4.30 (ChatGPT), and 5.13 (Gemini) by 2025. By that point, almost all firms in the sample mentioned AI somewhere in Item 1A, with only 4–7% still receiving a score of zero.

Inter-model agreement

Pairwise Spearman rank correlations for the AI risk score, pooled across all firm-year observations, are reported in Table 4. The correlations range from 0.844 (ChatGPT–Gemini) to 0.906

Table 3: AI Risk Exposure Ratings: Descriptive Statistics by Year

	Claude		ChatGPT		Gemini		Simple Avg		PC1
Year	Mean	% Zero	Mean	% Zero	Mean	% Zero	Mean	% Zero	Mean
2020	0.21	80	0.22	71	0.28	83	0.24	67	−0.63
2021	0.26	75	0.69	66	0.31	77	0.41	60	−0.54
2022	0.32	73	1.06	66	0.32	77	0.56	58	−0.46
2023	1.36	36	0.82	46	2.69	37	1.63	24	0.00
2024	2.06	14	2.34	13	3.16	15	2.52	8	0.46
2025	2.96	6	4.30	4	5.13	7	4.13	2	1.21
Pooled	1.19	47	1.57	44	1.97	50	1.58	37	0.00

Mean AI risk score and percentage of zero-score firms by fiscal year and rating model. Each model's score is the average of four sub-category ratings (displacement, operational, regulatory, cybersecurity), each on an integer 0–10 scale. Simple Avg is the arithmetic mean of the three model scores. PC1 is the first principal component of the three standardised scores, pooled across all years; its mean is approximately zero by construction.

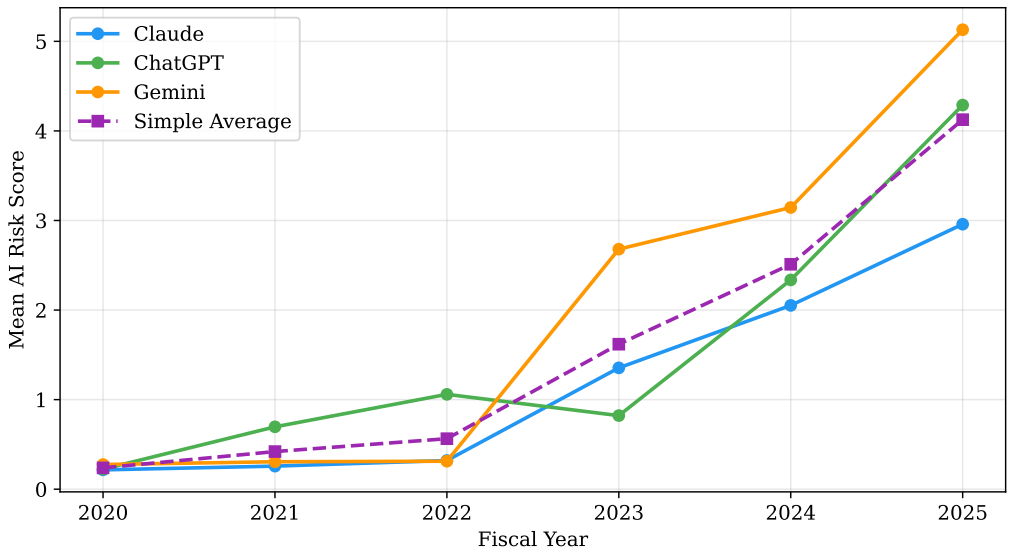

Figure 1: Mean overall AI risk exposure score by rating model, fiscal years 2020–2025. The sharp increase from fiscal year 2023 onward coincides with the broader adoption of generative AI and the resulting growth in AI-related risk disclosure.

(Claude–Gemini), indicating strong agreement on the relative ordering of firms. When computed year by year—which strips out the common upward trend—the correlations are lower but still substantial, averaging between 0.73 and 0.78 depending on the pair (Table 12 in the Appendix).

The models do differ in their levels. In the 2024–2025 subsample, Gemini assigns a mean score of 4.14, compared with 3.32 for ChatGPT and 2.51 for Claude. These differences are visible in Figure 1. Since the portfolio-based tests rely only on within-model rankings, these level differences have no bearing on the results.

Krippendorff's alpha (Krippendorff, 2011) is computed on the ordinal scale as a formal measure of inter-rater agreement. The coefficient is 0.851, above the conventional $\alpha \geq 0.800$ threshold for reliable coding (Krippendorff, 2004). Agreement is highest for the averaged score, which pools across all four subcategories (0.865). A full breakdown by subcategory and by year is provided in Table 12.

Table 4: Inter-Model Agreement on AI Risk Scores

	Claude	ChatGPT	Gemini
Claude	1.000	0.845	0.906
ChatGPT	0.845	1.000	0.844
Gemini	0.906	0.844	1.000
Krippendorff's α (ordinal, three coders): **0.851**			

Pooled Spearman rank correlations for the AI risk score (average of four subcategories), computed across all firm-year observations rated by all three models. Krippendorff's α is on the ordinal scale.

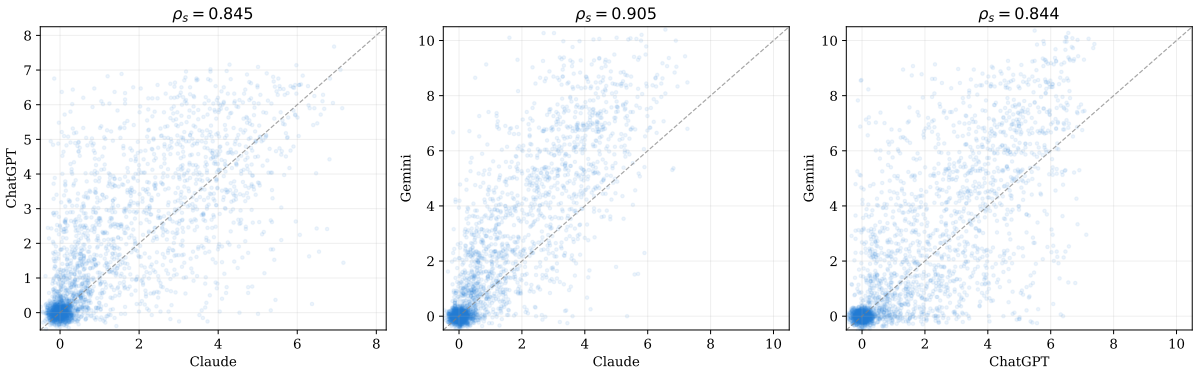

Figure 2: Pairwise scatter plots of overall AI risk scores across the three rating models, pooling all firm-year observations. Jitter added for visibility. Dashed line indicates perfect agreement.

Industry-level variation

If the risk scores capture meaningful differences in AI exposure, they should vary across industries in predictable ways. Figure 3 confirms this. In the 2024–2025 subsample, the highest average composite scores belong to Information Technology (4.8), Financials (4.3), and Communication Services (4.2)—sectors whose core operations are most directly exposed to AI-driven disruption and that count among the earliest adopters of AI tools. Health Care (3.5) also

ranks highly, consistent with the growing role of AI in diagnostics and drug discovery and the regulatory scrutiny that accompanies it. Energy (1.5) and Materials (1.7) sit at the other end. This ordering is broadly in line with the industry patterns documented by Cao et al. (2024) and Uberti-Bona Marin et al. (2025).

Figure 4 adds a time dimension. The sector ranking is relatively stable throughout the sample—Information Technology and Financials lead in every year, while Energy and Materials trail consistently. There is a difference not just in levels, but also in the timing of accelerated disclosures. Information Technology, Financials, and Health Care, for instance, saw their mean composite scores rise singificatly already in 2023. Consumer Discretionary and Energy show particularly steep accelerations in the 2024-2025 transition. By 2025 even the lowest-scoring sectors have mean values above 2.0, reflecting a broad fundamental shift in how firms discuss AI in their risk factor disclosures.

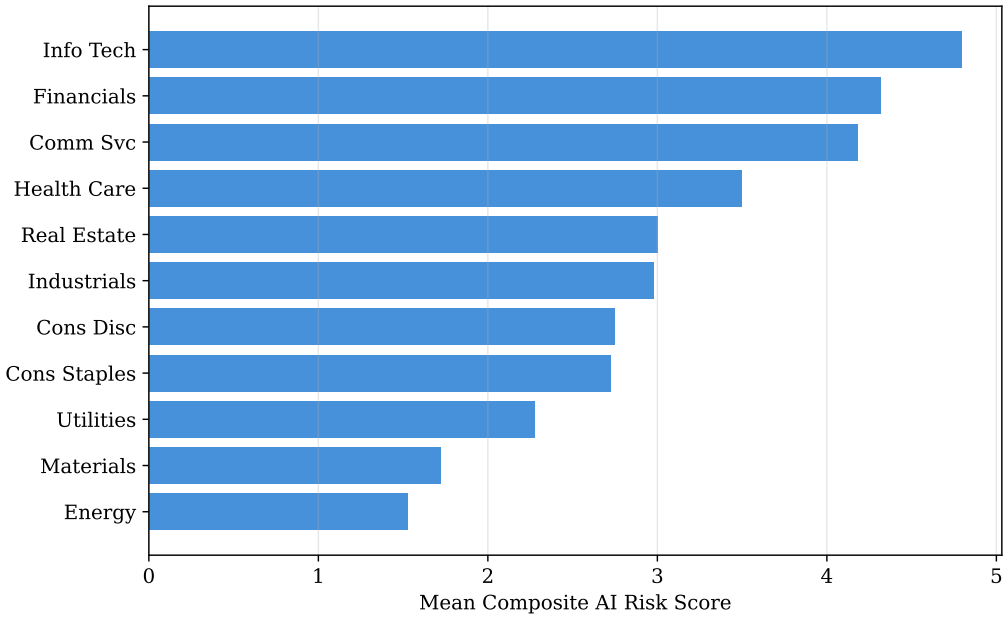

Figure 3: Mean composite AI risk score by GICS sector, fiscal years 2024–2025. The composite score is the average of the three models' overall exposure ratings.

Example disclosures

To illustrate what the scoring rubric captures in practice, Table 5 presents extracts from filings at both ends of the distribution. 5 high exposure and 5 low exposure (but non-zero) firm-year pairs are selected at random with the constraint that not more than 2 are from the same industry, and the most AI risk relevant passages are extracted. Panel A shows high-scoring disclosures. They typically name specific technologies (generative AI, large language models), identify concrete business risks (hallucination, bias, regulatory clearance), and discuss operational concerns in some detail. Panel B shows the opposite, with filings where AI appears once, almost always in a boilerplate cybersecurity paragraph, with no specifics about how AI affects the firm's business.

2.4 Empirical Methods

Following Florackis et al. (2023), portfolios are formed in April of each year. By that date the vast majority of firms have filed their 10-K for the previous fiscal year, so the risk disclosures

Table 5: Example Disclosure Extracts by AI Risk Exposure

Firm	FY	Score	Representative Extract
Panel A: High Exposure			
Meta (META)	2023	8.2	"There are significant risks involved in developing and deploying AI [...] our AI-related efforts, particularly those related to generative AI, subject us to risks related to harmful or illegal content, accuracy, misinformation [...] bias, discrimination, toxicity, intellectual property infringement [...] data privacy, cybersecurity [...] We face significant competition from other companies that are developing their own AI features and technologies."
NVIDIA (NVDA)	2025	7.1	"The computing industry is experiencing a broader and faster launch cadence of accelerated computing platforms to meet a growing and diverse set of AI opportunities [...] Governments and regulators [...] have imposed restrictions on the hardware, software, and systems used to develop frontier foundation models and generative AI. For example, the EU AI Act [...] may impact our ability to train, deploy, or release AI models."
Capital One (COF)	2025	6.3	"We rely on quantitative models and, in some cases, the use of AI [...] generative AI has been known to produce false or 'hallucinatory' inferences or output [...] AI may subject us to new or heightened legal, regulatory, ethical or other challenges [...] If the models or AI solutions [...] are deficient, inaccurate, biased, unethical or controversial—we could incur [...] legal liability, or brand or reputational harm."
Airbnb (ABNB)	2023	6.7	"The rapid evolution of AI and machine learning will require the application of resources to develop, test, and maintain our offerings [...] Uncertainty around new and emerging AI applications may require additional investment [...] AI technologies, including generative AI, may create content or information that appears correct but is factually inaccurate or flawed."
GE HealthCare (GEHC)	2025	6.2	"We are making significant investments in AI initiatives and are building AI into many of our digital offerings [...] Using AI in this manner presents risks [...] including flawed AI algorithms; insufficient, overly-broad, or biased datasets; unauthorized use or access to personal data [...] difficulties in obtaining [...] the necessary regulatory approvals or clearances."
Panel B: Low Exposure (Non-Zero)			
Coterra Energy (CTRA)	2024	0.3	"Cybersecurity risk is exacerbated with the advancement of technologies like artificial intelligence, which malicious third parties are using to create new, more sophisticated [attacks]."
NextEra Energy (NEE)	2024	0.3	"The advancement of artificial intelligence has given rise to added vulnerabilities and potential entry points for cyberattacks."
STERIS PLC (STE)	2024	0.2	"We rely on [...] technical applications and platforms, including some that employ artificial intelligence ('AI'), some of which are managed [...] by third-parties."
Amcor PLC (AMCR)	2024	0.3	"Emerging artificial intelligence technologies may intensify these cybersecurity risks."
Lennar Corp. (LEN)	2024	0.3	"Cyber intrusion efforts are becoming increasingly frequent and sophisticated, including as a result of the use of [artificial intelligence]."

Representative extracts from the Item 1A risk factor sections of selected firms. "Score" is the composite AI risk score (average of four sub-category scores, averaged across three models). High-exposure firms are drawn from the highest-scoring firm-years across diverse GICS sectors. Low-exposure firms are drawn from the lowest-scoring firm-years with nonzero composites. Brackets indicate omitted text.

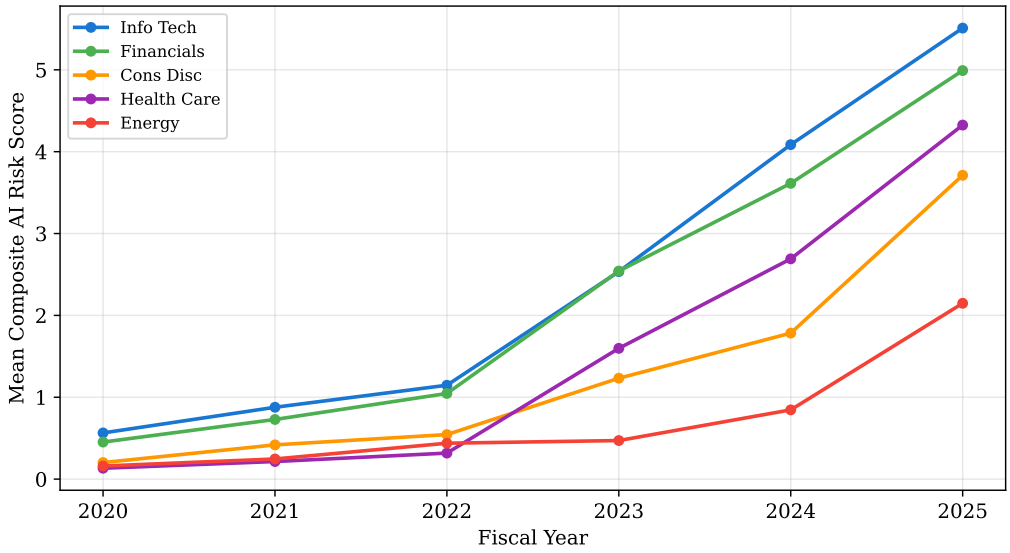

Figure 4: Evolution of mean composite AI risk scores for selected GICS sectors, fiscal years 2020–2025. Information Technology and Financials lead throughout; the gap narrows from 2023 onward as AI risk disclosure broadens across sectors.

are publicly available. AI risk ratings from fiscal year Y filings are used to sort firms into portfolios in April of year $Y+1$, and portfolio returns are tracked from May of $Y+1$ through April of $Y+2$. This timing ensures that all scores are known to investors before the return measurement period begins, avoiding look-ahead bias.

Each formation year, firms are ranked by their AI risk score and assigned to four equal-sized groups using unconditional breakpoints. Q1 (Low) contains the bottom 25% of firms by AI risk, Q4 (High) the top 25%, and Q2–Q3 the middle two quartiles. In the early years of the sample, when 65–84% of firms receive zero scores, Q1 and Q2 both consist primarily of zero-score firms while Q3–Q4 contain firms with positive scores. In the 2024–2025 subsample, where AI disclosure is more widespread, the breakpoints separate firms more evenly across the distribution.

Equal-weighted (EW) and value-weighted (VW) portfolio returns are computed within each quartile for each month, and the long-short return is Q4 minus Q1. Excess returns are the portfolio return minus the monthly risk-free rate. Long-short returns are evaluated against three factor models: the CAPM (market factor only), the Carhart (1997) four-factor model (adding SMB, HML, and momentum), and the Fama and French (2015) five-factor model (market, SMB, HML, RMW, and CMA). Alphas are the intercepts from time-series regressions of long-short returns on the respective factor sets.

All t-statistics on portfolio returns and alphas are based on Newey and West (1987) HAC standard errors with three lags, following the $T^{1/3}$ rule of thumb (approximately 2.7 for the 20-month subsample, rounded to 3, and 3.8 for the full 56-month sample). As a robustness check, the main tests are re-estimated on daily return data, where the longer time series (over 400 trading days in the subsample) renders the HAC standard errors insensitive to the lag choice; these results appear in Tables 14 and 16 in the Appendix.

Cross-sectional regressions following Fama and MacBeth (1973) are also estimated. In the first stage, individual firm excess returns are regressed cross-sectionally on lagged AI risk and control variables for each month, producing a set of monthly slope coefficients. In the second stage, the time-series average of these monthly coefficients serves as the estimate of

the risk premium, with inference based on the time-series variability of the estimates. This two-step approach avoids the cross-sectional dependence that would arise in a single pooled regression and is the standard method for testing whether a firm characteristic is priced in the cross-section (Cochrane, 2005). All variables are standardised to zero mean and unit variance. The baseline specification includes AI risk together with controls for log market capitalisation, book-to-market, return on assets (ROA), leverage, and R&D intensity. Additional specifications introduce GICS sector fixed effects, or exclude Information Technology and Communication Services firms entirely, to test whether the AI risk coefficient survives outside technology-oriented sectors. Formally, for each month t:

$$R_{i,t} - R_t^f \;=\; \gamma_{0,t} \;+\; \gamma_{1,t}\, AIRisk_{i,t-1} \;+\; \gamma_{2,t}'\, \mathbf{X}_{i,t-1} \;+\; \varepsilon_{i,t} \tag{1}$$

where $\mathbf{X}_{i,t-1}$ is the vector of standardised firm-level controls. The risk premium estimate is the time-series average of the monthly slope coefficients:

$$\hat{\gamma}_1 \;=\; \frac{1}{T}\sum_{t=1}^{T} \hat{\gamma}_{1,t} \tag{2}$$

with inference based on Newey–West standard errors of the $\{\hat{\gamma}_{1,t}\}$ series.

Bivariate portfolio sorts are used to examine whether the AI risk premium varies across firm types. In each formation year, firms are independently sorted into two groups (above and below the cross-sectional median) on a given characteristic and into AI risk quartiles as described above. The EW long-short AI risk spread is then computed separately within each characteristic half. The procedure is repeated for five characteristics: size, book-to-market, ROA, R&D intensity, and leverage.

An important concern is that the return spread may simply reflect between-sector variation. If high-AI-risk sectors such as Information Technology outperformed low AI risk sectors such as Utilities during the sample period, the portfolio sorts would register this as an AI risk premium even if AI risk played no role at the firm level. To address this, an exact between-sector versus within-sector decomposition is performed, alongside within-sector quartile sorts. In the decomposition, the total Q4$-$Q1 spread each month is split additively into a between-sector component (differences in sector-level mean returns across the quartiles) and a within-sector component (differences in firms' deviations from their sector means). For the within-sector sorts, firms are ranked into AI risk quartiles within each GICS sector in each year, and the long-short spread is computed within each sector-month pair. These spreads are then averaged across sectors with equal weight, producing a sector-neutral return. Sectors with fewer than eight firms in a given year are excluded. Formally, each firm's return is decomposed as $R_{i,t} = \bar{R}_{s(i),t} + (R_{i,t} - \bar{R}_{s(i),t})$, where $\bar{R}_{s(i),t}$ is the equal-weighted return of firm i's GICS sector. Averaging each component across Q4 and Q1 firms gives:

$$\underbrace{\bar{R}_t^{Q4} - \bar{R}_t^{Q1}}_{\text{Total spread}} \;=\; \underbrace{\bar{R}_{s,t}^{Q4} - \bar{R}_{s,t}^{Q1}}_{\text{Between-sector}} \;+\; \underbrace{\left(\bar{R}_t^{Q4} - \bar{R}_{s,t}^{Q4}\right) - \left(\bar{R}_t^{Q1} - \bar{R}_{s,t}^{Q1}\right)}_{\text{Within-sector}} \tag{3}$$

Given that asymptotic t-statistics may not be reliable with only 20 monthly observations, a permutation-based placebo test is also employed. In each of 1,000 iterations, AI risk scores are randomly shuffled across firms within each formation year, preserving the cross-sectional distribution of scores but breaking any link between a firm's AI risk and its subsequent returns. The EW long-short spread is recomputed for each permutation, building up a null distribution under the hypothesis of no relationship between AI risk and returns. The one-sided p-value is

the fraction of permuted spreads that equal or exceed the observed spread.

Two sample periods are used throughout. The full sample covers May 2021 through December 2025 (56 return months), based on ratings for fiscal years 2020 through 2024. However, meaningful AI-related disclosure only became widespread after the launch of ChatGPT in November 2022; the pre-2023 ratings are largely zeros, with a few exceptions. The subsample therefore restricts attention to fiscal year 2023–2024 ratings, where cross-sectional variation in AI risk scores is substantially greater. This gives portfolio returns from May 2024 through December 2025, covering 20 calendar months across two formation dates. Results are reported for both periods.

2.4.1 High-Attention Day Identification and SVI Regressions

The portfolio sorts and Fama-MacBeth regressions test whether AI exposure predicts the cross-section of average returns. A different but complementary question is whether AI-exposed firms react differently on days when AI-related risk becomes important to the public. To identify such days this paper draws inspiration from the Google Search Volume Index (SVI) methodology of Florackis et al. (2023).

SVI, accessed through Google Trends, measures the relative intensity of user search queries for a given term on a 0–100 scale. Daily SVI series are collected for two search terms, "AI regulation" and "AI safety", over the period May 2024 to December 2025. These search terms were chosen due to being a major concern to exposed firms and carrying a largely negative connotation. The terms also experienced a large volume of searches during the entire 20 month sample period. Because Google Trends returns daily-resolution data only for query windows of approximately eight months or fewer, three overlapping windows are constructed and stitched using ratio-of-medians scaling on the overlap days, following the general procedure described in Florackis et al. (2023). This produces a single continuous daily SVI series for each keyword.

For each keyword, abnormal SVI on day t is defined as the ratio of raw SVI to its trailing 14-day median:

$$ASVI_t = \frac{SVI_t}{\text{median}(SVI_{t-14}, \ldots, SVI_{t-1})}$$

A day is flagged as high-attention if abnormal SVI exceeds the prior 14-day rolling mean of abnormal SVI by more than two standard deviations. The aggregate event dummy $HighSVI_t$ equals one if either keyword triggers the flag on day t. This procedure identifies 22 high-attention days, all in 2025.

The SVI event dummy is used in a firm-day panel regression:

$$CAR_{i,[t,t+k]} = \alpha_i + \beta_1 HighSVI_t + \beta_2 HighSVI_t \times AIExposure_i + \varepsilon_{i,t} \tag{4}$$

where $CAR_{i,[t,t+k]}$ is the cumulative abnormal return for firm i over the window $[t, t+k]$, $AIExposure_i$ is the composite AI risk score, and α_i is firm (or sector) fixed effects. The coefficient of interest is β_2. A negative estimate implies that firms with higher AI exposure earn lower abnormal returns on days of increased public attention to AI risk.

Two definitions of abnormal returns are considered. The primary specification uses market-adjusted returns (firm return minus the S&P 500 return). As a robustness check abnormal returns are also computed from a Fama–French five-factor plus momentum model, with factor loadings estimated over calendar year 2024 (approximately 250 trading days). Two fixed-effects structures are employed. Firstly, firm fixed effects, which absorb all firm characteristics that don't vary over time, including the level of AI exposure, and secondly GICS sector fixed

effects, which allow the AI exposure rating to enter as a regressor while controlling for sector-wide return patterns. Standard errors are clustered by firm. Results are reported across three event windows: $[0]$, $[0,+1]$, and $[-1,+1]$. Like for the previous tests, AI risk ratings are lagged, so fiscal year t 10-K scores are matched to returns from May of year $t+1$ through April of year $t+2$, which ensures that all scores are publicly available before the return measurement period begins.

3 Results

3.1 Univariate Portfolio Sorts

The main results from the quartile portfolio sorts are presented in Table 6. Firms are sorted into four equal-sized groups based on their AI risk score, and monthly long-short (Q4−Q1) excess returns and risk-adjusted alphas are reported for all five measures.

Over the full 56-month sample, the EW long-short spread is positive for all five measures but reaches statistical significance on a risk-adjusted basis only for Gemini, which produces a five-factor alpha of 0.39% per month ($t = 2.45$). The remaining measures show positive but insignificant spreads. This is not unexpected, given that cross-sectional variation in AI risk scores was limited during the 2020–2022 period, when the majority of firms had no AI-related disclosure language at all.

The results are considerably stronger in the 20-month subsample. Statistically significant EW excess returns are obtained for four of the five measures: ChatGPT at 0.73% ($t = 2.22$), Gemini at 0.94% ($t = 2.92$), Simple Average at 0.83% ($t = 2.25$), and PC1 at 0.86% ($t = 2.43$). Claude is positive but insignificant at 0.52% ($t = 1.51$). In annualised terms, the significant spreads amount to approximately 8.8% to 11.2%. The five-factor alphas are significant for ChatGPT (0.59%, $t = 2.46$), Gemini (0.77%, $t = 3.09$), Simple Average (0.70%, $t = 2.47$), and PC1 (0.73%, $t = 2.60$), which indicates that the return spread is not explained by standard size, value, profitability, or investment factors. The return pattern across the four quartiles is perfectly monotonic (Q1 $<$ Q2 $<$ Q3 $<$ Q4) for Claude and ChatGPT. For Gemini and the two composite measures, Q3 dips slightly below Q2, though Q1 is always the lowest-returning portfolio and Q4 always the highest.

The value-weighted spreads in the subsample are even larger, with ChatGPT at 1.38% ($t = 1.83$) and Gemini at 1.36% ($t = 1.67$) per month. That VW spreads exceed EW spreads raises the possibility that a small number of large-cap, high-AI-risk firms drive the VW results disproportionately, and for this reason the EW results are treated as the primary specification throughout.

The consistency of the results across three independent LLMs and two aggregation methods is worth noting. As reported in Section 2.3, the Spearman rank correlations between the models exceed 0.84, and the PCA confirms that the three models load approximately equally on the first component, explaining 86.6% of the variance. This pattern is consistent with the return-relevant information being embedded in the 10-K text itself, rather than in any particular model's scoring tendencies. Figure 5 shows the time-series behaviour of the long-short portfolios. The spread occurs almost entirely after May 2024, which aligns with the notion that the detectable AI risk premium only materialised once firms began disclosing AI-related risks to a meaningful extent.

Table 6: Quartile Portfolio Sorts on AI Risk

		Full Sample (56 months)				Subsample (20 months)			
		Excess Ret	CAPM α	FFC α	FF5 α	Excess Ret	CAPM α	FFC α	FF5 α
Panel A: Equal-Weighted Portfolios									
Claude	Q4−Q1	0.268	0.117	0.190	0.174	0.519	0.272	0.375	0.511*
		[1.06]	[0.55]	[1.08]	[0.99]	[1.51]	[0.71]	[1.22]	[1.88]
ChatGPT	Q4−Q1	0.274	0.112	0.247	0.239	0.730**	0.402	0.440*	0.594**
		[1.04]	[0.52]	[1.52]	[1.46]	[2.22]	[1.14]	[1.91]	[2.46]
Gemini	Q4−Q1	0.427	0.266	0.383**	0.388**	0.936***	0.540	0.632***	0.769***
		[1.58]	[1.21]	[2.53]	[2.45]	[2.92]	[1.57]	[3.02]	[3.09]
Simple Avg	Q4−Q1	0.353	0.171	0.292	0.286	0.827**	0.455	0.567**	0.701**
		[1.23]	[0.74]	[1.58]	[1.48]	[2.25]	[1.11]	[2.14]	[2.47]
PC1	Q4−Q1	0.362	0.183	0.294	0.284	0.860**	0.506	0.615**	0.734***
		[1.27]	[0.79]	[1.59]	[1.49]	[2.43]	[1.26]	[2.35]	[2.60]
Panel B: Value-Weighted Portfolios									
Claude	Q4−Q1	0.501	0.240	0.434	0.363	0.810	0.088	0.585	0.859
		[0.91]	[0.54]	[1.28]	[0.94]	[1.04]	[0.12]	[1.32]	[1.46]
ChatGPT	Q4−Q1	0.661	0.382	0.659*	0.565	1.384*	0.580	1.139***	1.366***
		[1.12]	[0.80]	[1.94]	[1.55]	[1.83]	[0.90]	[3.61]	[2.81]
Gemini	Q4−Q1	0.617	0.357	0.603*	0.527	1.364*	0.556	1.082***	1.266**
		[1.04]	[0.73]	[1.73]	[1.40]	[1.67]	[0.83]	[2.77]	[2.04]
Simple Avg	Q4−Q1	0.618	0.327	0.595*	0.499	1.181	0.359	0.920**	1.101*
		[1.02]	[0.68]	[1.67]	[1.28]	[1.44]	[0.52]	[2.40]	[1.81]
PC1	Q4−Q1	0.637	0.349	0.608*	0.503	1.222	0.458	0.936**	1.079*
		[1.08]	[0.74]	[1.77]	[1.35]	[1.64]	[0.73]	[2.53]	[1.88]

Monthly long-short (Q4−Q1) portfolio returns from quartile sorts on AI risk. Firms are sorted into four equal-sized groups within each formation year. Returns are in monthly percentages. Alphas are intercepts from regressions on CAPM, Carhart four-factor (FFC), and Fama-French five-factor (FF5) models. Simple Avg is the arithmetic mean of the three model scores; PC1 is the first principal component of the three standardised model scores. Newey-West t-statistics (3 lags) in brackets. *, **, *** denote significance at the 10%, 5%, and 1% levels.

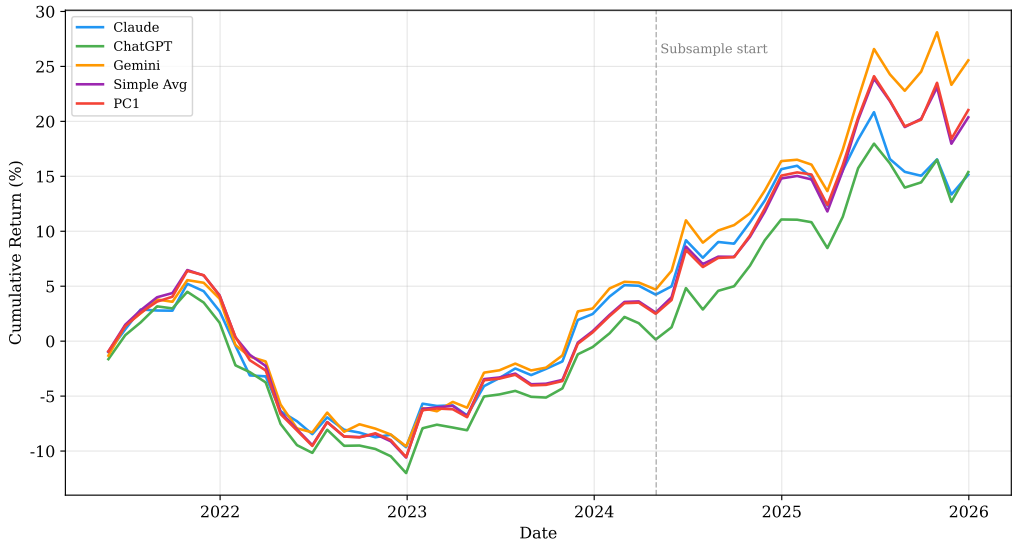

Figure 5: Cumulative return of the equal-weighted long-short (Q4−Q1) portfolios for all five AI risk measures, May 2021 through December 2025. The dashed vertical line marks the beginning of the 2024–2025 subsample.

3.2 Placebo Test

As noted above, Newey-West t-statistics rely on asymptotic properties that may not hold well with only 20 monthly observations. A permutation test that does not depend on distributional assumptions or the number of time periods is therefore used as a complement.

The results are reported in Table 7. In the 20-month subsample, the actual Q4−Q1 spread falls in the right tail of the permutation distribution for four of the five measures, with p-values of 0.005 for ChatGPT, 0.001 for Gemini, 0.003 for Simple Average, and 0.006 for PC1. Only Claude fails to reach significance ($p = 0.161$). Figure 6 illustrates this for Gemini. The 1,000 permuted spreads cluster tightly around zero while the actual spread lies well beyond the 99th percentile.

Over the full sample, Gemini remains significant ($p = 0.003$, $z = 2.78$), as do both composite measures (Simple Average $p = 0.014$; PC1 $p = 0.045$). That the composites are significant in the full sample where the individual LLMs are not likely reflects the noise-reduction benefit of averaging across models.

These permutation results confirm that the particular AI risk scores assigned to firms do predict subsequent returns. Because the test is distribution-free, it provides a useful complement to the parametric results in Table 6 and helps address the concern that the short sample could produce unreliable inference.

Table 7: Placebo Test: Permutation-Based Significance of EW Q4−Q1 Spread

Measure	Sample	Actual Spread (%)	Perm. Mean (%)	z-score	p-value
Claude	Full (56m)	0.168	≈ 0	1.34	0.098
ChatGPT	Full (56m)	0.167	≈ 0	1.54	0.060
Gemini	Full (56m)	0.319	≈ 0	2.78	0.003
Simple Avg	Full (56m)	0.273	≈ 0	2.23	0.014
PC1	Full (56m)	0.232	≈ 0	1.79	0.045
Claude	Subsample (20m)	0.229	≈ 0	0.99	0.161
ChatGPT	Subsample (20m)	0.574	≈ 0	2.42	0.005
Gemini	Subsample (20m)	0.691	≈ 0	2.92	0.001
Simple Avg	Subsample (20m)	0.648	≈ 0	2.70	0.003
PC1	Subsample (20m)	0.565	≈ 0	2.40	0.006

In each of 1,000 iterations, AI risk scores are randomly shuffled across firms within a given formation year. The EW long-short (Q4−Q1) spread is recomputed for each permutation. The z-score is (actual − permutation mean) / permutation standard deviation; the p-value is the fraction of permuted spreads $\geq$ the actual spread.

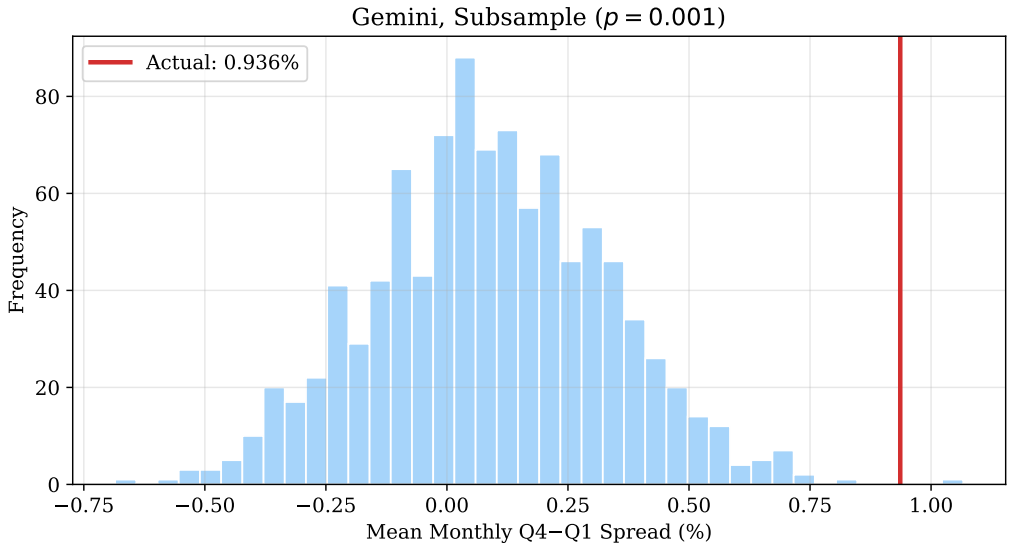

Figure 6: Distribution of 1,000 permuted EW Q4−Q1 spreads for Gemini ratings, 20-month subsample. The vertical line indicates the actual observed spread.

3.3 Within-Sector Evidence

The portfolio sorts reported above could, in principle, reflect nothing more than between-sector variation. Firms in sectors like Information Technology and Communication Services tend to receive high AI risk scores and performed well during the sample period, while sectors such as Utilities and Energy score lower and had weaker returns. If the observed premium merely reflects these sector-level return differences, it would say little about whether AI risk itself is priced at the firm level. Three sets of tests are conducted to address this concern.

First, the total Q4−Q1 spread is decomposed into an exact between-sector and within-sector component. For each month, each firm's excess return is split as $r_i = \bar{r}_{s(i)} + (r_i - \bar{r}_{s(i)})$, where $\bar{r}_{s(i)}$ is the equal-weighted return of the firm's GICS sector. The total spread equals the between-sector component (reflecting Q4 firms belonging to higher-return sectors) plus the within-sector component (reflecting Q4 firms outperforming their own sector peers). The results are reported in Table 8, Panel A. For Gemini, the between-sector component amounts to 0.51% ($t = 2.90$, 55% of total) and the within-sector component to 0.42% ($t = 1.80$, 45%). For the composite measures, the split is approximately 66/34. The between-sector channel is the dominant one for all measures, though Gemini retains a marginally significant within-sector component.

Second, firms are sorted into AI risk quartiles within each GICS sector in each formation year. The resulting within-sector Q4−Q1 spreads are reported for all eleven GICS sectors and all five measures in Panel B. The Financials sector results stand out with the within-sector spread for Gemini being 0.51% per month ($t = 2.35$), significant at the 5% level. Marginally significant spreads are also observed for Claude (0.43%, $t = 1.71$) and Simple Average (0.41%, $t = 1.90$). Consumer Discretionary produces a large spread for Gemini (1.33%, $t = 1.75$), and Energy shows a strong result for ChatGPT (1.53%, $t = 3.09$). Information Technology, notably, exhibits near-zero or slightly negative within-sector spreads across all measures, which confirms that the IT contribution to the total premium operates entirely through the between-sector channel. Health Care shows negative within-sector spreads for Claude (−0.92%, $t = -2.08$), meaning that within that sector the firms with higher AI risk disclosure actually earned lower returns.

Third, Fama-MacBeth regressions with alternative sector control specifications are reported in Panel C. The inclusion of GICS sector dummies eliminates significance for all measures. However, when Information Technology is excluded from the sample, Gemini's coefficient remains significant at the 5% level (0.21%, $t = 2.04$), indicating that the Gemini-based signal is not driven solely by IT sector membership. The other measures are weaker but remain positive. Taken together, these results show that the AI risk premium has both a between-sector and a within-sector component. The between-sector channel accounts for the majority of the observed spread, but meaningful within-sector variation is present in Financials, and Energy. A relatively significant spread can be observed in Consumer Discretionary though more obscure. It should however be noted that the results of the sector-level within-sector regressions are often contradictory, which is likely a consequence of both cross-model disagreement when it comes to more nuanced differences between firms in the same sector, and the limited power that can be achieved when the short sample period is paired with sector splits.

3.4 Subcategory Decomposition

The composite AI risk measure used in the preceding sections averages four subcategories: displacement risk, operational risk, regulatory risk, and cybersecurity risk. To identify which dimensions of AI exposure are most strongly associated with the return premium, firms are

Table 8: Between-Sector vs. Within-Sector Evidence and Fama-MacBeth with Sector Controls

	Claude		ChatGPT		Gemini		Simple Avg		PC1	
	Ret	Share	Ret	Share	Ret	Share	Ret	Share	Ret	Share
Panel A: Decomposition of EW Q4−Q1 Spread, Subsample										
Total	0.519	100%	0.730**	100%	0.936***	100%	0.827**	100%	0.860**	100%
	[1.51]		[2.22]		[2.92]		[2.25]		[2.43]	
Between	0.533***	103%	0.392***	54%	0.514***	55%	0.547***	66%	0.568***	66%
	[2.90]		[2.90]		[2.90]		[3.09]		[3.10]	
Within	−0.014	−3%	0.338	46%	0.422*	45%	0.280	34%	0.292	34%
	[−0.04]		[1.27]		[1.80]		[0.91]		[0.98]	
Panel B: Within-Sector Q4−Q1 Spread by GICS Sector, Subsample										
	Claude		ChatGPT		Gemini		Simple Avg		PC1	
Energy	−0.162		1.534***		−0.142		0.376		0.403	
	[−0.41]		[3.09]		[−0.31]		[0.97]		[1.06]	
Materials	−0.700		−0.512		−0.572		−0.574		0.456	
	[−1.16]		[−0.94]		[−1.03]		[−1.06]		[1.02]	
Industrials	0.078		−0.207		0.285		0.319		0.319	
	[0.13]		[−0.42]		[0.55]		[0.55]		[0.55]	
Cons. Disc.	0.612		0.392		1.328*		0.573		0.623	
	[0.67]		[0.47]		[1.75]		[0.65]		[0.68]	
Cons. Stap.	0.273		0.063		0.779		0.610		0.610	
	[0.37]		[0.08]		[1.03]		[0.75]		[0.75]	
Health Care	−0.923**		0.458		−0.430		−0.644		−0.493	
	[−2.08]		[0.95]		[−1.21]		[−1.33]		[−1.15]	
Financials	0.430*		−0.365		0.507**		0.413*		0.028	
	[1.71]		[−1.13]		[2.35]		[1.90]		[0.11]	
Info Tech	−0.167		−0.005		−0.081		−0.295		−0.416	
	[−0.12]		[−0.00]		[−0.07]		[−0.22]		[−0.30]	
Comm. Svc.	−0.575		0.424		0.436		−0.545		−0.297	
	[−0.70]		[0.64]		[0.59]		[−0.57]		[−0.30]	
Utilities	−0.215		0.075		0.012		0.095		−0.084	
	[−0.53]		[0.12]		[0.02]		[0.15]		[−0.14]	
Real Estate	−0.318		−0.390		−0.344		−0.135		0.188	
	[−0.64]		[−0.66]		[−0.57]		[−0.26]		[0.39]	
Panel C: Fama-MacBeth AI Risk Coef. (Subsample)										
M2: + controls	0.119		0.206		0.229*		0.212		0.203	
	[0.70]		[1.28]		[1.65]		[1.31]		[1.23]	
M3: + sect. FE	−0.074		0.043		0.072		0.031		0.018	
	[−0.42]		[0.35]		[0.64]		[0.21]		[0.12]	
M6: excl. IT	0.110		0.197		0.212**		0.193		0.185	
	[0.90]		[1.58]		[2.04]		[1.60]		[1.51]	

Panel A: Exact decomposition of the total EW Q4−Q1 spread into between-sector and within-sector components. For each month, firm returns are decomposed as $r_i = \bar{r}_{s(i)} + (r_i - \bar{r}_{s(i)})$. Share denotes the percentage of the total spread attributed to each component. Panel B: Within-sector Q4−Q1 spreads (%) from quartile sorts on AI risk within each GICS sector; sectors with fewer than 8 firms per formation year are excluded. Panel C: Fama-MacBeth coefficients ($\times 100$, monthly %) on standardised AI risk. M2 includes controls only; M3 adds GICS sector dummies; M6 excludes Information Technology. Newey-West t-statistics (3 lags) in brackets.

sorted separately on each subcategory. The EW Q4−Q1 results from quartile sorts on each subcategory for the 20-month subsample are reported in Table 9.

Displacement risk produces the largest and most consistent raw excess returns. Significant spreads of 0.82–1.00% are observed across Claude, ChatGPT, Gemini, and the composites. Five-factor alphas are significant for Claude (0.49%, $t = 1.80$), Gemini (0.62%, $t = 3.08$), and both composites (0.68% and 0.66%), though they are absorbed by standard factors for ChatGPT. For operational risk, the strongest risk-adjusted signal belongs to Gemini, with a five-factor alpha of 0.79% ($t = 3.31$), which is higher than the composite alpha. Firms integrating AI into their core operations face implementation failures, model path dependency, and adaptation costs, and this type of ongoing uncertainty appears not to be captured by standard factor models.

Regulatory risk is broadly significant for Gemini (five-factor alpha 0.61%, $t = 2.60$) and the composites. Cybersecurity risk shows more variation across models, with Gemini again producing the strongest result (0.61%, $t = 2.65$). The overall composite measure performs well but does not uniformly dominate the individual subcategories on a risk-adjusted basis, unlike in alternative quartile-based sorting approaches where the composite tends to dominate.

Table 9: Quartile Sorts by AI Risk Subcategory, 20-Month Subsample

	Claude		ChatGPT		Gemini		Simple Avg		PC1	
Subcategory	Ret	FF5 α	Ret	FF5 α	Ret	FF5 α	Ret	FF5 α	Ret	FF5 α
Displacement	0.818**	0.494*	0.676*	0.512	0.984***	0.620***	1.000***	0.675**	0.976***	0.664***
	[2.14]	[1.80]	[1.93]	[1.55]	[2.84]	[3.08]	[2.59]	[2.54]	[2.60]	[2.60]
Operational	0.238	0.294	0.731***	0.541**	0.894***	0.791***	0.645*	0.540**	0.677*	0.615***
	[0.82]	[1.29]	[2.67]	[2.27]	[2.60]	[3.31]	[1.80]	[2.20]	[1.88]	[2.78]
Regulatory	0.477	0.530	0.166	0.353	0.771**	0.614***	0.537	0.598**	0.565	0.614*
	[1.45]	[1.60]	[0.52]	[1.15]	[2.40]	[2.60]	[1.45]	[2.04]	[1.53]	[1.95]
Cybersecurity	0.244	0.449**	0.349**	0.199	0.726**	0.605***	0.513	0.418	0.428	0.353
	[0.89]	[2.29]	[2.21]	[1.30]	[2.36]	[2.65]	[1.59]	[1.52]	[1.48]	[1.42]
Composite	0.519	0.511*	0.730**	0.594**	0.936***	0.769***	0.827**	0.701**	0.860**	0.734***
	[1.51]	[1.88]	[2.22]	[2.46]	[2.92]	[3.09]	[2.25]	[2.47]	[2.43]	[2.60]

EW long-short (Q4−Q1) monthly returns (%) from quartile sorts on each AI risk subcategory individually, 20-month subsample. Newey-West t-statistics (3 lags) in brackets.

3.5 Bivariate Portfolio Sorts

Bivariate portfolio sorts are conducted to examine whether the AI risk premium survives after conditioning on firm characteristics. Firms are independently sorted into two groups (above and below the cross-sectional median) on each of five characteristics, and simultaneously into AI risk quartiles. The Q4−Q1 spread is then computed separately within each characteristic subgroup. Results for the 20-month subsample are reported in Table 10.

The most notable pattern is the asymmetry across book-to-market groups. Among high book-to-market (value) firms, the AI risk spread is large and significant for all five measures, with raw returns of 0.49–0.98% and five-factor alphas of 0.50–0.91%. Among low book-to-market (growth) firms, the spreads are positive but generally insignificant. This is consistent with the AI risk premium being concentrated among value firms, where AI-driven disruption may pose a more direct threat to established business models.

A similar asymmetry is observed for profitability. Among low-ROA firms, the AI risk spread exceeds 1% per month for Gemini, Simple Average, and PC1, with significant five-

factor alphas. Among high-ROA firms, the raw spreads are smaller and mostly insignificant, although the five-factor alphas remain significant for several measures. Less profitable firms that also disclose high AI risk thus appear to command the largest premium.

For the size sort, the pattern is more balanced. The AI risk spread is significant among large firms for ChatGPT (0.89%, $t = 1.96$), Gemini (1.07%, $t = 2.41$), and both composites, while among small firms the five-factor alphas are significant for Gemini (0.50%, $t = 1.97$) and both composites. The R&D sort produces significant spreads in both halves for most measures, with slightly larger spreads among high-R&D firms. The leverage sort shows that the premium concentrates among low-leverage firms, where the five-factor alphas reach 0.72–0.93% for the composites. Among high-leverage firms, the raw spreads are positive but the five-factor alphas are generally not significant.

Table 10: Bivariate Portfolio Sorts: AI Risk Premium by Firm Characteristics, Subsample

		Claude		ChatGPT		Gemini		Simple Avg		PC1	
Char.	Group	LS Ret	FF5 α	LS Ret	FF5 α	LS Ret	FF5 α	LS Ret	FF5 α	LS Ret	FF5 α
Size	Low	0.240	0.474*	0.310	0.207	0.557	0.502**	0.506	0.481**	0.531	0.542**
		[0.52]	[1.79]	[0.86]	[0.90]	[1.63]	[1.97]	[1.41]	[2.04]	[1.34]	[2.10]
	High	0.839**	0.692**	0.893*	0.838***	1.066**	0.903***	0.929**	0.706*	0.842*	0.638*
		[1.98]	[2.04]	[1.96]	[2.63]	[2.41]	[2.87]	[2.11]	[1.96]	[1.85]	[1.72]
B/M	Low	0.641	0.562	0.657	0.660	0.912*	0.679	0.636	0.540	0.618	0.560
		[1.21]	[1.34]	[1.02]	[1.50]	[1.67]	[1.53]	[0.98]	[1.18]	[0.94]	[1.28]
	High	0.485*	0.631**	0.851***	0.500*	0.952***	0.769***	0.984***	0.910***	0.953***	0.896***
		[1.78]	[2.29]	[3.06]	[1.83]	[4.04]	[3.30]	[3.48]	[2.88]	[3.41]	[2.90]
ROA	Low	0.560	0.413	0.836**	0.478	1.245***	0.840**	1.061***	0.637	0.951**	0.647*
		[1.61]	[1.01]	[2.44]	[1.64]	[3.90]	[2.07]	[2.99]	[1.63]	[2.55]	[1.75]
	High	0.502	0.677**	0.447	0.596*	0.592	0.634**	0.496	0.603*	0.572	0.676**
		[1.03]	[2.23]	[0.77]	[1.88]	[1.24]	[2.28]	[0.95]	[1.94]	[1.11]	[2.27]
R&D	Low	0.341	0.240	0.562*	0.440	0.569*	0.302	0.746**	0.534	0.636*	0.487
		[0.78]	[0.64]	[1.76]	[1.47]	[1.75]	[1.10]	[2.01]	[1.57]	[1.71]	[1.50]
	High	0.578	0.796***	0.603	0.526*	1.095**	0.962***	0.697	0.803***	0.727	0.817***
		[1.38]	[2.86]	[1.23]	[1.94]	[2.35]	[2.86]	[1.33]	[2.74]	[1.42]	[2.77]
Leverage	Low	0.539	0.868*	0.714	0.757	0.949**	0.932**	0.648	0.720*	0.722	0.793*
		[1.05]	[1.85]	[1.48]	[1.48]	[2.32]	[2.58]	[1.43]	[1.66]	[1.57]	[1.83]
	High	0.261	−0.016	0.703**	0.389	0.878**	0.462	0.709*	0.333	0.633	0.260
		[0.65]	[−0.05]	[1.99]	[1.13]	[2.21]	[1.07]	[1.79]	[0.79]	[1.59]	[0.66]

Independent double sorts on AI risk quartiles and firm characteristic median splits. EW long-short (Q4−Q1) monthly returns (%) within each characteristic subgroup, 20-month subsample. Newey-West t-statistics (3 lags) in brackets.

3.6 SVI Regressions

If the return premium documented in the portfolio sorts reflects compensation for bearing AI-related risk, firms with higher AI exposure should underperform on days when that risk becomes more salient to the market. This is tested by estimating the panel regression described in Section 2.4.1 on a firm-day panel, interacting the composite AI exposure score with a dummy for high-attention days identified systematically from public search interest in AI regulation and safety. The procedure yields 22 high-attention days.

The interaction coefficient β_2 is reported in Table 11 across three event windows, two abnormal return definitions, and two fixed-effects structures.

Under the market-adjusted specification, the results in columns (1) and (3) are uniformly significant at the 1% level. On the event day itself, each additional point of composite AI exposure is associated with -4.3 basis points of abnormal return under firm fixed effects ($t = -5.02$) and -4.0 basis points under sector fixed effects ($t = -4.64$). The estimates increase with the event window: the $[0,+1]$ coefficients are -6.5 and -5.9 basis points, respectively. In economic terms, a firm at the 75th percentile of composite exposure versus the 25th percentile (a gap of approximately 3.4 points) earns a differential abnormal return of roughly -15 basis points on the event day, growing to -22 basis points over $[0,+1]$. Moving from firm fixed effects to sector fixed effects changes the point estimate by less than 10% in every specification, which indicates that the differential reaction is not driven by sector-level differences in AI exposure.

The factor-adjusted results in columns (2) and (4) preserve the negative sign for the $[0]$ and $[0,+1]$ windows but are not statistically significant. This attenuation can be traced to the behaviour of the Fama–French factor portfolios on event days. The market factor averages -30 basis points and the HML factor averages $+22$ basis points on high-attention days, compared to $+9$ and -1 basis points on normal days. Because firms with high AI exposure tend to load as growth stocks (negative HML beta), the factor model attributes part of their underperformance to a value–growth difference rather than to AI-specific risk. The $[-1,+1]$ factor-adjusted coefficients flip sign, presumably as the pre-event day introduces factor noise unrelated to the attention spike.

Table 11: AI Exposure and Abnormal Returns on High AI-Attention Days

	Firm FE		Sector FE (11)	
	Mkt-adj (1)	FF5+Mom (2)	Mkt-adj (3)	FF5+Mom (4)
Window: $[0]$				
High Att. × AI Exp.	-4.329^{***}	-0.986	-3.964^{***}	-0.621
	(0.863)	(0.915)	(0.855)	(0.919)
High Attention	24.500^{***}	-2.351	23.447^{***}	-3.158
	(3.589)	(3.819)	(3.583)	(3.892)
Window: $[0,+1]$				
High Att. × AI Exp.	-6.518^{***}	-1.434	-5.912^{***}	-0.746
	(1.501)	(1.515)	(1.450)	(1.508)
High Attention	34.919^{***}	-6.423	33.231^{***}	-7.897
	(5.926)	(5.864)	(5.829)	(5.984)
Window: $[-1,+1]$				
High Att. × AI Exp.	-6.774^{***}	1.587	-5.882^{***}	2.609
	(2.047)	(2.018)	(1.955)	(2.003)
High Attention	33.819^{***}	-24.261^{***}	31.427^{***}	-26.341^{***}
	(7.694)	(7.523)	(7.495)	(7.708)
Fixed effects	Firm (496)	Firm (496)	Sector (11)	Sector (11)
AI Exposure level	Absorbed	Absorbed	Yes	Yes
Clustering	Firm (496 clusters)			
N	$\approx 201{,}000$			

Coefficients from firm-day panel regressions of cumulative abnormal returns (in basis points) on the interaction of a high-attention dummy with firm-level AI exposure, as specified in equation (4). AI Exposure is the simple average across three LLMs (Claude, ChatGPT, Gemini) of each model's mean of the four subcategory scores (displacement, operational, regulatory, cybersecurity; each scored 0–10). Ratings are lagged: fiscal year t 10-K scores are matched to returns from May of year $t+1$ through April of year $t+2$. Market-adjusted returns subtract the S&P 500 daily return; FF5+Momentum returns subtract predicted returns from a six-factor model estimated over calendar year 2024. High-attention days are identified as described in Section 2.4.1 (22 event days). Under firm FE, the AI Exposure main effect is absorbed; under sector FE it enters as a level control. Standard errors are clustered by firm (in parentheses). *, **, *** denote significance at the 10%, 5%, and 1% levels.

4 Discussion

The empirical evidence is broadly coherent, but careful interpretation is needed. In the full 56-month sample, only the Gemini-based long-short portfolio yields a significant five-factor alpha, reflecting the limited cross-sectional variation in AI risk disclosures during 2020–2022. In the post-2023 subsample, four of the five measures produce significant equal-weighted spreads with five-factor alphas of 0.59–0.77% per month, and the permutation test confirms that these are unlikely to be caused by the small sample. An exact decomposition shows that 54–66% of the spread operates through the between-sector channel, with the within-sector component concentrated in Financials and Consumer Discretionary. The SVI panel regressions show that high-exposure firms earn significantly lower abnormal returns on days of elevated public attention to AI regulation and safety, and bivariate sorts reveal that the premium concentrates among value and low-profitability firms. Together, these patterns are consistent with disclosed AI risk carrying return-relevant information since disclosure has become widespread, but the majority of that information operates through the between-sector channel.

For the SVI high attention days, we observe significantly lower returns per point on the AI risk exposure scale when adjusting for the market, with both firm level fixed effects as well as sector level. The results suggest that around days with high attention towards the negative topics chosen earlier, highly exposed firms experience a negative shift relative to less exposed firms. The factor-adjusted specifications attenuate this result, likely because high AI exposure firms tend to load as growth stocks (negative HML beta), and the HML factor averages +22 basis points on event days versus -1 basis points on normal days. The factor model therefore attributes part of the underperformance of high-exposure firms to a value–growth factor rather than to AI-specific risk, even though it is plausibly triggered by the same attention shock.

The SVI results demonstrate that the AI risk disclosure index we constructed is not limited to capturing solely the upside of firms classified as disclosing high levels of AI risk, which may be argued to be attributed to the ongoing AI related stock market boom. Instead, the results show that our measure is also able to capture the downside during days where there is high intensity of AI related Google searches with negative associations, and the market moves up regardless.

A related consideration is whether the results say something about AI risk at the firm level or merely capture the strong performance of technology and AI exposed sectors. It could be that firms in sectors like IT or Communication Services generally score high because their filings naturally discuss AI more thoroughly and with higher specificity, and these sectors did well for reasons that are unrelated to the AI exposure of any individual firm.

The decomposition shows that 54–66% of the observed differential returns operates through the between-sector channel, meaning firms in highly AI exposed sectors like IT and Financials score higher and earn higher returns. While this does not imply that an AI risk premium doesn't exist, it does limit the claim that it is priced at the firm level. As discussed below, this can be caused by multiple factors, including the early maturity of AI risk disclosure and a limited sample period.

The within-sector sorts, although hampered by limited power due to a short sample period, can help illuminate the issue. IT, despite being the highest-scoring sector, produces near-zero within-sector spreads across all five measures, suggesting that the index cannot differentiate meaningfully among firms in a sector where virtually every filing discusses AI extensively. The sectors where within-sector sorts do produce significant spreads across some models, like Financials and Consumer Discretionary, have concrete internal variation in AI adoption. Adding GICS sector fixed effects to the Fama-MacBeth regressions eliminates significance for all mea-

sures, which is the strongest evidence against a firm-level interpretation. However, excluding only IT leaves Gemini's coefficient significant at the 5% level, indicating that the signal is not solely an artefact of IT sector membership and that the index carries some firm-level information in sectors where AI risk is material but not yet fully incorporated through other channels.

A complementary form of heterogeneity emerges from the bivariate sorts. The AI risk premium concentrates among high book-to-market firms, with five-factor alphas of 0.50–0.91% across measures, while spreads in the growth half are positive but largely insignificant. A similar asymmetry holds for profitability, where raw spreads exceed 1% per month for several measures among low-ROA firms but are smaller in the high-ROA half. This is consistent with a risk-based reading, where legacy and lower-margin firms have the most to lose from AI-driven disruption and the least operating slack to fund offsetting investment, so disclosed exposure plausibly translates into more material economic risk for them. Read alongside the SVI evidence, the firms earning the largest unconditional premium are also those most punished on attention-driven downside days.

The AI exposure measure construction and validation contribute meaningfully to the literature on using large language models to evaluate corporate risk disclosures. Firstly, the high cross-model agreement, as shown by pooled Spearman coefficients of 0.84-0.91 and a first principal component that captures 86.6% of the variance with nearly equal loadings, suggests that the LLMs do not suffer to a high degree from idiosyncratic model biases but accurately incorporate the common information in the 10-K extracts.

Secondly, we observe that composite scores, whether arithmetic means or principal components, perform well when compared to individual model ratings. While significance levels may be lower in individual specifications due to a dilution of the idiosyncrasies of any single model, they are more consistent and generally relatively high across multiple specifications. This echoes the findings of the ensemble forecasting literature, which has found that combining multiple independent forecasters typically outperforms any single forecaster over time (for example, Timmermann (2006)).

There are a few limitations in the data and methodology that affect the possibility of drawing reliable conclusions. Firstly, because AI risk disclosure in 10-K filings are new and developing, the sample period with high quality data is necessarily short. Fiscal year 2023 is the first with significant non-zero disclosure, and all filings for that year weren't published until the spring of 2024. We therefore form portfolios on 1st of May, which leaves us with only 20 monthly observations until the end of 2025. Although steps are taken to mitigate the effects of a small sample size, including adjusting the NW error lag and confirming the robustness of key results using permutation-based approaches, the short period remains an impediment to the power of tests that are made on a subsection of the data. This puts into question the validity of the results, both null and otherwise, most notably of the within-sector spreads with the sample being reduced further.

Secondly, with AI still constituting a recent field, and more so a new risk category, it remains largely unregulated. Specifically to this paper, this matters as AI risk disclosures in 10-K filings remain entirely voluntary and unregulated by the SEC. This can mean that relative to cybersecurity risk, which has well defined regulations and reporting standards, disclosed AI risk cannot be trusted to the same extent as it does not carry the same legal backlash and implications. This fundamental difference could possibly explain the difference in statistical significance between our paper and Florackis et al. (2023) between mutual tests.

Third, there are also some inherent issues with the LLM rating methodology. While the prompt design attempts to limit the information scope of the LLMs to the section 1A extract at hand, there is no reliable way to verify compliance. The models may identify the firms

and rate their exposure using their general training, which could help to explain the less-than-perfect agreement. While this would not necessarily compromise the performance in terms of examining differences in the cross-section of stock returns, it could affect the comparability across models, as well as the general conclusions about LLM use in analyzing corporate risk statements that can be drawn from this study.

Also, it is important to acknowledge that the results observed from the LLMs may be skewed due to resource limitations. Under ideal conditions, all state of the art LLMs would have been used in the ratings process, and specific model configuration would have been accessible. More specifically, the ability to set the temperature setting, which controls the model's randomness and creativity, would have allowed tuning the model to be more conservative in its approach. This could have resulted in more grounded and reliable ratings from the LLMs and perhaps higher correlation between them as well.

Several extensions of this work appear natural. The first concerns the disclosure regime itself. Future SEC guidance on AI risk disclosure, analogous to the cybersecurity guidance issued from 2011 onward, would standardise both the structure and the substance of what firms report. Such standardisation should also sharpen the measure's sub-categories, allowing investors to distinguish more cleanly between displacement, operational, regulatory, and cybersecurity exposures rather than relying on the bundled signal that current voluntary disclosure provides. As the regulatory framework develops, more of the within-sector component of the spread, to the degree that it exists, should become identifiable. A second straightforward extension is sample length. Fiscal year 2023 is the first year of substantive AI disclosure, and as additional years accrue, the small-sample concerns documented above will dissipate, enabling tests of persistence and regime stability that the present sample cannot support.

A more substantive extension would link disclosed AI risk to realised AI-related events such as lawsuits, regulatory enforcement actions, AI-related restatements, or material reputational incidents, mirroring what Florackis et al. (2023)'s verified-attack sample did for cybersecurity. Such a test would provide external validation of the index against verifiable downside outcomes. We did not pursue this in the present thesis for two connected reasons. Negative AI related events lack the ex ante definition that cyberattacks have. Cybersecurity benefits from SEC-required 8-K disclosure of material breaches and several established incident databases, while AI does not. A defensible event sample would require hand-collection from press coverage, court filings, and enforcement registers, with substantial researcher judgement about what qualifies. Furthermore, the EU AI Act's enforcement provisions only began phasing in during 2025, so the population of formally actionable AI related events remains small. Both constraints should weaken as the regulatory infrastructure matures and as a longer track record of incidents accumulates.

A related extension is to broaden the sample beyond US-listed firms. EU AI Act (Regulation (EU) 2024/1689) imposes mandatory transparency and risk-management obligations on covered firms and thereby creates the kind of standardised disclosure environment that AI lacks in the US. Replicating the analysis on EU AI Act-covered firms, or on the US-listed subsidiaries of EU-covered groups, would help separate the disclosure-quality channel from the disclosure-intensity channel. If firm-level pricing emerges more cleanly under a mandatory regime, the implication is that the partial null we document in the US setting reflects unstandardised reporting rather than the absence of a priced AI risk factor at the firm level.

5 Conclusion

The post-2023 evidence is consistent with AI risk disclosure carrying information for equity returns. Long-short quartile portfolios deliver Fama-French five-factor alphas of 0.59% per month for ChatGPT and 0.77% for Gemini in the 20-month subsample, with composite measures producing similar magnitudes and consistent monotonicity across quartiles. The SVI regressions show that high-exposure firms underperform the market on days of elevated AI-related search activity, indicating that the index reflects downside exposure too and not just membership in sectors that have benefited from the strong recent tech and AI performance. Taken together, the cross-section and the event-style results suggest that firms whose Item 1A language signals greater AI risk exposure have, in the period where such disclosure has been substantive, earned higher average returns and borne greater losses on bad news days.

We are nonetheless cautious about pushing this into a Florackis-like claim that AI risk is priced at the firm level. Three structural features of our setting separate it from theirs and explain why we should not expect the same clarity. First, AI risk disclosure remains voluntary and unstandardised, whereas SEC cybersecurity guidance has shaped reporting language since 2011; firms therefore differ in how, not only how much, they discuss AI. Second, fiscal year 2023 is the first year of substantial disclosure, which leaves a post-formation window of only twenty months and limits the precision of any firm-level inference. Third, AI as a risk category is itself in formation. While the cost of cybersecurity failures have been thoroughly documented, the early stage of the widespread adoption of artificial intelligence means that AI risk still remains a comparatively nebulous concept when applied to the average firm. None of these features is permanent. The natural reading of our results is therefore that disclosure intensity already carries return-relevant information and that any pricing channel is likely to strengthen as the sample lengthens and as disclosure becomes more standardised.

A second contribution concerns the development and use of the AI risk scores themselves. Three frontier LLMs from different developers, given identical prompts and inputs, produce scores that agree to a high degree, and composite measures consistently outperform any single model. We read this as evidence that the rating procedure recovers a common signal from the underlying text rather than amplifying any one model's idiosyncrasies, and that ensemble rating across LLM coders is a useful discipline whenever a labeled training sample like that of Florackis et al. (2023) is not available.

What our results describe is an early stage of a process rather than its endpoint. AI risk has begun to appear in disclosed text and in equity returns, and these patterns are likely to become sharper as the disclosure regime, the possible sample, and the category itself continue to develop.

References

Barthwal, Devaang, 2024, S&P 500 holdings and weights (SPY) 2000–2024 (dataset), Kaggle, available at https://www.kaggle.com/datasets/devaangbarthwal/s-and-p-500-holdings-and-weights-spy-2000-2024.

Cao, Yang, Miao Liu, Jiaping Qiu, and Ran Zhao, 2024, Information in disclosing emerging technologies: Evidence from AI disclosure, SSRN Working Paper No. 4987085.

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

Center for Research in Security Prices (CRSP), 2026, CRSP US Stock Database, Wharton Research Data Services, available at https://wrds-www.wharton.upenn.edu/.

Cochrane, John H., 2005, *Asset Pricing*, revised edition (Princeton University Press, Princeton, NJ).

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, 2020, Lazy prices, *Journal of Finance* 75, 1371–1415.

European Union, 2024, Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), *Official Journal of the European Union* L 2024/1689, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

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

Florackis, Chris, Christodoulos Louca, Roni Michaely, and Michael Weber, 2023, Cybersecurity risk, *Review of Financial Studies* 36, 351–407.

French, Kenneth R., 2026, Data library: Fama/French factors, Tuck School of Business at Dartmouth, available at https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Google, 2025, Google Trends, available at https://trends.google.com/trends/, accessed 2025.

Kim, Alex G., Maximilian Muhn, and Valeri V. Nikolaev, 2024a, Bloated disclosures: Can ChatGPT help investors process information?, SSRN Working Paper No. 4425527.

Kim, Alex G., Maximilian Muhn, and Valeri V. Nikolaev, 2024b, Financial statement analysis with large language models, arXiv preprint No. 2407.17866.

Krippendorff, Klaus, 2004, Reliability in content analysis: Some common misconceptions and recommendations, *Human Communication Research* 30, 411–433.

Krippendorff, Klaus, 2011, Computing Krippendorff's alpha-reliability, Departmental Papers (ASC), University of Pennsylvania.

Lopez-Lira, Alejandro, and Yuehua Tang, 2023, Can ChatGPT forecast stock price movements? Return predictability and large language models, SSRN Working Paper No. 4412788.

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *Journal of Finance* 66, 35–65.

National Institute of Standards and Technology (NIST), 2024, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1 (U.S. Department of Commerce), available at https://doi.org/10.6028/NIST.AI.600-1.

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.

S&P Global Market Intelligence, 2026, Compustat North America Fundamentals Annual, Wharton Research Data Services, available at https://wrds-www.wharton.upenn.edu/.

Timmermann, Allan, 2006, Forecast combinations, in Graham Elliott, Clive W. J. Granger, and Allan Timmermann, eds.: *Handbook of Economic Forecasting*, vol. 1, chap. 4 (North-Holland, Amsterdam) 135–196.

Uberti-Bona Marin, Lucas G., Bram Rijsbosch, Gerasimos Spanakis, and Konrad Kollnig, 2025, Are companies taking AI risks seriously? A systematic analysis of companies' AI risk disclosures in SEC 10-K forms, SSRN Working Paper No. 5368860.

U.S. Securities and Exchange Commission, 2026, EDGAR full-text search: Company filings, available at https://efts.sec.gov/LATEST/search-index.

Appendix

A. AI Risk Scoring Prompt

Context Provided

You will receive:

- The company name, ticker symbol, and CIK number
- The full text of Item 1A (Risk Factors) from the firm's annual 10-K filing with the U.S. Securities and Exchange Commission

If any portion of the provided text is clearly not part of the Risk Factors section (e.g., table of contents fragments, headers from adjacent sections), ignore those portions entirely.

Your Task

Evaluate the AI-related risk exposure communicated in this firm's risk disclosure. Base your evaluation solely on the language, specificity, and framing present in the provided text. Do not use any prior knowledge about the company, its industry, competitors, share price, or macroeconomic context. You are measuring what the disclosure *says* about AI exposure — not what you believe the firm's true AI exposure to be.

If the text contains no discussion of AI, machine learning, automation, algorithmic systems, natural language processing, or related technologies, assign 0 to all scores and skip directly to the OUTPUT step.

Step 1: Identification

Identify all passages in the text that discuss artificial intelligence, machine learning, automation, algorithmic systems, natural language processing, or related technologies.

For each passage, note:

- The key phrase(s) (brief quote, not full paragraphs)
- Its approximate location in the text (e.g., "within the competition risk factor" or "in a standalone AI risk factor")

If no passages are found, state "No AI-related content identified" and skip to OUTPUT with all scores set to 0.

Step 2: Classification

Classify each identified passage into one or more of the following sub-categories:

(a) **Displacement Risk** — AI as a competitive threat. The firm's products, services, or business model could be disrupted or rendered obsolete by AI technologies deployed by competitors, new entrants, or substitute products. Score this based on what the disclosure states about competitive AI threats, not on your own assessment of whether the threat is real.

(b) **Operational Risk** — AI as an implementation challenge. Risks arising from the firm's own adoption or deployment of AI, including issues of reliability, accuracy, bias, workforce displacement, integration costs, or reputational damage from AI failures.

(c) **Regulatory/Legal Risk** — Risks from current or anticipated regulation of AI technologies that could affect the firm's operations, compliance burden, or cost structure. Includes intellectual property risks related to AI-generated content.

(d) **Cybersecurity & Data Privacy Risk** — AI-attributed expansion of the cyber threat landscape or data protection exposure. Includes AI-enabled cyberattacks, automated breaches, adversarial AI, misuse of personal data through AI systems, or failures in safeguarding sensitive information specifically linked to AI adoption or AI-powered threats.

A single passage may be classified under multiple sub-categories.

Step 3: Specificity Assessment

For each classified passage, assess how specific and substantive the discussion is. Use these criteria:

- Does the passage name specific AI technologies (e.g., large language models, computer vision, autonomous systems, generative AI)?
- Does the passage identify specific business lines, products, or operations affected?
- Does the passage quantify any AI-related impact (e.g., investment amounts, headcount, revenue exposure)?
- Does the passage describe concrete actions taken or planned in response?
- Or is the passage generic boilerplate that could apply to any firm in any industry?

These criteria inform the scoring in the next step. Higher specificity = higher score within a given sub-category.

Step 4: Scoring

Score each of the four sub-categories independently on a 0–10 scale. Then assign the `overall_exposure` score as the arithmetic mean of the four sub-category scores.

Scoring Scale (applies to each sub-category individually)

Score	Description
0	**No mention:** The disclosure contains no content relevant to this sub-category.
1–2	**Minimal / Boilerplate:** AI appears only in passing or generic language within this risk dimension. No specific technologies, business lines, or impacts are identified. Example: a single clause mentioning "technological change including AI" within a broader competition risk factor.
3–4	**Moderate / General:** One or more passages address this risk dimension with some relevance to the firm, but the discussion remains general. Limited concrete detail about specific technologies, business lines, or planned responses.
5–6	**Substantial / Specific:** Detailed, firm-specific discussion of AI exposure within this dimension. Specific technologies, business lines, or operational impacts are identified. The firm demonstrates awareness of how AI specifically affects this area of its business.
7–8	**Extensive / Detailed:** This risk dimension receives prominent treatment. The firm names concrete technologies, quantifies impacts or investments, identifies specific business lines at risk, and/or describes concrete strategic responses. Discussion goes well beyond boilerplate.
9–10	**Dominant / Comprehensive:** This risk dimension is a central theme in the disclosure, with extensive, specific, and multi-layered discussion. Reserved for cases where the firm treats this particular AI risk dimension as a defining strategic concern.

Key Scoring Rules

- **Absence = 0:** If the disclosure contains no content relevant to a sub-category, that sub-category receives 0. Do not infer risk from silence.
- **Boilerplate detection:** Generic language that could appear in any filing (e.g., "rapid technological change may adversely affect our business") should not score above 2 in any sub-category, regardless of how many times it appears.
- **Quantity vs. quality:** Five generic mentions of AI do not outweigh one highly specific, detailed passage. Weight specificity and substance over frequency.
- **Flexibility:** If a disclosure does not map cleanly to a single score level, use your judgment to assign the score that best reflects the weight and specificity of the content. The scale descriptions are guidelines, not rigid cutoffs.

Overall Score Computation Assign the `overall_exposure` score as the arithmetic mean of the four sub-category scores (`displacement_risk`, `operational_risk`, `regulatory_risk`, `cybersecurity_risk`).

Step 5: Output

Return **only** the following structured output. Do not repeat the analysis from Steps 1–3.

```
company_name:        [from input]
ticker:              [from input]
cik:                 [from input]
displacement_risk:   [0-10, integer]
operational_risk:    [0-10, integer]
regulatory_risk:     [0-10, integer]
cybersecurity_risk:  [0-10, integer]
overall_exposure:    [average of the four sub-category scores]
justification:       [2-3 sentences. State which sub-categories drove
                      the score, cite the most important passage(s) by
                      brief reference, and note whether the language was
                      specific or boilerplate. If the score is 0, state
                      "No AI-related content identified."]
```

B. Rank Correlations

Table 12: Year-by-Year Spearman Rank Correlations between Model AI Risk Scores

Fiscal Year	*N*	Claude–ChatGPT	Claude–Gemini	ChatGPT–Gemini
2020	501	0.644	0.695	0.725
2021	502	0.808	0.705	0.797
2022	501	0.788	0.834	0.743
2023	497	0.782	0.928	0.808
2024	497	0.805	0.859	0.856
2025	496	0.571	0.651	0.728
Average		0.733	0.779	0.776
Pooled	2,994	0.845	0.906	0.844

Spearman rank correlations between firm-level AI risk scores from the three LLMs, computed within each fiscal year. The pooled row reproduces the correlations reported in Table 4; the average row is the unweighted mean across the six annual correlations. Each annual correlation strips out the common upward trend in scores over time, isolating cross-sectional agreement on relative firm rankings within a single year. The lower 2025 correlations reflect the wider cross-sectional dispersion as more firms move away from zero scores.

C. Portfolio Characteristics by Quartile

Table 13: Portfolio Characteristics by AI Risk Quartile, Subsample

Quartile	*N* Firms	AI Risk	ln(Mktcap)	B/M	ROA	Leverage	R&D Int.
Claude							
Q1 (Low)	246	0.08	10.33	0.379	0.075	0.360	0.021
Q2	245	0.65	10.43	0.352	0.079	0.372	0.028
Q3	244	1.94	10.53	0.392	0.062	0.343	0.046
Q4 (High)	246	4.12	10.81	0.323	0.077	0.300	0.069
ChatGPT							
Q1 (Low)	246	0.11	10.40	0.387	0.073	0.370	0.017
Q2	246	0.83	10.31	0.379	0.071	0.347	0.041
Q3	245	1.84	10.53	0.364	0.073	0.353	0.050
Q4 (High)	246	3.53	10.87	0.316	0.076	0.304	0.055
Gemini							
Q1 (Low)	246	0.07	10.34	0.387	0.077	0.356	0.023
Q2	246	1.09	10.39	0.374	0.073	0.374	0.027
Q3	245	3.69	10.59	0.354	0.070	0.334	0.049
Q4 (High)	246	6.77	10.78	0.332	0.073	0.311	0.065
Simple Avg							
Q1 (Low)	246	0.11	10.34	0.385	0.077	0.366	0.017
Q2	245	0.99	10.36	0.356	0.076	0.373	0.032
Q3	244	2.54	10.58	0.377	0.069	0.331	0.048
Q4 (High)	246	4.61	10.82	0.329	0.071	0.304	0.066
PC1							
Q1 (Low)	246	−1.19	10.34	0.379	0.077	0.363	0.017
Q2	245	−0.47	10.38	0.357	0.077	0.380	0.033
Q3	244	0.77	10.54	0.387	0.066	0.325	0.045
Q4 (High)	246	2.46	10.84	0.323	0.072	0.306	0.068

Time-series averages of cross-sectional means within each quartile. *N* Firms is the total unique firms assigned to each quartile across both formation years.

D. Daily-Frequency Robustness

The following tables re-estimate the main portfolio-sort tests using daily return data. All other aspects of the methodology—quartile assignment, portfolio formation timing, and factor models—are identical to the monthly analysis. The daily series contains 1,173 trading days in the full sample and 419 in the subsample. Newey-West standard errors with six lags are used, well within the range supported by $T^{1/3} \approx 7.5$ for the subsample. The key advantage of the daily specification is that the HAC t-statistics are virtually invariant to the lag choice (Table 15), which confirms that the monthly results are not an artefact of a particular lag selection.

Table 14: Quartile Portfolio Sorts on AI Risk, Daily Data

		Full Sample (1,173 days)				Subsample (419 days)			
		Excess Ret	CAPM α	FFC α	FF5 α	Excess Ret	CAPM α	FFC α	FF5 α
Claude	Q4−Q1	0.013	0.006	0.008	0.009	0.023	0.012	0.008	0.003
		[1.09]	[0.59]	[0.87]	[0.93]	[1.05]	[0.58]	[0.43]	[0.15]
ChatGPT	Q4−Q1	0.014	0.006	0.010	0.010	0.032	0.023	0.018	0.014
		[1.09]	[0.57]	[0.98]	[1.04]	[1.64]	[1.24]	[1.11]	[0.89]
Gemini	Q4−Q1	0.021	0.014	0.017*	0.018*	0.043*	0.030	0.025	0.020
		[1.63]	[1.20]	[1.79]	[1.91]	[1.90]	[1.45]	[1.49]	[1.21]
Simple Avg	Q4−Q1	0.017	0.009	0.012	0.012	0.036	0.023	0.018	0.013
		[1.26]	[0.75]	[1.13]	[1.23]	[1.55]	[1.07]	[0.99]	[0.72]
PC1	Q4−Q1	0.018	0.009	0.012	0.013	0.038*	0.025	0.020	0.015
		[1.30]	[0.80]	[1.17]	[1.26]	[1.65]	[1.19]	[1.11]	[0.83]

EW long-short (Q4−Q1) daily returns (%). Returns are daily percentages; multiply by 252 for approximate annualisation. Newey-West t-statistics (6 lags) in brackets.

Table 15: NW Lag Sensitivity of Daily FF5 Alpha, Subsample

	NW(0)	NW(3)	NW(6)	NW(10)
Claude	0.003 [0.16]	0.003 [0.15]	0.003 [0.15]	0.003 [0.15]
ChatGPT	0.014 [0.95]	0.014 [0.90]	0.014 [0.89]	0.014 [0.89]
Gemini	0.020 [1.25]	0.020 [1.21]	0.020 [1.21]	0.020 [1.21]
Simple Avg	0.013 [0.75]	0.013 [0.72]	0.013 [0.72]	0.013 [0.71]
PC1	0.015 [0.86]	0.015 [0.84]	0.015 [0.83]	0.015 [0.83]

Daily FF5 alpha (%) and NW t-statistics under alternative lag specifications. The near-identical t-statistics across lag choices indicate that inference is robust to the HAC bandwidth parameter when the time series is sufficiently long.

Table 16: Between-Sector vs. Within-Sector Decomposition, Daily Data, Subsample

	Total		Between		Within	
	Ret	Share	Ret	Share	Ret	Share
Claude	0.023 [1.05]	100%	0.023* [1.65]	102%	−0.001 [−0.04]	−2%
ChatGPT	0.032 [1.64]	100%	0.017 [1.49]	51%	0.016 [1.21]	49%
Gemini	0.043* [1.90]	100%	0.023* [1.73]	53%	0.021 [1.46]	47%
Simple Avg	0.036 [1.55]	100%	0.024* [1.70]	67%	0.012 [0.88]	33%
PC1	0.038* [1.65]	100%	0.025* [1.73]	66%	0.013 [0.96]	34%

EW daily returns (%). The between/within decomposition follows the same methodology as the monthly analysis. Newey-West t-statistics (6 lags) in brackets.

E. AI Use Disclosure

Large Language Models have been used for assistance in the research process. LLM assistance has been used in data exploration, coding, as well as in editing and formatting of the thesis text.