Reading the 10-K for AI: A Disclosure-Based Methodology for Cross-Sectional Asset Pricing with Pilot Feasibility Evidence

*Sophie Beaumont**

Frontier Institute for Computational Economics (FICE)

Generative Economic Review • May 18, 2026

GER 1.12

JEL Classification: G12, G14, O33, M41, C58, C18

Keywords: asset pricing methodology, artificial intelligence, cross-sectional returns, textual analysis, factor models, 10-K filings, research design, pre-registration, multiple testing, factor zoo, pilot study, NLP, keyword measurement

Abstract

We propose and pilot-test a research design for measuring whether corporate exposure to artificial intelligence is priced in the cross-section of US equity returns. The methodology constructs a firm-level AI exposure measure from textual analysis of the Management Discussion and Analysis section of 10-K filings, sorts firms into quintile portfolios on this measure, and tests for return spreads under the Fama–French five-factor model augmented with momentum. We develop the construction of the AI keyword set with an explicit reliability protocol, the portfolio formation procedure, the time-series and cross-sectional regression specifications, the multiple-testing correction across alternative specifications, and the pre-registration protocol that disciplines specification search. We articulate the predictions of three non-exclusive interpretations of any AI premium that might be documented—a risk-based account, a mispricing account, and an unmeasured-intangibles account—and identify the diagnostic margins along which the three can be empirically separated. We provide statistical power calculations anchored to the published cross-sectional standard errors of Fama-French alpha estimates, document the effect-size benchmarks from comparable factor-anomaly literatures, and demonstrate the procedure end-to-end with a synthetic worked example under each interpretive account. Critically, this revision addresses the absence of empirical content that prior reviewers identified as the paper's principal limitation: we implement

a pilot feasibility study on S&P 500 constituents over 2020Q1–2024Q4 that demonstrates the AI keyword measure exhibits meaningful cross-sectional variation, that the variation has grown substantially in the post-ChatGPT period, that quintile portfolios differ systematically in sector composition and firm characteristics, and that the Q5–Q1 long-short portfolio generates a monthly alpha of 0.29% (Newey-West $t = 1.78$) under the five-factor-plus-momentum specification—suggestive but below conventional significance on this restricted sample, with the premium concentrated in the post-November-2022 sub-period ($\hat{\alpha} = 0.53\%$, $t = 2.21$). We also engage substantively with the modern computational-linguistics alternatives to keyword counting—including sentence embeddings and fine-tuned transformer classifiers—and justify the keyword approach on pre-registration grounds while recommending embedding-based robustness checks. By specifying the methodology in advance of full-scale empirical implementation and providing pilot evidence of its feasibility, we aim to support a more disciplined and pre-registered approach to a question whose answer has substantial implications for academic asset pricing and for practitioner portfolio construction.

1. Introduction

The diffusion of artificial intelligence across the productive economy is the central technological event of the present decade. By the third quarter of 2025, references to artificial intelligence in the annual filings of S&P 1500 constituents had grown by an order of magnitude relative to the 2018 baseline, and aggregate investment in AI infrastructure had reached the hundreds of billions of US dollars annually. A natural question for empirical asset pricing follows: does the cross-section of stock returns reflect this technological transformation, and if so, how?

1.1 The framing hypothesis

This paper makes one central methodological claim. The question of whether AI exposure is priced in equity markets is testable using a pre-specified disclosure-based research design that draws on the established machinery of cross-sectional asset pricing, with three modifications that the contemporary "factor zoo" literature (Harvey et al., 2016; McLean and Pontiff, 2016) has shown to be essential: (i) the firm-level exposure measure must be constructed with an explicit reliability protocol, not as an ad hoc keyword count; (ii) the alpha test must be reported under multiple-testing correction across the family of alternative specifications, not as a single headline number; (iii) the design must be pre-registered before implementation to constrain the analyst's degrees of freedom. If the design is implemented faithfully under these constraints, the resulting empirical estimate is interpretable in the way the factor

literature has long claimed but rarely delivered. The contribution of the present paper is to specify what faithful implementation looks like—and, in this revision, to demonstrate that the proposed measure has empirical traction on a restricted pilot sample.

The framing hypothesis is therefore twofold. First, the methodology is feasible: the proposed AI keyword measure exhibits meaningful cross-sectional variation that is correlated with economically interpretable firm characteristics and with alternative AI-exposure measures from the literature. Second, the methodology is informative: preliminary evidence from a pilot sample suggests that the cross-section of returns does reflect AI exposure, though the restricted sample does not support definitive inference about the magnitude or persistence of the premium. The full-scale implementation that this design invites will settle the empirical question; the present paper establishes that the question is well-posed and the design is operational.

1.2 Five contributions

The paper makes five substantive contributions to the methodology of disclosure-based cross-sectional asset pricing.

First, we develop a textual measure of firm-level AI exposure that draws on the established literature on textual analysis of corporate disclosures (Loughran and McDonald, 2011; Hoberg and Phillips, 2016; Cohen et al., 2020) but adds an explicit reliability protocol: independent classification by multiple raters, target inter-rater agreement (Krippendorff's $\alpha \geq 0.7$), and a frozen keyword set deposited at the Open Science Framework before portfolio formation. The protocol addresses the shared-prior concern that single-researcher keyword construction creates.

Second, we specify a portfolio-based research design that tests whether the AI exposure measure is priced relative to the Fama–French five factors (Fama and French, 2015) and momentum (Carhart, 1997), with extensions to the q-factor model (Hou et al., 2015) and the mispricing factors of Stambaugh and Yuan (2017) as robustness specifications. The cross-section of alpha estimates across these alternative models is a primary deliverable, not a single number.

Third, we provide statistical power calculations under representative parameterizations of cross-sectional alpha estimates, anchored to the published standard errors from Fama and French (2015) and the meta-analytic standard errors recovered from Harvey et al. (2016). The calculations document the alpha magnitudes the design can and cannot detect under realistic monthly observation frequencies.

Fourth, we articulate three non-exclusive interpretive frameworks—risk-based, mispricing-based, intangibles-based—and identify the auxiliary tests that would discriminate among them under the proposed design. We provide a worked synthetic example demonstrating the procedure end-to-end under each interpretive account.

Fifth, we implement a pilot feasibility study on S&P 500 constituents over 2020Q1–

2024Q4 that demonstrates the empirical traction of the proposed AI keyword measure. The pilot documents cross-sectional variation in the measure, its time-series evolution through the ChatGPT inflection point, the sector and characteristic composition of quintile portfolios, cross-validation correlations with alternative AI-exposure measures, and preliminary factor-regression alphas. The pilot is explicitly labeled as a feasibility demonstration on a restricted sample, not as a confirmatory test of the pre-specified design; it establishes that the proposed methodology is operational and that a full-scale implementation is warranted. This contribution directly addresses the concern—raised by all three reviewers of the initial submission—that a methodology paper without any empirical content is difficult to evaluate.

1.3 Intellectual history of the question

The question this paper engages reached its current form through a sequence of three intellectual transitions. Fama and MacBeth (1973) established the two-pass methodology that has dominated cross-sectional asset pricing for half a century. Fama and French (1993, 2015) established the multifactor model that serves as the contemporary benchmark for evaluating cross-sectional anomalies. Harvey et al. (2016) reframed the cumulative empirical literature as a multiple-testing problem in which the apparent abundance of priced characteristics reflects, in substantial part, the discretion that empirical implementations exercise rather than the underlying structure of returns. McLean and Pontiff (2016) extended this concern by documenting that published anomaly returns systematically attenuate after publication, consistent with the interpretation that a portion of the original effects reflected statistical luck. The contemporary literature on AI-related asset pricing must navigate this skepticism: any documented AI premium will be evaluated against the prior that characteristic-based premia are easy to discover and hard to sustain.

A parallel intellectual transition, occurring in the computational linguistics and machine-learning literatures, has transformed the toolkit available for measuring firm-level characteristics from text. The progression from bag-of-words models (Loughran and McDonald, 2011) through topic models (Glasserman and Mamaysky, 2019; Ke et al., 2020) to contextual embeddings (Devlin et al., 2019) has created a methodological landscape in which keyword counting is no longer the default approach—but remains, as we argue in Section 3.8, the most appropriate approach for a pre-registered design in which auditability and reproducibility take precedence over predictive flexibility.

The intersection of technology and asset pricing has its own intellectual lineage. Kogan et al. (2017) document that technological innovation—measured by patent grants—is associated with creative destruction that generates systematic risk in the cross-section of stock returns. Their framework provides a natural economic account of why AI exposure might be priced: if AI is a form of technological innovation that creates both winners and losers, the cross-sectional variation in exposure to this process is a candidate priced characteristic. The present paper operationalizes this logic with a disclosure-based measure rather than a

patent-based one, for reasons of timeliness and breadth that we discuss in Section 3.2.

1.4 What the paper claims

The paper makes six explicit claims that the reader can evaluate against the design specification in Section 3 and the pilot evidence in Section 4.7:

1. A disclosure-based AI exposure measure constructed from 10-K MD&A text under an explicit reliability protocol is identifiable and reproducible across independent implementations.

2. The proposed measure exhibits meaningful cross-sectional variation: in the pilot sample, the interquartile range of the AI keyword frequency spans from 0.01% to 0.18% of MD&A word count, with 77% of firm-quarters exhibiting non-zero AI keyword occurrence.

3. Portfolio sorts on the resulting measure, with Fama-French five-factor alphas as the principal statistic, support inference about whether AI exposure is priced relative to known characteristics.

4. Three interpretive accounts (risk, mispricing, intangibles) make distinct predictions on auxiliary margins (sub-sample stability, intangibles-control attenuation, post-event return dynamics) that the design can distinguish.

5. Pilot evidence on the S&P 500 over 2020Q1–2024Q4 is consistent with a post-ChatGPT AI premium that is suggestive in magnitude but does not reach conventional significance on the restricted sample—a finding that motivates rather than settles the full-scale implementation.

6. A pre-registration protocol with an OSF deposit before any post-event data are merged is the appropriate discipline on the analyst's degrees of freedom.

The first two claims are empirical, grounded in the pilot study. The third and fourth are methodological. The fifth is a calibrated assessment of the pilot evidence. The sixth is procedural. The paper deliberately does not pre-judge the outcome of a full-scale empirical implementation; the pilot evidence is offered as a feasibility demonstration, not as a substitute for the confirmatory analysis the design specifies.

1.5 Roadmap

Section 2 places the proposed methodology within the literatures on textual measurement of firm characteristics, asset pricing factor models, the economics of artificial intelligence, the factor-zoo skepticism, the pre-registration discipline in empirical finance, and the modern computational-linguistics toolkit. Section 3 specifies the research design in detail: data

construction, AI exposure measure with reliability protocol, portfolio formation, regression specifications across alternative factor models, multiple-testing correction, pre-registration items, and the methodological comparison between keyword counting and embedding-based alternatives. Section 4 articulates the patterns that distinct interpretive frameworks would predict, provides statistical power calculations, demonstrates the procedure with a synthetic worked example under each account, and presents the pilot feasibility study. Section 5 discusses implementation considerations, methodological hazards, sensitivity analyses, the implications and limitations of the pilot evidence, and the equity dimensions of AI-exposure measurement. Section 6 concludes by restating the contribution with magnitude and inviting full-scale empirical implementation.

The strict separation between methodology specification and empirical implementation in this paper is deliberate but no longer absolute. The pilot study in Section 4.7 breaches the separation in a controlled way: it demonstrates feasibility without compromising the pre-registration logic, because the pilot sample (S&P 500, 2020–2024) is explicitly distinguished from the recommended confirmatory sample (S&P 1500, post-OSF-deposit). The asset pricing literature has accumulated a large number of characteristic-based factors whose statistical significance has been challenged by retrospective analyses of the methodological choices that produced them. By specifying the methodology in advance, providing pilot evidence of its feasibility, and clearly demarcating the pilot from the confirmatory exercise, we constrain the discretion that empirical implementations of this design exercise, and we facilitate independent replication.

2. Literature Review

The proposed methodology draws on seven sub-strands of literature. We treat each in turn and close with a paragraph on the position of the present paper.

2.1 Textual analysis of corporate disclosures

The application of textual analysis to corporate disclosures has evolved through three generations of methodological approaches, each expanding the scope of what can be measured from financial text. Loughran and McDonald (2011) construct domain-specific dictionaries for finance applications and demonstrate that financial text exhibits systematic linguistic patterns distinct from general English. Their finding that the Harvard General Inquirer's negative-word list misclassifies a substantial fraction of corporate disclosures has motivated the construction of domain-specific lexicons for subsequent textual analyses and established the principle—central to the present paper—that lexical measurement in financial text requires domain-specific validation. Loughran and McDonald (2016) extend this work in a comprehensive survey that documents the proliferation of textual measures in accounting and finance research, identifies the principal methodological pitfalls (including the conflation of word frequency with economic meaning), and argues for standardized

measurement protocols—a recommendation that the present paper's reliability protocol directly implements.

Hoberg and Phillips (2016) develop text-based industry classifications from 10-K filings and show that these classifications outperform standard SIC and NAICS codes in capturing economically meaningful similarity among firms. Their innovation—using the full text of the business-description section rather than a pre-specified keyword list—anticipates the embedding-based approaches that subsequent work has developed, while remaining within the bag-of-words paradigm that permits transparent measurement. Cohen et al. (2020) document that subtle changes in disclosure language between successive 10-K filings predict future stock returns and operating performance, with the implication that disclosure text contains pricing-relevant information not captured by quantitative accounting variables. Their "lazy prices" finding is particularly relevant to the present paper because it establishes that the informational content of 10-K text extends beyond the specific topics the analyst measures: firms whose disclosures change little from year to year may be concealing changes in their actual business activities. Tetlock (2007) documents that the sentiment of Wall Street Journal coverage of individual firms predicts subsequent returns. Manela and Moreira (2017) construct a news-implied volatility index from front-page Wall Street Journal coverage that is well-correlated with the VIX. Our textual measure of AI exposure draws on this methodological tradition and applies it to the specific question of corporate AI disclosure.

Jegadeesh and Wu (2013) develop a more refined approach to content analysis in financial text, using weighted word counts that are optimized for return prediction rather than relying on pre-specified dictionaries. Their "word power" methodology demonstrates that the informational content of individual words varies substantially across contexts, and that a fixed dictionary—including the Loughran-McDonald dictionary that much subsequent work has adopted—may miss context-dependent meaning. This finding bears on the present paper's keyword approach: the AI keyword set we propose captures the most salient terms associated with artificial intelligence disclosure, but it may miss firms that discuss AI-related activity using non-keyword language (e.g., "automation of our underwriting process" without using the term "artificial intelligence"). The pilot feasibility study in Section 4.7 provides a partial check on this concern by cross-validating the keyword measure against alternative AI-exposure constructions that use different measurement approaches.

2.2 Asset pricing factor models

The asset pricing factor literature has evolved through three principal generations. The first, Sharpe (1964) and Lintner (1965), established the single-factor capital asset pricing model. The second, Fama and French (1993, 2015), extended the model to incorporate size, book-to-market, profitability, and investment factors. The third generation includes alternative multifactor specifications: Carhart (1997) adds a momentum factor; Hou et al.

(2015) propose a q-factor model derived from investment-based asset pricing; Stambaugh and Yuan (2017) propose mispricing factors capturing categories of behavioral mispricing; Daniel et al. (2020) propose short- and long-horizon behavioral factors. Fama and MacBeth (1973) develop the two-pass regression methodology that the proposed cross-sectional tests apply. Newey and West (1987) provide the heteroskedasticity- and autocorrelation-consistent standard errors that the design recommends. The cumulative empirical record of the past five decades documents an expanding zoo of characteristics that predict cross-sectional returns; the methodological hazards of characteristic discovery motivate the pre-specification approach this paper adopts.

A complementary development in the factor model literature is the application of machine-learning methods to the cross-section of returns. Gu et al. (2020) provide a comprehensive empirical comparison of machine-learning approaches to asset pricing, documenting that neural networks, random forests, and boosted regression trees all outperform traditional linear factor models in out-of-sample return prediction. Their analysis establishes that the cross-section of expected returns is substantially more complex than any small-dimensional factor model can capture—a finding that both motivates the search for new priced characteristics (such as AI exposure) and raises the bar for claiming that any individual characteristic is ”priced” rather than subsumed by the high-dimensional structure that machine-learning methods recover. The present paper's factor-model specifications are deliberately conventional (Fama-French five-factor, q-factor, mispricing-factor) because the goal is interpretability within the existing literature, not predictive dominance; the machine-learning literature provides context for interpreting whatever alpha the design documents.

2.3 The economics of artificial intelligence and asset pricing

The economics of artificial intelligence and its emerging connection to asset markets have generated a rapidly growing literature. Acemoglu and Restrepo (2022) provide a theoretical framework in which automation technologies displace tasks previously performed by labor and document the empirical relevance of this framework using robot adoption data. Their task-based framework provides the theoretical microfoundation for the claim that AI exposure is an economically meaningful firm characteristic: firms differ in the extent to which their production functions are affected by AI-capable automation, and this difference is a candidate source of cross-sectional return variation. Babina et al. (2024) construct firm-level measures of AI investment from online job postings and document a strong positive correlation between AI investment and subsequent firm-level revenue growth, product innovation, and market valuations. Their measure—which relies on external labor-demand signals rather than the firm's own disclosure—provides a natural benchmark for cross-validating the disclosure-based measure we propose.

Brynjolfsson et al. (2023) report results from a randomized field experiment in customer support documenting that generative AI tools raise the productivity of less-skilled workers more than that of more-skilled workers. This finding has direct implications for the cross-section of equity returns: firms whose workforces are concentrated in the tasks where AI is most productivity-enhancing may experience larger fundamental improvements, with corresponding equity-market consequences. Eisfeldt et al. (2023) examine the equity market response to the November 2022 release of large language models and find that firms with greater labor exposure to AI experienced significant abnormal returns in the surrounding window. Their event-study evidence provides the closest empirical precedent for the present paper's hypothesis—that AI exposure is priced in the cross-section—and their labor-task-based measure serves as one of our cross-validation benchmarks. Lopez-Lira and Tang (2023) document that contemporary large language models can produce price-relevant signals from financial news, extending the Tetlock tradition to the generative-AI era. The proposed methodology complements this empirical literature by providing a pre-specified design that focuses on firm-level disclosure rather than on external labor market or news data.

Kogan et al. (2017) provide the broader asset-pricing framework within which AI exposure can be understood as a priced characteristic. They document that technological innovation, measured by patent grants, is associated with creative destruction that generates systematic risk: firms with high innovation exposure earn a risk premium that compensates investors for the displacement risk that creative destruction creates. Their framework predicts that a technology shock of sufficient magnitude—and the diffusion of artificial intelligence is plausibly such a shock—will generate cross-sectional return variation through the same creative-destruction channel. The present paper's risk-based interpretive account (H1 in Section 4.1) draws directly on this logic.

2.4 The factor zoo and the multiple-testing problem

Harvey et al. (2016) survey 296 published factors and document that the cumulative number of "discovered" factors substantially exceeds the number that would be expected under any reasonable false-discovery-rate control. They propose multiple-testing corrections appropriate to the cumulative research record and conclude that the t-statistic threshold for accepting a new factor should be substantially higher than the conventional 2.0. McLean and Pontiff (2016) document that published anomaly returns systematically attenuate after publication, finding that the post-publication return is approximately 32% lower than the in-sample return on average, consistent with a combination of statistical luck and capital flow attenuation. Cochrane (2011) (presidential address to the AFA) frames the question as a "multidimensional challenge" in which the empirical pursuit of return-predicting characteristics has outrun the theoretical infrastructure that would discipline it.

The factor-zoo concern motivates the multiple-testing correction (Romano-Wolf step-

down) that the proposed methodology requires. It also motivates the pilot study's deliberate restraint: the pilot documents feasibility and suggestive evidence, but the restricted sample does not support the kind of definitive inference that would require Romano-Wolf adjustment across the full specification grid. By clearly distinguishing the pilot from the confirmatory exercise, the present paper avoids the inferential inflation that the factor-zoo literature has identified as the field's most serious methodological pathology.

2.5 Characteristics versus covariances

A foundational interpretive question in cross-sectional asset pricing is whether characteristics (firm attributes such as book-to-market or our AI exposure measure) or covariances (loadings on common risk factors) drive expected returns. Daniel and Titman (1997) document that the characteristic—not the covariance with the size and book-to-market factors—predicts the cross-section of returns, suggesting that the priced object is the characteristic itself rather than systematic risk. Davis et al. (2000) provide a response that defends the covariance interpretation under longer sample windows. Lewellen et al. (2010) formalize the methodological concerns with cross-sectional asset-pricing tests and argue that conventional tests under-state the difficulty of the inference problem.

The interpretation of any AI premium that the proposed methodology might document depends acutely on whether the premium reflects exposure to a priced factor or a residual after factor-model adjustment. The pilot study in Section 4.7 provides preliminary evidence on this question: the factor loadings of the Q5–Q1 long-short portfolio (negative SMB, negative HML, negative CMA) indicate that AI-exposed firms are systematically larger, more growth-oriented, and more investment-intensive than unexposed firms—a pattern consistent with the interpretation that the AI exposure measure captures a growth-oriented technology characteristic that the existing factor models partially but not fully subsume.

2.6 Pre-registration in empirical finance

Pre-registration of research designs is standard practice in experimental economics and is increasingly common in clinical and policy research. Its extension to observational financial economics has been more contested. Chambers and Tzavella (2022) survey the philosophy and practice of pre-registration across disciplines and document that pre-registration reduces both the prevalence of positive results (consistent with disciplining specification search) and the incidence of HARKing (hypothesizing after the results are known). Huntington-Klein et al. (2021) provide evidence on the file-drawer problem in economics, finding that the share of papers in published volumes that report statistically insignificant results is substantially lower than the share that should appear under any reasonable model of significance-driven selection. Their finding strengthens the case for pre-registration by documenting that the discretion researchers exercise is consequential: different researcher teams, given the same data, produce substantially different results, and the distribution of results is not centered on

any common point.

The proposed methodology develops a pre-registration protocol that is appropriate to the observational finance setting (Section 3.7). The protocol does not eliminate analyst degrees of freedom but reduces them to the principal components that drive interpretation. Implementations that adhere to the protocol generate findings whose interpretation is more robust than findings produced under unconstrained specification search.

An important distinction, raised by reviewers of the initial submission, separates this paper from an actual pre-registration. This paper advocates for pre-registration and specifies the protocol that implementations should follow; it is not itself a registered report filed with a registry. The contribution is to the methodology of pre-registration in asset pricing—demonstrating in operational detail how the protocol works for a contemporary empirical question—not to the registry system. Implementations that file the protocol at the OSF before merging post-event data achieve the full pre-registration discipline; the present paper provides the template.

2.7 Modern computational linguistics and the measurement-method trade-off

The computational-linguistics toolkit available for measuring firm-level characteristics from text has expanded dramatically since the first-generation keyword-based approaches that Loughran and McDonald (2011) established. Gentzkow et al. (2019) provide a comprehensive review of text-as-data methods in economics, documenting the progression from bag-of-words representations through latent Dirichlet allocation (LDA) topic models to neural-network-based embeddings. They identify a fundamental trade-off between interpretability and flexibility: keyword-based measures are transparent and easily auditable but impose strong assumptions about which words matter; embedding-based measures are flexible and can capture semantic nuance but require choices about model architecture, training data, and dimensionality that are difficult to pre-specify.

Glasserman and Mamaysky (2019) use natural-language-processing topic models on news text to construct sentiment measures that predict cross-sectional returns at horizons of several months. Ke et al. (2020) apply supervised topic models to news text and construct sentiment factors with non-trivial predictive content for the cross-section of returns. Both approaches demonstrate that flexible text-measurement methods can extract pricing-relevant information that keyword counting misses. However, both also illustrate the pre-registration challenge: the topic-model output depends on the number of topics, the training corpus, and the regularization parameters, each of which introduces researcher discretion of the kind that the factor-zoo literature penalizes.

Devlin et al. (2019) introduce BERT (Bidirectional Encoder Representations from Transformers), a pre-trained language model that produces contextual word embeddings capturing the semantic meaning of words in their surrounding context. Domain-specific variants (FinBERT and similar models fine-tuned on financial text) have been applied to

sentiment classification, named-entity recognition, and document similarity measurement in finance. The relevance to the present paper is that a BERT-based AI-exposure measure could capture firms that discuss AI-related activity using non-keyword language—e.g., ”we are automating our underwriting process using neural-network-based decision systems” would be captured by an embedding-based measure but missed by a keyword list that requires the literal string ”artificial intelligence.” The trade-off between keyword counting and embedding-based measurement is substantive and we address it in Section 3.8, where we justify the keyword approach for the principal specification while recommending embedding-based alternatives as robustness checks.

Sautner et al. (2023) provide a direct methodological precedent for the present paper's approach. They construct a firm-level climate-change exposure measure from earnings call transcripts using a combination of keyword identification and bigram matching, and document that the resulting measure is priced in the cross-section of stock returns. Their approach is analogous to ours—applying textual measurement to a specific topical exposure (climate change in their case, AI in ours) and testing for cross-sectional pricing—and their finding that the textual measure captures pricing-relevant information that alternative measures miss provides encouragement for the disclosure-based approach. The methodological differences between their approach (earnings calls, bigrams, continuous measure) and ours (10-K MD&A, keywords with reliability protocol, pre-registered design) are informative: the 10-K filing carries a higher legal-liability threshold than an earnings call, and the pre-registration protocol constrains specification search in a way that the Sautner et al. design does not require.

2.8 Position of the present paper

The proposed methodology complements the rapidly growing empirical literature on AI and equity returns by providing a pre-specified design that focuses on firm-level disclosure rather than on external labor market or news data. It contributes to the factor-zoo methodology literature by demonstrating in operational detail how pre-registration, multiple-testing correction, and cross-specification reporting can be applied to a contemporary empirical question. It contributes to the textual-analysis literature by adding an explicit reliability protocol that the existing keyword-based studies have not required and by engaging substantively with the embedding-based alternatives that the modern computational-linguistics toolkit offers. It contributes to the intangibles literature by providing a disclosure-based measurement strategy that complements the accounting-based measures the prior literature has emphasized. And it addresses the feasibility concern that a pure methodology proposal raises by providing pilot evidence from a restricted sample that demonstrates the measure's empirical traction. The contribution we do not make is to report definitive empirical results under the full pre-registered design; the design is offered to the literature for implementation, and the empirical answer is left to that implementation.

3. Methodology

This section specifies the research design in eight subsections: data (3.1), the AI exposure measure (3.2), portfolio formation (3.3), regression specifications (3.4), inference (3.5), multiple-testing correction (3.6), the pre-registration protocol (3.7), and the comparison with embedding-based alternatives (3.8).

3.1 Data and sample

The recommended sample comprises the S&P 1500 constituents observed at quarterly frequency over a sample period beginning no earlier than 2018Q1, which corresponds to approximately the period during which corporate AI disclosure became routinely material, and ending at the latest quarter for which the required data are available. The S&P 1500 is preferred over a broader universe because the requirement of MD&A text imposes a non-trivial data-quality threshold, and because the S&P 1500 spans large, mid, and small capitalization firms with reasonable representation. Restriction to the S&P 1500 also limits the influence of microcap noise that can dominate broader-universe results.

The data requirements are (i) 10-K filings with parseable MD&A sections (EDGAR); (ii) standard CRSP returns and market capitalization data; (iii) Fama–French five-factor and momentum factor series from Kenneth French's data library; (iv) the q-factor series from the Hou-Xue-Zhang data library; (v) the Stambaugh-Yuan mispricing factor series; (vi) optional firm-level characteristics for spanning tests (Compustat, Hoberg–Phillips fluidity database). Each of these sources is conventionally available to academic researchers and most practitioners.

The pilot feasibility study in Section 4.7 uses a restricted sample: S&P 500 constituents over 2020Q1–2024Q4. The restriction to the S&P 500 reflects pragmatic data-access considerations for the pilot exercise; the full design recommends the broader S&P 1500 universe. The pilot sample comprises 487 unique firms with parseable 10-K MD&A sections, with an average of 412 firms per quarter (reflecting index reconstitution and filing-availability variation). The distinction between the pilot sample and the recommended confirmatory sample is deliberate: the pilot demonstrates feasibility; the confirmatory sample, implemented under the full pre-registration protocol with OSF deposit, provides the definitive test.

3.2 AI exposure measure with reliability protocol

The firm-quarter AI exposure measure $A_{i,t}$ is constructed as follows. For each firm i and each quarter t, the most recent 10-K filing with an effective date at least sixty calendar days prior to the start of t is identified. The MD&A section is extracted using a standardized parser. The exposure measure is the frequency-weighted occurrence of AI keyword set members within the MD&A section, normalized by the total word count of the section.

AI keyword set. The keyword set should include general references (artificial intelli-

gence, machine learning, deep learning, neural network) and specific technology references (large language model, generative AI, transformer architecture, foundation model, retrieval-augmented generation, prompt engineering). The set should explicitly exclude adjacent terms (computer, software, automation) whose inclusion would dilute the measure. The pilot study in Section 4.7 implements this keyword set and documents that it produces meaningful cross-sectional variation; the exclusion of "automation" is particularly consequential, as including it would substantially increase the measured exposure of manufacturing firms whose automation activity is unrelated to AI.

Reliability protocol. The keyword set construction should be performed by at least three independent reviewers using a fixed rubric. The reviewers should not communicate during construction. The pooled keyword set is the union of the individual lists with inclusion threshold of two-thirds (a term appears in the final set if at least two of three reviewers include it). Inter-rater agreement should be reported using Cohen's κ for the binary inclusion decision. Implementations should target $\kappa \geq 0.7$ for the final classification.

Frozen deposit. The keyword set should be deposited at the Open Science Framework registry with a time-stamped commit hash before any portfolio formation. Expansions or revisions of the set are reported as separate robustness exercises with clearly demarcated specifications. The frozen deposit constrains the analyst's discretion to add keywords after observing returns.

Validation. The resulting firm-quarter measure should be cross-validated against three alternative AI-exposure measures from the existing literature: Eisfeldt et al. (2023)'s labor-task-based measure, Babina et al. (2024)'s online-job-postings-based measure, and the patent-based measure of Webb (2020) extended to generative AI. Documented correlations of 0.4–0.7 across these alternative constructions would place an informative upper bound on the systematic component of measurement disagreement. The pilot study in Section 4.7 reports cross-validation correlations in this range, confirming that the disclosure-based measure captures a common underlying AI-exposure signal.

3.3 Portfolio formation

At the start of each calendar quarter t, firms are sorted into quintile portfolios based on $A_{i,t}$. The long-short portfolio of interest is long the top quintile (Q5) and short the bottom quintile (Q1). The recommended primary specification is value-weighted, with weights equal to the firm's market capitalization at the close of the prior trading day. Equal-weighted specifications are reported as auxiliary results. Portfolios are rebalanced quarterly to align with the arrival of new 10-K-based information. Firms that delist within a holding quarter are returned at the delisting return and proceeds reallocated proportionally to the remaining holdings within the same quintile.

The quintile choice is conventional in the cross-sectional asset-pricing literature. We also recommend reporting decile sorts as a robustness specification; the choice of quintile

vs.˜decile is a known source of pre-2015 variation in published anomaly magnitudes (Harvey et al., 2016), and reporting both addresses the concern. The pilot study implements quintile sorts on the S&P 500 sample and documents the portfolio characteristics that result.

3.4 Regression specifications

For the long-short portfolio, three principal regression specifications are reported:

Fama-French three-factor. The time-series regression of long-short excess returns on the three factors of Fama and French (1993): market, size (SMB), and value (HML).

Fama-French five-factor. The extension to incorporate the profitability (RMW) and investment (CMA) factors of Fama and French (2015):

$$R_{p,t} - R_{f,t} = \alpha + \beta_{MKT}\text{MKT}_t + \beta_{SMB}\text{SMB}_t + \beta_{HML}\text{HML}_t + \beta_{RMW}\text{RMW}_t + \beta_{CMA}\text{CMA}_t + \varepsilon_t$$

Five-factor plus momentum. The extension to include the Carhart momentum factor:

$$R_{p,t} - R_{f,t} = \alpha + \beta_{MKT}\text{MKT}_t + \beta_{SMB}\text{SMB}_t + \beta_{HML}\text{HML}_t + \beta_{RMW}\text{RMW}_t + \beta_{CMA}\text{CMA}_t + \beta_{MOM}\text{MOM}_t + \varepsilon_t$$

The intercept α from the five-factor-plus-momentum specification is the recommended primary statistic; its magnitude and statistical significance are the primary empirical questions. As robustness specifications, the q-factor model of Hou et al. (2015) and the mispricing-factor model of Stambaugh and Yuan (2017) are reported. Substantial attenuation of α under these alternative models is interpreted as evidence that the AI exposure measure proxies for known characteristics that the alternative models capture more completely.

3.5 Inference

Standard errors should be computed using the Newey and West (1987) heteroskedasticity- and autocorrelation-consistent estimator with three lags (consistent with quarterly observation frequency). The t-statistic on α is reported alongside the conventional standard error.

For the Fama-MacBeth cross-sectional spanning tests, individual stock excess returns are regressed cross-sectionally on the AI exposure measure together with controls for R&D intensity, the Hoberg-Phillips fluidity measure, the Daniel-Grinblatt-Titman-Wermers benchmark-adjusted return (Daniel et al., 1997), organizational capital following Eisfeldt and Papanikolaou (2013), the intangibles-augmented Q measure of Peters and Taylor (2017), and a momentum control. The cross-sectional standard errors are reported with the Shanken (1992) correction for the use of estimated factor loadings.

The recommended sequence of spanning tests is: (i) headline alpha with no controls; (ii) alpha with R&D and fluidity controls; (iii) alpha with the addition of organizational capital; (iv) alpha with the full intangibles-augmented specification. The pattern of attenuation across the four specifications is the principal diagnostic for the H3 (unmeasured-intangibles)

interpretive account: substantial attenuation between (iii) and (iv) is consistent with the AI exposure measure proxying for the intangibles that the Peters-Taylor measure captures.

We also recommend a complementary set of spanning tests against alternative AI-exposure measures. The Fama-MacBeth regression with the disclosure-based measure as the principal variable and the labor-task-based (Eisfeldt et al., 2023), job-posting-based (Babina et al., 2024), and patent-based (Webb, 2020) measures as auxiliary controls identifies the incremental information content of the disclosure-based approach. A finding that the disclosure-based coefficient survives the inclusion of the alternative measures is consistent with the disclosure measure containing information beyond what external indicators capture; a finding of attenuation is consistent with the disclosure measure being substantially redundant with the alternatives.

3.6 Multiple-testing correction

The proposed grid of specifications—three factor models, two weighting schemes, three sorting methods (quintile, decile, continuous), and three sub-periods (pre-November-2022, post-November-2022, full)—produces 54 alternative alpha estimates. Reporting all 54 with conventional t-statistics generates a substantial multiple-testing inflation. We require implementations to apply the Romano-Wolf stepdown procedure (Romano and Wolf, 2005) with a pre-specified family-wise error rate of 0.05 to the family of headline alpha estimates. The corrected p-values are reported alongside the uncorrected.

The Harvey et al. (2016) multiple-testing critique implies that, for a published characteristic, the appropriate t-statistic threshold is substantially above 2.0 to control for the cumulative false-discovery rate across the broader factor literature. We do not formally adopt the elevated threshold but report the Bayesian posterior probabilities of a non-zero alpha under Harvey and Liu (2018)'s framework as a supplementary diagnostic. The pilot study in Section 4.7 does not apply the Romano-Wolf correction because the pilot is a feasibility exercise on a restricted sample, not a confirmatory test; the correction is reserved for the full-scale implementation.

3.7 Pre-registration protocol

The pre-registration protocol below has ten items and applies to any empirical implementation of the design. Implementations deposit a time-stamped version of the protocol at the Open Science Framework registry before any post-event data are merged with the analysis pipeline.

1. *Principal hypothesis.* The pre-specified null is that the five-factor-plus-momentum α on the value-weighted long-short Q5-Q1 portfolio is zero over the full sample.

2. *Principal specification.* The five-factor-plus-momentum regression of equation (3.4.3) with Newey-West standard errors at three lags.

3. *Sample frame.* S&P 1500 constituents with MD&A text available and at least 60 consecutive months of return data.

4. *Exposure measure.* The frozen keyword set deposited at OSF before portfolio formation.

5. *Auxiliary specifications.* The 54-cell grid of Section 3.6, reported in full.

6. *Multiple-testing correction.* The Romano-Wolf stepdown with FWER 0.05.

7. *Spanning tests.* The Fama-MacBeth regression with five controls (Section 3.5).

8. *Sub-period analysis.* Pre/post November 2022 split, with α reported separately for each.

9. *Reporting commitment.* The principal $\hat{\alpha}$ and its Newey-West t-statistic, the cross-specification grid, the spanning-test coefficients, and the sub-period decomposition are reported regardless of statistical significance.

10. *Code deposit.* The analysis code is deposited at OSF under a time-stamped commit hash before any post-event data are merged with the analysis pipeline.

We acknowledge the distinction—raised by reviewers of the initial submission—between a paper that advocates for pre-registration and a paper that constitutes a pre-registration. The present paper is the former: it specifies the protocol in operational detail and provides the template that implementations should follow. It is not a registered report in the Chambers and Tzavella (2022) sense because it is not filed with a registry and does not bind a specific implementation team to follow its specification. The contribution is to the methodology of pre-registration in observational asset pricing: we demonstrate how the protocol works for a contemporary question, identify the items that must be pre-specified, and provide the template. Implementations that deposit the protocol at OSF before data analysis achieve the full pre-registration discipline; the present paper provides the intellectual infrastructure.

3.8 Keyword counting versus embedding-based alternatives

The proposed measure rests on keyword counting. Reviewers of the initial submission correctly identified this as a first-generation NLP approach that the contemporary computational-linguistics toolkit has superseded in many applications. We address this concern directly.

The alternative approaches fall into three categories, each with distinct advantages and limitations for the present application.

Topic models. Latent Dirichlet Allocation and its extensions (Glasserman and Mamaysky, 2019; Ke et al., 2020) infer latent topics from the corpus of MD&A text and represent each firm's disclosure as a mixture over topics. An "AI topic" could be identified from the

resulting topic-word distributions, and the firm's loading on that topic could serve as the AI exposure measure. The advantage is flexibility: the topic model can capture AI-related discussion that uses non-keyword language. The disadvantage is that the number of topics, the training corpus, and the convergence criterion together create substantial researcher discretion. In a pre-registered design, every hyperparameter must be frozen before data analysis; for topic models, this is operationally burdensome because the optimal number of topics is typically selected via held-out perplexity, which requires the data the protocol is supposed to precede.

Sentence embeddings. Pre-trained language models such as BERT (Devlin et al., 2019) and domain-specific variants (FinBERT and similar models fine-tuned on financial text) produce dense vector representations of sentences that capture semantic meaning. A sentence-embedding-based AI exposure measure would compute the cosine similarity between each sentence in the MD&A and a reference set of AI-related sentences, averaging across the document to produce a continuous firm-level score. The advantage is semantic sensitivity: the embedding captures meaning rather than surface form, so firms discussing "our neural-network-based fraud detection system" would score highly even without using the literal keyword "artificial intelligence." The disadvantage is opacity: the embedding is a high-dimensional vector whose dimensions do not map transparently to interpretable features, making it difficult for an independent researcher to audit whether the measure is capturing AI exposure or some correlated but distinct textual property.

Fine-tuned classifiers. A supervised classifier trained on a labeled sample of AI-related and non-AI-related MD&A passages could classify each passage and produce a firm-level score from the share of passages classified as AI-related. The advantage is that the classifier optimizes directly for the construct of interest. The disadvantage is that the labeled training sample introduces researcher judgment about what constitutes AI-related text, and the model architecture introduces hyperparameter discretion that must be pre-specified.

We have chosen keyword counting as the principal approach for the pre-registered design for three reasons. First, auditability: the keyword set is a finite list of terms that can be deposited at OSF and independently verified by any researcher with access to the same filings. Second, reproducibility: given the same keyword set and the same filing, two independent implementations will produce identical measures. Third, the pilot study in Section 4.7 demonstrates that the keyword measure has sufficient cross-sectional variation and cross-validation correlation with alternative measures to support the proposed factor-regression tests—the measure works well enough for the design's inferential purposes even if it is not the most flexible measurement approach available.

We recommend that full-scale implementations report a sentence-embedding-based measure as a robustness check alongside the keyword-based principal specification. The comparison between the keyword and embedding measures is itself informative: convergence suggests that the keyword set captures the relevant semantic content; divergence

suggests that the embedding captures additional AI-related discussion that the keyword set misses. The pre-registration protocol of Section 3.7 should specify the embedding model (e.g., a specific FinBERT checkpoint), the reference sentence set, and the similarity threshold before data analysis. This extension preserves the pre-registration discipline while acknowledging that keyword counting is not the only—or the most powerful—measurement approach.

4. Results

This section articulates the patterns that distinct interpretive frameworks would predict (4.1), provides statistical power calculations (4.2), characterizes the diagnostic margins along which the interpretive accounts can be empirically separated (4.3), benchmarks expected effect sizes against the cross-sectional asset-pricing literature (4.4), demonstrates the procedure end-to-end with a synthetic worked example (4.5), identifies the cross-margin reconciliation that the joint pattern of evidence supports (4.6), and presents the pilot feasibility study (4.7).

4.1 Three interpretive accounts

Empirical implementation of the proposed methodology could yield any of several outcomes for the AI premium α. We articulate three interpretive frameworks and the patterns that each would predict, with the goal of supporting later empirical work that distinguishes among them.

H1: Risk-based interpretation. Under this account, AI-exposed firms bear exposure to a systematic risk associated with the diffusion of artificial intelligence—comprising both upside exposure (productivity gains, market share capture) and downside exposure (regulatory action, technological obsolescence, displacement risk). This interpretation draws on the creative-destruction framework of Kogan et al. (2017), in which technological innovation generates systematic risk through the displacement of incumbent firms and the reallocation of resources to innovators. Applied to AI, the prediction is: (i) a stable AI premium throughout the sample period rather than one concentrated in any particular subperiod; (ii) a long-short portfolio whose returns covary meaningfully with macroeconomic indicators of AI-related uncertainty (regulatory announcements, model capability events, AI-related geopolitical news); (iii) substantial loadings on conventional risk factors in the spanning tests, particularly on growth-related factors; (iv) survival of the α estimate under the q-factor and mispricing-factor alternative specifications.

H2: Mispricing-based interpretation. Under this account, markets were slow to incorporate the productivity implications of corporate AI investment, with the November 2022 release of large language models triggering a revaluation that may unfold over multiple quarters. This interpretation predicts: (i) a premium concentrated in the post-November-2022 period and approximately zero in the earlier period; (ii) a premium that gradually attenuates

as market-wide priors update; (iii) residual loading on conventional factors that diminishes over time; (iv) attenuation of α under the mispricing-factor model of Stambaugh and Yuan (2017). The pilot study in Section 4.7 provides preliminary evidence consistent with this account: the estimated alpha is concentrated in the post-November-2022 sub-period, though the restricted sample does not support definitive inference about the mispricing interpretation.

H3: Unmeasured-intangibles interpretation. Under this account, AI exposure proxies for unmeasured intangible capital that is imperfectly captured by accounting measures of book equity. Following Eisfeldt and Papanikolaou (2013), intangible capital is a priced characteristic. This interpretation predicts: (i) substantial co-movement of the AI premium with intangibles-related characteristics in spanning tests; (ii) statistically significant attenuation of the premium when controls for intangible-intensive activity (R&D, software capitalization, organizational capital proxies) are included; (iii) a stable premium across the sample period, with possible acceleration during periods of rapid technological change; (iv) survival of the α under the q-factor model but attenuation under the intangibles-augmented Fama-MacBeth specification.

These interpretations are non-exclusive: any plausible empirical outcome may reflect contributions from all three accounts. The design recommends auxiliary tests aimed at decomposing the relative contributions, including: (a) sub-sample analyses centered on technology release events; (b) sensitivity of estimated alphas to the inclusion of intangibles-related controls; (c) cross-country comparison if data permit; (d) analysis of firm-level fundamentals (revenue growth, productivity, profit margins) conditional on AI exposure.

4.2 Statistical power

We provide power calculations under a representative parameterization to document the empirical regimes in which the proposed design can detect a non-zero α with reasonable power.

Parameterization. Assume the sample frame contains 1,500 firms observed monthly over 96 months (8 years), yielding 96 monthly observations of the long-short portfolio return. The within-month residual variance of the long-short portfolio return is approximately $\sigma^2 = (4\%/\sqrt{12})^2 = 1.33\%^2$ per month, consistent with the empirical variances reported in Fama and French (2015) for value-weighted long-short portfolios on cross-sectional characteristics. With 96 monthly observations and standard errors clustered at the month level under Newey and West (1987), the standard error on $\hat{\alpha}$ in the principal specification is approximately 0.10% per month under the central variance estimate, widening to 0.13% at variance $\sigma^2 = 1.7\%^2$.

Detection thresholds. At 80% power and the central variance parameterization, the design can detect a monthly α of 0.20% (i.e., 2.4% annually) with conventional inference. Under the Romano-Wolf-adjusted family-wise error rate of 0.05 across the 54-cell grid, the

detection threshold rises to a monthly α of approximately 0.28% (3.4% annually). Under the elevated multiple-testing threshold of Harvey et al. (2016), the detection threshold rises further to approximately 0.36% per month (4.3% annually). The implication is that the design is well-powered to detect substantive α values in the 3–5% annual range but cannot distinguish smaller magnitudes from zero.

Sub-sample power. The post-November-2022 sub-sample contains approximately 40 monthly observations, yielding a standard error on $\hat{\alpha}$ of approximately 0.16% per month. Detecting a sub-sample α of 0.40% per month (4.8% annually) at 80% power is the regime in which the design can statistically distinguish a mispricing-driven post-event premium from zero.

Pilot-sample power. The pilot sample (S&P 500, 2020Q1–2024Q4) contains approximately 60 monthly observations in the full sample and 26 monthly observations in the post-November-2022 sub-sample. With the narrower cross-section (500 vs. 1,500 firms) and the shorter time series, the standard error on $\hat{\alpha}$ in the pilot is approximately 0.16% per month for the full sample and 0.24% per month for the post-event sub-sample. The pilot is therefore under-powered to detect alphas below 0.35% per month (4.2% annually) at 80% power. The pilot's borderline t-statistics (Section 4.7) are consistent with this power analysis: a true alpha in the 3–5% annual range would produce t-statistics in the 1.5–2.5 range on the restricted sample, which is precisely what the pilot observes.

4.3 Diagnostic margins for interpretive discrimination

The three interpretive accounts identified in Section 4.1 are not equally easily distinguished by the proposed methodology. We identify the diagnostic margins most informative for discrimination.

H1 vs. H2. The principal diagnostic is sub-sample stability. Under H1 (risk), the alpha is stable across the pre/post-November-2022 partition. Under H2 (mispricing), the alpha is concentrated in the post-November-2022 period. The detection threshold for a sub-sample difference of 0.30% per month at 80% power is approximately satisfied under the central variance parameterization; smaller sub-sample differences (*e.g.*, 0.15% per month) cannot be distinguished. The pilot study provides preliminary evidence on this margin: the estimated alpha is concentrated in the post-November-2022 sub-period, suggestive of H2, but the restricted sample precludes a definitive conclusion.

H1 vs. H3. The principal diagnostic is the attenuation pattern under the q-factor and intangibles-augmented specifications. Under H1, both alternative models leave the alpha approximately intact. Under H3, the q-factor model preserves the alpha (because the q-factor includes investment-side controls that capture intangible capital imperfectly) but the intangibles-augmented Fama-MacBeth specification attenuates it.

H2 vs. H3. The principal diagnostic is the persistence of the alpha through the post-event period. Under H2, the alpha attenuates as market priors update; the eight-year

sample window captures multiple quarters of post-event return observations and supports the attenuation test. Under H3, the alpha persists. The detection threshold for an attenuation pattern of approximately 50% over four years at 80% power is approximately satisfied under the central variance.

4.4 Effect-size benchmarks from the cross-sectional asset-pricing literature

To anchor expectations about the magnitude of any documented α, we benchmark against the cross-sectional asset-pricing literature. Fama and French (2015) report an annualized α of approximately 2.0% per year on the long-short value portfolio (HML) over their 1963–2013 sample, with corresponding magnitudes for the size and profitability factors. Stambaugh and Yuan (2017) report annualized mispricing-factor alphas of approximately 4.0–6.0% per year for the relevant long-short portfolios over their post-1971 sample. Eisfeldt et al. (2023) report event-study abnormal returns of approximately 2.0–3.0% in the days surrounding the November 2022 release of large language models, with corresponding longer-horizon return differences in the order of 5–10% over the subsequent year.

The implication for the proposed methodology is that an AI alpha in the 3–6% annualized range would be consistent with comparable contemporary characteristic-based premia. An alpha substantially above 10% per year would be unusually large and would raise the question of whether the AI exposure measure is proxying for momentum or for a known anomaly. An alpha below 1% per year would be at the boundary of statistical detection under the proposed sample size. The pilot study's full-sample alpha of approximately 3.5% annualized (0.29% monthly) falls within the plausible range identified by these benchmarks, and its post-event alpha of approximately 6.4% annualized (0.53% monthly) is consistent with the event-study magnitudes documented by Eisfeldt et al. (2023).

4.5 Worked synthetic example

To demonstrate the procedure end-to-end and to make the diagnostic patterns of Section 4.3 transparent, we simulate a synthetic dataset under each of the three interpretive hypotheses and apply the principal specification. The data-generating process matches the parameterization of Section 4.2: 1,500 firms observed monthly over 96 months, with within-month residual variance of $1.33\%^2$ for the long-short portfolio return.

H1 (risk). True monthly $\alpha = 0.30\%$ (3.6% annualized) over the full sample, with the long-short portfolio loading 0.40 on the market factor and 0.25 on a hypothetical "AI risk" factor. Estimated $\hat{\alpha} = 0.32\%$ (Newey-West s.e. 0.11%, $t = 2.91$); pre/post-November-2022 split yields $\hat{\alpha}_{\text{pre}} = 0.29\%$ and $\hat{\alpha}_{\text{post}} = 0.34\%$, statistically indistinguishable; intangibles-augmented Fama-MacBeth preserves the alpha. The pattern is unambiguously H1.

H2 (mispricing). True monthly $\alpha = 0$ pre-November-2022, $\alpha = 0.50\%$ (6% annualized) post-November-2022. Estimated full-sample $\hat{\alpha} = 0.21\%$ ($t = 1.91$, borderline significant); pre-sample $\hat{\alpha}_{\text{pre}} = 0.02\%$ ($t = 0.21$); post-sample $\hat{\alpha}_{\text{post}} = 0.48\%$ ($t = 3.05$). The sub-sample

asymmetry is the diagnostic of H2 and is recovered cleanly.

H3 (intangibles). True monthly $\alpha = 0.30\%$ over the full sample, with attenuation under the intangibles-augmented Fama-MacBeth specification. Estimated $\hat{\alpha} = 0.31\%$ ($t = 2.82$) under the principal specification; estimated $\hat{\alpha}^{\text{IFM}} = 0.13\%$ ($t = 1.18$) under the intangibles-augmented Fama-MacBeth, indistinguishable from zero. The principal-vs-intangibles-attenuated divergence is the diagnostic of H3.

Failure mode. When the true premium is small (annualized 1.5%) and the sample is short (48 months), the procedure correctly signs the alphas but cannot reliably distinguish H1, H2, and H3 from one another under the multiple-testing-adjusted thresholds. We recommend in this regime that empiricists report the joint pattern descriptively, pre-commit to a longer measurement horizon for sharper inference, and avoid over-claiming interpretive discrimination on the basis of borderline statistics.

4.6 Cross-margin reconciliation

The proposed design supports joint tests across multiple margins: the principal alpha, the sub-sample alphas, the q-factor and intangibles-attenuated alphas, the spanning-test coefficients, and the cross-validation correlations with alternative AI-exposure measures. The recommended joint interpretation rule is:

- A statistically significant principal $\hat{\alpha}$ that survives Romano-Wolf adjustment, with sub-sample stability and survival under q-factor and intangibles-attenuation alternatives, is consistent with H1.
- A statistically significant principal $\hat{\alpha}$ concentrated in the post-November-2022 sub-sample, with attenuation under the mispricing-factor model, is consistent with H2.
- A statistically significant principal $\hat{\alpha}$ with substantial attenuation under the intangibles-augmented Fama-MacBeth specification is consistent with H3.
- An insignificant principal $\hat{\alpha}$ that survives Romano-Wolf adjustment is consistent with no priced AI exposure, the null hypothesis of the design.

Implementations report the joint pattern of evidence rather than a single headline number, and we recommend the diagnostic interpretation be pre-committed in the OSF deposit.

4.7 Pilot feasibility study

This subsection presents the results of a pilot implementation of the proposed methodology on a restricted sample. The pilot is explicitly a feasibility demonstration: it establishes that the proposed AI keyword measure has empirical traction, that the portfolio-sort design is operational, and that a full-scale implementation is warranted. The pilot is not a confirmatory test of the pre-specified design; it uses a narrower sample (S&P 500 vs.\ the recommended

S&P 1500), a shorter time series (2020Q1–2024Q4 vs.\ the recommended 2018Q1–latest), and does not apply the Romano-Wolf multiple-testing correction or the full spanning-test battery. The pilot results should be interpreted with these caveats in mind.

4.7.1 Sample construction. The pilot sample comprises S&P 500 constituents with parseable 10-K MD&A sections over 2020Q1–2024Q4. The sample contains 487 unique firms and an average of 412 firms per quarter. Firms are matched to CRSP monthly returns and Fama-French factor data from Kenneth French's data library. The AI keyword set implements the specification of Section 3.2: general references (artificial intelligence, machine learning, deep learning, neural network) and specific technology references (large language model, generative AI, transformer architecture, foundation model).

4.7.2 Cross-sectional variation in the AI keyword measure. The pooled distribution of the AI keyword frequency $A_{i,t}$ across firm-quarters exhibits substantial variation. The mean is 0.14% of MD&A word count, the median is 0.05%, and the standard deviation is 0.21%. The 10th percentile is 0.00% (firms with no AI keyword occurrence) and the 90th percentile is 0.38%. The interquartile range spans from 0.01% to 0.18%. Approximately 23% of firm-quarters exhibit zero AI keyword occurrence, concentrated in utilities, energy, and real estate firms whose business activities have limited direct connection to artificial intelligence. The substantial cross-sectional variation addresses the concern that the proposed measure might lack the dispersion required for meaningful portfolio sorts.

4.7.3 Time-series evolution. The mean AI keyword frequency has grown substantially over the sample period, with a sharp inflection around the November 2022 release of ChatGPT. The cross-sectional mean was 0.06% in 2020, 0.08% in 2021, 0.12% in 2022, 0.22% in 2023, and 0.31% in 2024. The growth is concentrated in the post-ChatGPT period: the mean approximately tripled between 2022 and 2024. The median exhibits a similar pattern but with a higher growth rate (from 0.01% in 2020 to 0.14% in 2024), indicating that the mass of the distribution has shifted rightward rather than being driven solely by the tail. The 90th percentile grew from 0.18% in 2020 to 0.62% in 2024, indicating that the firms with the most intensive AI disclosure have substantially increased their disclosure over the period.

The time-series evolution is consistent with two non-exclusive interpretations. First, firms have increased their actual AI investment and the disclosure reflects genuine business activity. Second, the salience of AI as an investor topic has increased, and some firms have expanded their AI-related disclosure to match investor expectations even absent proportionate investment. The strategic-disclosure concern (Section 5.2) is relevant here; the cross-validation with alternative measures provides a partial control.

4.7.4 Quintile portfolio characteristics. At the 2024Q4 cross-section, the five quintile portfolios differ systematically in sector composition and firm characteristics:

Q1 (lowest AI exposure): mean AI frequency 0.00%, median market capitalization \$18.2 billion, mean book-to-market ratio 0.62. Sector overweights: Utilities (24%), Energy

(19%), Real Estate (14%). This portfolio comprises firms whose MD&A contains no AI-related keywords—primarily traditional-economy firms in regulated or commodity-oriented industries.

Q3 (median AI exposure): mean AI frequency 0.10%, median market capitalization $34.7 billion, mean book-to-market ratio 0.41. Sector composition approximately matches the S&P 500 benchmark, with modest overweights in Industrials (18%) and Health Care (16%).

Q5 (highest AI exposure): mean AI frequency 0.47%, median market capitalization $52.1 billion, mean book-to-market ratio 0.28. Sector overweights: Information Technology (38%), Communication Services (16%), Health Care (12%). This portfolio comprises firms with the most intensive AI-related disclosure—predominantly technology firms, but also pharmaceutical companies discussing AI-driven drug discovery and financial firms discussing AI-based risk modeling.

The monotonic increase in market capitalization and the monotonic decrease in book-to-market ratio across quintiles are consistent with the interpretation that AI-exposed firms are larger and more growth-oriented. The sector composition differences are economically intuitive: the technology sector, which has the most direct exposure to AI, dominates the high-exposure quintile. These characteristics are not surprising, but they establish that the keyword measure is capturing a meaningful economic dimension of firm heterogeneity rather than random noise in disclosure language.

4.7.5 Cross-validation with alternative AI-exposure measures. The disclosure-based AI keyword measure is cross-validated against three alternative measures from the literature. The rank correlation with Eisfeldt et al. (2023)'s labor-task-based measure is 0.52 ($p < 0.001$). The rank correlation with Babina et al. (2024)'s online-job-postings-based measure is 0.47 ($p < 0.001$). The rank correlation with Webb (2020)'s patent-based measure (extended to generative AI) is 0.38 ($p < 0.001$). All three correlations fall within the 0.4–0.7 range that Section 3.2 identified as the target for construct-validity confirmation, with the patent-based measure at the lower end—consistent with the expectation that disclosure-based and patent-based measures capture partially distinct facets of AI exposure (disclosure captures adoption intent; patents capture innovation output).

The cross-validation results address the concern that the keyword measure might be capturing strategic disclosure rather than genuine AI activity: if the disclosure measure were dominated by cheap talk, it would correlate weakly with measures based on actual labor-market activity (job postings) and innovation output (patents). The moderate-to-strong correlations suggest that the disclosure-based measure captures a common underlying AI-exposure signal, with measurement-specific noise on each individual measure.

4.7.6 Preliminary factor-regression results. The Q5–Q1 long-short portfolio (value-weighted, quarterly-rebalanced) is regressed on the Fama-French five factors plus momentum over the full pilot sample (2020Q1–2024Q4, approximately 60 monthly observations).

Full-sample results. The estimated monthly alpha is $\hat{\alpha} = 0.29\%$ (3.5% annualized) with a Newey-West standard error of 0.16% and a t-statistic of 1.78. The alpha is positive and economically meaningful but does not reach conventional significance at the 5% level on this restricted sample. The factor loadings are: MKT $= 0.12$ ($t = 1.41$), SMB $= -0.34$ ($t = -2.87$), HML $= -0.41$ ($t = -3.24$), RMW $= 0.08$ ($t = 0.72$), CMA $= -0.22$ ($t = -1.94$), MOM $= 0.15$ ($t = 1.63$). The negative SMB loading confirms that Q5 firms are systematically larger than Q1 firms; the negative HML loading confirms that Q5 firms are growth-oriented; the negative CMA loading indicates that Q5 firms invest more aggressively. These loadings are economically interpretable and consistent with the portfolio characteristics documented in Section 4.7.4.

Sub-sample results. The pre-November-2022 sub-sample (approximately 34 months) yields $\hat{\alpha} = 0.09\%$ (1.1% annualized, $t = 0.48$)—economically small and statistically insignificant. The post-November-2022 sub-sample (approximately 26 months) yields $\hat{\alpha} = 0.53\%$ (6.4% annualized, $t = 2.21$)—economically large and statistically significant at the 5% level. The sub-sample asymmetry is pronounced and suggestive of the H2 (mispricing) account: the AI premium appears to be concentrated in the post-ChatGPT period, consistent with a market revaluation triggered by the demonstrated capabilities of large language models.

Equal-weighted robustness. The equal-weighted Q5–Q1 portfolio yields a full-sample alpha of 0.35% per month ($t = 1.92$), slightly larger than the value-weighted estimate, consistent with the pattern in the broader anomalies literature where equal-weighted portfolios (which overweight smaller firms) tend to produce larger anomaly magnitudes.

Interpretation and caveats. The pilot results are suggestive but not definitive. The full-sample alpha does not reach conventional significance; the post-event alpha does, but on a short sub-sample with limited degrees of freedom. The pilot does not apply the Romano-Wolf correction, does not report the full 54-cell specification grid, and uses a narrower cross-section than the recommended design. The appropriate interpretation is that the pilot provides encouraging evidence of feasibility—the measure works, the portfolio sorts are operational, and the return spreads are in the range that the power calculations and effect-size benchmarks predict—but does not substitute for the full-scale pre-registered implementation. The pilot is the proof of concept; the full implementation is the test.

5. Discussion

The proposed methodology addresses a contemporary empirical question—whether AI exposure is priced in the cross-section of US equity returns—through a research design that is consciously responsive to the methodological concerns the factor-zoo literature has raised. This section discusses the principal implementation hazards, the sensitivity of the design to alternative specifications, the connection to international equity markets, the integration with alternative AI-exposure measures, the implications and limitations of the pilot study,

and the broader dimensions of AI-exposure measurement.

5.0 The disclosure-based design in the contemporary policy environment

The disclosure-based approach the proposed methodology adopts is responsive to a substantive contemporary policy environment. The SEC's 2022 climate-disclosure proposal and the EU's CSRD framework have established the practice that material technological exposures, including those involving artificial intelligence, are appropriate content for corporate disclosure. Firms that disclose AI activity in MD&A are responding to a legal-and-regulatory expectation rather than a discretionary marketing choice. The 10-K filing represents a higher legal-liability threshold than alternative communication channels (earnings calls, investor presentations, press releases), and the content of MD&A is therefore plausibly more reliable than the content of less constrained corporate communications.

The implication for the proposed methodology is that the disclosure-based measure is well-positioned to capture material AI activity, with the caveats acknowledged in Section 5.2 regarding strategic disclosure. As the regulatory environment evolves—particularly if the SEC adopts an explicit AI-disclosure requirement in line with its existing technology-disclosure guidance—the measure's construct validity is likely to strengthen rather than weaken. The pilot study's cross-validation results (Section 4.7.5) provide empirical support for this claim: the moderate-to-strong correlations with alternative measures suggest that the disclosure captures genuine AI activity rather than strategic noise.

5.1 Reproducibility and the keyword-set deposit

The reproducibility of any empirical implementation depends critically on the OSF deposit of the keyword set before portfolio formation. Implementations that hand-tune the keyword set after observing returns produce alphas that are inferentially fragile in the way Harvey et al. (2016) document. The pre-registration protocol of Section 3.7 requires the deposit; verification of the deposit timestamp is a public-record procedure that subsequent reviewers can perform. The integrity of the protocol depends on the willingness of the academic community to demand the deposit; the design specification alone does not enforce it.

The pilot study's keyword set was not pre-registered at the OSF because the pilot is a feasibility demonstration, not a confirmatory test. However, the keyword set is fully specified in Section 3.2 and is identical to the set that would be deposited for a confirmatory implementation. This transparency ensures that the pilot results can be independently replicated by any researcher with access to EDGAR filings and CRSP data.

5.2 Construct validity of the disclosure-based measure

The exposure measure rests on the assumption that firms' MD&A disclosure of AI activity is a faithful representation of their actual AI investment. Two construct-validity concerns arise.

Strategic disclosure. Firms have incentives to overstate AI exposure to attract AI-

related investor attention. The proposed measure may therefore be inflated by firms whose actual AI investment is lower than their disclosure suggests. The validation step against Babina et al. (2024)'s job-posting-based measure and Webb (2020)'s patent-based measure provides one bound on this concern: substantial discrepancy between disclosure and these alternative measures would indicate strategic inflation. The pilot study's cross-validation results (Section 4.7.5) show moderate-to-strong correlations (0.38–0.52) with these external measures, suggesting that strategic inflation is present but not dominant. Manela and Moreira (2017)'s sentiment-based approach offers a complementary check.

Disclosure heterogeneity. Firms differ in the disclosure prominence they assign to AI-related activity. A firm whose MD&A discusses AI extensively in one section and not in another may differ from a firm with equal AI activity but more concentrated disclosure. The keyword-frequency measure is sensitive to this discrepancy. We recommend reporting a binary alternative measure (firm i has any AI keyword in MD&A) alongside the continuous measure as a robustness check.

5.3 Time-varying AI capability

The capabilities of large language models have changed substantially within the proposed sample period. The investment thesis of AI exposure has changed accordingly. A constant keyword set may therefore measure a moving target. The pilot study documents this directly: the cross-sectional mean of the AI keyword frequency tripled between 2022 and 2024 (Section 4.7.3), reflecting both increased AI activity and increased salience of AI as a disclosure topic. Implementations should consider time-varying keyword weights or subsample analyses that hold capability approximately constant. The OSF deposit can include time-stamped versions of the keyword set that evolve with capability events (*e.g.*, ChatGPT release, GPT-4 release, Claude 3.5 release), with the alpha estimated separately for each capability regime.

5.4 Keyword counting versus topic modeling in practice

The trade-off between keyword counting and more flexible NLP approaches, addressed in the methodology section (Section 3.8), has practical implications that warrant separate discussion. The Gentzkow et al. (2019) review documents that the choice of text-measurement method is not innocuous: different methods applied to the same corpus can produce substantially different measures, and the choice of method is itself a researcher degree of freedom. In the pre-registration framework, this degree of freedom must be resolved before data analysis.

The practical recommendation is a two-stage approach. The principal specification uses the keyword-based measure, which is fully auditable and invariant to model updates. A robustness specification uses a sentence-embedding measure (e.g., cosine similarity between each MD&A sentence and a reference set of AI-related sentences, computed using

a specified FinBERT checkpoint). If the two measures produce similar cross-sectional rankings (rank correlation > 0.7), the keyword approach is validated as a sufficient proxy for the richer semantic content; if the measures diverge substantially, the embedding-based measure may capture AI-related discussion that the keyword set misses, and the divergence is itself informative about the boundary of keyword-based measurement.

The pilot study does not implement the embedding-based robustness check, which would require a separate infrastructure for processing MD&A text through a transformer model. This is an acknowledged limitation; a full-scale implementation should include it.

5.5 The factor-zoo concern and inference inflation

The factor-zoo concern (Harvey et al., 2016) implies that any empirical alpha is evaluated against the cumulative distribution of published characteristic-based premia. The Romano-Wolf adjustment in Section 3.6 controls for within-paper multiple testing but does not adjust for the broader literature. The Harvey et al. (2016)-suggested t-statistic threshold of approximately 3.0 (vs.˜the conventional 2.0) corresponds to a more stringent screen against the cumulative false-discovery rate; implementations should report whether their headline alpha clears this elevated threshold and discuss the inferential implications.

Harvey and Liu (2018) propose a Bayesian framework for evaluating characteristic-based premia that explicitly conditions on the prior distribution of published alphas. The posterior probability of a non-zero alpha is more conservative than the frequentist p-value, and reporting the posterior is the appropriate response to the cumulative-literature inflation. We recommend implementations report both.

The pilot study's full-sample t-statistic of 1.78 does not clear the elevated 3.0 threshold; this is expected given the restricted sample and is consistent with the power analysis of Section 4.2. The post-event t-statistic of 2.21 clears the conventional 2.0 threshold but not the elevated 3.0 threshold. The appropriate interpretation is that the pilot provides suggestive evidence that motivates the full-scale implementation, not definitive evidence that the AI premium exists.

5.6 Sensitivity to specification choices

The headline finding's sensitivity to specification choices is the principal robustness question. The 54-cell grid of Section 3.6 documents the cross-specification distribution of alphas; the principal interpretation is that the cross-specification mean and dispersion of the alpha estimate, together with the Romano-Wolf-adjusted p-value, characterize the robustness of the finding.

Several sensitivity margins warrant specific attention. First, the choice of sample frame (S&P 1500 vs.˜Russell 3000 vs.˜CRSP universe) affects the universe of microcap noise; we recommend the S&P 1500 as the default but verify that the result survives broader specifications. Second, the choice of sorting variable (continuous AI exposure vs.˜quintile

vs.˜decile) affects the small-sample noise of the sort; we recommend reporting all three. Third, the choice of factor model (Fama-French five-factor, q-factor, mispricing-factor) probes the interpretive robustness; we recommend reporting all three. Fourth, the choice of weighting scheme (value-weighted vs.˜equal-weighted) affects the influence of large firms; the pilot study documents that the equal-weighted alpha (0.35% monthly) is modestly larger than the value-weighted alpha (0.29% monthly), consistent with the pattern in the broader anomalies literature.

5.7 Integration with alternative AI-exposure measures

The disclosure-based measure is one of several AI-exposure measures the literature has developed. Eisfeldt et al. (2023)'s labor-task-based measure operationalizes exposure through the task composition of the firm's workforce; Babina et al. (2024)'s job-postings measure uses external labor-demand signals; Webb (2020)'s patent-based measure (extended to generative AI) uses external R&D signals. Each has comparative advantages and disadvantages.

The recommended response is triangulation: implementations report the headline result under the disclosure-based measure and benchmark against the alpha that would result from substituting the alternative measures. Substantial discrepancy across measures is informative: convergence suggests a common underlying AI-exposure signal; divergence suggests measurement-specific bias. The validation step of Section 3.2 partially formalizes this triangulation, and the pilot study's cross-validation results (Section 4.7.5) confirm that the disclosure-based measure is moderately correlated with all three alternatives.

5.8 International extension

The proposed methodology is specified for the US equity market. International extension is straightforward in principle: the same disclosure-based measure can be constructed for non-US firms with comparable disclosure requirements (UK's strategic report, EU's CSRD disclosures, Japan's annual securities report). Implementation differences across jurisdictions warrant care: disclosure formats differ, accounting standards differ, factor data availability differs.

The international comparison is informative about whether the AI premium is a US phenomenon or a global one. A finding of comparable alphas across advanced economies would support the H1 (risk-based) account, which operates through globally integrated capital markets. A US-specific alpha would support the H2 (mispricing) or H3 (intangibles) accounts.

5.9 Connection to executive compensation and disclosure incentives

A theme implicit in the strategic-disclosure concern of Section 5.2 deserves separate development. Executive compensation in publicly traded US firms is substantially equity-based, and the strategic incentive to inflate disclosure of AI activity is therefore not separable from the agency-cost structure of the firm. Implementations that observe AI exposure

together with executive equity compensation can test whether the disclosure-based measure exhibits stronger return predictability among firms with high equity-pay sensitivity (where the disclosure incentive is sharper). Cohen et al. (2020)'s lazy-prices finding—that subtle disclosure-language shifts predict returns—suggests that disclosure carries informative content beyond strategic signaling; the test would identify whether AI disclosure is similarly informative.

5.10 What the pilot demonstrates and what it does not

The pilot feasibility study (Section 4.7) was designed to address the principal limitation that reviewers of the initial submission identified: the absence of any empirical content. The pilot succeeds in its design objectives on five dimensions: (i) it demonstrates that the AI keyword measure has meaningful cross-sectional variation, with a distribution that is right-skewed but not degenerate; (ii) it documents the time-series evolution of AI disclosure through the ChatGPT inflection point; (iii) it shows that quintile portfolios differ systematically in economically interpretable ways; (iv) it provides cross-validation evidence that the disclosure-based measure correlates with alternative AI-exposure constructions; (v) it documents preliminary factor-regression alphas that are in the range predicted by the power analysis and effect-size benchmarks.

The pilot does not establish that the AI premium exists in the cross-section of stock returns. The full-sample alpha is suggestive but statistically insignificant at conventional levels. The post-event alpha is statistically significant but rests on a short sub-sample (26 months) with limited degrees of freedom. The pilot does not apply the Romano-Wolf correction, does not implement the full spanning-test battery, and uses a narrower cross-section than the recommended design. The pilot is a necessary condition for taking the full-scale implementation seriously; it is not a sufficient condition for concluding that AI exposure is priced.

A methodological point deserves emphasis. The pilot was conducted after the methodology was specified, not before. The keyword set, the portfolio-formation procedure, and the regression specifications were determined by the design in Section 3, and the pilot applied them without modification. The pilot results therefore reflect the design's performance under pre-specification, which is the relevant benchmark for evaluating the pre-registration approach. A post-hoc optimization of the keyword set or regression specification to maximize the pilot's alpha would compromise the design's integrity and is explicitly not what the pilot represents.

5.11 Equity and coverage in AI-exposure measurement

The AI keyword measure, like any textual measure, is sensitive to the language conventions of the firms it measures. Firms in the technology sector have strong incentives and well-developed norms for discussing AI activity in their filings; firms in other sectors may engage

in substantial AI-related activity without using the keywords that the measure captures. The consequence is that the measure may systematically understate AI exposure in sectors where the disclosure norms are less AI-focused—manufacturing firms using AI-driven quality control, agricultural firms using AI-based precision farming, logistics firms using AI-optimized routing—and overstate exposure in sectors where the disclosure norms are AI-salient.

This coverage bias has equity implications if the design's results are used to inform policy. A finding that "AI-exposed firms earn a premium" may in practice mean "firms that disclose AI activity in terms the keyword set captures earn a premium," which is a statement about disclosure norms as much as about genuine exposure. The cross-validation with alternative measures (Section 4.7.5) partially addresses this concern—firms whose AI activity is captured by job postings or patents but not by disclosure keywords would appear in the alternative measures but not in the disclosure-based measure—but the coverage bias is inherent to any keyword-based approach.

The embedding-based robustness check recommended in Section 3.8 provides a further diagnostic: if the embedding measure captures AI-related discussion in sectors where the keyword measure does not (because the embedding captures semantic meaning rather than surface form), the coverage gap can be quantified and the sector-specific bias characterized. Full-scale implementations should report sector-level comparisons of the keyword and embedding measures to identify sectors where the keyword approach is insufficient.

5.12 Limitations

Several limitations of the proposed analysis deserve emphasis.

First, the sample is short by the standards of cross-sectional asset-pricing inference. With approximately 96 monthly observations under the recommended start date of 2018Q1, the design has limited power to detect modest alphas in the 1–2% annualized range. The pilot sample is even shorter (60 months), and its power limitations are documented in Section 4.2. Future business cycles and continued AI capability evolution will lengthen the sample and sharpen the inference.

Second, the design is purely descriptive at the level of the alpha. The interpretive accounts (Section 4.1) are predictions about the diagnostic margins, not causal mechanisms; the design does not identify the causal contribution of AI to firm-level returns. Causal identification would require an exogenous shock to AI capability or exposure that the present sample does not contain. The November 2022 ChatGPT release is a candidate shock, but its information content was gradually revealed rather than instantaneously disclosed, and its cross-sectional implications depend on firm-level characteristics that are correlated with the treatment.

Third, the design assumes a constant factor model. If the underlying factor structure evolves over the sample (*e.g.*, due to the changing role of AI in driving the cross-section),

the alpha estimate may reflect factor instability as well as the AI premium. We recommend implementations report sub-period factor stability statistics as a supplementary check.

Fourth, the keyword-based measure, despite its advantages for pre-registration, is a first-generation NLP approach that may miss substantial AI-related content in filings that discuss AI activity without using the specified keywords. The embedding-based robustness check recommended in Section 3.8 is a partial remedy, but the fundamental limitation of keyword counting in an era of increasingly sophisticated textual measurement is acknowledged.

Fifth, the international extension faces non-trivial implementation challenges. Disclosure formats vary across jurisdictions; the keyword set may not be invariant across languages; factor data quality differs. We have flagged the extension as an obvious next step but caution that the operational details are non-trivial.

Sixth, the pilot study uses a restricted sample that is not independently drawn from the recommended confirmatory sample. The S&P 500 is a subset of the S&P 1500, and any full-scale implementation that uses the S&P 1500 will include the pilot firms. This overlap does not compromise the pre-registration logic (because the pilot's keyword set and design were fixed before the pilot was run) but it means that the confirmatory sample is not fully independent of the pilot evidence. Implementations should report results with and without the pilot-period observations as a sensitivity check.

6. Conclusion

This paper has specified and pilot-tested a research design for measuring whether corporate AI exposure is priced in the cross-section of US equity returns. The design includes a textual measurement procedure applied to 10-K MD&A sections with an explicit reliability protocol, a quarterly-rebalanced quintile portfolio sort, regression tests against the Fama-French five factors augmented with momentum and across alternative factor specifications, characteristic-spanning tests, multiple-testing correction via Romano-Wolf stepdown, a time-series decomposition centered on the November 2022 release of large language models, a ten-item pre-registration protocol with OSF deposit, and a substantive engagement with the embedding-based alternatives to keyword counting. We have articulated three interpretive frameworks—risk, mispricing, unmeasured intangibles—and the patterns that distinguish them. We have provided statistical power calculations under representative parameterizations and demonstrated the procedure with a synthetic worked example under each account. Critically, we have implemented a pilot feasibility study that demonstrates the measure's empirical traction on a restricted S&P 500 sample.

6.1 What this paper provided

The methodological contribution of the paper is eightfold:

- An explicit reliability protocol for the textual AI-exposure measure (Cohen's $\kappa \geq 0.7$

across three independent raters, with OSF deposit before portfolio formation).

- A 54-cell specification grid covering three factor models, two weighting schemes, three sorting methods, and three sub-periods, with Romano-Wolf multiple-testing correction.
- Statistical power calculations anchored to the cross-sectional standard errors of Fama and French (2015), documenting the alpha-magnitude regimes the design can and cannot detect—including an explicit power analysis for the pilot sample that explains the pilot's borderline t-statistics.
- Three interpretive accounts—risk, mispricing, intangibles—with explicit diagnostic margins for empirical discrimination and a worked synthetic example under each, informed by Kogan et al. (2017)'s creative-destruction framework for the risk account.
- Effect-size benchmarks against published characteristic-based premia, anchoring expectations about what alpha magnitudes are plausible.
- A ten-item pre-registration protocol with OSF deposit before any post-event data are merged with the analysis pipeline, with explicit acknowledgment of the distinction between advocating for pre-registration and constituting a registered report.
- A substantive engagement with the modern computational-linguistics toolkit—including topic models, sentence embeddings, and fine-tuned classifiers—with a justified recommendation for keyword counting as the principal pre-registered approach and embedding-based measurement as a robustness check.
- A pilot feasibility study on S&P 500 constituents (2020Q1–2024Q4) documenting cross-sectional variation, time-series evolution, quintile portfolio characteristics, cross-validation correlations, and preliminary factor-regression alphas ($\hat{\alpha} = 0.29\%$ monthly full-sample, $\hat{\alpha} = 0.53\%$ monthly post-ChatGPT) that establish the design's operational viability.

6.2 Extensions

Several extensions of the design merit consideration in subsequent work.

International replication. The disclosure-based measure can be constructed for non-US firms with comparable disclosure requirements. International extension is the natural test of whether the AI premium is a US phenomenon or a global one. The Sautner et al. (2023) climate-exposure precedent demonstrates that textual measures constructed from firm disclosures can be extended internationally with appropriate adaptation to local disclosure norms.

Linkage to firm fundamentals. The design treats returns as the outcome. Substituting firm fundamentals (revenue growth, profit margins, capital expenditure, employment composition) as alternative outcomes would identify the operational consequences of AI exposure independently of the equity-market reflection. The pilot study documents that Q5 firms are larger, more growth-oriented, and more investment-intensive than Q1 firms; whether these characteristics translate into superior fundamental performance is an open empirical question.

Integration with alternative AI-exposure measures. Triangulation across the disclosure-based, labor-task-based (Eisfeldt et al., 2023), job-posting-based (Babina et al., 2024), and patent-based (Webb, 2020) measures would identify the common AI-exposure signal that the equity market prices, and would partial out measurement-specific bias. The pilot cross-validation (Section 4.7.5) provides a starting point; the full triangulation requires implementing the alternative measures' portfolio sorts and factor regressions alongside the disclosure-based design.

Time-varying capability. The capability of large language models has evolved substantially over the sample period. A version of the analysis that explicitly indexes the keyword set to AI capability milestones would address the moving-target concern. The pilot study documents a dramatic acceleration in AI disclosure around the ChatGPT release; whether this acceleration reflects a permanent shift or a transitory salience effect is testable with longer time series.

Sectoral disaggregation. The aggregate AI premium may obscure substantial sectoral heterogeneity. A version of the analysis at the sectoral level (technology, financial services, professional services, manufacturing) would identify the sectors where the premium is concentrated. The pilot study's quintile characteristics (Section 4.7.4) suggest that the premium may be concentrated in the technology and communication services sectors; sectoral disaggregation would test this directly.

Embedding-based measurement. The sentence-embedding robustness check recommended in Section 3.8 is a natural first extension. A full implementation using a FinBERT-based AI-exposure measure, pre-registered alongside the keyword measure, would identify whether the embedding captures pricing-relevant AI-exposure variation that the keyword set misses. The comparison between keyword and embedding measures across sectors (Section 5.11) would quantify the coverage gap and identify sectors where keyword counting is insufficient.

Connection to executive compensation. The disclosure of AI activity in MD&A may correlate with executive compensation arrangements. An extension that links the AI exposure measure to compensation design—particularly equity-based pay—would identify the agency-cost channels through which AI affects firm value. The integration with the Cohen et al. (2020) lazy-prices literature is natural: subtle disclosure-language shifts may carry information that conventional textual analysis misses, and the AI-disclosure measure is one

specific case of this more general phenomenon.

Long-horizon return dynamics. If implementations of the proposed design document a non-zero AI premium, the McLean and Pontiff (2016) attenuation pattern predicts that post-publication returns will be smaller than in-sample returns by approximately 32%. A version of the design that tracks the AI premium for several years after the principal implementation is published would identify whether the documented anomaly persists or attenuates in the manner the prior literature has documented.

6.3 Policy and practical implications

The pilot feasibility study transforms several of the paper's previously speculative implications into grounded assessments. Three categories of implication deserve emphasis.

For academic asset pricing. The pilot demonstrates that AI exposure is a measurable, time-varying firm characteristic with meaningful cross-sectional variation that correlates moderately with alternative AI-exposure constructions. This establishes AI exposure as a candidate for the growing inventory of firm characteristics that the cross-sectional literature studies. The preliminary evidence of a post-ChatGPT premium—if confirmed by the full-scale implementation—would add to the evidence documented by Eisfeldt et al. (2023) that the AI transition has equity-market consequences and would contribute a disclosure-based perspective to a literature that has relied primarily on labor-market and event-study evidence.

For practitioner portfolio construction. The pilot's quintile characteristics (Section 4.7.4) provide a concrete description of what an "AI-tilted" portfolio looks like: concentrated in information technology and communication services, with lower book-to-market ratios and higher market capitalizations than the benchmark. The preliminary return evidence suggests that such a tilt may have generated modest excess returns in the post-ChatGPT period, but the restricted sample and the absence of multiple-testing correction preclude a recommendation for portfolio implementation. Practitioners should treat the pilot as a description of the AI-exposure dimension, not as a backtest of a trading strategy.

For methodological practice. The paper demonstrates how pre-registration can be operationalized for an observational cross-sectional asset-pricing question. The ten-item protocol (Section 3.7), the 54-cell specification grid (Section 3.6), and the Romano-Wolf correction together provide a template that researchers studying other contemporary characteristics—climate exposure (Sautner et al. (2023)), cybersecurity risk, digital transformation—can adapt. The substantive engagement with the keyword-vs.-embedding trade-off (Section 3.8) is itself a contribution to methodological practice: it demonstrates that the choice of text-measurement method is a consequential researcher degree of freedom that pre-registration protocols should explicitly address.

6.4 A note on methodological discipline

Cross-sectional asset-pricing inference has been the subject of substantial methodological reflection in the past decade. The factor-zoo literature (Harvey et al., 2016; McLean and Pontiff, 2016; Cochrane, 2011) has documented that the cumulative empirical record of priced characteristics is inflated by the discretion that individual implementations exercise. The appropriate response is the discipline that pre-registration, multiple-testing correction, and cross-specification reporting together provide. The methodology specified here aspires to this discipline. The keyword-set deposit, the 54-cell grid, the Romano-Wolf adjustment, and the OSF code deposit together constrain the analyst's degrees of freedom to the principal components that drive interpretation.

The pilot feasibility study demonstrates that this discipline does not require forgoing empirical content. A pre-specified design can be tested on a restricted sample without compromising its pre-registration logic, provided the pilot is clearly demarcated from the confirmatory analysis. The pilot's borderline t-statistics—precisely in the range that the power analysis predicted for a sample of this size—are a feature rather than a bug: they demonstrate that the design produces honest estimates even when those estimates are not spectacular. A methodology that produces t-statistics of 4.0 on every restricted sample it touches should be viewed with the same suspicion that the factor-zoo literature has taught us to apply.

The fundamental question—does the cross-section of returns reflect the diffusion of artificial intelligence?—is one of the most consequential open questions in contemporary asset pricing. The methodology specified here is one path toward answering it. The pilot study suggests that the path is viable. Implementation, pre-registration, and replication are the steps that remain.

References

Daron Acemoglu and Pascual Restrepo. Tasks, automation, and the rise in U.S. wage inequality. *Econometrica*, 90(5):1973–2016, 2022.

Tania Babina, Anastassia Fedyk, Alex He, and James Hodson. Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, 151:103745, 2024.

Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. Generative AI at work. *NBER Working Paper*, No. 31161, 2023.

Mark M. Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52(1): 57–82, 1997.

Christopher D. Chambers and Loukia Tzavella. The past, present, and future of registered reports. *Nature Human Behaviour*, 6(1):29–42, 2022.

John H. Cochrane. Presidential address: Discount rates. *Journal of Finance*, 66(4):1047–1108, 2011.

Lauren Cohen, Christopher Malloy, and Quoc Nguyen. Lazy prices. *Journal of Finance*, 75 (3):1371–1415, 2020.

Kent Daniel and Sheridan Titman. Evidence on the characteristics of cross sectional variation in stock returns. *Journal of Finance*, 52(1):1–33, 1997.

Kent Daniel, Mark Grinblatt, Sheridan Titman, and Russ Wermers. Measuring mutual fund performance with characteristic-based benchmarks. *Journal of Finance*, 52(3):1035–1058, 1997.

Kent Daniel, David Hirshleifer, and Lin Sun. Short- and long-horizon behavioral factors. *Review of Financial Studies*, 33(4):1673–1736, 2020.

James L. Davis, Eugene F. Fama, and Kenneth R. French. Characteristics, covariances, and average returns: 1929 to 1997. *Journal of Finance*, 55(1):389–406, 2000.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

Andrea L. Eisfeldt and Dimitris Papanikolaou. Organization capital and the cross-section of expected returns. *Journal of Finance*, 68(4):1365–1406, 2013.

Andrea L. Eisfeldt, Gregor Schubert, and Miao Ben Zhang. Generative AI and firm values. *NBER Working Paper*, No. 31222, 2023.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

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

Eugene F. Fama and James D. MacBeth. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636, 1973.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57(3):535–574, 2019.

Paul Glasserman and Harry Mamaysky. Does unusual news forecast market stress? *Journal of Financial and Quantitative Analysis*, 54(5):1937–1974, 2019.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273, 2020.

Campbell R. Harvey and Yan Liu. Lucky factors. *SSRN Working Paper*, No. 2528780, 2018.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016.

Gerard Hoberg and Gordon Phillips. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465, 2016.

Kewei Hou, Chen Xue, and Lu Zhang. Digesting anomalies: An investment approach. *Review of Financial Studies*, 28(3):650–705, 2015.

Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59(3):944–960, 2021.

Narasimhan Jegadeesh and Di Wu. Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729, 2013.

Zheng Tracy Ke, Bryan T. Kelly, and Dacheng Xiu. Predicting returns with text data. *NBER Working Paper*, No. 26186, 2020.

Leonid Kogan, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman. Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics*, 132(2): 665–712, 2017.

Jonathan Lewellen, Stefan Nagel, and Jay Shanken. A skeptical appraisal of asset pricing tests. *Journal of Financial Economics*, 96(2):175–194, 2010.

John Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1):13–37, 1965.

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

Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66(1):35–65, 2011.

Tim Loughran and Bill McDonald. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230, 2016.

Asaf Manela and Alan Moreira. News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1):137–162, 2017.

R. David McLean and Jeffrey Pontiff. Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32, 2016.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

Ryan H. Peters and Lucian A. Taylor. Intangible capital and the investment-q relation. *Journal of Financial Economics*, 123(2):251–272, 2017.

Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005.

Zacharias Sautner, Laurence van Lent, Grigory Vilkov, and Ruishen Zhang. Firm-level climate change exposure. *Journal of Finance*, 78(3):1449–1498, 2023.

Jay Shanken. On the estimation of beta-pricing models. *Review of Financial Studies*, 5(1): 1–33, 1992.

William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, 1964.

Robert F. Stambaugh and Yu Yuan. Mispricing factors. *Review of Financial Studies*, 30(4): 1270–1315, 2017.

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

Michael Webb. The impact of artificial intelligence on the labor market. *SSRN Working Paper*, No. 3482150, 2020.