Distant Investments: Decoding Mutual Fund Skill through Fund-Firm Semantic Alignment

Xiyuan Ma[*] Matthew Spiegel[†] Hong Zhang[‡] Yijun Zhou[§]

Abstract

Traditional measures of mutual fund skill offer little insight into stock selection mechanisms. To fill this gap, we employ Large Language Models to measure the semantic distance between fund prospectuses and firms' 10-K strategic priorities, capturing difficult-to-understand information whose interpretation requires the specialized expertise revealed in each fund's own prospectus. High-activeness funds outperform when they invest heavily in such "distant" stocks. Moreover, the trading activities of distant-investing funds predict future stock returns and enhance market efficiency. Our findings reveal a novel mechanism behind managerial skill through a methodological refocus of textual analysis—from "what is written" to "who is reading."

JEL Codes: G14; G23

Keywords: Mutual Funds, Skill, Large Language Models (LLMs)

[*] Singapore Management University, 50 Stamford Road, Singapore 178899, xiyuanma@smu.edu.sg
[†] School of Management, Yale University, P.O. Box 208200 New Haven, CT 06520-8200, Email: matthew.spiegel@yale.edu
[‡] Singapore Management University, 50 Stamford Road, Singapore 178899, hongzhang@smu.edu.sg
[§] Florida State University, 821 Academic Way, Tallahassee, FL 32306, yzhou@business.fsu.edu

1. Introduction

Institutional investors, such as mutual funds, play a pivotal role in modern financial markets. In theory, skillful fund managers' adept processing of firm-level information underpins both informational and allocational efficiency in the securities market (Gârleanu and Pedersen 2018; Gervais and Strobl 2023). Yet existing measures of skill largely focus on realized fund performance (e.g., Carhart 1997) or portfolio activeness (e.g., Kacperczyk, Sialm, and Zheng 2005, 2008; Cremers and Petajisto 2009; Doshi, Elkamhi, and Simutin 2015), while offering little insight into the rationale behind these fund activities: *how* do skilled managers actually select promising firms? Are active funds inherently skillful, or does skill manifest through more nuanced channels—such as the processing of firm-level information through the lens of their own expertise? These questions are critical to deciphering managerial skill and its connection to market efficiency.

This paper introduces a novel framework to explore these questions by leveraging Large Language Models (LLMs) to examine the semantic alignment between mutual fund prospectuses and the strategic priorities outlined in firms' 10-K filings. While 10-Ks contain rich, evolving narratives regarding valuable information such as new investment opportunities, such information is often unstructured and linguistically complex (e.g., Basu et al., 2022).[1] Consequently, stock prices adjust sluggishly to these qualitative disclosures (Cohen et al., 2020), creating an opportunity for skilled fund managers who can process such difficult-to-understand information (DUI) more effectively than the market.

To identify this managerial skill, we propose a new measure: the semantic distance between mutual fund prospectuses—which articulate investment strategies (e.g., Kostovetsky and Warner 2020; Abis and Lines 2024)—and firms' strategic priorities as described in Item 1 of their 10-K filings.[2] A fund's prospectus serves as a semantic anchor that reflects its domain of expertise, allowing semantic distance to capture the difficulty of applying this expertise to evaluate new investment opportunities disclosed by firms. We hypothesize that investing in stocks with a large semantic gap between these two texts—what we term "distant investment"—reflects managerial skill in overcoming this difficulty and extracting valuable signals from firm disclosures.

[1] Other disclosed information in 10-K filings includes financial constraints (Hoberg and Maksimovic, 2015; Buehlmaier and Whited, 2018), earnings management practices (Frankel, Jennings, and Lee, 2016), and market competition dynamics (Li, Lundholm, and Minnis, 2013; Hoberg and Phillips, 2016).

[2] As detailed in later sections, the distance is calculated as Euclidean distance between the embeddings of the firm's strategic priorities (10-K Item 1) and the strategy description extracted from the mutual fund prospectus, calculated using SBERT.

To build intuition, consider the Putnam VT Sustainable Future Fund. Its prospectus indicates a focus on "impact companies" whose products and services "contribute to sustainable social, environmental and economic development." Now consider two potential investment targets.

- *Xylem Inc.* describes itself in its 2020 10-K (Item 1) as "a leading global water technology company" addressing "customer needs across the water cycle, from the delivery, measurement and use of drinking water to the collection, testing, analysis and treatment of wastewater to the return of water to the environment." This description is easy to understand and aligns with sustainability goals, yielding a low semantic distance to the fund (0.6006, near the 10th percentile of the distribution).
- *Seattle Genetics, Inc.* (now *Seagen*) presents a more challenging case in its filings. It highlights commercializing ADCETRIS® for certain CD30-expressing lymphomas and PADCEV™ (enfortumab vedotin-ejfv) for metastatic urothelial cancers—technical oncology advancements. While cancer treatment supports sustainable social development, the specialized nature of antibody-drug conjugates makes evaluation demanding, resulting in high semantic distance even after industry adjustment (0.9597, near the 99th percentile; industry-adjusted, still near the 90th percentile).

The Putnam VT fund invested in both. Notably, the fund initiated a position in Seattle Genetics after PADCEV™ first appeared in its 10-K in 2019—suggesting active processing of this emerging information. This investment proved highly profitable: Pfizer acquired Seagen in 2023 for approximately $43 billion, nearly double its valuation in April 2020.

This example illuminates the core insight of our argument: valuable information is revealed not only by *what is written* in firm disclosures, but also by *who is reading* them. Classical theories suggest that the value of information depends on the reader's expertise (e.g., Van Nieuwerburgh and Veldkamp 2009; Farboodi et al. 2025). Our semantic distance measure operationalizes this insight by quantifying the "interpretative fit" between a fund and a firm.

Investing in a "distant" firm like Seagen signals substantial managerial effort and skill, because the firm's new investment opportunities may appear semantically distant and thus difficult to evaluate based on the fund's domain of expertise. However, as detailed shortly, a skilled manager may still be able to recognize its economic value by *extending* her existing expertise to analyzing new and unfamiliar subjects. Moreover, such new opportunities are likely difficult to understand for the general public as well, creating opportunities for skilled managers to capitalize on its DUI before the market fully digests it. In contrast, semantically "close" firms—like Xylem—pose low interpretative demands to both managers and the public. Their lack of DUI implies more efficient market prices and thus diminished opportunities for skilled managers.

The above discussion suggests a novel mechanism of managerial skill: the ability to apply domain-specific expertise—as revealed by the fund's prospectus—to learn from semantically distant but processable firm signals. This ability is essential for extracting valuable information from newly emerging investment opportunities, which often appear "distant" or "unfamiliar" relative to the manager's existing expertise. Cognitive science has long demonstrated that people can learn about unfamiliar subjects by identifying a deep *relational structure mapping* from existing knowledge (Gentner 1983; Gentner and Smith 2012). By using existing expertise as a cognitive anchor, this mapping enables the *analogical transfer* of knowledge for new problem solving (Gick and Holyoak 1980) and the *absorptive capacity* to recognize and commercialize the value of new information (Cohen and Levinthal 1990). In our setting, a fund's prospectus reflects its domain expertise, which serves as a cognitive anchor for processing firm disclosures—mapping new information onto the deep structure of existing knowledge to extract valuable signals.

In the Online Appendix, we discuss these classical cognitive foundations in more detail and formulate this mechanism as a simple extension of Kyle (1985) model with distance eroding signal extraction and heterogeneous structural mapping abilities. Skilled investors benefit most from stocks at a processable distance, neither too close nor too distant. While Seagen and Xylem illustrate cases of processable and close distance, the "too distant" case can also arise when an asset lies outside the fund's domain of expertise, even after extension—such as value stocks for a growth manager. Even when value stocks contain DUI from a growth fund's perspective, its manager may lack the ability to process that information effectively. Consequently, a skilled growth manager may venture beyond well-recognized growth stocks (i.e., "close" firms) to invest in companies with uncapitalized growth potential—firms with processable or "good" distance—but not in value stocks, which represent non-processable or "bad" distance.

This mechanism has direct implications for measuring fund skill. Because it takes time for the market to fully digest DUI, the value of distant firms is not properly reflected in their market capitalization and thus in most benchmarks used by the mutual fund industry. Consequently, a skill-based distant investment naturally leads the fund to actively deviate from its benchmark—a deviation captured by traditional measures of "activeness." However, neither activeness nor distance is a sufficient proxy for skill on its own. Indeed, some deviations (e.g., tracking errors due to gambling preferences) or distance (e.g., passive investments in stocks unrelated to the fund's mandate) could reflect governance issues rather than trading skill. Instead, managerial skill should

be revealed at the intersection of the two: it requires both a willingness to deviate from a benchmark (activeness) and a valid economic rationale for doing so. Distant investment provides precisely such a rationale. This framework yields a list of testable predictions as follows.

Fund Performance: If distant investment represents a concrete mechanism to complement fund activeness in revealing managerial skill, then the interaction between traditional active measures and a fund's propensity for distant investment should predict future fund performance. Active funds should outperform when they also engage in distant investment.

Information Discovery: If distant-investing managers are indeed skilled at processing DUI, their collective trades should predict future stock returns, as such trading helps incorporate their processed information into stock prices (e.g., Kyle 1985 and our extension).

Market Efficiency: An increase in the presence of skilled "distant" managers should lead to a more informationally efficient market, reducing the time it takes for qualitative 10-K disclosures to be reflected in prices.

Conversely, if "bad distance" dominates, then distant investments should be associated with poor future performance and reduced price informativeness. The null hypothesis is that distant investments could be pure noise and thus unrelated to performance.

We test these hypotheses using a comprehensive sample of active U.S. equity mutual funds from 2011 to 2023. To measure traditional activeness, we employ two well-established and complementary metrics: active weight (Doshi, Elkamhi, and Simutin 2015), which captures end-of-quarter holding deviations from value-weighted benchmarks, and return gap (Kacperczyk, Sialm, and Zheng 2008), which infers intra-quarter trading activity. Our primary proxy for conventional skill, *fund activeness*, is the average rank of these two measures.

To capture distant investment, we calculate the semantic distance for each fund-firm pair. For each fund, we then define its *holding distance* as the average of industry-adjusted distances to the stocks it holds, weighted by the fund's active weights in these stocks.[3] Industry adjustment isolates the impact of firm-specific DUI (rather than sector-level commonalities), while weighting by active weight captures the component of semantic distance associated with active investments.

[3] More specifically, for any firm invested by the fund, we first adjust pairwise fund-firm distance by the average distance between the fund and all firms in the same industry. Next, we value-weight this adjusted distance by the difference between the actual and market-cap-based weights—the latter provides a counterfactual benchmark of the fund following Doshi, Elkamhi, and Simutin (2015). These two steps reduce the impact of industry and passive investments, which could possibly allow a fund to exhibit a high holding distance without really processing firm-level information.

Finally, we interact fund activeness with holding distance to construct our main proxy for managerial skill, which we refer to as *Skilled Distant Investment* (*SDI*).

Before turning to our main tests, we conduct a diagnostic analysis to understand the economic content of our semantic distance measure. We compute the distance between all funds and all S&P 500 stocks—regardless of whether funds actually invest in these stocks—and explore the potential drivers. These include changes in fund prospectuses, shifts in firms' 10-K strategic priorities, and the emergence of new investment opportunities—proxied by newly added keywords related to AI or ESG topics, or any of the 43 machine-learning-derived categories of investment opportunities identified by Basu, Ma, and Briscoe-Tran (2022). Individually, each driver explains distance variation. Jointly, emerging opportunities dominate 10-K shifts, confirming that our measure captures the "frontier" of firm-level developments that likely contains DUI.

Building on this diagnostic evidence, we next more formally examine the three predictions of distant investments. We begin with the performance implications. Using independent portfolio sorts, we find that conventional skill—fund activeness—predicts outperformance only among funds with high holding distance. Within the top distance quintile, the most active funds outperform the most inactive by a monthly Fama-French-Carhart's (1997) four-factor-adjusted before-fee return of 0.48% (t = 3.15). In the bottom distance quintile, the same spread becomes negligible (0.14%). The difference between the two spreads, 0.34% (t = 2.06), highlights the economic impact of distant investments. After-fee performance yields similar patterns.

Multivariate Fama-MacBeth (1973) quarterly regressions confirm the performance impact while controlling for fund characteristics. *Fund activeness* alone predicts four-factor alphas (before and after fees) and value-added (Berk and van Binsbergen 2015), albeit with weaker power for value-added. *SDI* significantly augments predictability across all metrics. Economically, a one-standard-deviation increase in *fund activeness* is associated with a $2.17 million higher quarterly value-added. *SDI* further enhances this by $1.84 million, an 84.8% relative increase.

Subsample tests provide further insights. The predictive power of *SDI* concentrates among funds with above-median exposure to emerging AI and machine-learning-identified investment opportunities, underscoring the role of managerial expertise in processing related DUI. This predictivity is also stronger for larger funds and those with lower turnover, suggesting that distant investment alleviates diseconomies of scale without evoking excessive trading. Finally, the effect

is more pronounced among funds charging higher expenses, consistent with the notion that managers capture economic rents from their skill (Berk and Green 2004).

Thus far, our fund-level results affirm that distant investment gives rise to performance gains. But if the fund-level results reflect genuine information processing, then the trades of distant-investing managers should predict future stock returns. Therefore, we next examine stock-level return predictability. To this end, we adapt mutual fund order imbalance (e.g., Da, Gao, and Jagannathan 2011; Jones et al. 2025) to distant investments. Specifically, we aggregate each fund's order imbalance on a given stock, weighting by the fund's skilled distant investment (*SDI*) measure. This weighted measure, which we term as *skilled distant investment order imbalance* (*SDI_OI*), captures the net buys of distant-investing funds. We then use it to forecast out-of-sample Daniel, Grinblatt, Titman, and Wermers (1997; hereafter DGTW) characteristic-adjusted stock returns.

We find that *SDI_OI* strongly predicts out-of-sample DGTW-adjusted returns. In our baseline specification, a one-standard-deviation increase in *SDI_OI* is associated with a 0.313% higher out-of-sample quarterly return (t=2.71). The economic magnitude further increases to 0.328% when focusing on above-median *SDI_OI.* The return predictability of SDI_OI also remains highly robust when controlling for the linguistic complexity of 10-K filings (Loughran and McDonald 2024) and the geographic proximity between funds and stocks, suggesting that distant investment captures independent sources of fund skill. To compare this predictive power to that related to fund activeness alone, we also construct fund activeness--weighted order imbalance (*AWOI*). Controlling for *AWOI* does not change the return predictive power of *SDI_OI*.

Our final prediction concerns market-level efficiency. If distant-investing managers accelerate the incorporation of DUI into prices, their trading should attenuate the inefficiency of "lazy prices" (Cohen, Malloy, and Nguyen 2020, hereafter CMN), where firms with material changes in their 10-Ks ("Changers") experience persistent price declines up to six-month post-release. We observe that this inefficiency is sharply attenuated for stocks heavily traded by distant-investing managers during the filing quarter. Indeed, the interaction between Changers and high distant-manager trading carries a coefficient opposite in sign and similar in magnitude to the baseline drift, implying that these managers' trades effectively incorporate the new information into prices.

Complementing this price-based test, we also follow Brogaard et al. (2021) to decompose return variance into four components based on their economic origins: firm-specific private information, firm-specific public information, market-wide information, and noise. Linking these

components to distant-investment trading intensity (which sums up the magnitude of both buy and sell orders), we find the latter elevates private-information variance while curbing noise—consistent with skilled managers revealing firm-specific DUI to enhance price efficiency.[4]

Having established fund-level and stock-level evidence on distant investment, we consider a list of additional analyses to deepen our understanding of its economic insights. First, our distance measure is conceptually and empirically distinct from traditional text-based proxies. Standalone metrics—whether drawn from 10-K filings (linguistic complexity, Loughran and McDonald 2024; textual changes, Cohen et al. 2020; or the emergence of AI, ESG, and ML-identified investment opportunities) or from fund prospectuses (textual changes)—fail to replicate our results.

This divergence arises because information value is often investor-specific (e.g., Van Nieuwerburgh and Veldkamp 2009; Farboodi et al. 2025): fund managers may optimize performance by developing expertise in specialized information domains where they hold an initial comparative advantage. In our framework, the fund prospectus serves as a revealed benchmark of this specialization. By conditioning the firm's 10-K strategic priorities on the fund's stated expertise (i.e., its prospectus), our measure moves beyond *unconditional* textual analysis to capture the *relative interpretability* of firm disclosure—and thus DUI—from the managers' perspective. This methodological refocus on relational alignment, rather than document-centric properties, explains the superior predictive power of our measure relative to traditional textual metrics.

Next, we report a striking contrast: some low-skill fund managers also appear to adopt distant investment—yet with negative performance. What explains their incentives? We find that these managers are flow-motivated. They buy distant stocks precisely when such stocks attract retail attention (as proxied by high Google search volume). Such investments help buffer outflows but may harm future performance[5]—a form of window dressing that benefits managers but harms investors. By contrast, high-skill managers are immune to this motive.

Furthermore, we show that our results remain highly robust when using alternative measures of fund activeness or alternative natural language processing (NLP) methods. Specifically, using active weight or return gap alone, or incorporating additional activeness proxies like industry concentration (Kacperczyk, Sialm, and Zheng 2005), active share (Cremers and Petajisto 2009),

[4] It is worth noting that the private information component does not imply insider information. As we will discuss in later sections, skilled investors can process information better than the market by converting a firm's noisy public signals (e.g., 10-K in our study) into more accurate information (e.g., Kim and Verrecchia 1994; Kandel and Pearson 1995).

[5] Since a high sentiment likely indicates that the market has digested a firm's information or even overpriced its value, such sentiment-driven distant investments is likely to give rise to underperformance.

or deviations from a factor model (Amihud and Goyenko 2013) do not change our results. To alleviate the black-box nature of LLM, we also verify that our main conclusions remain robust—albeit with a smaller economic magnitude—based on the more traditional but interpretable bag-of-words (BoW) approach. Moreover, while our main LLM analysis focuses on the first chunk of Item 1, averaging across all chunks produces consistent, albeit weaker, results.[6]

Last but not least, although other parts of the 10-Ks, such as risk factors (Item 1A), may provide additional information (Bai et al. 2024), we observe insignificant results when we calculate the semantic distance using the risk narrative section. This placebo test confirms that our findings are specific to strategic, forward-looking information rather than generic risk disclosures.

Collectively, our results support the notion that distant-investing fund managers can deliver superior performance by processing firm-specific DUI. This enables them to predict stock returns on the one hand while enhancing market efficiency on the other. These findings underscore the economic origins and significance of distant investments in the modern economy. Since our measure captures the *relative interpretability* of firm-specific information and thus DUI, our findings also provide a new economic interpretation of managerial skill: it is not merely the replication of expertise within familiar ("close") firms, but the ability to apply that expertise to exploit DUI and novel opportunities in processable firms—i.e., firms with "good distance."

Our results speak to several strands of literature. First, we build on and extend studies of mutual funds that seek to identify and quantify managerial skill. Numerous works infer skill from observed fund information, such as returns and holdings, by measuring the activeness of portfolio management (e.g., among others, Kacperczyk, Sialm, and Zheng 2005, 2008; Cremers and Petajisto 2009; Amihud and Goyenko 2013; Doshi, Elkamhi, and Simutin 2015; Jones and Mo 2021 provide a recent empirical analysis of existing measures). The underlying intuition is that skilled managers must deviate from benchmarks—such as market indices, value-weighted industry allocations, or factor models—to outperform them. We fill the remaining missing link between a manager's stated investment mandate and their actual portfolio choices, offering a new economic rationale—grounded in classical cognitive science—for how skill manifests as activeness and, ultimately, alpha.

[6] Our main text embedding analysis adopts Sentence-BERT following the literature (e.g., Acemoglu, Mühlbach, and Scott 2022; Liu et al. 2019; Reimers and Gurevych 2019). SBERT has a limited capacity to process information. We focus on the first chunk of Item 1 since it describes the most important information of the firm, such as its overall strategy, operational priorities, and most valuable investment opportunities.

Our analysis also contributes to research on the social value of mutual funds in promoting market efficiency. Early studies often infer market efficiency from fund performance.[7] However, Berk and Green (2004) demonstrate that fund performance reflects investor competition for scarce managerial skill rather than asset price efficiency. Gârleanu and Pedersen (2018) indicate that a more efficient mutual fund industry can enhance asset price informativeness, while Gervais and Strobl (2023) show how mutual funds improve resource allocation through feedback effects, benefiting investors even if funds underperform benchmarks. We differ by offering direct evidence and a concrete, fund-specific mechanism for how skilled managers process firm-level 10-K information that the market struggles to digest (e.g., Cohen, Malloy, and Nguyen 2020), which boosts market efficiency. However, we also find evidence that unskilled managers may exploit retail investors, indicating persistent frictions in the mutual fund industry.

Finally, our work extends recent applications of machine learning, particularly LLMs, in asset management and financial markets. Studies such as Li and Rossi (2020), DeMiguel et al. (2023), and Kaniel et al. (2023) employ machine learning to predict mutual fund performance. Cong et al. (2021) develop deep reinforcement learning models for portfolio management, and Sheng et al. (2024) document hedge funds' growing reliance on generative AI. Recent work also explores the textual content of mutual fund disclosures (e.g., Kostovetsky and Warner 2020; Abis and Lines 2024; Cao, Yang, and Zhang 2024 a,b; Gao, Xiong, and Yuan 2024).[8]

In particular, a nascent literature employing textual similarity to quantify economic linkages and exposures. Acemoglu, Mühlbach, and Scott (2022) and Seegmiller, Papanikolaou, and Schmidt (2023) pioneer this approach, using semantic overlap between job tasks and "age-friendly" dictionaries or patent texts to measure occupational exposure to aging and innovation, respectively. Recent studies also use textual similarity to identify societal sentiment in historical books (Jha, Liu, and Manela 2025) and measure the innovation-driven displacement risk by comparing rival patents to the focal firm's 10k (Kakhbod et al. 2025). A common property of these studies is their *document-centric focus*: they treat textual similarity as an unconditional metric that applies uniformly to all observers. Indeed, this document-centric perspective is at the core of traditional

[7] See, among others, Fama 1970; Malkiel 1995; Gruber 1996; Carhart 1997; Zheng 1999; Wermers 2000; Bollen and Busse 2001; Christoffersen and Musto 2002; Gil-Bazo and Ruiz-Verdú 2009; and Fama and French 2010.

[8] Other textual analysis contents range from news and social media (Bybee 2023; Lopez-Lira and Tang 2025; Chen, Peng, and Zhou 2024), firms' 10-Ks (Choi and Kim 2024; Kim and Nikolaev 2025; Kim, Muhn, and Nikolaev 2025; Breitung and Müller 2025), conference calls (Bai et al. 2023; Jha et al. 2024, 2025; Kim, Muhn, and Nikolaev 2024, 2025; Siano 2025), patent descriptions (Zheng 2025), to published books (Jha, Liu, and Manela 2025).

textual analysis, which aims to provide linguistic metrics—such as complexity, sentiment, or similarity—that implicitly assume a homogeneous interpretation of document properties across all readers.

The novelty of our distance measure lies in its *agent-specific* nature to capture the relational interpretation of text from the agent's perspective. It is well documented theoretically that investors interpret the same news differently (Rubinstein 1993; Kim and Verrecchia 1994; Kandel and Pearson 1995), rendering the value of information inherently investor-specific (e.g., Van Nieuwerburgh and Veldkamp 2009; Farboodi et al. 2025). Our measure operationalizes this classical insight. By introducing fund prospectuses as a semantic anchor—the revealed benchmark of managers' specialized domains—our distance measure moves beyond unconditional signals to capture investor-specific information value: the relative interpretability of firm disclosures from each manager's unique perspective.

In this regard, we are most closely related to Ash, Chen, and Naidu (2026), who measure the changes in judicial ideology based on shifts in similarity between judges' writing and a lexicon of law and economics terminology. However, their evaluation applies a uniform and unconditional benchmark—i.e., the universal concept of "Law and Economics"—to all judges. We differ by focusing on a two-way relational measure that reveals the conditional, investor-specific value of firm disclosures. By using a fund's own prospectus as the reference point, our approach shifts the focus from "what is written" to "who is reading," offering a heuristic for future research in settings where investor heterogeneity and domain expertise play a central role.

2. Data and Main Variables

2.1 Main Data

In this section, we first describe the main sources of data used in our analysis.

Mutual Fund Prospectuses: A mutual fund prospectus is a document that provides essential information about a fund, including details on its principal investment strategy, primary risks, past performance, management, and other key aspects relevant to investors. We collect all mutual fund prospectuses from the SEC's Mutual Fund Prospectus Risk/Return Summary Data Sets. Although the SEC began requiring mutual fund prospectuses in 2009, the number of funds publishing

prospectuses was extremely small before 2011. Therefore, our data sample spans from the first quarter of 2011 to the end of 2023.

Other Mutual Fund Data: From the CRSP Survivorship Bias-Free Mutual Fund Database, we obtain fund-level information such as historical performance, total net assets (TNA), and expense fee data, among others. Since most mutual funds offer multiple share classes that mainly differ in fee structures and target investors, we consolidate them into a single fund. We calculate each fund's TNA by summing the TNAs of its share classes and determine the fund age based on the oldest share class. For other fund characteristics, we calculate their TNA-weighted averages across the share classes.

To obtain information on fund holdings, we link the CRSP database to Thomson Financial Mutual Fund Holdings via MFLINKS files from WRDS. Additionally, to match the corresponding mutual fund prospectuses, we match another fund-level SEC-assigned identifier—Central Index Key (CIK)—using CIK-MAP files from WRDS.

The Main Mutual Fund Sample: Our main testing sample includes actively managed U.S. equity mutual funds. Following the literature (e.g., Kacperczyk, Sialm, and Zheng 2008; Doshi, Elkamhi, and Simutin 2015), we first select domestic equity mutual funds classified under the following Lipper classification codes: EIEI, LCCE, LCGE, LCVE, MCCE, MCGE, MCVE, MLCE, MLGE, MLVE, SCCE, SCGE, and SCVE. If Lipper classification codes are unavailable, we include funds with Strategic Insight target codes of AGG, GMC, GRI, GRO, ING, and SCG. If these are also missing, we consider the funds with Wiesenberger target codes of G, G-I, GCI, IEQ, LTG, MCG, and SCG. We require that at least 50% of the fund's assets be invested in common equities. We then follow the methodologies of Ben-David, Rossi, and Song (2022) and Dannhauser and Pontiff (2019) to identify and exclude passive funds. A fund is classified as passive if it is flagged as passive by CRSP or its name contains a relevant keyword.[9] Additionally, we exclude international funds, balanced funds, bond/preferred funds, target date funds, and leveraged/inverse funds.[10]

[9] A fund is classified as passive if the CRSP index_fund_flag or et_flag is not missing, or if the CRSP fund name contains the following strings: SP, DOW, Dow, DJ, ETF, ETN or if the lowercase version of the CRSP fund name contains: index, indx, idx, composite, nyse, nasdaq, s&p, s and p, s & p, ishares, exchange traded, exchange-traded, 50, 100, 200, 400, 500, 600, 1000, 1500, 2000, 2500, 3000. These numbers are selected based on major U.S. stock indices. We manually check some funds whose names include 'Morningstar', 'Wilshire', 'Bloomberg', 'FTSE', etc., and find that almost all can be absorbed by existing filters.

[10] Target date funds are flagged if the lowercase version of the CRSP fund name contains target, retirement, 2010, 2015, 2020, 2025, 2030, 2035, 2040, 2045, 2050, 2055, 2060, 2065. These numbers are selected based on S&P target date indices. Inverse and leveraged funds are identified if the lowercase version of their name contains the following strings: inverse, ultra, 1.5x, 2x, 2.5x.

To mitigate the impact of small funds, we exclude funds that have fewer than 5 invested stocks or below-$15 million TNAs in our main analysis. In addition, to mitigate the incubation bias (Evans, 2010), we also remove the first 2 years of fund history.

Firm's 10-K Filings: The 10-K filings are regulatory documents that provide a comprehensive overview of a firm's business, strategies, financial performance, and operations. These filings are available on the SEC's EDGAR system. As mentioned earlier, the SEC uses the identifier CIK to mark both funds and stocks. We link individual stocks to CIK using GVKEY through the linking tables provided by the SEC and WRDS. In cases where multiple versions are filed within the same month due to revisions, we retain only the latest filing for each CIK each month.

Other Stock-level Information: For stock returns and accounting data, we use the CRSP and Compustat/NA databases. We extract monthly stock data from CRSP for all common stocks with codes 10 or 11 listed on the NYSE, AMEX, and NASDAQ. Industry classifications follow the FFI-48 definitions from Kenneth French's data library. Stocks with prices below $5 and those in the financial sector are excluded from stock-level analyses. Stock characteristics are obtained from Chen and Zimmermann (2020), the IBES database, and other public sources. Missing values are filled using the stock's own lagged values.

2.2 Main Variables Construction

We next describe the construction of our main variables.

Pairwise Fund-firm Distance: We use large language models (LLMs) to compute the pairwise distance between mutual fund prospectuses and firms' 10-K filings. LLMs convert a text document into embeddings, which are high-dimensional numerical vectors that encode the semantic and syntactic information of the text, such as meaning of words, phrases, and the structure features of the text.[11] Words with similar meanings tend to have similar embeddings, meaning they are located closer together in the vector space. In addition to semantics, these vectors also encode syntactic roles and contextual relationships, reflecting how words are used in sentences and within surrounding contexts.

[11] Embeddings lie at the core of LLMs and have been leveraged by researchers for tasks such as text clustering, semantic search, sentiment analysis, similarity estimation, classifications, or as features in machine learning models. For LLMs built with the transformer architecture (such as BERT or models developed OpenAI), text is processed through multiple layers, each capturing different linguistic properties and dependencies. The common practice is to extract embeddings from the final layer of the models, as these vectors provide the most refined and contextually informed presentation of the text.

Following Acemoglu, Mühlbach, and Scott (2022), we apply a specific LLM called Sentence-BERT (Liu et al. 2019; Reimers and Gurevych 2019) to compute embeddings for each text document.[12] This algorithm generates a 768-dimensional embedding vector that encodes the semantic and syntactic information of the input text. Based on these embeddings, we compute the pairwise fund-firm distance as the Euclidean distance between the two vectors:

$$Distance_{f,s} = \| \boldsymbol{V}_f - \boldsymbol{V}_s \| = \sqrt{\sum_{i=1}^{n} \left(V_f^i - V_s^i\right)^2}, \qquad (1)$$

where $Distance_{f,s}$ is the pairwise semantic distance between fund f and firm s, $\boldsymbol{V}_f$ and $\boldsymbol{V}_s$ refer to the embedding vectors of the prospectus for mutual fund f and the 10-K filing (Item 1) for firm s, respectively, and V_f^i and V_s^i refer to the i^{th} element of the corresponding embedding vectors. The Euclidean distance captures the semantic (dis)alignment between the mutual fund's stated mandate and the firm's strategy—using Cosine distance will yield to similar results.

Fund prospectuses are typically updated on an annual basis. For each prospectus filing, we pair the fund's disclosure with the most recent Form 10-K filed by each firm as of that specific date. This allows us to compute a pairwise semantic distance for every fund-firm combination, resulting in an annual panel of distance measures.[13]

To capture the potential skill of a fund to process firm-specific, difficult-to-understand information (DUI), we further adjust, for each firm in a fund's portfolio, the pairwise distance by subtracting the average distance between the fund and all firms in the same industry.

As an illustration, Figure 1 plots the industry-adjusted pairwise distances between the prospectus of the Putnam VT Sustainable Future Fund and firms' 10-K filings in 2019. The left panel shows the fund's actual portfolio holdings, including Xylem Inc. and Seattle Genetics. Notably, Seattle Genetics continues to stand out as a "distant investment" even after adjusting for industry differences. The right panel includes all S&P 500 firms that the fund could potentially invest in. Comparing the two panels reveals that the Putnam fund is highly selective in its realized

[12] The Sentence-BERT model (SBERT) is built on the BERT model by modifying the original BERT architecture to produce semantically meaningful sentence embeddings that are optimized for semantic similarity tasks. It uses siamese and triplet network architectures to fine-tune BERT, enabling efficient computation of similarity metrics like cosine similarity. This fine-tuning ensures that embeddings from Sentence-BERT are consistent for semantic similarity computations. Specifically, sentences with similar meanings are placed closer together in the embedding space.

[13] For example, the prospectus released by the Putnam VT Fund in 2019 April is paired with the 10-k filing released by *Seattle Genetics* in 2019 February to calculate their pairwise distance.

holdings. Taken together, these patterns suggest that the fund may have actively processed information to select a subset of firms for investment.

Figure 2 provides a more general illustration by plotting the industry-adjusted distances between 100 randomly selected equity mutual funds and S&P 500 firms. No fund is particularly closer to, or clustered around, any single industry, confirming that this measure captures within-industry variation in fund–firm distance rather than an industry-level effect.

Holding Distance: We next compute the holding distance for each fund in a given quarter, which captures the extent to which a fund manager tilts their portfolio toward semantically distant firms. To this end, we aggregate the pairwise fund-firm distances within the holding portfolio of the fund. We further value-weight each fund-firm distance by the stock's active weight, which is the deviation of the fund's actual portfolio weight from a passive market-cap-weighted benchmark, following Doshi, Elkamhi, and Simutin (2015). As these authors show, active weights reflect managers' stock-picking ability. Hence, this weighting scheme ensures that our measure captures the component of semantic distance associated with skill-based active stock selections.

Applying these adjustments, we compute fund f's *holding distance* in quarter t as follows:

$$Holding\ Distance_{f,t} = \sum_{s} (w_{f,t}^{s} - w_{f,t}^{sm}) \times (Distance_{f,s,t} - \overline{Distance}_{f,s\in j,t}), \quad (2A)$$

where $w_{f,t}^{s}$ is the actual portfolio weight of stock s in fund f's portfolio at the end of quarter t, and $w_{f,t}^{sm}$ is the weight that this stock would have received had the manager market cap-weighted the equity portfolio. $Distance_{f,s,t}$ is the most recent pairwise fund-firm distance between fund f and firm s available prior to the end of quarter t (we require the fund's most recent prospectus to be filed within two years prior to quarter-end), and $\overline{Distance}_{f,s\in j,t}$ is the average distance between fund f and all firms (including those not held by the fund) in industry j.

Fund Activeness: To proxy for traditional skill, we employ two well-established measures of fund activeness: active weight (Doshi, Elkamhi, and Simutin 2015) and return gap (Kacperczyk, Sialm, and Zheng 2008). We focus on these metrics because they draw on complementary sources of information. Active weight captures end-of-quarter holding deviations from a passive benchmark, while return gap—the difference between a fund's reported return and the return implied by its disclosed holdings—reflects within-quarter interim trading strategies.

To combine these complementary signals, we compute fund activeness as the average of their cross-sectional ranks:

$$Fund\ Activeness_{f,t} = Avg\left(Rank_{f,t}^{RG}, Rank_{f,t}^{AW}\right), \qquad (2B)$$

where $Rank_{f,t}^{RG}$ and $Rank_{f,t}^{AW}$ refer to the cross-sectional ranks of fund *f* in *return gap* and *active weight* in quarter t. Both ranks are normalized to follow a uniform distribution between 0 and 1. Consequently, the composite *fund activeness* measure ranges from 0 to 1, with a median close to 0.5. Later sections will show that incorporating more activeness measures does not change our main conclusions.

Skilled Distant Investment (SDI): Our main proxy for managerial skill is the interaction between fund activeness and holding distance. This interaction captures not only the willingness to deviate from a benchmark (activeness) but also the economic rationale for doing so—namely, the ability to process DUI for distant stocks and invest accordingly. Hence, we construct *Skilled Distant Investment* (*SDI*) as follows:

$$SDI_{f,t} = Fund\ Activeness_{f,t} \times Holding\ Distance_{f,t}. \qquad (3)$$

Order Imbalance: To test whether the trades of distant-investing funds predict stock returns, we construct two stock-level order imbalance measures that capture aggregate mutual fund trading, building on Da, Gao, and Jagannathan (2011) and Jones et al. (2024). We first infer each fund's buy and sell orders for a given stock from quarterly changes in holdings. Specifically, the buy order, $Buy_{f,s,t}$, is defined as the percentage increase in the invested value of fund f in stock s in quarter t (zero if no purchase). Similarly, the sell order, $Sell_{f,s,t}$, is defined as the percentage decrease in the invested value of fund f in stock s (zero if no sales).[14]

We then construct our main measure, skilled distant investment order imbalance (SDI_*OI)*, which captures net buying pressure from funds engaging in skilled distant investments. *SDI_OI* weights each fund's net trade by its SDI measure, and is computed as:

[14] We first calculate the percentage change in the invested value of fund f in stock s in quarter t as $\Delta_{f,s,t} = \frac{Holding\ value_{f,s,t} - Holding\ value_{f,s,t-1} \times R_{s,t}}{Holding\ value_{f,s,t-1}}$, where $Holding\ value_{f,s,t}$ is the dollar value of holding of fund f in stock s by the end of quarter t, and $R_{s,t}$ is the realized stock price appreciation in the same quarter. Using price appreciation rather than total return assumes funds distribute dividends to investors, which is standard practice—our results are robust to using total return instead, indicating that the option of reinvesting dividends by fund investors does not affect our main conclusions. The buy and sell orders are then constructed as $Buy_{f,s,t} = max\{0, \Delta_{f,s,t}\}$ and $Sell_{f,s,t} = -min\{0, \Delta_{f,s,t}\}$.

$$SDI\_OI_{s,t} = \frac{\sum_f SDI_{f,t} \times (Buy_{f,s,t} - Sell_{f,s,t})}{\sum_f (Buy_{f,s,t} + Sell_{f,s,t})}. \quad (4A)$$

Following the definition of traditional order imbalance, *SDI_OI* is also scaled by the total value of buy and sell orders to ensure that its value is distributed between -1 and 1. As such, *SDI_OI* indicates the direction of mutual fund trades. A positive *SDI_OI* can be interpreted as a net buy by mutual funds originating from their skilled distant investments.

For comparison, we can also construct a similar proxy for mutual fund trading related to the traditional measure of fund activeness, activeness-weighted order imbalance (*AWOI*) as follows:

$$AWOI_{s,t} = \frac{\sum_f Fund\ Activeness_{f,t} \times (Buy_{f,s,t} - Sell_{f,s,t})}{\sum_f (Buy_{f,s,t} + Sell_{f,s,t})}. \quad (4B)$$

Positive *AWOI* values indicate net buying pressure from more active funds.

2.3 Summary Statistics

Our final sample includes about 3,000 unique actively managed U.S. equity mutual funds and 73,064 fund-quarter observations from 2011 to 2023.

Table 1 reports the summary statistics of mutual fund variables. All main variables have reasonable distribution. For instance, *Holding Distance* has a mean of 0.009 and a median of 0.007, indicating a fairly symmetric distribution. *Fund Activeness*, by contrast, shows greater variability ranging from 0.381 at the 25th percentile to 0.638 at the 75th percentile. Other fund characteristics, such as the total net assets (TNA), age, expense ratio, turnover, past flows, past returns, and return volatility of our sample, are comparable to the literature, though fund size and age are slightly higher than reported in the literature, as our sample period is more recent.

Mutual funds in our sample collectively hold over 5800 unique stocks in the CRSP/Compustat merged database. The average fund holds approximately 126 stocks in a given quarter. We provide the summary statistics for firm-level measures and characteristics in Online Appendix Table A1. These firm-level variables also exhibit reasonable distribution.

3. Distance and Fund Performance

This section provides empirical analysis that can shed light on the economic interpretation of fund-firm distance and its implications for fund performance.

3.1 Economic Interpretations of Pairwise Fund-firm Distance

Before examining performance, we conduct a diagnostic analysis to validate the economic content of our semantic distance measure. We explore whether the distance between a fund's prospectus and a firm's 10-K reflects meaningful shifts in investment mandates, corporate strategies, or the emergence of new investment opportunities that are likely to contain "difficult-to-understand" information (DUI).

To ensure a broad and representative interpretation, we calculate the pairwise semantic distance between all active mutual funds and the universe of S&P 500 firms, regardless of whether a fund actually holds the stock. This sample ensures that variation in distance reflects the genuine textual and informational dynamics rather than endogenous investment decisions. We then link it to three sets of economic drivers as follows.

Changes in Fund's Stated Expertise (Fund Change): Changes in a fund's prospectus may consequently change its semantic proximity to firms. We measure this economic driver using the embedding distance between the fund's current and previous prospectuses.

Changes in Firm's Disclosed Strategy (Firm Change): Shifts in firms' strategic priorities may affect their alignment with funds. We measure such shifts by the embedding distance between a firm's current and prior Item 1 section of its 10-K filing.

Emergence of New Investment Opportunities: To capture the appearance of DUI at the firm level, we construct three dummy variables based on the appearance of novel keywords in the current 10-K related to: (i) Artificial Intelligence (AI), (ii) Environmental, Social, and Governance (ESG) topics, and (iii) any of the 43 comprehensive, machine-learning-identified categories from Basu, Ma, and Briscoe-Tran (2022). [15] These dummy indicators proxy for the frontier of firm-level developments likely to contain DUI.

Specifically, we estimate the following panel specification:

$$Distance_{f,s,t} = \alpha + \beta_1 \times Distance_{f,s,t-1} + \beta_2 \times Fund\ Change_{f,t} + \beta_3 \times Firm\ Change_{s,t} + \beta_4 \times NewInvOpp_{s,t} + \boldsymbol{\Gamma_1}' \times \boldsymbol{M}_{f,t-1} + \boldsymbol{\Gamma_2}' \times \boldsymbol{M}_{s,t-1} + \varepsilon_{f,s,t}, \quad (5)$$

[15] AI and ESG-related keywords are generated by ChatGPT. The ESG-related keywords include: "sustainability", "green investments", "social responsibility", "corporate governance", "environmental impact", "ethical investing", "climate change", "renewable energy", "diversity and inclusion", "carbon footprint", "impact investing", "fair trade", "stakeholder engagement", "corporate social responsibility", "greenwashing", "social equity", "climate risk", "environmental stewardship", "sustainable finance", "gender equality", "responsible investing", "supply chain sustainability", "governance practices", "transparency", "shareholder activism", "human rights", "waste reduction", "green bonds", "net-zero emissions", "environmental disclosure". The AI-related keywords include: "artificial intelligence", "machine learning", "deep learning", "neural networks", "data science", "natural language processing", "computer vision", "reinforcement learning", "automation", "big data", "predictive analytics", "AI ethics", "AI governance", "speech recognition", "chatbot", "robotics", "cognitive computing", "AI algorithms", "self-driving cars", "facial recognition", "generative AI", "algorithmic bias", "AI model", "augmented intelligence", "AI-driven", "cloud computing", "edge computing", "AI in healthcare", "AI in finance", "AI in education", "AI-powered", "autonomous systems".

where $Distance_{f,s,t}$ refers to the pairwise semantic distance between the prospectus of fund f and the 10-K of firm s in quarter t, $Fund\ Change_{f,t}$ and $Firm\ Change_{s,t}$ refer to the textual changes in the fund prospectus and firm 10-K filings, and $NewInvOpp_{s,t}$ refers to the dummy indicators for the emergence of AI, ESG, and machine learning-identified new investment opportunities. The vector $\boldsymbol{M}_{f,t-1}$ stacks a set of lagged fund-level control variables, including total net assets, age, turnover, expense ratio, past flows, past returns, and return volatility. The vector $\boldsymbol{M}_{s,t-1}$ stacks a set of lagged firm-level control variables, including market capitalization, book-to-market ratio, past returns, asset growth, operating profitability, illiquidity, and analyst coverage.

The results are presented in Table 2. In Columns (1) through (4), we introduce economic drivers progressively, culminating in a joint specification in Column (5). All specifications include Fund, Stock, and Quarter fixed effects. To address potential endogeneity and unobserved heterogeneity at the pair level, Column (6) employs a more stringent Fund × Stock fixed effects to absorb any time-invariant pair-specific factors.

The findings provide clear evidence on the economic content of our distance measure. In Column (1), the positive and significant coefficient on the lagged dependent variable confirms the persistence of the fund-firm distance. More importantly, the coefficients on both fund changes and firm changes are significantly positive, indicating that updates to a fund's investment mandate or a firm's strategic narrative are both reflected in the semantic alignment between them.

Columns (2) through (4) further show that the emergence of new investment opportunities—AI, ESG, and the broader ML-identified topics—is each associated with a significant *increase* in distance. This indicates that when firms begin to discuss novel and complex topics, they become semantically more distant from the average fund's stated expertise.

The joint model in Column (6) offers the most insightful result. With the inclusion of all drivers and the most restrictive fixed effects, the coefficients on the $NewInvOpp_{s,t}$ dummies remain economically and statistically significant. In contrast, the explanatory power of general textual changes in the 10-K (*Firm Change*) becomes insignificant. This pattern suggests that our semantic distance measure is not merely capturing general changes in a firm's 10-K, but is specifically sensitive to the very essence of DUI related to new, complex, and potentially valuable investment opportunities.

In summary, this diagnostic analysis validates that our pairwise distance measure is driven by the confluence of a fund's evolving expertise and, critically, the emergence of novel investment

opportunities in a firm's disclosures. This confirmation provides the necessary foundation for our main tests, which investigate whether skilled funds will act on this "good distance".

3.2 Fund Performance Prediction

We now formally test whether the interaction of fund activeness and holding distance—our proxy for distant investment—jointly predicts future fund performance. We begin with portfolio sorts and Fama-MacBeth regressions to establish the performance implications, followed by an examination of the persistence of our measure to confirm that it captures enduring skill rather than transient luck.

3.2.1 Portfolio Sorts

Each quarter, we independently sort funds into quintiles based on the most recent values of *Fund Activeness* and *Holding Distance* and form value-weighted portfolios. Table 3 then reports the out-of-sample performance of these portfolios using the Fama-French-Carhart (1997) four-factor model, with Newey-West (1987) adjusted t-statistics with four months of lags. Panel A presents the results for before-fee returns, and Panel B for after-fee returns.

The results reveal a striking pattern: conventional skill—proxied by *Fund Activeness*—predicts outperformance *only* among funds that also exhibit a high propensity for *Holding Distance*. Within the highest *Holding Distance* quintile, the most active funds outperform the least active ones by a monthly before-fee risk-adjusted return of 0.48% (5.8% annualized), with a t-statistic of 3.15. In sharp contrast, the high-minus-low performance spread shrinks to a statistically insignificant 0.14% among funds in the lowest *Holding Distance* quintile. The difference in these two spreads is both economically and statistically significant, amounting to 0.34% per month ($t = 2.06$). This spread difference highlights the role of distant investments in unlocking the value of active management. Panel B confirms that these patterns hold for after-fee returns, alleviating concerns that the results are driven by expenses.[16]

Figure 3 visualizes the benefits of distant investment by plotting the cumulative after-fee returns of two extreme groups: high-activeness/high-distance funds and low-activeness/low-distance funds, as well as the cumulative difference between them. Throughout our sample period, the high/high portfolio consistently outperforms its low/low counterpart. Notably, the cumulative

[16] Another notable finding in Table 3 is that funds in the lowest activeness quintile and highest holding-distance quintile exhibit significantly negative performance in the subsequent quarter. We provide a flow-based explanation for this result in Section 6.1.

return difference exhibits a steady upward trend with relatively low volatility, suggesting that the performance of investing in distant firms through an active mandate does not entail commensurately higher risk.

3.2.2 Fama-MacBeth Regressions

To control for a broader set of fund characteristics and to quantify economic magnitudes more precisely, we estimate quarterly Fama-MacBeth (1973) regressions of the following form:

$$Perf_{f,t} = \alpha + \beta_1 \times Holding\ Distance_{f,t-1} + \beta_2 \times Fund\ Activeness_{f,t-1} + \beta_3 \times SDI_{f,t-1} + \boldsymbol{\Gamma}' \times \boldsymbol{M}_{f,t-1} + \varepsilon_{f,t}, \qquad (6)$$

where $Perf_{f,t}$ refers to the performance of fund f in quarter t, $Holding\ Distance_{f,t-1}$ and $Fund\ Activeness_{f,t-1}$ are, respectively, its holding distance and fund activeness in quarter *t–1*, and $SDI_{f,t-1}$is the interaction term describing the level of skilled distant investment. $\boldsymbol{M}_{f,t-1}$ stacks a set of lagged fund-level control variables, including total net assets, age, turnover, expense ratio, past flows, past returns, and return volatility.

For ease of interpretation, both holding distance and fund activeness are standardized by the standard deviation of the distribution of the respective variable. As a result, the regression coefficient can be directly interpreted as the performance impact introduced by a one-standard-deviation increase in the independent variable. The coefficient of interest is β_3, which captures the effect for distant investment. Because *SDI* is the product of these standardized variables, this coefficient represents the incremental performance impact of a one-standard-deviation increase in *Holding Distance* on top of a one-standard-deviation increase in *Fund Activeness*.

The results are reported in Table 4, which employs three performance metrics. Columns (1) to (3) use the quarterly Carhart four-factor alpha estimated before fees; Columns (4) to (6) use the after-fee equivalent. To compute these quarterly fund alphas, we estimate each fund's factor betas using monthly returns over the prior 24 months (requiring at least 12 months of data) and compute alphas as the difference between realized returns and the implied risk premium. Columns (7) to (9) use the value-added measure of Berk and van Binsbergen (2015), which captures the dollar amount of the value created by fund managers. We report Newey-West (1987) adjusted t-stat with three quarters of lags.

The regression results both confirm and extend the portfolio sort evidence. First, *Fund Activeness* alone positively predicts future performance when measured by the two alphas, but its coefficient is statistically insignificant in the value-added specification. *Holding Distance* on its own exhibits no significant predictive power, consistent with the notion that distance without activeness may not reflect skill.

More importantly, the interaction term, *SDI*, is positive and statistically significant across all specifications, indicating the performance-enhancing effect when fund activeness is directed toward distant firms. The economic magnitudes are substantial. Focusing on value-added in Column (9), a one-standard-deviation increase in *Fund Activeness* is associated with a $2.17 million increase in quarterly value-added. The coefficient on *SDI* implies that a one-standard-deviation increase in *Holding Distance*—on top of the same increase in *Fund Activeness*—contributes an additional $1.84 million, representing an 84.8% relative improvement. This incremental effect provides direct economic quantification of the value created by skilled managers applying their expertise to process DUI in distant firms.

Although the above conclusions are obtained from the Fama-MacBeth specification, using panel regressions with fund and quarter fixed effects leads to qualitatively and quantitatively similar results. These additional results are reported in Online Appendix Table A2.

3.2.3 Persistence of *Distant Investment* and *Fund Activeness*

Thus far, our fund-level tests suggest that the combination of activeness and holding distance predicts superior performance. If such *SDI* genuinely captures managerial skill, we should observe stability in these attributes—skilled funds should remain skilled. Conversely, if our results were driven by transient luck, we would expect little persistence.

Online Appendix Table A3 reports transition matrices, where funds are sorted into quintiles each quarter based on their most recent *Holding Distance* (Panel A), *Fund Activeness* (Panel B) and *SDI* (Panel C). We then compute the probability that a fund classified in one quintile transitions to another quintile in the following quarter. Our main finding is that the diagonal elements are substantially larger than the off-diagonal elements, indicating a high degree of persistence. For instance, a fund in the highest *Holding Distance* quintile (Q5) has a 77.80% probability of remaining in Q5 in the following quarter, which is even higher than that of *Fund Activeness* (68.50%). The persistence of *SDI* lies between these two measures. This pattern suggests that a fund's propensity to invest in distant firms—a portfolio characteristic we argue

reflects specialized expertise—is an even more stable than its general level of fund activeness. These observations are consistent with the interpretation that distant investments may reflect enduring skill rather than ephemeral factors.

3.3 Sub-sample Analyses

To further illuminate the economic drivers of *SDI*, we conduct several subsample analyses. These tests examine whether the complementarity between holding distance and fund activeness is stronger in environments characterized by high information complexity and whether it aligns with established theories of delegated money management.

3.3.1 Exposure to Emerging Investment Frontiers

We first investigate whether the predictive power of *SDI* is contingent on the fund's exposure to "frontier" topics examined in our diagnostic analysis. We sort funds into subsamples based on the aggregate portfolio weight of firms whose 10-Ks contain novel keywords related to (i) AI, (ii) ESG, and (iii) the 43 machine-learning-derived categories of investment opportunities.

Panel A of Table 5 reports the results. Columns (1) and (2) partition funds by their exposure to emerging AI topics. The coefficient on *SDI* is positive and statistically significant only for funds with above-median AI exposure (0.067, $t = 2.76$), while it is indistinguishable from zero for those with below-median exposure. Similarly, the predictive strength of *SDI* is heavily concentrated among funds with above-median exposure to comprehensive ML-identified investment categories (Column 6). Since AI and the broader set of ML-identified categories represent novel, complex, and rapidly evolving domains, they precisely contain the type of information that is difficult for the general market to process but may be interpretable by skilled managers with relevant expertise. The concentration of predictive power in high-exposure funds is consistent with the notion that *SDI* may indeed capture the successful application of specialized skill to exploit related DUI.

Interestingly, *SDI* remains a significant predictor of performance regardless of a fund's exposure to ESG topics (Columns 3 and 4). This divergence suggests that while ESG disclosures are frequent, they may be less linguistically complex or more "standardized" than AI-related technical advancements or machine-learning-derived comprehensive categories of investment opportunities. The specialized expertise required to exploit "distance" in the latter two cases represents a more valuable form of managerial skill.

3.3.2 Fund Characteristics and Economic Rents

In Panel B of Table 5, we examine how the predictive power of *SDI* varies with three fund characteristics: size, turnover, and expense ratio. Each characteristic speaks to a different aspect of the skill hypothesis.

The first two columns show that the performance-predictive power of *SDI* concentrates among funds with above-median total net assets (TNA). This finding is consistent with the idea that distant investment helps mitigate diseconomies of scale. As funds grow, traditional active strategies may face diminishing returns due to price impact and liquidity constraints (Berk and Green 2004). By investing in firms with DUI—where information is complex and competition is lower—skilled managers can deploy larger capital bases without fully impounding their informational advantage, thereby preserving alpha.

Columns (3) and (4) indicate that the predictive power of *SDI* is concentrated among funds with below-median turnover ratios. This result suggests that the superior performance associated with distant investment is not achieved through frequent trading. Instead, it is consistent with a longer-term investment horizon in which managers patiently hold positions in firms whose DUI they have processed, allowing time for the market to gradually incorporate this information into prices. High turnover, by contrast, may erode the gains from information processing through transaction costs or may reflect a less thoughtful, more speculative approach.

Finally, Columns (5) and (6) reveal that *SDI* predicts performance more strongly among funds charging above-median expense ratios. This pattern aligns with the rent-capture mechanism of Berk and Green (2004): skilled managers who generate alpha through distant investments can command higher fees, and investors, in turn, are willing to pay these fees to access the managers' superior information-processing abilities. The concentration of predictive power in high-expense funds suggests that at least a portion of the economic rents from skilled distant investment accrues to fund families and managers.

Collectively, these subsample results paint a coherent portrait of the skilled distant-investing fund managers. From the fund's perspective, distant investment appears to be a scalable, low-turnover strategy that can manifest through the processing of complex, novel, and difficult-to-understand information emerging from firms' strategic frontiers.

4. Asset Pricing Implications for Distant Investment

Having established that *SDI* captures managerial skill and predicts superior fund-level performance, we now investigate its asset pricing implications. If skilled managers generate superior performance by processing DUI, their trading should accelerate the incorporation of such information into stock prices, thereby predicting stock returns on one hand and enhancing market efficiency on the other. This section empirically examines these asset pricing implications.

4.1 Return Predictive Power for Stocks

When distant investments reflect skilled fund managers processing superior firm-level information (especially DUI), their trading should predict future stock returns. This is because such informed trading helps incorporate their processed information into stock prices (e.g., Kyle 1985). To test this implication, we conduct stock-level tests to investigate the return predictive power of distant investments, linking out-of-sample DGTW-adjusted abnormal stock returns to distant-investment motivated mutual fund trading. Specifically, we estimate the following quarterly Fama-MacBeth (1973) regression:

$$DGTW_{s,t} = \alpha + \beta \times SDI\_OI_{s,t-1} + \boldsymbol{\Gamma}' \times \boldsymbol{M}_{s,t-1} + \varepsilon_{s,t}, \quad (7)$$

where $DGTW_{s,t}$ refers to the DGTW-adjusted abnormal return for stock *s* in quarter *t*, $SDI\_OI_{s,t-1}$ is the *SDI*-weighted order imbalance for stock s in quarter $t-1$, which is our main proxy for the trading originated from skilled distant investment. In our baseline specification, the vector $\boldsymbol{M}_{s,t-1}$ stacks a set of lagged firm-level control variables, including market capitalization (*Size*), book-to-market ratio (*BM*), past returns, asset growth (*Investment*), operating profitability, illiquidity, and analyst coverage. For robustness checks, we also add additional controls, including idiosyncratic volatility (estimated from daily Fama–French three-factor model regressions), investor attention (log Google search volume), geographic proximity to funds (holding value-weighted indicators of whether mutual funds are located in the same state as the firm), and the firm's 10-K complexity (Loughran and McDonald 2024). We again report Newey-West (1987) adjusted t-stat with three quarters of lags.

Table 6 presents the results. Column (1) shows that *SDI_OI* predict significant out-of-sample DGTW-adjusted returns. A one-standard-deviation increase in *SDI_OI* is associated with a 0.313% higher quarterly DGTW return. To simplify the economic interpretation, Column (2) further replaces *SDI_OI* with a dummy variable, *High SDI_OI*, which takes the value of one when *SDI_OI*

is above median in the cross section. This dummy variable indicates high net buying pressure from skilled distant-investing funds. When this occurs, the underline stock will deliver 0.328% out-of-sample DGTW-adjusted returns.

Columns (3) and (4) test whether the predictivity of skilled trading identified from distant investment (*SDI_OI and High SDI_OI*) could be absorbed by alternative explanations. First, geographic proximity may allow mutual funds to collect soft information and thus better predict stock returns. Second, maybe it is the complexity of firms' 10-K filing itself, rather than mutual funds' processing of related information, that predicts stock returns. Additionally, distant investment might simply load on characteristics such as idiosyncratic risk or investor attention, rather than processing DUI from firms' 10-Ks. Our empirical results indicate that the return predictivity of *SDI_OI* and *High SDI_OI* remain highly robust, suggesting that our distant investment measures capture independent economic sources of fund skills unrelated to these alternative explanations.

Columns (5) and (6) further compare the predictivity of skilled trading identified from distant investment (*SDI_OI* and *High SDI_OI*) to that from fund activeness alone (*AWOI*). We observe that the predictive power of *SDI_OI* and *High SDI_OI* remains highly significant, while *AWOI* does not significantly predict returns. Online Appendix Table A4 further controls for the trading originated from holding distance (*DWOI*), alone or together with *AWOI*. The predictivity of *SDI_OI* remains highly robust. Indeed, neither AWOI nor DWOI exhibits significantly predict power, confirming that only *SDI_OI*, which captures the managerial skill revealed by active distant investment, has the proper power to predict stock returns.

To see whether the return predictivity of *SDI_OI* and *High SDI_OI* concentrates only on small stocks, we excluding the bottom 20% of stocks with the smallest market capitalization. The results, reported in Online Appendix Table A5, demonstrate that our main findings are robust to this subsample of stocks. Hence, the return predictivity is not entirely driven by small-cap stocks. This result is reasonable because DUI could emerge from not only small stocks but also large, innovative companies.

4.2 Correcting the "Lazy Price" Inefficiency

Cohen, Malloy, and Nguyen (2020; hereafter CMN) document a striking market inefficiency: firms that materially change their 10-K filings ("Changers") experience prolonged price declines for up to eighteen months following the filing. This "lazy price" effect suggests that 10-Ks contain

information the market struggles to process promptly. If distant-investing managers are skilled at processing DUI contained in 10-K filings, their trading should attenuate this inefficiency.

To test this implication, we examine whether trading by distant-investing managers during the 10-K release quarter attenuates post-filing price drifts. We construct two dummy indicators. The first, labelled *Changers*, equals one if the embedding change between a firm's current Item 1 and its prior-year Item 1 exceeds the cross-sectional median. This variable is constructed to capture the lazy-price effect of CMN.

The second variable, *High SDI_OI*, equals one if *SDI_OI* exceeds the cross-sectional median in the filing quarter. We have used this variable in the previous table. The difference is that we only focus on trading of skilled distant-investing funds during the 10-K filing quarter in the current test (as opposed to all quarters in the previous test).

We then estimate annual Fama-MacBeth (1973) regressions linking post-release cumulative DGTW abnormal returns to *Changers*, *High SDI_OI*, and their interaction (along with controls including standardized unexpected earnings, SUE, following CMN). The results are reported in Table 7.

Since CMN indicates that the laze price effect could last for approximately six months after the 10-K releasing quarter (e.g., their Figure 7), we focus on the cumulative returns measured over the subsequent quarter (Columns 1–3) and the following two quarters (Columns 4–6). Consistent with the lazy price effect, Columns (1) and (4) show that *Changers* experience significant price declines of approximately 1.4% over the subsequent quarter and 2.8% over the following two quarters after the release quarter. All these price declines are highly significant.

Column (2) introduces the interaction with distant-investment skilled trading. While *High SDI_OI* alone is insignificant, its interaction with *Changers* is positive and significant (0.027), with a magnitude that can almost offset the *Changers* coefficient (-0.031). This implies that among *Changers* with high distant-investment skilled trading during the filing quarter, the subsequent price decline is mostly eliminated. In other words, the trading of distant-investing funds during the 10-K release quarter properly incorporates the 10-K information of *Changers* into stock prices. The remaining *Changers* experience an even larger 3% decline. Column (3) confirms robustness after including SUE.

Similar patterns emerge over longer horizons (Columns 5–6): distant-investment trading during the release quarter largely eliminates the lazy-prices drift for *Changers*, reducing

cumulative declines to near zero over two quarters. For instance, Column (5) indicates that during the six-month period after the 10-K filing quarter, the *Changers* experience a 4.7% decline in price, while high distant-investment trading almost completely offset this price impact (3.3%).

Combined with the return-predictability evidence, these results indicate that distant-investing managers process the DUI contained in10-K filings and incorporate it into prices via trading, thereby reducing inefficiencies associated with qualitative disclosures that the market struggles to digest promptly.

4.3 Economic Sources of Information: Variance Decomposition

To provide a structural estimation of the information channel, we follow Brogaard et al. (2021) and decompose stock return variance into four economically distinct components: (i) firm-specific private information, (ii) firm-specific public information, (iii) market-wide information, and (iv) noise. If distant-investing managers are skilled at processing firm-specific DUI, their trading should be associated with a larger private information component and a smaller noise component.[17]

We estimate the four variance components annually for each stock using daily returns. Because variance is affected by both buy and sell orders, we construct trading intensity measures rather than using the directional order imbalances.[18] Specifically, we construct *SDI*-Weighted Trading Intensity (*SDI_TI*) as the *SDI*-weighted sum of the absolute values of buy and sell orders, scaled by total buy and sell orders. The measure is averaged annually to match the frequency of the dependent variables. We then regress each variance component on this indicator in panel specifications with year fixed effects.

Table 8 reports the results, with columns (1)-(4) corresponding to firm-specific private information, firm-specific public information, market-wide information, and noise, respectively. The most striking patterns emerge for private information. Column (1) and (2) show that *SDI_TI* is significantly positively associated with the firm-specific private-information component, but has no significant effect on the firm-specific public-information component. In contrast, Columns (3)

[17] It is worth noting that in the first component, "private information" is not equivalent to insider information. This component is estimated from trading volume—Brogaard et al. (2021) interpret it as "private information" based on the assumption that trading volume contains private information. In theory, superior private information may arise when fund managers use their skill to process public information. Kim and Verrecchia (1994) and Kandel and Pearson (1995), for instance, point out that skilled investors can obtain superior information by converting a firm's noisy public signals (e.g., its 10-K filing) into more accurate information based on their experience or skill. Engelberg, Reed, and Ringgenberg (2012) and Lin, Massa, and Zhang (2014) further provide empirical evidence on short sellers and mutual fund managers.

[18] Order imbalance reflects the directions of trading (i.e., net buying), whereas trading intensity adds up the magnitude of both buy and sell orders to capture the price impact of both buy and sell orders.

and (4) show that *SDI_TI* is significantly negatively associated with the market-wide information and noise components.

Overall, these findings indicate that distant-investment trading enhances price informativeness by revealing firm-specific private information while reducing noise, consistent with skilled managers processing DUI and improving allocational efficiency. The fact that these effects obtain only for the interaction term reinforces our central thesis: skill manifests at the *intersection* of active management and a focus on semantically distant firms containing DUI.

5. Additional Tests for Economic Insights

Our main analyses demonstrate that distant investments have important implications for both mutual fund performance and stock market efficiency. In this section, we conduct additional tests to deepen our understanding of the underlying mechanisms, contrast skilled versus unskilled distant investment, and establish the robustness of our findings.

5.1 The Power of Relational Fund-Firm Distance

A central contribution of this paper is the move from unconditional textual analysis to relational metrics. Rather than relying on document-centric properties of fund prospectuses or firms' 10-Ks alone, it captures the semantic alignment between these two texts. A natural question is whether fund prospectuses or firms' 10-Ks in isolation can reveal similar insights into skilled distant investments—in other words, whether the power of our measure stems from the fund-firm relationship or merely reflects fund-specific or firm-specific information.

To address this question, we construct alternative measures based solely on textual information from fund prospectuses or firms' 10-Ks. For fund-specific information, we compute *Fund Change* as the semantic distance between a fund's current and prior prospectus. For firm-specific information, we use *Firm Change*, *New AI*, *New ESG*, and *New ML-identified Investment Opportunities* (as defined in Section 3.1), along with *Complexity* (Loughran and McDonald 2024), which quantifies the linguistic complexity of firms' 10-K filings. We aggregate firm-specific measures to the fund level using portfolio weights.

We then revisit the quarterly Fama-MacBeth (1973) regressions from Equation (6), replacing *Holding Distance* (which reflects the fund-firm semantic relationship) with these fund- or firm-specific metrics:

$$Perf_{f,t} = \alpha + \beta_1 \times TextInfo_{f,t-1} + \beta_2 \times Fund\ Activeness_{f,t-1} + \beta_3 \times TextInfo_{f,t-1} \times Fund\ Activeness_{f,t-1} + \boldsymbol{\Gamma'} \times \boldsymbol{M}_{f,t-1} + \varepsilon_{f,t}, \quad (8)$$

where $TextInfo_{f,t-1}$ denotes one of the fund- or firm-specific textual measures. We apply each measure individually to compare predictive power. The coefficient of interest is β_3, which captures the interaction analogous to *SDI*. For example, using *Complexity* alone, a significant β_3 would imply that investing in firms with unconditionally complex 10-Ks enables active funds to generate performance (presumably by the processing of unconditional DUI).

Table 9 reports the results. Column (1) uses the fund-specific measure; Columns (2) through (6) use the firm-specific measures. Strikingly, β_3 is insignificant across most specifications and performance metrics. In some cases, the coefficient is even negative (though insignificant). These results stand in sharp contrast to the strong and robust predictive power of our relational measure of *SDI*.

Why does standalone textual analysis fail? Classical theories offer an explanation: different investors interpret the same information differently (Rubinstein 1993; Kim and Verrecchia 1994; Kandel and Pearson 1995). The application of an unconditional measure of linguistic complexity or textual change essentially assumes that "difficulty" is a fixed property of the text itself. Yet theory shows that investors optimally specialize in domains where they possess an initial information advantage (Van Nieuwerburgh and Veldkamp 2009), rendering the value of information inherently investor-specific (Farboodi et al. 2025; Massa, Zhang, and Zhou 2024).

In our setting, these arguments imply that whether a 10-K filing contains DUI depends critically on who is reading it. A growth fund manager is best positioned to judge whether a growth firm's disclosures contain DUI, but not a value manager. Conversely, a value stock's complexity is best assessed by a value specialist. Unconditional linguistic measures, by treating complexity as a fixed document attribute applying to all investors, cannot capture this heterogeneity.

Our distance measure addresses this limitation by using the fund prospectus as a revealed benchmark of managerial expertise—a semantic anchor. This allows us to identify conditional, fund-specific DUI based on the relative interpretability of firm disclosures from each manager's unique perspective. In doing so, we essentially quantify the "difficulty" of a text conditional on "who is reading," which provides a powerful framework in settings characterized by investor heterogeneity and domain specialization, such as the professional mutual fund industry.

5.2 Distant Investments by Low-skill Managers

We next revisit a puzzling observation from our earlier tests: some low-skill fund managers also appear to adopt distant investment strategies, yet with negative performance and stock return predictability. For instance, Table 3 Panel A shows that funds in the lowest quintile of activeness but the highest quintile of holding distance underperform by 0.37% monthly. How can we reconcile this with our evidence that distant investments are, on average, informed?

We hypothesize that fund flows provide an explanation. The literature documents that investors chase past performance (Sirri and Tufano 1998) and that funds engage in window dressing to mitigate outflows following poor returns (Agarwal, Gay, and Ling 2014; Xin, Yeung, Zhang 2024). When new investment opportunities—such as AI technologies—eventually become popular and capture public attention, retail investors may favor funds that invest in these attention-grabbing firms. Low-skill managers, lacking the ability to genuinely process DUI, may nonetheless window dress their portfolios toward distant, high-attention stocks to buffer outflows.

To test this mechanism, we follow Da, Engelberg, and Gao (2011) and construct the Abnormal Search Volume Index (ASVI) as a proxy for retail attention. A stock is classified as attention-grabbing if its ASVI exceeds the cross-sectional median in a given quarter. We then compute, for each fund-quarter, the proportion of holdings in attention-grabbing stocks.

Online Appendix Table A6 revisits the double-sorting portfolio from Table 3 and reports the value-weighted ASVI for each portfolio. Strikingly, the puzzling funds—those in the lowest activeness quintile and highest distance quintile—exhibit significantly positive ASVI. Indeed, among the 25 double-sorted portfolios, they are the only group with significant ASVI. This suggests that distant investments by low-skill managers could indeed be motivated by capturing or catering to investor's attention, consistent with window dressing.

Window dressing could also explain why these funds have low activeness. On one hand, window dressing does not require large deviations from benchmarks: even a small overweight in attention-grabbing stocks could suffice to attract unsophisticated retail investors. On the other hand, window dressing has a cost—too high a deviation could lead to higher tracking error and more underperformance. As a result, low-skill funds keep their fund activeness at a low level.

To formally test this mechanism, we conduct two tests. The first one examines the determinants of fund trading decisions using the following panel specification:

$$
\begin{aligned}
Buy\ (Sell)_{f,s,t} = \alpha &+ \beta_0 \times Distance_{f,s,t-1} + \beta_1 \times Distance_{f,s,t-1} \times LowActive_{s,t-1} \\
&+ \beta_2 \times Distance_{f,s,t-1} \times HighActive_{s,t-1} + \cdots \\
&+ \gamma_1 \times Distance_{f,s,t-1} \times ASVI_{s,t-1} \times LowActive_{s,t-1} \\
&+ \gamma_2 \times Distance_{f,s,t-1} \times ASVI_{s,t-1} \times HighActive_{s,t-1} + \cdots \quad (9)
\end{aligned}
$$

where $Buy\ (Sell)_{f,s,t}$ is a dummy indicator equal to one if fund f increases (decreases) its position in stock s in quarter t. $Distance_{f,s,t-1}$ is the pairwise semantic distance between the prospectus of fund *f* and the 10-K of firm *s* in quarter $t-1$; $ASVI_{s,t-1}$ measures the retail attention, proxied by the Abnormal Google Search Volume Index; and $HighActive_{f,t-1}$ and $LowActive_{f,t-1}$ are dummy indicators equal to 1 for funds in the top 20% and bottom 20% of the cross-sectional distribution of *Fund Activeness* in quarter $t-1$, respectively. For brevity, Equation (9) displays only the key terms; the full specification includes all other variables, interactions, and controls. We also include Fund × Quarter and Stock × Quarter fixed effects to absorb the impact of fund or stock characteristics. Robust standard errors are clustered at the fund and stock level.

The results are reported in Table 10 Panel A, columns (1) to (2) for buying decisions and (3) to (4) for sales. The first column links buys to distance-activeness interactions (without ASVI). Among the three relevant coefficients, only β_2 (for $Distance_{f,s,t-1} \times HighActive_{s,t-1}$) is significant. Its positive sign suggests that the buying decision of only high-activeness funds—but not other funds—is positively influenced by distance, consistent with skilled distant investment.

Column (2) introduces the triple interactions with ASVI. The coefficient γ_1 (for low-active funds × distance × attention) is positive and significant. Combined with the insignificant β_1, this reveals that low-active funds buy distant stocks *only* when they become attention-grabbing. This behavior is precisely what window dressing predicts: lacking the skill to process DUI, they cater to investor attention.

In contrast, γ_2 is insignificant, indicating that high-activeness funds buy distant stocks unconditionally (positive β_2) but do not amplify their buying when attention is high. This is consistent with informed trading. To the extent that high retail attention signals that the market has likely digested—or even overpriced—the opportunity, skilled managers act before retail attention.

Columns (3)–(4) show similar patterns for sells: insignificant β_2 and γ_2 imply that high-activeness funds are unaffected by attention. For low-active funds, negative $\gamma_1$indicate that they sell less attention-grabbing distant stocks. Overall, they buy more attention-grabbing distant stocks and hold them longer.

Collectively, our first test indicates that high-activeness funds' decisions may reflect skill: they buy distant stocks and sell close ones, while ignoring investor attention. Low-active funds appear uninformed. They tend to buy and hold attention-grabbing distant stocks, which explains the underperformance of their (window dressing) distant investments.

Our second test examines whether low-skill funds' distant investing can indeed buffer outflows by reducing flow-performance sensitivity. We regress quarterly flows on lagged returns and their interactions with holding distance in the following quarterly Fama-MacBeth (1973) specification:

$$Flow_{f,t} = \alpha + \beta_1 \times Ret_{f,t-1} + \beta_2 \times Holding\ Distance_{f,t-1} + \beta_3 \times Ret_{f,t-1} \times Holding\ Distance_{f,t-1} + \boldsymbol{\Gamma'} \times \boldsymbol{M_{f,t-1}} + \varepsilon_{f,t} \quad (10)$$

where $Flow_{f,t}$ refers to the flow of fund f in quarter t, calculated as $\frac{TNA_{f,t} - TNA_{f,t-1} \times (1+r_{f,t})}{TNA_{f,t-1}}$, and $Ret_{f,t-1}$ refers to one -quarter lagged fund returns.

We report the results in Table 10 Panel B, with Columns (1) to (2) for the sample of low-active (bottom 20%) funds and (3) to (4) for high-active (top 20%) funds. Columns (1) and (3) indicate that, for both types of funds, fund flows are positively related to at least two quarters of past returns, confirming the notion of mutual fund investors chasing past performance.

More interestingly, Columns (2) and (4) reveal that β_3—the interaction between one-quarter-lagged return and holding distance—is significantly negative for low-active funds but insignificant for high-active funds. A negative coefficient implies reduced flow-performance sensitivity, allowing lagged poor performance to trigger *less* outflow when low-skill funds have invested in distance stocks. In other words, investing in distant and attention-grabbing stocks could help buffer outflows for these funds to some extent.

Taken together, these results paint a coherent picture of two distinct types of distant investment. For high-skill managers, distant investment reflects genuine information processing and generates alpha. For low-skill managers, it reflects window dressing—buying attention-grabbing distant

stocks to mitigate outflows—and predicts poor performance. This duality underscores the importance of conditioning distant investment on fund activeness to isolate genuine skill.

5.3 Alternative Measures and Placebo Tests

Lastly, we assess the robustness of our results by using alternative measures of fund activeness and alternative natural language processing (NLP) methods. We also provide a placebo test using alternative 10-K sections.

Our main proxy for fund activeness is constructed as the average rank of *return gap* (Kacperczyk, Sialm, and Zheng, 2008) and *active weight* (Doshi, Elkamhi, and Simutin, 2015). Online Appendix Table A7 shows that using either measure individually yields qualitatively similar double-sorting results, albeit with smaller magnitudes. In Panel C of Table A7, we further construct a more composite skill measure ($Skill^C$) based on the average rank of *return gap, active weight, active share* (Cremers and Petajisto 2009), *industry concentration index* (Kacperczyk, Sialm and Zheng 2005), *abnormal cash holdings* (Simutin 2014), *1 minus R-square* (Amihud and Goyenko 2013), and *skill index* (Kacperczyk, Nieuwerburgh and Veldkamp 2014). Overall, our fund performance results on distant investment remain highly robust across these alternative measures of fund activeness, suggesting that holding distance can indeed be viewed as a general trading strategy that complements traditional activeness-based measures of fund skill.

The second set of robustness checks focuses on alternative NLP methods. In our main analysis, we adopt the Sentence-BERT embedding-based NLP technique, increasingly used in the finance and economics literature. However, given the less interpretable nature of LLMs, we use the more traditional yet straightforward bag-of-words (BoW) approach. This method compares the word distribution between mutual fund prospectuses and firms' 10-Ks and then calculates the distance between their corresponding word-frequency vectors. As shown in Online Appendix Table A8 Panel A, our double sorting results remain robust using the more traditional but interpretable BoW distance. However, the economic magnitude is smaller, suggesting that, perhaps not surprisingly, LLM provides a more powerful NPL method to analyze the relationship between fund prospects and firms' 10-K filings.

One limitation of S-BERT is that it can only process 512-token at a time. To address this limitation, researchers commonly split each 10-K into chunks (i.e., individual sections or segments) and take the average embedding value across all chunks to represent the full text. For our purposes, not all chunks are of equal importance. The first chunk, which typically describes the firm's overall

strategy, operational priorities, and potential planned investment opportunities, contains the most critical information relevant to the fund's expertise outlined in the fund prospectus. As a result, we use the first chunk to compute our main proxy of distance. In Online Appendix Table A8 Panel B, we show that the average distance of the full texts also produces consistent results, though the economic significance and statistical power are slightly weaker.

The third test examines alternative contents of 10-K filings. In our main analysis, we focus on the fund strategy narrative section of the prospectus and the business description part of 10-K filings (Item 1). While these parts of descriptions have been used intensively to interpret fund strategies and firm-level information, other parts of the 10-Ks, such as risk factors (Item 1A), may also provide additional information (Bai et al. 2024). Thus, we test whether risk-related information in fund prospectuses and Item 1A of 10-K filings can help infer distant investments in the same way as strategy descriptions. As shown in Online Appendix Table A8 Panel C, while fund activeness continues to predict fund performance, risk-based holding distance fails to offer significant incremental explanatory value. This insignificance indicates that our empirical strategies have adequate power to filter out firm information that does not complement fund activeness.

6. Conclusion

Traditional measures of mutual fund skill focus on activeness or performance but offer little insight into how managers actually select stocks. We fill this gap by introducing a framework that leverages Large Language Models to analyze the semantic alignment between fund prospectuses and firms' 10-K strategic priorities. We hypothesize that investing in stocks with a large semantic distance—distant investment—reveals the processing of fund-specific difficult-to-understand information (DUI) requiring specialized expertise. Because DUI is not immediately reflected in prices, skilled managers deviate from benchmarks to invest in distant stocks, allowing holding distance to complement traditional activeness measures in capturing skill.

Our empirical analysis supports this hypothesis. Portfolio sorts reveal that funds with both high activeness and high holding distance deliver superior out-of-sample performance. Multivariate regressions confirm that skilled distant investment (SDI)—the interaction between activeness and holding distance—predicts future fund performance, particularly among funds exposed to firms with frontier investment opportunities.

Distant investment also has important asset pricing implications. The trading of distant-investing funds predicts stock returns, attenuates the "lazy price" anomaly (Cohen, Malloy, and Nguyen 2020), and enhances price efficiency by increasing firm-specific private information while reducing noise. Conversely, unskilled managers exploit retail attention for window dressing, highlighting the presence of industry frictions.

Overall, we bridge the gap between investment mandates and portfolio choices, providing a new economic rationale for skill as active distant investment. By using fund prospectuses as a semantic anchor to quantify investor-specific information, we demonstrate that relational linguistic metrics can operationalize the classical insight that investors interpret the same news differently based on their unique expertise. This approach shifts the structural focus of textual analysis from "what is written" to "who is reading," providing a general method to capture conditional information that traditional linguistic metrics overlook. We call for further research to apply this relational lens to other domains of financial decision-making where investor heterogeneity plays a central role.

Reference

Abis, S. and Lines, A., 2024. Broken promises, competition, and capital allocation in the mutual fund industry. *Journal of Financial Economics*, *162*, p.103948.

Acemoglu, D., Mühlbach, N.S. and Scott, A.J., 2022. The rise of age-friendly jobs. *The Journal of the Economics of Ageing*, *23*, p.100416.

Agarwal, V., Gay, G.D. and Ling, L., 2014. Window dressing in mutual funds. *The Review of Financial Studies*, *27*(11), pp.3133-3170.

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), pp.31-56.

Amihud, Y. and Goyenko, R., 2013. Mutual fund's R 2 as predictor of performance. *The Review of Financial Studies*, *26*(3), pp.667-694.

Amihud, Y. and Goyenko, R., 2015. How to measure the skill of your fund manager. *American Association of Individual Investors Journal*, *37*, pp.27-31.

Ang, A., Hodrick, R.J., Xing, Y. and Zhang, X., 2006. The cross-section of volatility and expected returns. *Journal of Finance*, *61*(1), pp.259-299.

Bai, J.J., Boyson, N.M., Cao, Y., Liu, M. and Wan, C., 2023. Executives vs. chatbots: Unmasking insights through human-AI differences in earnings conference Q&A. *Northeastern U. D'Amore-McKim School of Business Research Paper*, (4480056).

Bai, J.J., Tang, Y., Wan, C. and Yuksel, H.Z., 2024. Thematic Concentration and Mutual Fund Performance. *Zafer, Thematic Concentration and Mutual Fund Performance (May 1, 2024). Northeastern U. D'Amore-McKim School of Business Research Paper*, (4164823).

Bali, T.G., Cakici, N. and Whitelaw, R.F., 2011. Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics*, *99*(2), pp.427-446.

Bali, T.G., Engle, R.F. and Murray, S., 2016. *Empirical asset pricing: The cross section of stock returns*. John Wiley & Sons.

Basu, S., Ma, X. and Briscoe-Tran, H., 2022. Measuring multidimensional investment opportunity sets with 10-K text. *The Accounting Review*, *97*(1), pp.51-73.

Ben-David, I., Li, J., Rossi, A. and Song, Y., 2022. What do mutual fund investors really care about?. *The Review of Financial Studies*, *35*(4), pp.1723-1774.

Berk, J.B. and Green, R.C., 2004. Mutual fund flows and performance in rational markets. *Journal of political economy*, *112*(6), pp.1269-1295.

Berk, J.B. and Van Binsbergen, J.H., 2015. Measuring skill in the mutual fund industry. *Journal of Financial Economics*, *118*(1), pp.1-20.

Bollen, N.P. and Busse, J.A., 2001. On the timing ability of mutual fund managers. *Journal of Finance*, 56(3), pp.1075–1094.

Breitung, C. and Müller, S., 2025. Global Business Networks. *Journal of Financial Economics*, *166*, p.104007.

Brogaard, J., Nguyen, T.H., Putnins, T.J. and Wu, E., 2022. What moves stock prices? The roles of news, noise, and information. *The Review of Financial Studies*, *35*(9), pp.4341-4386.

Buehlmaier, M.M. and Whited, T.M., 2018. Are financial constraints priced? Evidence from textual analysis. *The Review of Financial Studies*, *31*(7), pp.2693-2728.

Bybee, J.L., 2023. The ghost in the machine: Generating beliefs with large language models. *arXiv preprint arXiv:2305.02823*.

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

Cao, S., Yang, B. and Zhang, A.L., 2024. Beyond the lines: Uncovering private information from fund managers' disclosures. *Available at SSRN 3713966.*

Cao, S., Yang, B. and Zhang, A.L., 2025. Managerial risk assessment and fund performance: Evidence from textual disclosure. *Available at SSRN 4060307.*

Chen, A.Y. and Zimmermann, T., 2020. Publication bias and the cross-section of stock returns. *The Review of Asset Pricing Studies*, *10*(2), pp.249-289.

Chen, S., Peng, L. and Zhou, D., 2024. Wisdom or Whims? Decoding Investor Trading Strategies with Large Language Models. *Decoding Investor Trading Strategies with Large Language Models (June 19, 2024)*.

Choi, G.Y. and Kim, A., 2024. Firm-level tax audits: A Generative AI-based measurement. *Chicago Booth Research Paper*, (23-23).

Christoffersen, S.E. and Musto, D.K., 2002. Demand curves and the pricing of money management. *Review of Financial Studies*, 15(5), pp.1499–1524.

Cohen, L., Malloy, C. and Nguyen, Q., 2020. Lazy prices. *Journal of Finance*, *75*(3), pp.1371-1415.

Cong, L.W., Tang, K., Wang, J. and Zhang, Y., 2021. AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI. Available at SSRN, 3554486.

Cooper, M.J., Gulen, H. and Schill, M.J., 2008. Asset growth and the cross-section of stock returns. *Journal of Finance*, *63*(4), pp.1609-1651.

Cosemans, M. and Frehen, R., 2021. Salience theory and stock prices: Empirical evidence. *Journal of Financial Economics*, *140*(2), pp.460-483.

Cremers, K.M. and Petajisto, A., 2009. How active is your fund manager? A new measure that predicts performance. *The Review of Financial Studies*, *22*(9), pp.3329-3365.

Da, Z., Gao, P. and Jagannathan, R., 2011. Impatient trading, liquidity provision, and stock selection by mutual funds. *The Review of Financial Studies*, *24*(3), pp.675-720.

Daniel, K., Grinblatt, M., Titman, S. and Wermers, R., 1997. Measuring mutual fund performance with characteristic-based benchmarks. *Journal of Finance*, *52*(3), pp.1035-1058.

Dannhauser, C.D. and Pontiff, J., 2024. Flow. Available at SSRN 3428702.

DeMiguel, V., Gil-Bazo, J., Nogales, F.J. and Santos, A.A., 2023. Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics*, 150(3), p.103737.

Doshi, H., Elkamhi, R. and Simutin, M., 2015. Managerial activeness and mutual fund performance. *The Review of Asset Pricing Studies*, *5*(2), pp.156-184.

Evans, R.B., 2010. Mutual fund incubation. *Journal of Finance*, *65*(4), pp.1581-1611.

Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2), pp.383–417.

Fama, E.F. and French, K.R., 2006. Profitability, investment and average returns. *Journal of Financial Economics*, *82*(3), pp.491-518.

Fama, E.F. and French, K.R., 2010. Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance*, 65(5), pp.1915–1947.

Fama, E.F. and MacBeth, J.D., 1973. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3), pp.607-636.

Farboodi, M., Singal, D., Veldkamp, L. and Venkateswaran, V., 2025. Valuing financial data. *The Review of Financial Studies*, *38*(3), pp.938-980.

Frankel, R., Jennings, J. and Lee, J., 2016. Using unstructured and qualitative disclosures to explain accruals. *Journal of Accounting and Economics*, *62*(2-3), pp.209-227.

Gârleanu, N. and Pedersen, L.H., 2018. Efficiently inefficient markets for assets and asset management. *Journal of Finance*, *73*(4), pp.1663-1712.

Gao, Z., Xiong, W. and Yuan, J., 2024. Structured beliefs and fund performance: An LLM-based approach. *Available at SSRN*.

Gervais, S. and Strobl, G., 2023. Money management and real investment. Available at SSRN 4325133.

Gil-Bazo, J., Ruiz-Verd´u, P., 2009. The relation between price and performance in the mutual fund industry. *Journal of Finance*, 64(5), pp.2153–2183. 49.

Gruber, M.J., 1996. Another puzzle: The growth in actively managed mutual funds. *Journal of Finance*, 51(3), pp.783–810.

Hoberg, G. and Maksimovic, V., 2015. Redefining financial constraints: A text-based analysis. *The Review of Financial Studies*, *28*(5), pp.1312-1352.

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

Jha, M., Liu, H. and Manela, A., 2025. Does finance benefit society? A language embedding approach. *The Review of Financial Studies*, p.hhaf012.

Jha, M., Qian, J., Weber, M. and Yang, B., 2024. *ChatGPT and corporate policies* (No. w32161). National Bureau of Economic Research.

Jha, M., Qian, J., Weber, M. and Yang, B., 2024. Harnessing Generative AI for Economic Insights. *arXiv preprint arXiv:2410.03897*.

Jones, C.M., Shi, D., Zhang, X. and Zhang, X., 2025. Retail trading and return predictability in China. *Journal of Financial and Quantitative Analysis*, *60*(1), pp.68-104.

Jones, C.S. and Mo, H., 2021. Out-of-sample performance of mutual fund predictors. *The Review of Financial Studies*, *34*(1), pp.149-193.

Kacperczyk, M., Nieuwerburgh, S.V. and Veldkamp, L., 2014. Time-varying fund manager skill. *Journal of Finance*, *69*(4), pp.1455-1484.

Kacperczyk, M., Sialm, C., and Zheng, L. 2005. On the Industry Concentration of Actively Managed Equity Mutual Funds. *Journal of Finance* 60:1983–2011.

Kacperczyk, M., Sialm, C., and Zheng, L. 2008. Unobserved actions of mutual funds. *Review of Financial Studies* 21:2379–2416.

Kandel, E. and Pearson, N.D., 1995. Differential interpretation of public signals and trade in speculative markets. *Journal of Political Economy*, *103*(4), pp.831-872.

Kaniel, R., Lin, Z., Pelger, M. and Van Nieuwerburgh, S., 2023. Machine-learning the skill of mutual fund managers. *Journal of Financial Economics*, 150(1), pp.94-138.

Kim, A., Muhn, M., Nikolaev, V. 2023. Bloated disclosures: can ChatGPT help investors process information?. *arXiv preprint arXiv:2306.10224.*

Kim, A., Muhn, M. and Nikolaev, V., 2024. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

Kim, A., Muhn, M., Nikolaev, V. and Zhang, Y., 2024. Learning Fundamentals from Text. *Chicago Booth Accounting Research Center Research Paper, Fama-Miller Working Paper*.

Kim, A.G. and Nikolaev, V., 2024. Context-Based Interpretation of Financial Information. *Journal of Accounting Research*.

Kim, O. and Verrecchia, R.E., 1994. Market liquidity and volume around earnings announcements. *Journal of Accounting and Economics*, *17*(1-2), pp.41-67.

Kostovetsky, L. and Warner, J.B., 2020. Measuring innovation and product differentiation: Evidence from mutual funds. *Journal of Finance*, *75*(2), pp.779-823.

Kyle, A.S. 1985. Continuous auctions and insider trading. *Econometrica*, 53(6), 1315–1335.

Li, F., Lundholm, R. and Minnis, M., 2013. A measure of competition based on 10-K filings. *Journal of Accounting Research*, *51*(2), pp.399-436.

Li, B. and Rossi, A.G., 2020. Selecting mutual funds from the stocks they hold: A machine learning approach. Available at SSRN 3737667.

Lin, C., Massa, M. and Zhang, H., 2014. Mutual funds and information diffusion: The role of country-level governance. *Review of Financial Studies*, *27*(11), pp.3343-3387.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lopez-Lira, A., Tang, Y. and Zhu, M., 2025. The Memorization Problem: Can We Trust LLMs' Economic Forecasts?. *arXiv preprint arXiv:2504.14765*.

Loughran, T., McDonald, B., 2024. Measuring firm complexity. *Journal of Financial and Quantitative Analysis*, 59(6), pp.2487–2514.75.

Malkiel, B.G., 1995. Returns from investing in equity mutual funds 1971 to 1991. *Journal of Finance*, 50(2), pp.549–572.

Massa, M., Zhang, H. and Zhou, Y., 2024. Data Specialists and Market Efficiency. *Available at SSRN.*

Reimers, N. and Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rubinstein, A., 1993. On price recognition and computational complexity in a monopolistic model. *Journal of Political Economy*, 101(3), pp.473–484

Seegmiller, B., Papanikolaou, D. and Schmidt, L.D., 2023. Measuring document similarity with weighted averages of word embeddings. *Explorations in Economic History*, *87*, p.101494.

Sheng, J., Sun, Z., Yang, B. and Zhang, A.L., 2024. Generative AI and asset management. Available at SSRN 4786575.

Siano, F., 2025. The news in earnings announcement disclosures: Capturing word context using LLM methods. *Management Science*, *71*(11), pp.9831-9855.

Simutin, M., 2014. Cash holdings and mutual fund performance. *Review of Finance*, *18*(4), pp.1425-1464.

Sirri, E.R. and Tufano, P., 1998. Costly search and mutual fund flows. *Journal of Finance*, 53(5), pp.1589-1622.

Van Nieuwerburgh, S. and Veldkamp, L., 2009. Information immobility and the home bias puzzle. *Journal of Finance*, 64(3), pp.1187-1215.

Wermers, R., 2000. Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses. *Journal of Finance*, 55(4), pp.1655–1695.

Wu, W., Chen, J., Yang, Z. and Tindall, M.L., 2021. A cross-sectional machine learning approach for hedge fund return prediction and selection. *Management Science*, 67(7), pp.4577-4601.

Xin, X., Yeung, P.E. and Zhang, Z., 2024. Wrong kind of transparency? mutual funds' higher reporting frequency, window dressing, and performance. *Journal of Accounting Research*, 62(2), pp.737-781.

Zheng, L., 1999. Is money smart? A study of mutual fund investors' fund selection ability. *Journal of Finance*, 54(3), pp.901–933.

Zheng, Y., 2025. Can investors learn from patent documents? Evidence from textual analysis. *Contemporary Accounting Research*, 42(2), pp.1331-1358.

Appendix: Variable Definitions

	Variable Definitions
Fund-level Variables	
Holding Distance	Active weight- and industry-adjusted firm distance to the fund, calculated as follows: $$Holding\ Distance_{f,t} = \sum_{s} (w_{f,t}^{s} - w_{f,t}^{sm}) \times (Distance_{f,s,t} - \overline{Distance}_{f,s\in j,t})$$ where $w_{f,t}^{s}$ is the actual portfolio weight of stock *s* in fund *f*'s portfolio and $w_{f,t}^{sm}$ is the weight that this stock would receive had the manager market cap-weighted the equity portfolio at the end of quarter *t*. $Distance_{f,s}$ represents the pairwise semantic fund-firm distance between fund *f* and firm *s* in quarter *t*. $\overline{Distance_{f,s\in J}}$ represents the average fund-firm distance between fund *f* and all firms in industry *j* in quarter *t*.
Fund Activeness	Average rank of *active weight* and *return gap* of fund, calculated as follows: $$Fund\ Activeness_{f,t} = Avg\left(Rank_{f,t}^{RG}, Rank_{f,t}^{AW}\right)$$ where $Rank_{f,t}^{RG}$ and $Rank_{f,t}^{AW}$ refer to the cross-sectional ranks of fund *f* in *return gap* and *active weight* in quarter t.
SDI	The product of *Holding Distance* and *Fund Activeness.*
TNA ($ million)	Total net assets under fund management, in millions.
Month Age	Number of months since the fund's inception date.
Expense (%)	Ratio of total investment that shareholders pay for the fund's operating expenses, in percent.
Turnover	Minimum of aggregated sales or aggregated purchases of securities, divided by the average 12-month TNA of the fund.
Prior 1Q Flow (%)	Fund's net flow, in percent, calculated as follows: $$Flow_{f,t} = \frac{TNA_{f,t} - TNA_{f,t-1} \times (1 + r_{f,t})}{TNA_{f,t-1}}$$ where $TNA_{f,t}$ ($TNA_{f,t-1}$) is the total net assets of fund *f* at the end of quarter *t (t-1)*, and $r_{f,t}$ is the cumulative net return of fund *f* in quarter *t*.
Prior 1Q Return (%)	Fund's cumulative net return over the past 3 months, i.e., the current quarter, in percent.
Prior 1Y Return (%)	Fund's cumulative net return over the past 12 months, i.e., the current and prior three quarters, in percent.
Return Volatility (%)	Fund return volatility, measured as the standard deviation of monthly fund returns over the prior 24 months, requiring at least 18 months of data, in percent.
Fund Change	Euclidean embedding distance between the strategy description in fund 's current prospectus and that in its prior prospectus.
Stock-level Variables	
SDI_OI	SDI-weighted order imbalance, calculated as follows: $$SDI\_OI_{s,t} = \frac{\sum_f SDI_{f,t} \times \left(Buy_{f,s,t} - Sell_{f,s,t}\right)}{\sum_f \left(Buy_{f,s,t} + Sell_{f,s,t}\right)}$$ where the buy order, $Buy_{f,s,t}$, is defined as the percentage increase in the invested value of fund *f* in stock *s* in quarter *t* (zero if no purchase). Similarly, the sell order, $Sell_{f,s,t}$, is defined as the percentage decrease in the invested value of fund *f* in stock *s* (zero if no sales).
AWOI	Activeness-weighted order imbalance, calculated as follows: $$AWOI_{s,t} = \frac{\sum_f Fund\ Activeness_{f,t} \times \left(Buy_{f,s,t} - Sell_{f,s,t}\right)}{\sum_f \left(Buy_{f,s,t} + Sell_{f,s,t}\right)}$$ where $Buy_{f,s,t}$ and $Sell_{f,s,t}$ are the same as in *SDI_OI*.

DWOI	Distance-weighted order imbalance, calculated as follows: $$DWOI_{s,t} = \frac{\sum_f Holding\ Distance_{f,t} \times \left(Buy_{f,s,t} - Sell_{f,s,t}\right)}{\sum_f \left(Buy_{f,s,t} + Sell_{f,s,t}\right)}$$ where $Buy_{f,s,t}$ and $Sell_{f,s,t}$ are the same as in *SDI_OI*.
SDI_TI	SDI-weighted trading intensity, calculated as follows: $$SDI\_TI_{s,t} = \frac{\sum_f SDI_{f,t} \times \left(Buy_{f,s,t} + Sell_{f,s,t}\right)}{\sum_f \left(Buy_{f,s,t} + Sell_{f,s,t}\right)}$$ where $Buy_{f,s,t}$ and $Sell_{f,s,t}$ are the same as in *SDI_OI*.
AWTI	Activeness-weighted trading intensity, calculated as follows: $$AWTI_{s,t} = \frac{\sum_f Fund\ Activeness_{f,t} \times \left(Buy_{f,s,t} + Sell_{f,s,t}\right)}{\sum_f \left(Buy_{f,s,t} + Sell_{f,s,t}\right)}$$ where $Buy_{f,s,t}$ and $Sell_{f,s,t}$ are the same as in *SDI_OI*.
DWTI	Distance-weighted trading intensity, calculated as follows: $$DWTI_{s,t} = \frac{\sum_f Holding\ Distance_{f,t} \times \left(Buy_{f,s,t} + Sell_{f,s,t}\right)}{\sum_f \left(Buy_{f,s,t} + Sell_{f,s,t}\right)}$$ where $Buy_{f,s,t}$ and $Sell_{f,s,t}$ are the same as in *SDI_OI*.
Size ($million)	Market capitalization, in millions.
BM	Book-to-market ratio.
Prior 1-Month Return	Stock return in the prior month.
Prior 12-Month Return	Stock cumulative return over the prior 12 months.
Investment	Annual growth rate of total assets (Cooper, Gulen, and Schill 2008).
Profitability	Annual operating profitability (Fama and French 2006).
Illiquidity	Daily ratio of the absolute stock returns to its dollar volume (in millions), following Amihud (2002). Monthly values are first computed and then averaged within each quarter.
Analyst Coverage	Number of analysts issuing valid EPS forecasts. Monthly values are first computed and then averaged within each quarter.
Idiosyncratic Volatility	The standard deviation of residuals from Fama-French three factor regressions estimated using daily returns over the past month. Monthly values are first computed and then averaged within each quarter.
ASVI	Abnormal Google search volume, which is defined as the difference between the logarithm of one plus the average search volume in the current quarter and that over the previous eight quarters, following Da, Engelberg, and Gao (2011).
Geographic Proximity	The holding-value-weighted average of a same-state indicator between the firm and its mutual fund shareholders. The indicator equals one if the firm and the mutual fund are located in the same state, and zero otherwise.
Complexity	Proportion of financial-complexity words in the firm's 10-K filing, following Loughran and McDonald (2024).
Firm Change	Euclidean embedding distance between the business description (Item 1) in stock's current 10-K filing and that in its prior 10-K filing.
New AI Topics	A dummy equal to 1 if AI-related keywords appear in firm s's current 10-K but do not appear in its prior 10-K. AI-related keywords are generated using ChatGPT.
New ESG Topics	A dummy equal to 1 if ESG-related keywords appear in firm s's current 10-K but do not appear in its prior 10-K. ESG-related keywords are generated using ChatGPT.
New Investment Opportunities	A dummy equal to 1 if keywords representing emerging investment opportunities appear in firm s's current 10-K but do not appear in its prior 10-K. The keywords are identified from 43 comprehensive categories using machine learning, following Basu, Ma, and Briscoe-Tran (2022).
SUE	Standardized unexpected earnings, constructed using analyst earnings forecasts and realized earnings from IBES. Monthly values are first computed and then averaged within each quarter.

Table 1. Summary Statistics (Fund-Level Variables)

This table presents the number of observation, mean, standard deviation, 25th percentile, median, and 75th percentile of fund characteristics at the quarterly frequency. *Holding Distance* is the active weight- and industry-adjusted firm distance to the fund. *Fund Activeness* is the average rank of *active weight* (Doshi, Elkamhi, and Simutin 2009) and *return gap* (Kacperczyk, Sialm, and Zheng 2008). See Section 2.2 for details. *TNA ($million)* is the total net asset under management at the end of the quarter. *Month Age* is the number of months since the fund's inception date. *Expenses (%)* is the ratio of total investment that shareholders pay for the fund's operating expenses. *Turnover* is the minimum of aggregated sales or aggregated purchases of securities, divided by the fund's average 12-month TNA. *Prior 1Q Return (%)* is the fund's cumulative return over the past 3 months. *Prior 1Y Return (%)* is the fund's cumulative return over the past 12 months. *Prior 1Q Flow (%) is* the fund's net flow in a given quarter, calculated as the difference between the current quarter's TNA and the prior quarter's TNA multiplied by one plus the fund's return over the period, scaled by the prior quarter's TNA. *Return Volatility (%)* is the fund return volatility, measured as the standard deviation of monthly fund returns over the prior 24 months, requiring at least 18 months of data. All variables are winsorized at the 1st and 99th percentiles, except *Holding Distance* and *Fund Activeness.* The sample period is from 2011Q1 to 2023Q4.

	N	Mean	Std	P25	Median	P75
Holding Distance	73064	0.009	0.015	0.000	0.007	0.017
Fund Activeness	73064	0.512	0.191	0.381	0.508	0.638
TNA ($ million)	73064	2148.93	4815.53	129.70	497.50	1708.50
Month Age	73064	255.72	169.98	145.00	227.00	318.00
Turnover	73064	0.600	0.545	0.250	0.450	0.760
Expense (%)	73064	1.012	0.326	0.828	1.000	1.200
Prior 1Q Flow (%)	73064	-1.140	10.000	-4.440	-2.024	0.562
Prior 1Q Return (%)	73064	2.860	9.046	-0.730	3.621	7.887
Prior 1Y Return (%)	73064	11.381	18.317	-0.681	11.172	21.687
Return Volatility (%)	73064	0.245	0.174	0.112	0.201	0.330

Table 2. Economic Drivers of Pairwise Fund-firm Distance

This table reports panel regression results examining how pairwise fund-firm distance is associated with its potential determinants.

$$Distance_{f,s,t} = \alpha + \beta_1 \times Lagged\ Distance_{f,s,t} + \beta_2 \times Fund\ Change_{f,t} + \beta_3 \times Firm\ Change_{s,t} + \beta_4 \times NewInvOpp_{s,t} + \boldsymbol{\Gamma_f}' \times \boldsymbol{M}_{f,t-1} + \boldsymbol{\Gamma_s}' \times \boldsymbol{M}_{s,t-1} + \varepsilon_{f,s,t}$$

where $Distance_{f,s,t}$ refers to the pairwise semantic distance between the prospectus of fund f and the 10-K of firm s in quarter t. $Fund\ Change_{f,t}$ and $Firm\ Change_{s,t}$ refer to the textual changes in the fund prospectus and firm 10-K filings, respectively. $NewInvOpp_{s,t}$ refers to the dummy indicators for the emergence of AI, ESG, and machine learning-identified new investment opportunities. Specifically, $NewInvOpp_{s,t}$ equals 1 if keywords, representing emerging investment opportunities related to AI, ESG, and machine-learning-identified 43 comprehensive categories, appear in the firm s's current 10-K. AI and ESG-related keywords are generated using ChatGPT, and investment opportunity-related keywords are from Basu, Ma, and Briscoe-Tran (2022). We focus on our fund sample and S&P 500 stocks. Here t refers to filing-update dates. The vector $\boldsymbol{M_{f,t-1}}$ stacks a set of lagged fund-level control variables, including the total net assets, age, turnover, expense ratio, past flows, past returns, return volatility. The vector $\boldsymbol{M_{s,t-1}}$ stacks a set of lagged firm-level control variables, including market value, book-to-market ratio, past 1- and 12-month returns, growth rate of total assets (Investment), operating profitability, illiquidity, and analyst coverage. Fund, Stock and Quarter fixed effects are included in Column (1) to Column (5). Fund × Stock and Quarter fixed effects are included in Column (6). Robust standard errors are clustered at the fund × stock level. T-statistics are presented in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	***Distance $_{f,s,t}$***					
Lagged Distance $_{f,s,t}$	0.564***	0.564***	0.564***	0.564***	0.564***	0.322***
	(899.93)	(900.28)	(901.30)	(900.93)	(900.43)	(396.69)
Fund Change $_{f,t}$	0.001***				0.001***	0.001***
	(3.88)				(3.89)	(10.74)
Firm Change $_{s,t}$	0.001***				0.000***	0.000
	(12.94)				(5.13)	(0.24)
New AI Topics $_{s,t}$		0.002***			0.002***	0.001***
		(23.23)			(20.66)	(14.74)
New ESG Topics $_{s,t}$			0.003***		0.003***	0.002***
			(26.45)		(24.06)	(15.91)
New Investment Opportunities $_{s,t}$				0.001***	0.000***	0.001***
				(15.93)	(8.13)	(17.00)
Fund, Firm, Quarter FE	Y	Y	Y	Y	Y	
Fund × Firm, Quarter FE						Y
Controls	Y	Y	Y	Y	Y	Y
N	5511237	5512413	5512413	5512413	5511237	5375974
Adj. R^2	0.757	0.757	0.757	0.757	0.757	0.759

Table 3. Fund Performance Under Portfolio Sorting

This table presents the value-weighted future returns of mutual funds sorted into 5 × 5 portfolios based on *Holding Distance* and *Fund Activeness*. We independently sort funds into quintiles based on the most recent value of each measure and rebalance the portfolios quarterly. Panel A reports the gross (before expense fees) alphas estimated using Carhart (1997) four-factor model. Panel B reports the net (after expense fees) alphas estimated using Carhart (1997) four-factor model. The final row of each panel shows the performance spreads between funds in the highest and lowest *Holding Distance* quintiles across different quintiles of *Fund Activeness*, as well as the long-short portfolio between funds with highest and lowest *Fund Activeness*. Newey-West (1987) t-statistics with four lags are reported in parentheses. All returns are expressed in percentage points. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

Panel A. Fund Carhart 4-factor alphas (Before Fee, %/month)

	Low Active	Active 2	Active 3	Active 4	High Active	HML	T
Low Distance	-0.22***	-0.10	-0.11**	-0.06	-0.08	0.14	[1.56]
Distance 2	-0.16***	-0.04	-0.13**	-0.05	-0.14	0.02	[0.23]
Distance 3	-0.10*	-0.11*	0.02	0.04	-0.02	0.08	[0.83]
Distance 4	-0.18	-0.11*	-0.08	0.12**	0.02	0.20	[1.44]
High Distance	-0.37***	-0.12	-0.10*	0.04	0.11	0.48***	[3.15]
HML	-0.16	-0.02	0.01	0.09	0.19**	0.34**	
T	[-1.15]	[-0.18]	[0.10]	[0.92]	[2.40]	[2.06]	

Panel B. Fund Carhart 4-factor alphas (After Fee, %/month)

	Low Active	Active 2	Active 3	Active 4	High Active	HML	T
Low Distance	-0.28***	-0.17**	-0.18***	-0.13	-0.15*	0.14	[1.52]
Distance 2	-0.22***	-0.09*	-0.19***	-0.11	-0.20**	0.02	[0.17]
Distance 3	-0.16***	-0.16**	-0.03	-0.02	-0.07	0.09	[0.86]
Distance 4	-0.23*	-0.17***	-0.14**	0.06	-0.03	0.20	[1.46]
High Distance	-0.44***	-0.18**	-0.17***	-0.03	0.05	0.48***	[3.16]
HML	-0.16	-0.01	0.01	0.10	0.19**	0.35**	
T	[-1.15]	[-0.15]	[0.19]	[0.98]	[2.47]	[2.10]	

Table 4. Fund Performance under Fama-MacBeth's (1973) Test

This table presents the results from quarterly Fama-MacBeth (1973) regressions of future fund performance on *Holding Distance*, *Fund Activeness* and their interaction *SDI*.

$$Perf_{f,t} = \alpha + \beta_1 \times Holding\ Distance_{f,t-1} + \beta_2 \times Fund\ Activeness_{f,t-1} + \beta_3 \times SDI_{f,t-1} + \boldsymbol{\Gamma'} \times \boldsymbol{M_{f,t-1}} + \varepsilon_{f,t}$$

where $Perf_{f,t}$ refers to the performance of fund f in quarter t. $Holding\ Distance_{f,t-1}$ is the active weight- and industry-adjusted fund distance to holding firms in quarter $t-1$, and $Fund\ Activeness_{f,t-1}$ measures the level of fund activeness, defined as the average rank of *active weight* and *return gap* of fund f in quarter $t-1$. The vector $\boldsymbol{M_{f,t-1}}$ stacks a set of lagged fund characteristics as controls variables, including total net assets (*TNA*), age, turnover, expense ratio, past flows, past returns, and return volatility. Fund performance measures include four-factor-adjusted alphas (both before and after expense fees) as well as the value created by fund managers, measured as *value-added* following Berk and van Binsbergen (2015). Control variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized to have a mean of zero and a standard deviation of one. Newey-West (1987) t-statistics with three lags are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable	***Carhart 4-factor Alphas $_{f,t}$ (%/quarter)***						***Value Added $_{f,t}$ ($million/quarter)***		
	Before Fee			***After Fee***					
Holding Distance $_{f,t-1}$	0.005		-0.046	0.004		-0.047	-0.008		-1.441
	(0.12)		(-1.06)	(0.10)		(-1.07)	(-0.01)		(-1.21)
Fund Activeness $_{f,t-1}$		0.092**	0.096**		0.091**	0.096**		1.890	2.170
		(2.55)	(2.61)		(2.52)	(2.59)		(1.38)	(1.52)
SDI $_{f,t-1}$			0.051**			0.051**			1.838***
			(2.49)			(2.49)			(2.75)
Log TNA $_{f,t-1}$	-0.009	-0.006	-0.008	-0.008	-0.005	-0.007	-6.602**	-6.528*	-6.513*
	(-0.32)	(-0.21)	(-0.28)	(-0.29)	(-0.19)	(-0.26)	(-2.04)	(-2.01)	(-2.01)
Month Age $_{f,t-1}$	-0.009	-0.012	-0.008	-0.010	-0.014	-0.010	-0.016	-0.089	0.033
	(-0.65)	(-0.91)	(-0.66)	(-0.77)	(-1.04)	(-0.79)	(-0.03)	(-0.14)	(0.05)
Turnover $_{f,t-1}$	-0.069*	-0.067*	-0.068*	-0.067*	-0.066*	-0.067*	-0.439	-0.427	-0.467
	(-1.96)	(-1.94)	(-1.97)	(-1.92)	(-1.89)	(-1.93)	(-1.12)	(-1.22)	(-1.30)
Expense $_{f,t-1}$	-0.009	-0.016	-0.016	-0.133***	-0.140***	-0.141***	-2.291**	-2.435**	-2.414**
	(-0.21)	(-0.39)	(-0.39)	(-3.20)	(-3.43)	(-3.44)	(-2.27)	(-2.38)	(-2.36)
Prior 1Q Flow $_{f,t-1}$	-0.001	-0.003	-0.002	-0.001	-0.003	-0.002	-0.267	-0.312	-0.339
	(-0.05)	(-0.13)	(-0.12)	(-0.05)	(-0.13)	(-0.11)	(-0.62)	(-0.69)	(-0.78)
Prior 1Q Return $_{f,t-1}$	0.296	0.297	0.301	0.295	0.296	0.300	7.297	7.787	7.580
	(0.98)	(0.96)	(1.01)	(0.97)	(0.96)	(1.00)	(1.30)	(1.37)	(1.37)
Prior 1Y Return $_{f,t-1}$	0.308	0.291	0.299	0.306	0.290	0.298	4.936	4.163	4.246
	(1.54)	(1.38)	(1.48)	(1.54)	(1.38)	(1.48)	(0.88)	(0.72)	(0.76)
Return Volatility $_{f,t-1}$	-0.310*	-0.313*	-0.331*	-0.312*	-0.315*	-0.332*	-1.528	-1.621	-1.904
	(-1.85)	(-1.87)	(-1.95)	(-1.86)	(-1.89)	(-1.97)	(-0.49)	(-0.52)	(-0.58)
N	73064	73064	73064	73064	73064	73064	73064	73064	73064
Adj. R^2	0.159	0.156	0.167	0.161	0.157	0.168	0.070	0.070	0.073

Table 5. Sub-sample Analysis on Fund Performance

This table presents the results from quarterly Fama-MacBeth (1973) regressions of future fund performance on *Holding Distance*, *Fund Activeness* and their interaction *SDI,* estimated across subsamples defined by different fund characteristics.

$$Perf_{f,t} = \alpha + \beta_1 \times Holding\ Distance_{f,t-1} + \beta_2 \times Fund\ Activeness_{f,t-1} + \beta_3 \times SDI_{f,t-1} + \boldsymbol{\Gamma'} \times \boldsymbol{M_{f,t-1}} + \varepsilon_{f,t}$$

where $Perf_{f,t}$ refers to the performance of fund f in quarter t, $Holding\ Distance_{f,t-1}$ is the active weight-weighted industry-adjusted fund distance to holding firms in quarter $t-1$, and $Fund\ Activeness_{f,t-1}$ describes the level of fund activeness, which is average rank of *active weight* and *return gap* of fund f in quarter $t-1$. The vector $\boldsymbol{M_{f,t-1}}$ stacks a set of lagged fund characteristics as firm-level control variables, as listed in Table 4. Panel A splits the sample into two halves based on the median of fund characteristics (*TNA, Turnover, and Expense*). Panel B splits the sample into two halves based on the median of weighted exposure of fund holdings to *New AI Topics*, *New ESG Topics*, and *New Investment Opportunities*. The weighted exposure is computed as the sum of portfolio weights assigned to holdings classified under each category—that is, those for which *New #* equals 1. *C*ontrol variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized to have a mean of zero and a standard deviation of one. Newey-West (1987) t-statistics with three lags are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	***Carhart 4-factor Before-fee Alphas $_{f,t}$ (%/quarter)***					
Panel A: Funds Holdings with	***New AI Topics***		***New ESG Topics***		***New Investment Opportunity***	
	Below Median	*Above Median*	*Below Median*	*Above Median*	*Below Median*	*Above Median*
Holding Distance $_{f,t-1}$	-0.017	-0.075*	-0.034	-0.053	-0.055	-0.022
	(-0.37)	(-1.72)	(-0.63)	(-1.39)	(-1.00)	(-0.51)
Fund Activeness $_{f,t-1}$	0.084**	0.085**	0.127***	0.057	0.124***	0.068
	(2.17)	(2.11)	(2.93)	(1.29)	(3.09)	(1.60)
SDI $_{f,t-1}$	0.019	0.067***	0.068**	0.041**	0.019	0.066**
	(0.74)	(2.76)	(2.17)	(2.21)	(0.77)	(2.44)
Controls	Y	Y	Y	Y	Y	Y
N	38926	34138	36936	36128	36321	36743
Adj. R^2	0.192	0.146	0.170	0.188	0.157	0.196
Panel B: Funds with	***TNA***		***Turnover***		***Expense***	
	Below Median	*Above Median*	*Below Median*	*Above Median*	*Below Median*	*Above Median*
Holding Distance $_{f,t-1}$	-0.074	-0.020	-0.059	-0.027	-0.042	-0.043
	(-1.64)	(-0.38)	(-1.31)	(-0.54)	(-0.76)	(-0.98)
Fund Activeness $_{f,t-1}$	0.078**	0.125***	0.070*	0.108**	0.104**	0.092**
	(2.12)	(2.73)	(1.78)	(2.65)	(2.18)	(2.58)
SDI $_{f,t-1}$	0.034	0.098***	0.069**	0.015	0.028	0.054**
	(1.47)	(3.12)	(2.59)	(0.47)	(1.02)	(2.23)
Controls	Y	Y	Y	Y	Y	Y
N	36001	37063	36918	36146	29632	43432
Adj. R^2	0.148	0.207	0.178	0.167	0.209	0.159

Table 6. Stock-level Return Predictive Power

This table presents the results from quarterly Fama-MacBeth (1973) regressions of future stock performance on *SDI_OI*.

$$DGTW_{s,t} = \alpha + \beta \times SDI\_OI_{s,t-1} + \boldsymbol{\Gamma'} \times \boldsymbol{M_{s,t-1}} + \varepsilon_{s,t}$$

where $DGTW_{s,t}$ denotes the performance of stock s in quarter t, measured using characteristic-adjusted returns following Daniel, Grinblatt, Titman, and Wermers (1997). $SDI\_OI_{s,t-1}$ is the *SDI*-weighted order imbalance for stock s in quarter $t-1$. $High\ SDI\_OI_{s,t-1}$ equals to 1 if SDI_OI is above the cross-sectional median in quarter t, and 0 otherwise. We also include the measure of $AWOI_{s,t-1}$ in Columns (3)-(6), which denotes the fund-activeness-weighted order imbalance for stock s in quarter $t-1$. The vector $\boldsymbol{M_{s,t-1}}$ stacks a set of lagged stock characteristics as control variables, including market capitalization (*Size*), book-to-market ratio (*BM*), past returns, asset growth (*Investment*), operating profitability, illiquidity, and analyst coverage. Additional controls include *Idiosyncratic Volatility* (defined as the standard deviation of residuals from Fama–French three-factor regressions estimated using daily returns), *ASVI* (defined as the change in log-transformed Google search volume from the previous quarters), *Geographic Proximity* (defined the holding-value-weighted average of a same-state indicator, which is equal to one if the firm and the mutual fund are located in the same state, and zero otherwise) and *Complexity* (defined as the proportion of financial-complexity words in the firm's 10-K filing, following Loughran and McDonald 2024). Control variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized to have a mean of zero and a standard deviation of one. Newey-West (1987) t-statistics with three lags are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	***DGTW-adjusted Returns $_{s,t}$ (%/quarter)***					
SDI_OI $_{s,t-1}$	0.313***		0.273**		0.381***	
	(2.71)		(2.45)		(2.81)	
High SDI_OI $_{s,t-1}$		0.328***		0.281**		0.376***
		(3.14)		(2.71)		(2.89)
AWOI $_{s,t-1}$					-0.244	-0.193
					(-1.60)	(-1.32)
Log Size $_{s,t-1}$	0.099	0.099	-0.085	-0.088	-0.113	-0.113
	(0.54)	(0.54)	(-0.40)	(-0.40)	(-0.55)	(-0.54)
Log BM $_{s,t-1}$	-0.154	-0.153	-0.190	-0.189	-0.177	-0.176
	(-0.68)	(-0.68)	(-0.95)	(-0.95)	(-0.89)	(-0.89)
Prior 1-Month Return $_{s,t-1}$	-0.293	-0.295	-0.302	-0.304	-0.307	-0.311
	(-1.12)	(-1.12)	(-1.09)	(-1.09)	(-1.11)	(-1.13)
Prior 12-Month Return $_{s,t-1}$	0.257	0.271	0.277	0.289	0.357	0.356
	(1.08)	(1.14)	(1.15)	(1.19)	(1.63)	(1.63)
Investment $_{s,t-1}$	-0.255	-0.260	-0.245	-0.249	-0.240	-0.245
	(-1.44)	(-1.46)	(-1.56)	(-1.58)	(-1.53)	(-1.56)
Profitability $_{s,t-1}$	0.195	0.193	0.137	0.137	0.134	0.133
	(1.24)	(1.21)	(1.02)	(1.01)	(0.99)	(0.98)
Illiquidity $_{s,t-1}$	-0.358	-0.355	-0.285	-0.283	-0.275	-0.276
	(-1.24)	(-1.23)	(-0.97)	(-0.96)	(-0.93)	(-0.94)
Log Analyst Coverage $_{s,t-1}$	-0.320	-0.322	-0.213	-0.213	-0.204	-0.205
	(-1.18)	(-1.18)	(-0.91)	(-0.91)	(-0.88)	(-0.88)
Idiosyncratic Volatility $_{s,t-1}$			-0.518	-0.520	-0.548	-0.541
			(-1.29)	(-1.28)	(-1.37)	(-1.35)
ASVI $_{s,t-1}$			0.136	0.135	0.140	0.138
			(1.16)	(1.15)	(1.20)	(1.18)
Geographic Proximity $_{s,t-1}$			-0.105	-0.104	-0.113	-0.110
			(-1.15)	(-1.12)	(-1.22)	(-1.18)
Complexity $_{s,t-1}$			-0.243	-0.245	-0.246	-0.246
			(-1.41)	(-1.43)	(-1.44)	(-1.44)
N	79710	79710	78949	78949	78949	78949
Adj. R^2	0.025	0.025	0.033	0.032	0.033	0.033

Table 7. Distant Investments vs. Lazy Price

This table examines how the cumulative abnormal returns of *Changers* and *Non-Changers* evolve following the release of 10-K filings. To study the effect of distant investment, we estimate the following quarterly Fama-MacBeth (1973) regression:

$$Cum\ DGTW_{s,t+n} = \alpha + \beta_1 \times Changer_{s,t} + \beta_2 \times High\ SDI\_OI_{s,t} + \beta_3 \times Changer_{s,t} \times High\ SDI\_OI_{s,t} + \boldsymbol{\Gamma'} \times \boldsymbol{M_{s,t-1}} + \varepsilon_{s,t}$$

where $Cum\ DGTW_{s,t+n}$ denotes the cumulative DGTW-adjusted return of stock s over the n-quarter horizon following the 10-K filing release quarter t. $Changer_{s,t}$ equals to 1 if the *Firm Change* between stock s's current 10-K filing and its prior-year 10-K filing is above the cross-sectional median in quarter t; and 0 otherwise. $High\ SDI\_OI_{s,t-1}$ equals to 1 if SDI_OI is above the cross-sectional median in quarter t, and 0 otherwise. The release quarter refers to the quarter-end corresponding to the disclosure of the firm's 10-K filing. The vector $\boldsymbol{M_{s,t-1}}$ stacks a set of control variables used in Cohen et al. (2020). $SUE_{s,t-1}$ denotes standardized unexpected earnings constructed using analyst earnings forecasts and realized earnings from IBES. Control variables are winsorized at the 1st and 99th percentiles. Newey-West (1987) t-statistics three lags are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	***DGTW*** $_{s,\,t+1}$			***DGTW*** $_{s,t+2}$		
Changers $_{s,t}$	-0.014***	-0.031***	-0.030***	-0.028***	-0.047***	-0.046***
	(-2.76)	(-2.83)	(-2.76)	(-2.92)	(-3.74)	(-3.59)
High SDI OI $_{s,t}$		-0.007	-0.007		-0.009	-0.008
		(-0.86)	(-0.79)		(-0.79)	(-0.75)
Changers $_{s,t}$ × *High SDI OI* $_{s,t}$		0.027**	0.027**		0.033***	0.034***
		(2.06)	(2.07)		(2.97)	(2.92)
SUE $_{s,t}$			0.009**			0.015**
			(2.26)			(2.24)
Controls	Y	Y	Y	Y	Y	Y
N	30390	30390	30390	30200	30200	30200
Adj. R^2	0.047	0.073	0.088	0.034	0.055	0.067

Table 8. Trading Intensity vs. Information Shares of Return Variance

This table presents the results from yearly panel regressions of different components of stock return variance on the trading intensity of distant investment as follows:

$$InfoShare_{s,t} = \alpha + \beta_1 \times SDI\_TI_{s,t} + \boldsymbol{\Gamma'} \times \boldsymbol{M_{s,t-1}} + \varepsilon_{s,t}$$

where $InfoShare_{s,t}$ represents the different information components of stock-return for stock s in year t. Following Brogaard et al. (2021), we decompose the return variance using daily stock returns each year. The variance is decomposed into market-wide information, firm-specific information revealed from private and public sources, and noise. *SDI_TI* are defined as *SDI*-weighted summation of buy and sell orders scaled by the summation of total buy and sell orders. Specifically,

$$SDI\_TI_{s,t} = \frac{\sum_f SDI_{f,t} \times (Buy_{f,s,t} + Sell_{f,s,t})}{\sum_f (Buy_{f,s,t} + Sell_{f,s,t})}$$

where $Buy_{f,s,t}$ and $Sell_{f,s,t}$ are defined in the same way as in order imbalance. The vector $\boldsymbol{M_{s,t-1}}$ stacks a set of lagged stock characteristics as listed in Table 6. To match the dependent variable, *SDI_TI* and controls are calculated as the average quarterly values within each year. Year fixed effects are included. Control variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized to have a mean of zero and a standard deviation of one. Robust standard errors are clustered at the stock level. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

	(1)	(2)	(3)	(4)
Dependent Variable	***Firm-specific Private Info*** $_{s,t}$	***Firm-specific Public Info*** $_{s,t}$	***Market Info*** $_{s,t}$	***Noise*** $_{s,t}$
SDI_TI $_{s,t}$	0.019***	-0.004	-0.013***	-0.003**
	(5.31)	(-1.39)	(-4.53)	(-2.16)
Controls	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
N	18671	18671	18671	18671
Adj. R^2	0.181	0.156	0.355	0.147

Table 9. Fund Performance Tests Using Fund- or Firm-specific Text Alone

This table presents the results from quarterly Fama-MacBeth (1973) regressions of future fund performance on fund- or firm-specific textual measures, *Fund Activeness* and their interactions.

$$Perf_{f,t} = \alpha + \beta_1 \times TextInfo_{f,t-1} + \beta_2 \times Fund\ Activeness_{f,t-1} + \beta_3 \times TextInfo_{f,t-1} \times Fund\ Activeness_{f,t-1} + \boldsymbol{\Gamma'} \times \boldsymbol{M_{f,t-1}} + \varepsilon_{f,t}$$

where $Perf_{f,t}$ refers to the performance of fund f in quarter t. $Fund\ Activeness_{f,t-1}$ measures the level of fund activeness, which is the average rank of *active weight* and *return gap* of fund f in quarter $t-1$. $TextInfo_{f,t-1}$ denotes one of the fund- or firm-specific textual measures. Fund-specific information includes *Fund Changes* as the semantic distance between the fund's current and prior prospectuses. Firm-specific information includes *Firm Changes*, *New AI*, *New ESG*, and *New ML-identified Investment Opportunities* (as defined in Section 3.1), along with *Complexity* (proxied by the proportion of financial-complexity words in the firm's 10-K filing, following Loughran and McDonald 2024). The vector $\boldsymbol{M_{f,t-1}}$ stacks a set of lagged fund characteristics as listed in Table 4. Fund performance measures include four-factor-adjusted before-fees alphas (Panel A), after-fees alphas (Panel B), and *value-added* (Panel C). Control variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized to have a mean of zero and a standard deviation of one. Newey-West (1987) t-statistics with three lags are reported in parentheses. ***, **, * denote statistical significance at the 1%, 5%, or 10% level, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
	Fund-specific Info	***Firm-specific Info***				
	Fund Change	***Firm Change***	***New AI***	***New ESG***	***New Investment Opportunity***	***Complexity***
Panel A: Dependent Var. = Carhart 4-factor Alphas $_{f,t}$ (Before Fee)						
TextInfo $_{f,t-1}$	-0.080	-0.072*	0.004	-0.068*	0.030	0.029
	(-1.19)	(-1.98)	(0.13)	(-1.83)	(0.71)	(0.81)
Fund Activeness $_{f,t-1}$	0.149**	0.115***	0.106***	0.117***	0.103***	0.105***
	(2.54)	(3.06)	(2.97)	(3.25)	(2.78)	(2.85)
Interaction	0.133	0.023	-0.011	-0.010	-0.021	-0.015
	(0.90)	(0.90)	(-0.47)	(-0.27)	(-0.80)	(-0.54)
Controls	Y	Y	Y	Y	Y	Y
N	70363	70363	70363	70363	70363	70363
Adj. R^2	0.174	0.178	0.175	0.179	0.179	0.179
Panel B: Dependent Var. = Carhart 4-factor Alphas $_{f,t}$ (After Fee)						
TextInfo $_{f,t-1}$	-0.080	-0.071*	0.003	-0.068*	0.030	0.030
	(-1.19)	(-1.98)	(0.10)	(-1.84)	(0.71)	(0.82)
Fund Activeness $_{f,t-1}$	0.148**	0.114***	0.105***	0.116***	0.102***	0.104***
	(2.52)	(3.04)	(2.94)	(3.22)	(2.75)	(2.83)
Interaction	0.134	0.024	-0.012	-0.009	-0.021	-0.014
	(0.91)	(0.92)	(-0.50)	(-0.26)	(-0.79)	(-0.51)
Controls	Y	Y	Y	Y	Y	Y
N	70363	70363	70363	70363	70363	70363
Adj. R^2	0.175	0.179	0.176	0.180	0.180	0.181
Panel C: Dependent Var. = Value Added $_{f,t}$						
TextInfo $_{f,t-1}$	-2.335*	-1.562*	-0.491	-0.774	0.058	0.002
	(-1.90)	(-1.85)	(-0.61)	(-1.11)	(0.08)	(0.00)
Fund Activeness $_{f,t-1}$	2.520*	2.158	2.060	1.955	1.844	1.800
	(1.69)	(1.53)	(1.46)	(1.36)	(1.32)	(1.11)
Interaction	2.124	1.220*	0.874	-0.185	-0.231	-0.142
	(0.90)	(1.76)	(1.29)	(-0.27)	(-0.34)	(-0.16)
Controls	Y	Y	Y	Y	Y	Y
N	70363	70363	70363	70363	70363	70363
Adj. R^2	0.091	0.092	0.091	0.091	0.091	0.092

Table 10. Flow-driven Distant Investment by Low-Skilled Managers

Panel A. Fund Trading and Distant Investment

This panel presents results from panel regressions examining how mutual fund trading behavior relates to distant investment.

$$\begin{aligned} Buy\ (Sell)_{f,s,t} = {} & \alpha + \beta_0 \times Distance_{f,s,t-1} + \beta_1 \times Distance_{f,s,t-1} \times LowActive_{s,t-1} \\ & + \beta_2 \times Distance_{f,s,t-1} \times HighActive_{s,t-1} + \cdots \\ & + \gamma_1 \times Distance_{f,s,t-1} \times ASVI_{s,t-1} \times LowActive_{s,t-1} \\ & + \gamma_2 \times Distance_{f,s,t-1} \times ASVI_{s,t-1} \times HighActive_{s,t-1} + \delta_{s,t} + \delta_{f,t} \\ & + \varepsilon_{f,s,t} \end{aligned}$$

where $Distance_{f,s,t-1}$ refers to the pairwise semantic distance between the prospectus of fund f and the 10-K of firm s in quarter $t-1$. $HighActive_{f,t-1}$ and $LowActive_{f,t-1}$ are dummy indicators equal to 1 for funds in the top 20% and bottom 20% of the cross-sectional *Fund Activeness* distribution in quarter $t-1$, respectively. $ASVI_{s,t-1}$ measures the abnormal investor attention for stock s in quarter $t-1$, calculated as the change in log-transformed Google search volume from the previous quarters. Columns (1)–(2) focus on fund purchases, where the dependent variable $Buy_{f,s,t}$ equals one if the fund increases its position in the stock and zero otherwise. Columns (3)–(4) examine fund sales of existing positions, where the dependent variable $Sell_{f,s,t}$ equals one if the fund decreases its position in the stock and zero otherwise. Fund × Quarter and Stock × Quarter fixed effects are included. Robust standard errors are clustered at the fund and stock level. T-statistics are presented in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Panel A Fund Trading and Distant Investment

	(1)	(2)	(3)	(4)
Dependent Variable	***Buy*** $_{f, s, t}$		***Sells*** $_{f, s, t}$	
Distance $_{f, s, t-1}$	-0.021**	-0.019**	0.024**	0.024**
	(-2.15)	(-1.97)	(2.29)	(2.19)
Distance $_{f, s, t-1}$ × *Low Active* $_{f, t-1}$	-0.002	-0.007	-0.027**	-0.024*
	(-0.24)	(-0.69)	(-2.16)	(-1.95)
Distance $_{f, s, t-1}$ × *High Active* $_{f, t-1}$	0.032***	0.030***	-0.013	-0.010
	(3.20)	(2.98)	(-1.04)	(-0.86)
Distance $_{f, s, t-1}$ × *ASVI* $_{s,t-1}$		-0.031		0.011
		(-1.61)		(0.56)
Low Active $_{f, t-1}$ × *ASVI* $_{s,t-1}$		-0.074***		0.055**
		(-2.92)		(2.02)
High Active $_{f, t-1}$ × *ASVI* $_{s,t-1}$		-0.028		0.039
		(-1.04)		(1.36)
Distance $_{f, s, t-1}$ × *ASVI* $_{s,t-1}$ × *Low Active* $_{f, t-1}$		0.064***		-0.048**
		(2.99)		(-2.10)
Distance $_{f, s, t-1}$ × *ASVI* $_{s,t-1}$ × *High Active* $_{f, t-1}$		0.027		-0.035
		(1.23)		(-1.47)
Fund × Quarter, Stock × Quarter FE	Y	Y	Y	Y
N	5539658	5538872	5539658	5538872
Adj. R^2	0.239	0.239	0.308	0.308

Panel B Flow-Performance Sensitivity

This panel reports quarterly Fama-MacBeth (1973) regressions of future fund flows on lagged fund returns and *Holding Distance* based on high and low levels of *Fund Activeness.*

$$Flow_{f,t} = \alpha + \beta_1 \times Ret_{f,t-1} + \beta_2 \times Holding\ Distance_{f,t-1} + \beta_3 \times Ret_{f,t-1} \times Holding\ Distance_{f,t-1} + \boldsymbol{\Gamma'} \times \boldsymbol{M_{f,t-1}} + \varepsilon_{f,t}$$

where $Flow_{f,t}$ refers to the flow of fund f in quarter t, and $Ret_{f,t-1}$ refer to one-quarter lagged fund returns. $Holding\ Distance_{f,t-1}$ is the active weight- and industry-adjusted fund distance to holding firms in quarter $t-1$. $HighActive_{f,t-1}$ and $LowActive_{f,t-1}$ are dummy indicators equal to 1 for funds in the top 20% and bottom 20% of the cross-sectional *Fund Activeness* distribution in quarter $t-1$, respectively. The vector $\boldsymbol{M_{f,t-1}}$ stacks a set of lagged fund characteristics as listed in Table 4. All variables are winsorized at the 1st and 99th percentiles. Independent variables are standardized to have a mean of zero and a standard deviation of one. Newey-West (1987) t-statistics with three lags are reported in parentheses. ***, **, * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

Dependent Variable	**(1)**	**(2)**	**(3)**	**(4)**
	Fund Flow $_{f,t}$			
	Low Active Funds		***High Active Funds***	
Return $_{f,t-1}$	0.018***	0.015***	0.014***	0.015***
	(3.70)	(3.06)	(4.15)	(4.71)
Holding Distance $_{f,t-1}$		0.000		0.002
		(0.06)		(1.57)
Return $_{f,t-1}$ × *Holding Distance* $_{f,t-1}$		-0.011**		-0.002
		(-2.33)		(-1.22)
Controls	Y	Y	Y	Y
N	11175	11175	14022	14022
Adj. R^2	0.173	0.178	0.239	0.242

Figure 1. Industry-Adjusted Pairwise Distance (Putnam Example)

This figure plots the industry-adjusted pairwise distances between the prospectus of the Putnam VT Sustainable Future Fund and the 10-K filings of individual firms in 2019. For each fund–firm pair, the industry-adjusted distance is computed as the semantic distance between the fund's prospectus and the firm's 10-K, minus the average semantic distance between the fund and all firms in the same Fama-French 48 (FF-48) industry (including firms not held by the fund). The left panel shows all firms in the Putnam VT Sustainable Future Fund's portfolio holdings; the right panel shows all S&P 500 firms. Radial distance from the center represents the industry-adjusted semantic distance between the fund and each firm. Nodes are colored by FF-48 industry; portfolio holdings are displayed as dots and S&P 500 firms as squares.

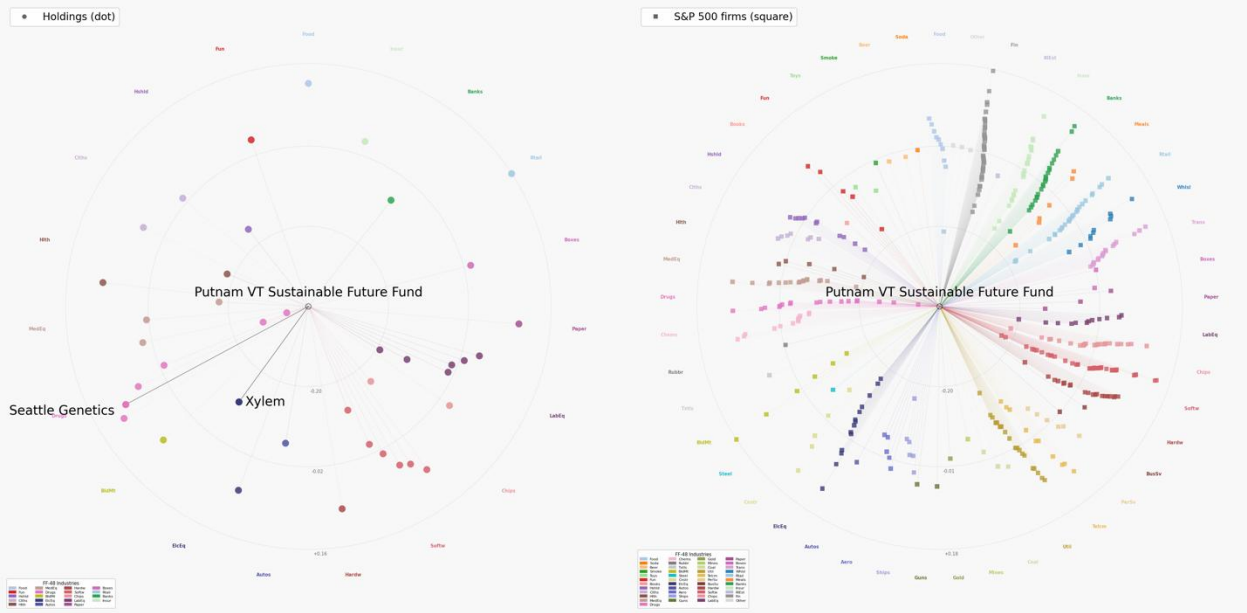


Figure 2. Fund-Firm Network

This figure plots the industry-adjusted pairwise distances between the prospectuses of 100 randomly selected equity mutual funds and the 10-K filings of S&P 500 firms in 2019. For each fund–firm pair, the industry-adjusted distance is computed as the semantic distance between the fund's prospectus and the firm's 10-K, minus the average semantic distance between the fund and all firms in the same Fama-French 48 (FF-48) industry (including firms not held by the fund). Hollow circles represent mutual funds and squares represent firms, with firm colors indicating FF-48 industries. Edges connect funds to firms, with edge length reflecting the industry-adjusted pairwise distance.

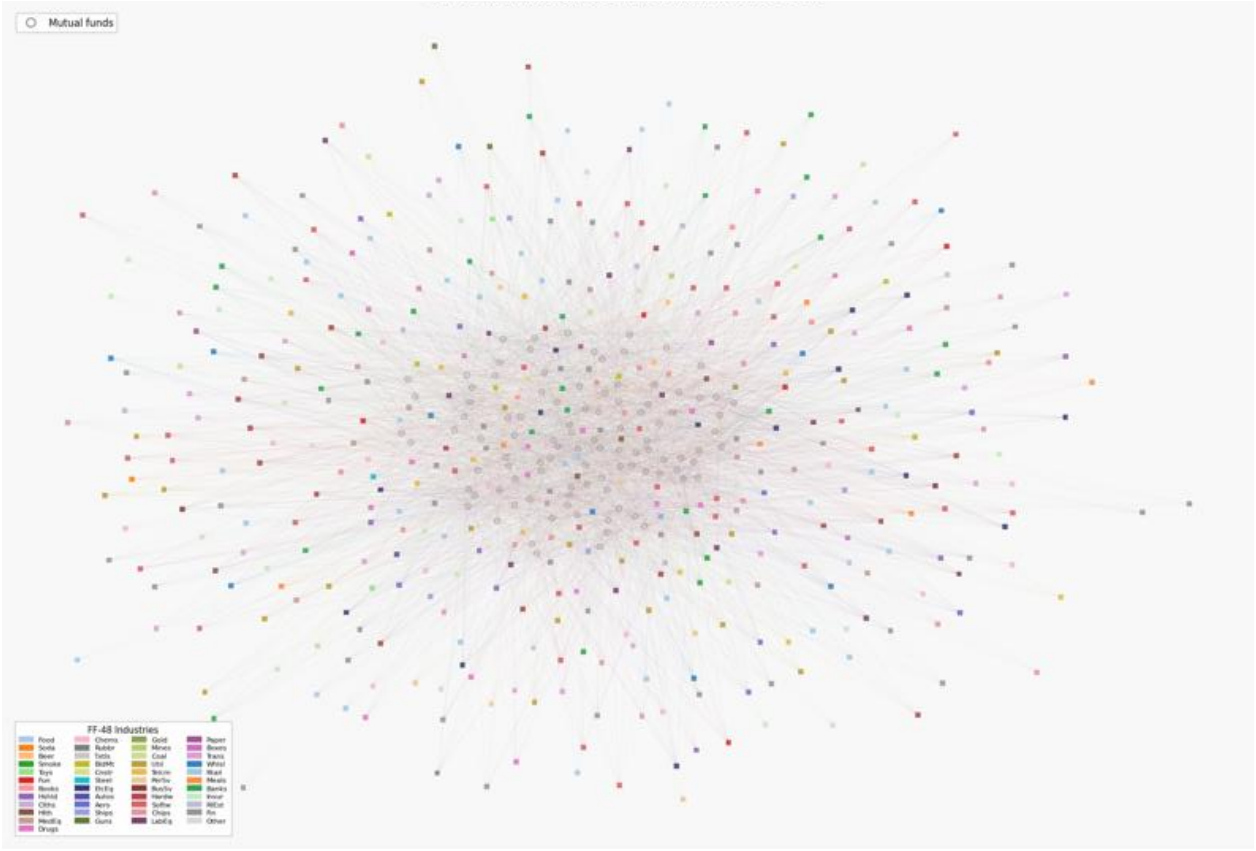
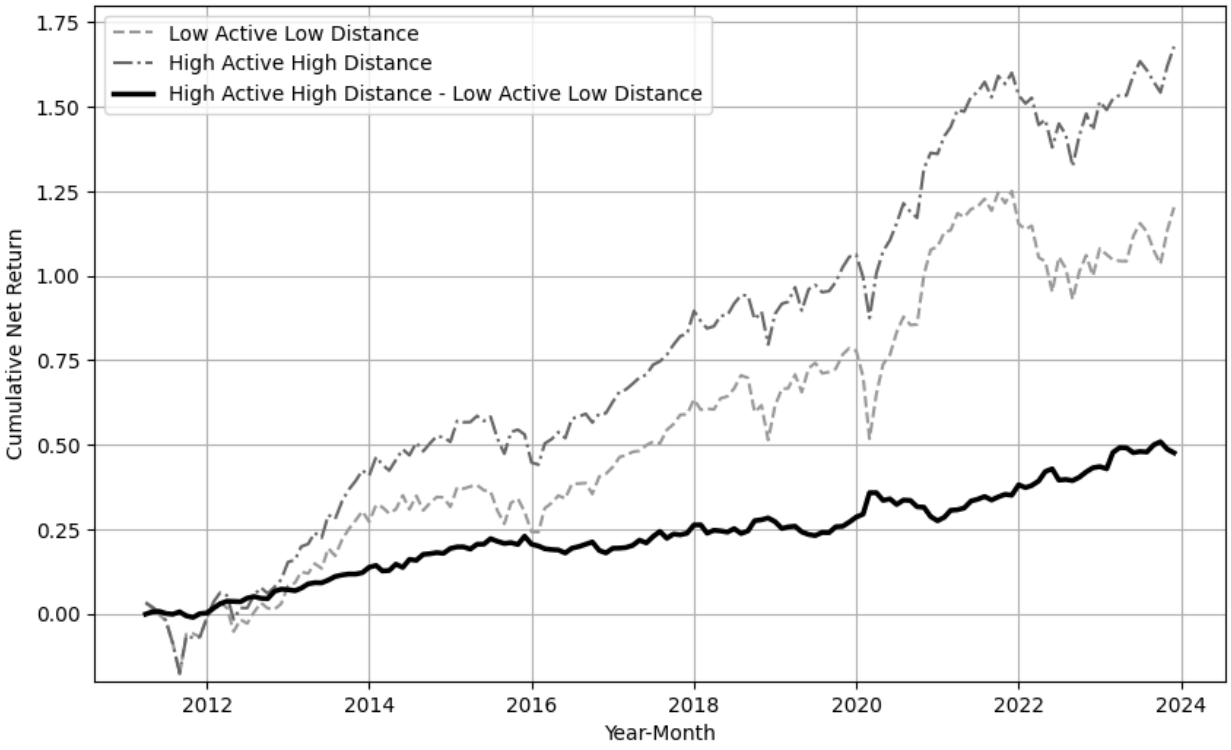


Figure 3. Fund Long-Short Portfolio Cumulative Returns

This figure plots the cumulative after-fee returns of long-short strategies from 2011 to 2023 based on double-sorted fund portfolios. Each quarter, mutual funds are sorted into 5 × 5 portfolios by *Holding Distance* and *Fund Activeness*, and value-weighted net returns are computed. Dashed lines represent the long position in the *High Active–High Distance* portfolio and the short position in the *Low Active–Low Distance* portfolio. The solid black line shows the spread between these positions, reflecting the return differential over the sample period.



Online Appendix

Part 1 of this online appendix provides the classical cognitive foundations in more detail and a simple extension of Kyle (1985) model to illustrate the mechanism of distant investment. Part 2 contains a list of tables for additional empirical analysis.

Online Appendix Part 1 (A Simple Conceptual Framework of Distant Investment)

Our main text argues that managerial skill lies in applying domain-specific expertise—revealed by a fund's prospectus—to interpret firm disclosures that are semantically distant yet still tractable. This appendix formalizes that idea in two steps. We begin by summarizing the cognitive foundations of such "expertise extension." We then formulate this intuition into a simple extension of Kyle (1985), in which investors differ in their ability to extract information from a common public disclosure. The framework yields a central implication: more skilled investors benefit most at an intermediate, or "good," distance—rather than when disclosures are either too close to, or too far from, their domain of expertise.

The Cognitive Foundation of Managerial Skill

The first cognitive foundation is structure-mapping theory (Gentner 1983; see also Gentner and Smith 2012 for a more recent survey), which explains how individuals use existing knowledge to learn about unfamiliar subjects. The central insight is that learning hinges on identifying deep, latent relational correspondences between a familiar "base" domain and a less familiar "target." This mapping allows individuals to project their prior expertise onto the target, facilitating the comprehension of unfamiliar contexts.

In our setting, the fund prospectus provides a revealed representation of the manager's domain-specific expertise (the base), while the firm's disclosure serves as the target. A key determinant of managerial skill is thus the ability to map this disclosure onto the manager's underlying relational knowledge structure. Our empirical measure of semantic distance serves as a reduced-form proxy for the difficulty of this mapping process. Economically, a larger distance indicates that the firm's disclosure contains more difficult-to-understand information (DUI) relative to the manager's perspective. Consequently, *distant investment* requires the specific skill of constructing relational correspondences between novel firm information and the manager's existing expertise.

How, then, do skilled managers employ relational mapping? The mechanism is analogical transfer, defined as "the use of an analogy from a semantically distant domain to guide the problem-solving process" (Gick and Holyoak 1980). A central challenge in this process is the "access failure": the inability to notice a distant analogy's relevance without external guidance. We posit that high-skill

managers overcome this obstacle by spontaneously identifying and applying structural alignments to interpret new investment opportunities. In contrast, low-skill managers are likely limited by semantic distance, failing to recognize the underlying patterns that connect their expertise to unfamiliar firm disclosures.

Finally, the bridge between relational mapping and value creation is absorptive capacity. Cohen and Levinthal (1990) define this as the ability to "recognize the value of new, external information, assimilate it, and apply it to commercial ends." This ability enables agents to synthesize novel data by establishing logical linkages with pre-existing concepts. Within our framework, a manager's specialized expertise serves as a cognitive anchor. This foundation allows them to evaluate and creatively utilize "distant" signals that investors lacking such an anchor may fail to recognize or effectively interpret.

Together, these three theoretical pillars formalize a novel mechanism of managerial skill: the capacity to extract valuable signals from semantically distant firms by extending specialized expertise. Because distant investment necessitates these cognitive abilities, observing a fund's capital allocation to such firms provides a revealed preference approach to identifying managerial skill.

These cognitive foundations also highlight a critical limitation of traditional textual analysis, which typically relies on a document-centric focus. By measuring surface-level attributes (e.g., sentiment), traditional methods implicitly assume a homogeneous interpretation by all readers. In contrast, structure-mapping is inherently *observer-specific*. This cognitive property aligns with classical economic theories suggesting that investors interpret the same news differently (Rubinstein 1993; Kim and Verrecchia 1994; Kandel and Pearson 1995), rendering the value of information inherently investor-specific (e.g., Van Nieuwerburgh and Veldkamp 2009; Farboodi et al. 2025). As such, our distance measure operationalizes this classic insight: it captures not merely what is written, but who is reading.

While these cognitive theories explain how managers process distant information, they do not determine how far they should go. Intuitively, the benefit of a "distant" signal disappears if it falls too far outside a manager's specialized expertise—for instance, a deep-value stock may be too "distant" for a growth-oriented manager to accurately decode. This motivates us to formalize the

trade-offs of distant investment within a framework based on Kyle (1985). The central takeaway is that skilled managers benefit most when signal distance is intermediate: neither too close, where the skill wedge is small, nor too far, where even skilled managers lose processing accuracy.

A Simple Model of Distant Investment

Setup of the model

A firm's terminal value is determined by:

$$v = \gamma' z, \tag{A1}$$

where $z \in \mathbb{R}^K$ is the vector of publicly observed firm attributes, such as the firm's strategic priorities, technologies, or business lines, and $\gamma \in \mathbb{R}^K$ denotes the latent economic structure that maps those observable features into the final value created by the firm.

Further assume that the latent economic structure follows a joint normal distribution, $\gamma \sim N(\bar{\gamma}, \Sigma_\gamma)$, where Σ_γ is positive definite. Without loss of generality, we normalize $\mu_v = z'\bar{\gamma} = 0$ by recentering γ around its mean. Hence, conditional on z, firm value is normally distributed with mean $\mu_v \equiv \mathbb{E}[v \mid z] = \bar{\gamma}' z$ and variance $\sigma_v^2 \equiv \mathrm{Var}\,(v \mid z) = z'\Sigma_\gamma z$. Investors do not observe firm value directly. But before trading, the firm releases a public disclosure, which can be thought of as a noisy signal about the latent economic structure: $s = \gamma + \varepsilon$, where $\varepsilon \sim N(0, \Sigma_s)$ is the noise term independent of γ.

The market has one noise trader and two informed investors. The demand of the noise trader follows a normal distribution, $u \sim N(0, \sigma_u^2)$, which is orthogonal to firm value and to the demand of informed investors. In contrast, the two informed investors infer the latent structure, γ, from this public disclosure based on their existing domain expertise. This allows them to extract the value of γ from the public disclosure, which they use to estimate the firm's value and determine their trading demand. Importantly, both investors observe the same public disclosure s. However, heterogeneity can arise from how investors process this common disclosure. Investor i's extracted signal (defined below) should therefore be interpreted as a sufficient-statistic representation of her posterior belief about the latent structure γ, conditional on s and her structural-mapping ability.

We further assume that the two informed investors differ in their structural-mapping abilities, denoted as θ_H and θ_L for high and low abilities, with $\theta_H > \theta_L$. The extracted value of the firm's disclosure is influenced by both its semantic distance from the domain of expertise of the investors, denoted as d, and the investors' structural-mapping skills. For tractability, we assume both informed investors face the same semantic distance d to the firm's disclosure, isolating heterogeneity in structural-mapping ability. In the empirical setting, however, distance is fund-specific. The model therefore focuses on one margin—heterogeneous skill conditional on a given level of distance—while abstracting from cross-investor variation in d.

Economically, greater distance should make interpretation harder, while greater skill mitigates this deterioration. For expositional clarity, we assume that the two informed investors face the same semantic distance $d \geq 0$ to the firm's disclosure. In economic terms, this setup resembles an economy with two growth fund managers (with heterogeneous skills). Depending on the value of distance d, the stock may range from an obvious growth firm (small distance) to a value stock (corresponding to very large distance).

Specifically, conditional on observing the public disclosure s, investor i forms an estimate of the latent structure γ. We represent this estimation in reduced form by a scalar sufficient statistic as follows:

$$\tilde{s}_i = \gamma + \xi_i, \tag{A2}$$

where $\xi_i \sim N(0, \Omega_i(d))$ is the remaining processing noise of the investor i ($i \in \{H, L\}$), which is independent of γ. Economically, $\tilde{s}_i$ is **not** a primitive signal disclosed by the firm. Rather, it provides a reduced form representation of investor i's extraction of the latent economic structure from firm disclosure based on their skills.

To obtain a tractable benchmark, we assume that investor i's residual uncertainty about the latent structure γ is proportional to the prior covariance matrix, with its scale governed by learning frictions that depend on semantic distance and managerial skill:

$$\Omega_i(d) = \omega_i(d, \theta_i)\Sigma_\gamma, \tag{A3}$$

Here, $\omega_i(d, \theta_i)$ captures learning frictions in extracting the latent structure from the disclosure. It increases with distance and decreases with managerial skill, i.e., $\frac{\partial \omega_i(d,\theta_i)}{\partial d} > 0$, $\frac{\partial \omega_i(d,\theta_i)}{\partial \theta_i} < 0$. This proportionality assumption is a tractable way to model heterogeneous learning while preserving the directional structure of uncertainty about γ. It implies that semantic distance and managerial skill scale the overall precision of inference without altering the relative importance of different components of the latent structure. This formulation allows us to capture distance-dependent learning frictions in a parsimonious manner, while maintaining closed-form solutions in a Kyle-style equilibrium.

For tractability, we assume both informed investors face the same semantic distance d to the firm's disclosure, isolating heterogeneity in structural-mapping ability. In the empirical setting, however, distance is fund-specific. The model therefore focuses on one margin—heterogeneous skill conditional on a given level of distance—while abstracting from cross-investor variation in d.

Under the above assumptions, informed investors can estimate the firm's value from their extracted signals, as summarized by the following lemma.

Lemma 1 (Posterior Firm Value): To investor i, the posterior valuation of the firm becomes

$$\mathbb{E}[v \mid \tilde{s}_i] = \rho_i(d, \theta_i)\, z' \tilde{s}_i. \tag{A4}$$

where $\rho_i(d, \theta_i) \equiv \frac{1}{1+\omega_i(d,\theta_i)} \in (0,1)$ is the posterior weight derived from the investor's precision in learning the latent structure γ.

If we define $y_i \equiv z' \tilde{s}_i$ to denote the extracted value of the firm, then

$$\mathbb{E}[\, v \mid y_i\,] = \rho_i(d, \theta_i)\, y_i. \tag{A5}$$

Proof (Lemma 1): Under normality and (A3), investor i's posterior mean of γ is

$$\mathbb{E}[\gamma \mid \tilde{s}_i] = \bar{\gamma} + \Sigma_\gamma(\Sigma_\gamma + \Omega_i(d))^{-1}(\tilde{s}_i - \bar{\gamma}) = \bar{\gamma} + \frac{1}{1+\omega_i(d,\theta_i)}(\tilde{s}_i - \bar{\gamma}).$$

Because firm value is $v = \gamma' z$, investor i's posterior valuation is

$$\mathbb{E}[\, v \mid \tilde{s}_i\,] = z' \mathbb{E}[\, \gamma \mid \tilde{s}_i\,] = \frac{1}{1+\omega_i(d)}\, z' \tilde{s}_i, \tag{A6}$$

where the last equality uses the normalization $z'\bar{\gamma} = 0$. This establishes (A4), and (A5) follows immediately. Note that y_i is a sufficient statistic for v given $\tilde{s}_i$ by the structure of our normal model.

Lemma 1 reveals the logic of our model. Semantic distance and managerial skill determine the precision of learning about the latent structure γ, which in turn governs the quality of the extracted scalar signal y_i. This signal determines posterior valuation $\mathbb{E}[v \mid y_i]$, and their trading demand (detailed below). Thus, distance and skill affect trading outcomes through their impact on the precision of structural learning and the resulting valuation signal.

To characterize how precision varies with distance and skill, note that $\rho_i(d, \theta_i)$is decreasing in semantic distance and increasing in managerial skill, i.e., $\partial\rho_i/\partial d < 0$ and $\partial\rho_i/\partial\theta_i > 0$. Intuitively, greater distance makes it harder to extract the latent structure from the disclosure, while higher structural-mapping ability mitigates this deterioration.

We further impose economically natural boundary conditions on $\rho_i(d, \theta_i)$. First, without loss of generality, we normalize zero distance as the case in which both investors perfectly recover the latent structure, so that $\rho_H(0, \theta_H) = \rho_L(0, \theta_L) = 1$. Second, for any positive distance, the high-skill investor maintains (weakly) higher precision, $0 < \rho_L(d, \theta_L) \leq \rho_H(d, \theta_H) < 1$, which defines skill as greater resilience in extracting information from semantically distant disclosures. Finally, both precisions vanish as distance becomes arbitrarily large, $\lim_{d\to\infty} \rho_H(d, \theta_H) = \lim_{d\to\infty} \rho_L(d, \theta_L) = 0$, so that sufficiently distant disclosures eventually become uninformative for all investors. We also impose a regularity condition that the profits of the more skilled investor increase near zero distance (i.e., $\Pi'_H(0) > 0$), which we will discuss in details in the proof of Proposition A2.

Under these conditions, the equilibrium profit function reflects a tradeoff between the high-skill investor's relative informational advantage and the overall deterioration of signal quality. To illustrate these conditions in a tractable parametric setting, we consider the exponential specification

$$\rho_i(d, \theta_i) = \exp\left(-\frac{d}{\theta_i}\right), \theta_i > 0, i \in \{H, L\}. \tag{A7}$$

This specification ensures $\rho_H(0,\theta_H) = \rho_L(0,\theta_L) = 1$, monotone decay in distance, and $\rho_H(d,\theta_H) \geq \rho_L(d,\theta_L)$ whenever $\theta_H \geq \theta_L$. Moreover, as shown below, the local condition $\Pi'_H(0) > 0$reduces to a simple restriction on the relative magnitude of $\theta_H$and θ_L.

Under these assumptions, each informed investor i then submits a linear order, x_i, proportional to his extracted signal:

$$x_i = \beta_i y_i, i \in \{H, L\}, \tag{A8}$$

where β_i is a constant parameter chosen to maximize expected profits conditional on extracted firm value from disclosure, y_i:

$$\max_{x_i} \ \mathbb{E}[(v-p)x_i \mid y_i]. \tag{A9}$$

Following Kyle (1985), we assume a competitive market maker observes the total order flow from all investors, denoted as $q = x_H + x_L + u$. We focus on the linear equilibrium, in which the competitive market maker determines the price of the assets based on the observed total order flows:

$$p = \mathbb{E}[\, v \mid x_H + x_L + u \,] = \mathbb{E}[\, v \mid q \,] = \lambda q, \tag{A10}$$

where λ is a constant parameter of market depth to be determined in equilibrium.

The following proposition proves that a standard Kyle-style linear equilibrium exists in this economy. But the informational heterogeneity arises from differential learning about the latent structure γ.

Proposition A1 (Linear Equilibrium)

For any $(\rho_H, \rho_L) \in (0,1)^2$, there exists a unique linear equilibrium of the form (A8)–(A10). The informed trading coefficients and the market depth parameter satisfy:

$$\beta_H = \frac{\rho_H(2-\rho_L)}{\lambda(4-\rho_H\rho_L)}, \beta_L = \frac{\rho_L(2-\rho_H)}{\lambda(4-\rho_H\rho_L)}, \tag{A11}$$

$$\lambda = \sqrt{\frac{\sigma_v^2(\rho_H^2\rho_L+\rho_H\rho_L^2-8\rho_H\rho_L+4\rho_H+4\rho_L)}{\sigma_u^2(4-\rho_H\rho_L)^2}}. \tag{A12}$$

Proof (Proposition A1)

Based on equation (A10), investors' problem becomes

$$\max_{x_i} \ \mathbb{E}[v \mid y_i]\, x_i \ - \ \lambda x_i^2 \ - \ \lambda x_i\, \mathbb{E}[x_j \mid y_i].$$

The first-order condition of informed investors is

$$\rho_i(d) y_i - 2\lambda x_i - \lambda \mathbb{E}[x_j \mid y_i] = 0. \tag{A13}$$

To calculate $\mathbb{E}[x_j \mid y_i]$, we notice that, since $\text{Cov}\,(y_j, y_i) = \text{Var}\,(v) = \sigma_v^2$, and $\text{Var}\,(y_i) = \frac{\sigma_v^2}{\rho_i(d)}$, we have $\mathbb{E}[y_j \mid y_i] = \frac{\text{Cov}\,(y_j, y_i)}{\text{Var}\,(y_i)} y_i = \rho_i(d) y_i$. Hence

$$\mathbb{E}\big[\, x_j \mid y_i \,\big] = \beta_j \rho_i(d) y_i. \tag{A14}$$

In the above equation, because the projection is conditional on y_i, the slope is governed by $\text{Var}\,(y_i)$, hence by ρ_i.

Substituting (A7) and (A14) into (A13) yields

$$2\lambda\beta_i + \lambda\beta_j \rho_i(d) = \rho_i(d), \tag{A15}$$

or,

$$2\lambda\beta_H + \lambda\beta_L \rho_H = \rho_H, \tag{A16}$$

$$\lambda\beta_H \rho_L + 2\lambda\beta_L = \rho_L, \tag{A17}$$

Solving the above two equations yields

$$\beta_H = \frac{\rho_H(2-\rho_L)}{\lambda(4-\rho_H\rho_L)}, \beta_L = \frac{\rho_L(2-\rho_H)}{\lambda(4-\rho_H\rho_L)}. \tag{A18}$$

To compute the price coefficient, note that under linear strategies, the market maker forms expectations based on the projection of v onto total order flow q. By normality, $p = E[v \mid q] = \frac{\text{Cov}(v,q)}{\text{Var}(q)} q$. Hence,

$$\lambda = \frac{\text{Cov}\,(v,q)}{\text{Var}\,(q)}. \tag{A19}$$

Using $\text{Cov}\,(v, y_i) = \sigma_v^2$, $\text{Var}\,(y_i) = \sigma_v^2/\rho_i$, and $\text{Cov}\,(y_H, y_L) = \sigma_v^2$, we obtain

$$\lambda = \frac{\sigma_v^2(\beta_H+\beta_L)}{\sigma_v^2\left(\beta_H^2/\rho_H+\beta_L^2/\rho_L+2\beta_H\beta_L\right)+\sigma_u^2}. \quad \text{(A20)}$$

Substituting (A18) into (A20) yields

$$\lambda^2 = \frac{\sigma_v^2(\rho_H^2\rho_L+\rho_H\rho_L^2-8\rho_H\rho_L+4\rho_H+4\rho_L)}{\sigma_u^2(4-\rho_H\rho_L)^2}. \quad \text{(A21)}$$

Because $\rho_H, \rho_L \in (0,1)$, the numerator in (A21) is strictly positive, so λ has a unique positive solution as specified in A12.

Proposition A1 characterizes the linear equilibrium for any given pair of signal precisions (ρ_H, ρ_L). We now ask how the high-skill investor's equilibrium profits vary with semantic distance. In general, the model contains two opposing forces. As distance rises, the low-skill investor's precision may deteriorate faster than the high-skill investor's precision, which widens the latter's relative informational advantage. But if distance becomes too large, even the high-skill investor's own signal becomes too imprecise for profitable extraction. The next proposition confirms that this "good distance" intuition holds valid under reasonable conditions.

Proposition A2 (Interior Good Distance)

Let $\rho_i(d, \theta_i)$ denote the signal precisions of high- and low-skill investors as functions of semantic distance d. Under the conditions discussed after Lemma 1, there exists an interior distance $d^* > 0$that maximizes the high-skill investor's expected profit.

Proof (Proposition A2)

The conditions for $\rho_i(d, \theta_i)$ are $\partial\rho_i/\partial d < 0$ and $\partial\rho_i/\partial\theta_i > 0$ and that (i) $\rho_H(0, \theta_H) = \rho_L(0, \theta_L) = 1$; (ii) $0 < \rho_L(d, \theta_L) \leq \rho_H(d, \theta_H) < 1$ for all $d > 0$; (iii) $\lim_{d\to\infty} \rho_H(d, \theta_H) = \lim_{d\to\infty} \rho_L(d, \theta_L) = 0$; and $\Pi_H'(0) > 0$.

Given $x_H = \beta_H y_H$ and $p = \lambda q$, expected profits of H can be written as $\Pi_H(d) \equiv \mathbb{E}[(v-p)x_H]$. Using the first-order condition, it becomes $\Pi_H = \mathbb{E}[(v-p)x_H] = \beta_H\sigma_v^2\left(1-\frac{\lambda\beta_H}{\rho_H}-\lambda\beta_L\right)$. Then from (A23): $2\lambda\beta_H = \rho_H(1-\lambda\beta_L)$, so $\lambda\beta_H/\rho_H = (1-\lambda\beta_L)/2$, giving $\Pi_H = \lambda\beta_H^2\sigma_v^2/\rho_H = \lambda\beta_H^2\mathrm{Var}(y_H)$. Since $\mathrm{Var}(y_H) = \sigma_v^2/\rho_H(d)$, substituting (A28) gives

$$\Pi_H(d) = \frac{\sigma_v^2}{\lambda(d)} \cdot \frac{\rho_H(d,\theta_H)(2-\rho_L(d,\theta_L))^2}{(4-\rho_H(d,\theta_H)\rho_L(d,\theta_L))^2}. \quad \text{(A22)}$$

Substituting $\lambda(d)$from Proposition A1 into (A22), and taking derivatives of log $\Pi_H(d)$, yields $\frac{\Pi'_H(0)}{\Pi_H(0)} = \frac{19\,\rho'_H(0,\theta_H) - 17\,\rho'_L(0,\theta_L)}{12}$. Hence, the regularity condition $\Pi'_H(0) > 0$ is equivalent to the technical restriction of $19\,\rho'_H(0,\theta_H) - 17\,\rho'_L(0,\theta_L) > 0$. For instance, under the exponential specification, this reduces to $\theta_H/\theta_L > 19/17$, a mild skill differential that is always satisfied when H is meaningfully more skilled than L. Under this condition, there exists $\varepsilon > 0$ such that around $d = 0$,

$$\Pi_H(d) > \Pi_H(0) \text{ for all } d \in (0, \varepsilon). \quad \text{(A23)}$$

By condition (iii), $\rho_H(d,\theta_H), \rho_L(d,\theta_L)$ also converges to zero. From Proposition A1,

$$\lambda(d) \sim \frac{\sigma_v}{2\sigma_u}\sqrt{\rho_H(d,\theta_H) + \rho_L(d,\theta_L)}. \quad \text{(A24)}$$

Substituting into (A22) yields

$$\Pi_H(d) \sim \frac{\sigma_v\sigma_u}{2} \cdot \frac{\rho_H(d,\theta_H)}{\sqrt{\rho_H(d,\theta_H)+\rho_L(d,\theta_L)}} \;\to\; 0. \quad \text{(A25)}$$

Thus, $\Pi_H(d)$ is continuous, exceeds its value at $d = 0$ for some positive distance, and converges to zero as $d \to \infty$. Therefore, it attains a maximum at some interior point $d^* \in (0, \infty)$.

The model yields three implications that clarify the mechanism underlying Propositions A1 and A2. First, investors observe the same public disclosure but differ in how effectively they extract the latent structure γ. The relevant heterogeneity is therefore not differential access to information, but differential interpretation of a common disclosure. This feature distinguishes the model from standard private-information settings and aligns with the idea that information is investor-specific in value.

Second, the mechanism operates in a clear sequence. Semantic distance and managerial skill jointly determine the precision of learning about γ, captured by $\rho_i(d,\theta_i)$. This precision governs the quality of the extracted signal y_i, which determines posterior valuation $\mathbb{E}[v \mid y_i]$, and ultimately trading demand. Informational advantage therefore arises from superior extraction of the same underlying disclosure, rather than from observing different signals.

Third, and most importantly, the value of distance is non-monotone. When distance is very low, both investors interpret the latent structure similarly, so informational rents are compressed by competition. When distance is very high, even the high-skill investor cannot recover the latent structure precisely enough to form accurate valuations, and profits vanish. Between these extremes lies an interior "good distance," at which the high-skill investor's relative informational advantage is strongest, and equilibrium profits are maximized.

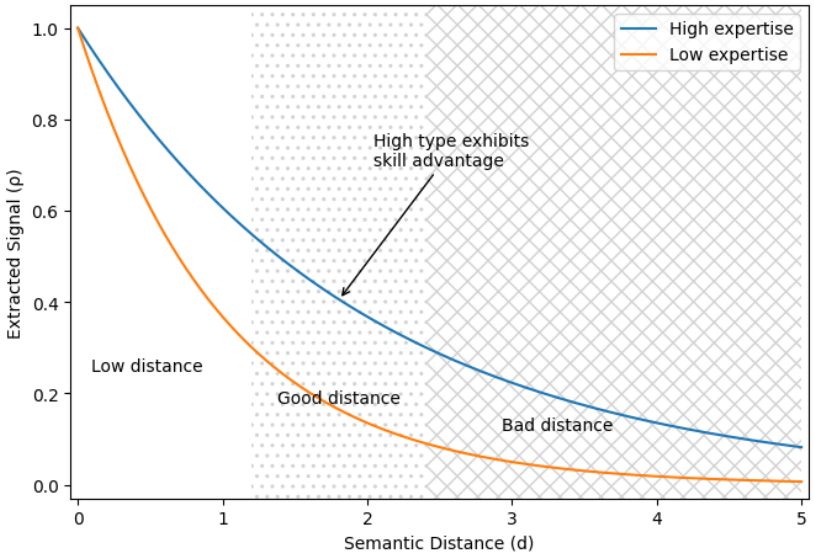
To illustrate this mechanism, Figure A1 presents a parametric example based on the exponential specification $\rho_i(d, \theta_i) = \exp(-d/\theta_i)$. Panel A plots signal precision as a function of distance for high- and low-skill investors. As distance increases, both precisions decline, but the high-skill investor's precision deteriorates more slowly, generating a widening gap in informational advantage over an intermediate range of distances.

Panel B plots the corresponding expected utility of the high-skill investor. Consistent with Proposition A2, utility is hump-shaped in distance: it is low when distance is near zero due to strong competition, rises as distance creates informational differentiation, and eventually declines as signal precision deteriorates for all investors. For illustration, the figure partitions distance into "low," "good," and "bad" regions using a cutoff on precision; however, the model's key implication is the existence of an interior optimum rather than any specific threshold.

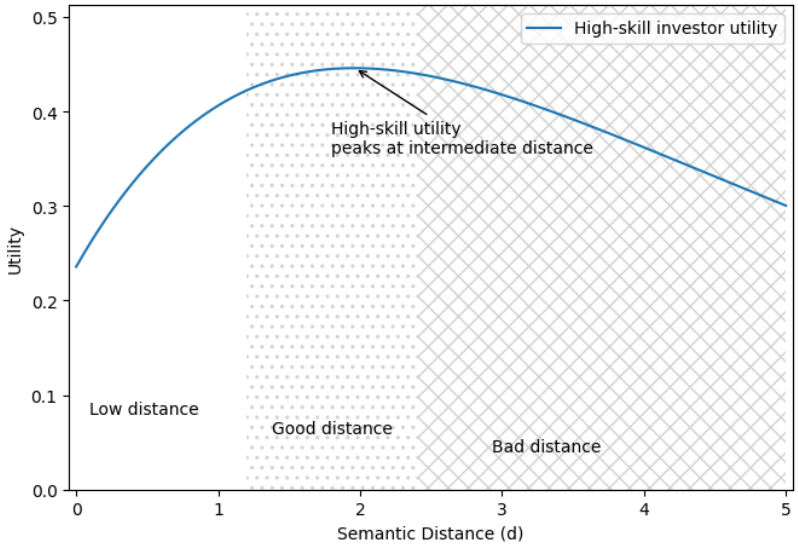
Figure A1. Extracted signal **and high-skill utility across semantic distance.**

This figure illustrates how signal extraction and expected utility vary with semantic distance in a parametric example. Distance is defined over $d \in [0,5]$. Signal precision follows $\rho_i(d) = e^{-d/\theta_i}$, with $\theta_H = 2.0$ (high expertise) and $\theta_L = 1.0$ (low expertise). Panel A plots extracted signal precision for high- and low-skill investors. Panel B plots the corresponding expected utility of the high-skill investor under $\sigma_v^2 = 1$, $\sigma_s^2 = 0$, and $\sigma_u = 1$. The figure highlights the high-skill investor's relative informational advantage by partitioning distance into "low," "good," and "bad" regions using a cutoff ρ=0.3. This partition is for illustrative purposes and depends on the chosen cutoff. The model's main implication is the existence of an interior optimum, rather than specific threshold values.

Panel A. Signal extraction vs. distance for high- and low-skill investors



Panel B. Expected utility for the high-skill investor



Online Appendix Part 2 (Additional Empirical Analysis)

Table A1. Additional Summary Statistics for Firm-Level Variables

This table presents the number of observations, mean value, standard deviation, 25th percentile, median, and 75th percentile of firm characteristics at the quarterly frequency. *SDI_OI*, *AWOI* and *DWOI* represent the SDI-weighted, activeness-weighted and distance-weighted order imbalance, respectively. See Section 2.2 for details. *Size ($million)* is the market capitalization at the end of each quarter. *BM* is the book to market ratio at the end of quarter. *Prior 1-Month Return* is the stock return in the prior month. *Prior 12-Month Return* is the stock's cumulative return over the prior 12 months. *Investment* is the annual growth rate of total assets (Cooper, Gulen, and Schill 2008). *Profitability* is the annual operating profitability (Fama and French 2006). *Illiquidity* is the daily ratio of absolute stock return to its dollar volume, averaged quarterly (Amihud 2002). *Analyst Coverage* is the number of analysts with valid forecast, averaged quarterly. *Idiosyncratic Volatility* is the standard deviation of residuals from Fama–French three-factor regressions estimated using daily returns, averaged quarterly. *ASVI* is as the change in log-transformed Google search volume from the previous quarters. *Geographic Proximity* is the holding-value-weighted average of a same-state indicator. *Complexity* is the proportion of financial-complexity words in the firm's 10-K filing (Loughran and McDonald 2024). We exclude stocks with a price below $5, those in the financial sector. We additionally exclude stocks with the top 10% *Illiquidity*, as this measure contains extreme values. The results remain robust without this additional filtering. All variables are winsorized at the 1st and 99th percentiles, except *SDI_OI*, *AWOI* and *DWOI*. The sample period is from 2011Q1 to 2023Q4.

	N	Mean	Std	P25	Median	P75
SDI_OI	79710	0.000	0.002	-0.001	0.000	0.001
AWOI	79710	0.028	0.186	-0.079	0.031	0.139
DWOI	79710	0.000	0.003	-0.001	0.000	0.002
Size ($million)	79710	7800.72	18129.22	636.13	1822.77	5855.80
BM	79710	0.525	0.441	0.227	0.412	0.690
Prior 1-Month Return	79710	0.005	0.114	-0.056	0.004	0.063
Prior 12-Month Return	79710	0.168	0.517	-0.138	0.100	0.365
Investment	79710	0.148	0.367	-0.012	0.059	0.174
Profitability	79710	0.243	0.645	0.119	0.228	0.366
Illiquidity	79710	0.007	0.017	0.000	0.001	0.005
Analyst Coverage	79710	8.014	6.167	3.333	6.333	11.333
Idiosyncratic Volatility	78949	0.018	0.010	0.011	0.016	0.022
Geographic Proximity	78949	0.062	0.450	-0.097	0.006	0.134
ASVI	78949	0.054	0.099	0.000	0.007	0.062
Complexity	78949	0.005	0.002	0.003	0.004	0.006

Table A2. Fund Performance Using Panel Regressions

This table presents the results from quarterly panel regressions of future fund performance on *Holding Distance*, *Fund Activeness*, and their interaction *SDI*.

$$Perf_{f,t} = \alpha + \beta_1 \times Holding\ Distance_{f,t-1} + \beta_2 \times Fund\ Activeness_{f,t-1} + \beta_3 \times SDI_{f,t-1} + \boldsymbol{\Gamma'} \times \boldsymbol{M_{f,t-1}} + \varepsilon_{f,t}$$

where $Perf_{f,t}$ refers to the performance of fund f in quarter t, $Holding\ Distance_{f,t-1}$ is the active weight-weighted industry-adjusted fund distance to holding firms in quarter $t-1$, and $Fund\ Activeness_{f,t-1}$ describes the level of fund activeness, which is average rank of *active weight* and *return gap* of fund f in quarter $t-1$. The vector $\boldsymbol{M_{f,t-1}}$ stacks a set of lagged fund characteristics as listed in Table 4. Fund performance measures include four-factor-adjusted alphas (both before and after expense fees) as well as the value created by fund managers, measured as *value-added* following Berk and van Binsbergen (2015). Control variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized to have a mean of zero and a standard deviation of one. Fund and Quarter fixed effects are included. Robust standard errors are clustered at the fund level. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	***Carhart 4-factor Alpha*** $_t$ ***(TNA-weighted)***				***Value Added*** $_t$	
	Before Fee		***After Fee***			
Holding Distance $_{f,t-1}$	-0.082*	-0.084**	-0.082*	-0.084**	-1.807**	-1.983**
	(-1.86)	(-2.03)	(-1.86)	(-2.03)	(-2.34)	(-2.57)
Fund Activeness $_{f,t-1}$	0.070*	0.092**	0.070*	0.092**	1.625*	1.680*
	(1.73)	(2.21)	(1.72)	(2.21)	(1.77)	(1.82)
SDI $_{f,t-1}$	0.134***	0.136***	0.133***	0.136***	1.740***	1.699**
	(2.79)	(2.95)	(2.79)	(2.95)	(2.60)	(2.55)
Controls	Y	Y	Y	Y	Y	Y
Fund FE	Y	Y	Y	Y	Y	Y
Quarter FE		Y		Y		Y
N	73011	73011	73011	73011	73011	73011
Adj. R²	0.027	0.103	0.029	0.105	0.003	0.019

Table A3. Transition Matrices of Sorted Mutual Funds

This table reports transition matrices of mutual funds across quintiles. In each quarter, funds are sorted into to quintiles based on their most recent holding distance (Panel A) or fund activeness (Panel B). The reported transition rates represent the probability that a fund classified in quintile *i* in a given quarter is classified in quintile *j* in the subsequent quarter.

Panel A. Transition rates of quintiles based on Holding Distance

Current Quintile	Quintile after One Quarter				
	1	**2**	**3**	**4**	**5**
1	71.56	17.20	3.94	1.27	0.96
2	17.01	57.25	16.80	2.29	0.82
3	3.94	16.15	57.13	16.35	1.84
4	1.31	2.37	15.79	62.70	13.43
5	0.89	0.88	2.14	13.10	77.80

Panel B. Transition rates of quintiles based on Fund Activeness

Current Quintile	Quintile after One Quarter				
	1	**2**	**3**	**4**	**5**
1	71.33	16.70	4.42	1.33	0.54
2	16.17	51.92	20.18	5.96	1.59
3	4.38	19.74	45.61	20.56	4.70
4	1.24	6.05	20.38	48.96	18.77
5	0.50	1.64	5.11	18.67	68.50

Panel C. Transition rates of quintiles based on SDI

Current Quintile	Quintile after One Quarter				
	1	**2**	**3**	**4**	**5**
1	71.52	15.72	4.95	1.66	1.08
2	15.83	60.36	15.45	1.82	0.60
3	4.84	14.80	58.14	16.36	1.64
4	1.73	1.93	15.42	61.79	14.66
5	1.06	0.74	1.98	14.19	76.66

Table A4. Stock Return Predictivity Controlling for both AWOI and DWOI

This table presents the results from quarterly Fama-MacBeth (1973) regressions of future stock performance on *SDI_OI*.

$$DGTW_{s,t} = \alpha + \beta_1 \times SDI\_OI_{s,t-1} + \beta_2 \times AWOI_{s,t-1} + \beta_3 \times DWOI_{s,t-1} + \boldsymbol{\Gamma'} \times \boldsymbol{M_{s,t-1}} + \varepsilon_{s,t}$$

where $DGTW_{s,t}$ denotes the performance of stock s in quarter t, measured using characteristic-adjusted returns following Daniel, Grinblatt, Titman, and Wermers (1997). $SDI\_OI_{s,t-1}$ is the *SDI*-weighted order imbalance for stock s in quarter $t-1$. $AWOI_{s,t-1}$ and $DWOI_{s,t-1}$ denotes the *fund activeness* and *holding distance*-weighted order imbalance for stock s in quarter $t-1$. The vector $\boldsymbol{M_{s,t-1}}$ stacks a set of lagged stock characteristics as listed in Table 6. Control variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized to have a mean of zero and a standard deviation of one. Newey-West (1987) t-statistics with three lags are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	***DGTW-adjusted Returns $_{s,t}$ (%/quarter)***					
SDI_OI $_{s,t-1}$	0.000		-0.115	-0.115		-0.233
	(0.00)		(-0.75)	(-0.86)		(-1.52)
AWOI $_{s,t-1}$		0.260**	-0.745		0.208*	-0.841*
		(2.25)	(-1.54)		(1.87)	(-1.85)
DWOI $_{s,t-1}$			1.137**			1.256***
			(2.42)			(2.89)
Log Size $_{s,t-1}$	0.089	0.097	0.094	-0.109	-0.088	-0.106
	(0.49)	(0.53)	(0.52)	(-0.52)	(-0.41)	(-0.51)
Log BM $_{s,t-1}$	-0.150	-0.152	-0.145	-0.186	-0.188	-0.177
	(-0.67)	(-0.67)	(-0.65)	(-0.93)	(-0.95)	(-0.90)
Prior 1-Month Return $_{s,t-1}$	-0.305	-0.292	-0.309	-0.311	-0.302	-0.317
	(-1.16)	(-1.11)	(-1.17)	(-1.12)	(-1.08)	(-1.14)
Prior 12-Month Return $_{s,t-1}$	0.316	0.258	0.317	0.359	0.280	0.362
	(1.46)	(1.09)	(1.46)	(1.64)	(1.16)	(1.65)
Investment $_{s,t-1}$	-0.260	-0.255	-0.251	-0.249	-0.244	-0.240
	(-1.47)	(-1.44)	(-1.43)	(-1.59)	(-1.56)	(-1.54)
Profitability $_{s,t-1}$	0.191	0.197	0.191	0.133	0.138	0.132
	(1.19)	(1.25)	(1.21)	(0.97)	(1.03)	(0.98)
Illiquidity $_{s,t-1}$	-0.359	-0.356	-0.349	-0.276	-0.282	-0.264
	(-1.26)	(-1.23)	(-1.21)	(-0.94)	(-0.96)	(-0.89)
Log Analyst Coverage $_{s,t-1}$	-0.316	-0.317	-0.318	-0.203	-0.209	-0.204
	(-1.15)	(-1.16)	(-1.17)	(-0.87)	(-0.89)	(-0.88)
Idiosyncratic Volatility $_{s,t-1}$				-0.539	-0.518	-0.545
				(-1.34)	(-1.29)	(-1.37)
ASVI $_{s,t-1}$				0.140	0.134	0.138
				(1.20)	(1.15)	(1.18)
Geographic Proximity $_{s,t-1}$				-0.107	-0.104	-0.109
				(-1.15)	(-1.14)	(-1.18)
Complexity $_{s,t-1}$				-0.246	-0.244	-0.249
				(-1.43)	(-1.42)	(-1.46)
N	79710	79710	79710	78949	78949	78949
Adj. R^2	0.026	0.025	0.027	0.033	0.033	0.034

Table A5. Stock Return Predictivity Excluding Small Stocks

This table presents the results from quarterly Fama-MacBeth (1973) regressions of future stock performance on *SDI_OI*, excluding the bottom 20% of stocks by market capitalization in each quarter.

$$DGTW_{s,t} = \alpha + \beta \times SDI\_OI_{s,t-1} + \boldsymbol{\Gamma'} \times \boldsymbol{M_{s,t-1}} + \varepsilon_{s,t}$$

where $DGTW_{s,t}$ denotes the performance of stock s in quarter t, measured using characteristic-adjusted returns following Daniel, Grinblatt, Titman, and Wermers (1997). $SDI\_OI_{s,t-1}$ is the *SDI*-weighted order imbalance for stock s in quarter $t-1$. $High\ SDI\_OI_{s,t-1}$ equals to 1 if SDI_OI is above the cross-sectional median in quarter t, and 0 otherwise. We also include the measure of $AWOI_{s,t-1}$ in Columns (3)-(6), which denotes the fund-activeness-weighted order imbalance for stock s in quarter $t-1$. The vector $\boldsymbol{M_{s,t-1}}$ stacks a set of lagged stock characteristics as listed in Table 6. Control variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized to have a mean of zero and a standard deviation of one. Newey-West (1987) t-statistics with three lags are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	***DGTW-adjusted Returns $_{s,t}$ (%/quarter)***					
SDI_OI $_{s,t-1}$	0.274**		0.251**		0.309***	
	(2.58)		(2.54)		(2.79)	
High SDI_OI $_{s,t-1}$		0.260**		0.228**		0.264*
		(2.34)		(2.13)		(1.84)
AWOI $_{s,t-1}$					-0.095	-0.013
					(-0.80)	(-0.10)
Log Size $_{s,t-1}$	-0.119	-0.117	-0.239	-0.236	-0.257	-0.256
	(-0.65)	(-0.64)	(-1.44)	(-1.40)	(-1.59)	(-1.57)
Log BM $_{s,t-1}$	-0.193	-0.192	-0.220	-0.218	-0.210	-0.209
	(-0.90)	(-0.90)	(-1.16)	(-1.14)	(-1.11)	(-1.11)
Prior 1-Month Return $_{s,t-1}$	-0.426	-0.426	-0.439	-0.438	-0.443	-0.446
	(-1.65)	(-1.65)	(-1.62)	(-1.61)	(-1.64)	(-1.65)
Prior 12-Month Return $_{s,t-1}$	0.284	0.301	0.265	0.281	0.313	0.312
	(1.35)	(1.43)	(1.23)	(1.29)	(1.57)	(1.56)
Investment $_{s,t-1}$	-0.229	-0.233	-0.207	-0.211	-0.204	-0.208
	(-1.19)	(-1.21)	(-1.24)	(-1.26)	(-1.22)	(-1.25)
Profitability $_{s,t-1}$	0.125	0.122	0.066	0.065	0.067	0.068
	(0.83)	(0.82)	(0.49)	(0.48)	(0.50)	(0.50)
Illiquidity $_{s,t-1}$	-0.291*	-0.287*	-0.232	-0.229	-0.216	-0.226
	(-1.94)	(-1.93)	(-1.48)	(-1.47)	(-1.31)	(-1.39)
Log Analyst Coverage $_{s,t-1}$	-0.264	-0.260	-0.168	-0.163	-0.163	-0.163
	(-0.96)	(-0.94)	(-0.74)	(-0.72)	(-0.73)	(-0.72)
Idiosyncratic Volatility $_{s,t-1}$			-0.479	-0.477	-0.496	-0.490
			(-1.28)	(-1.27)	(-1.33)	(-1.31)
ASVI $_{s,t-1}$			0.151*	0.150*	0.157*	0.158*
			(1.74)	(1.74)	(1.80)	(1.83)
Geographic Proximity $_{s,t-1}$			-0.033	-0.031	-0.038	-0.033
			(-0.35)	(-0.32)	(-0.40)	(-0.35)
Complexity $_{s,t-1}$			-0.286*	-0.287*	-0.286*	-0.286*
			(-1.71)	(-1.72)	(-1.72)	(-1.72)
N	76112	76112	75462	75462	75462	75462
Adj. R^2	0.026	0.026	0.034	0.033	0.035	0.034

Table A6. Fund Portfolio Sentiment of Mutual Funds

This table presents the value-weighted portfolio sentiment for mutual funds sorted into 5 × 5 portfolios according to *Holding Distance* and *Fund Activeness*. Fund-level sentiment is constructed as

$$Fund\ Sentiment_{f,t} = \sum_{s} \left(w_{f,t}^{s} - w_{f,t}^{sm}\right) \times 1\{ASVI_s \geq P_{50}\}$$

where $w_{f,s}$ is portfolio weight of stock s in fund f's portfolio and $w_{f,s}^{mkt}$ is the weight that this stock would have been assigned had the manager market cap-weighted the equity portfolio. $1\{\cdot\}$ is an indicator function equal to one if stock-level *ASVI* is above the cross-sectional percentile threshold P_τ in each quarter (τ=50). Newey-West (1987) t-statistics with four three lags are reported in parentheses. Fund sentiment is expressed in percentage points. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

Fund Sentiment (%)							
	Low Active	Active 2	Active 3	Active 4	High Active	HML	T
Low Distance	-0.12	0.00	-0.23	0.29	0.13	0.26	[0.34]
Distance 2	0.02	-0.01	0.19	0.44	-0.15	-0.18	[-0.29]
Distance 3	0.26	0.07	0.85	0.51	0.45	0.19	[0.39]
Distance 4	0.42	0.73	0.80	0.50	0.49	0.07	[0.12]
High Distance	1.48*	1.01	0.95	0.12	0.45	-1.04*	[-1.84]
HML	1.60**	1.01	1.19	-0.18	0.31	-1.29	
T	[2.10]	[1.25]	[1.31]	[-0.23]	[0.28]	[-1.54]	

Table A7. Double Sorting Using Alternative Measures

This table presents the value-weighted future returns of mutual funds sorted into 5 × 5 portfolios according to *Holding Distance* and different skill measures (proxied by *Return Gap-RG* in Panel A, *Active Weight-AW* in Panel B and more composite metrics-$Skill^C$ in Panel C). $Skill^C$ is the average rank of more comprehensive composite metrics, including *return gap* (Kacperczyk, Sialm and Zheng 2008), *active weight* (Doshi, Elkamhi and Simutin 2009), *active share* (Cremers and Petajisto 2009), *industry concentration index* (Kacperczyk, Sialm and Zheng 2005), *abnormal cash holdings* (Simutin 2014), *1 minus R-square* (Amihud and Goyenko 2013) and *skill index* (Kacperczyk, Nieuwerburgh and Veldkamp 2014). We independently sort funds into quintiles based on the most recent value of each measure and rebalance the portfolios quarterly. Fund performance is measured by four-factor-adjusted alphas (before expense fees). Newey-West (1987) t-statistics with four lags are reported in parentheses. All returns are expressed in percentage points. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

Panel A. Fund Carhart 4-factor alphas (Before Fee, %/month)

	Low *RG*	*RG* 2	*RG* 3	*RG* 4	High *RG*	HML	T
Low Distance	-0.25***	-0.16**	-0.06	0.03	-0.12	0.14	[1.51]
Distance 2	-0.14*	-0.15*	-0.05	-0.06	-0.17**	-0.03	[-0.31]
Distance 3	-0.03	-0.11	-0.10*	-0.02	0.07	0.09	[0.89]
Distance 4	-0.20***	-0.07	0.12**	0.12**	-0.06	0.14	[1.37]
High Distance	-0.12**	-0.10	-0.01	0.14**	0.10	0.22**	[2.08]
HML	0.13	0.06	0.05	0.11	0.21**	0.08	
T	[1.60]	[0.65]	[0.48]	[1.52]	[2.14]	[0.74]	

Panel B. Fund Carhart 4-factor alphas (Before Fee, %/month)

	Low *AW*	*AW* 2	*AW* 3	*AW* 4	High *AW*	HML	T
Low Distance	-0.16**	-0.08	-0.10	-0.03	-0.23**	-0.06	[-0.51]
Distance 2	-0.12***	-0.06	-0.09	-0.06	-0.05	0.07	[0.69]
Distance 3	-0.09*	-0.07	-0.09	0.00	0.03	0.12*	[1.80]
Distance 4	-0.30*	-0.02	0.05	-0.02	-0.10*	0.20	[1.16]
High Distance	-0.24*	-0.02	-0.04	-0.01	0.02	0.26*	[1.82]
HML	-0.08	0.06	0.06	0.02	0.25***	0.32	
T	[-0.47]	[0.63]	[0.73]	[0.21]	[3.01]	[1.63]	

Panel C. Fund Carhart 4-factor alphas (Before Fee, %/month)

	Low $Skill^C$	$Skill^C$ 2	$Skill^C$ 3	$Skill^C$ 4	High $Skill^C$	HML	T
Low Distance	-0.13	-0.18**	-0.21***	-0.06	-0.08	0.04	[0.37]
Distance 2	-0.20***	-0.07	-0.06	-0.12	-0.07	0.13*	[1.74]
Distance 3	-0.13**	-0.14**	-0.05	0.04	0.06	0.19*	[1.85]
Distance 4	-0.11	-0.10	0.10*	-0.07	-0.03	0.08	[0.83]
High Distance	-0.18*	-0.08	-0.06	0.01	0.15	0.33**	[2.20]
HML	-0.06	0.10	0.15*	0.07	0.23**	0.29**	
T	[-0.58]	[1.05]	[1.90]	[0.87]	[2.55]	[2.30]	

Table A8. Double Sorting Using Alternative Methods and Contents

This table presents the value-weighted future returns of mutual funds sorted into 5 × 5 portfolios according to *Holding Distance* and *Fund Activeness* under different methods and text sources. In Panel A, *Holding Distance* is computed using Bag-of-Words (BoW)-based distance between the full texts of fund strategy narratives and firm 10-K Item 1. Specifically, we represent each document as a word-frequency vector using the BoW approach and compute the cosine distance between the corresponding vectors as the BoW-based pairwise distance. In Panel B, *Holding Distance* is computed using SBERT based distance—defined in Section 2.2— between the full texts of fund strategy narratives (embeddings averaged across all chunks) and firm 10-K Item 1 (embeddings averaged across all chunks). In Panel C, *Holding Distance* is computed using SBERT based distance between the 1st chunk of the fund risk narratives and the 1st chunk of firm 10-K Item 1. Fund performance measures include four-factor-adjusted alphas (both before and after expense fees) as well as the value created by fund managers, measured as *value-added* following Berk and van Binsbergen (2015). We independently sort funds into quintiles based on the most recent value of each measure and rebalance the portfolios quarterly. Fund performance is measured by four-factor-adjusted alphas (before expense fees). Newey-West (1987) t-statistics with four lags are reported in parentheses. All returns are expressed in percentage points. ***, **, and * denote statistical significance at the 1%, 5%, or 10% levels, respectively.

Panel A. BoW with Full Strategy Text

	Low Active	Active 2	Active 3	Active 4	High Active	HML	T
Low Distance	-0.10	-0.16**	-0.06	0.04	0.03	0.14	[1.38]
Distance 2	-0.19*	-0.07	-0.03	0.06	0.05	0.24**	[2.22]
Distance 3	-0.16***	-0.07	-0.11**	0.03	-0.07	0.09	[1.17]
Distance 4	-0.15**	-0.24***	-0.10*	0.09	-0.03	0.11	[1.46]
High Distance	-0.33***	-0.05	-0.03	0.08	0.05	0.38***	[2.98]
HML	-0.22*	0.11	0.03	0.04	0.02	0.24*	
T	[-1.71]	[1.22]	[0.45]	[0.58]	[0.23]	[1.72]	

Panel B. SBERT-distance with Full Strategy Text

	Low Active	Active 2	Active 3	Active 4	High Active	HML	T
Low Distance	-0.17**	-0.19**	-0.01	0.00	-0.08	0.09	[1.05]
Distance 2	-0.12	-0.17**	-0.11**	0.08	-0.01	0.12	[1.14]
Distance 3	-0.20***	0.01	-0.13**	0.02	-0.01	0.18*	[1.80]
Distance 4	-0.22***	-0.08	0.03	0.06	0.15***	0.38***	[3.85]
High Distance	-0.18**	-0.18***	-0.12**	0.14**	0.03	0.22**	[2.21]
HML	-0.01	0.01	-0.11	0.14*	0.11*	0.12	
T	[-0.12]	[0.13]	[-1.61]	[1.86]	[1.78]	[1.11]	

Panel C. SBERT-distance with 1st Chunk Risk Text

	Low Active	Active 2	Active 3	Active 4	High Active	HML	T
Low Distance	-0.25***	-0.15*	-0.08	0.01	-0.05	0.21**	[2.14]
Distance 2	-0.13**	-0.15**	-0.09	-0.04	-0.06	0.07	[0.74]
Distance 3	-0.14**	-0.08*	-0.02	0.03	0.05	0.19**	[2.11]
Distance 4	-0.09	-0.07	-0.09*	0.09*	0.16**	0.25***	[2.68]
High Distance	-0.34***	-0.06	-0.05	0.13*	0.04	0.38***	[3.18]
HML	-0.08	0.09	0.03	0.11	0.09	0.18	
T	[-0.69]	[1.08]	[0.34]	[1.44]	[1.46]	[1.23]	