The First University of Maryland/Singapore Management University/UBS Quant Investment Forum: AI and Finance

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

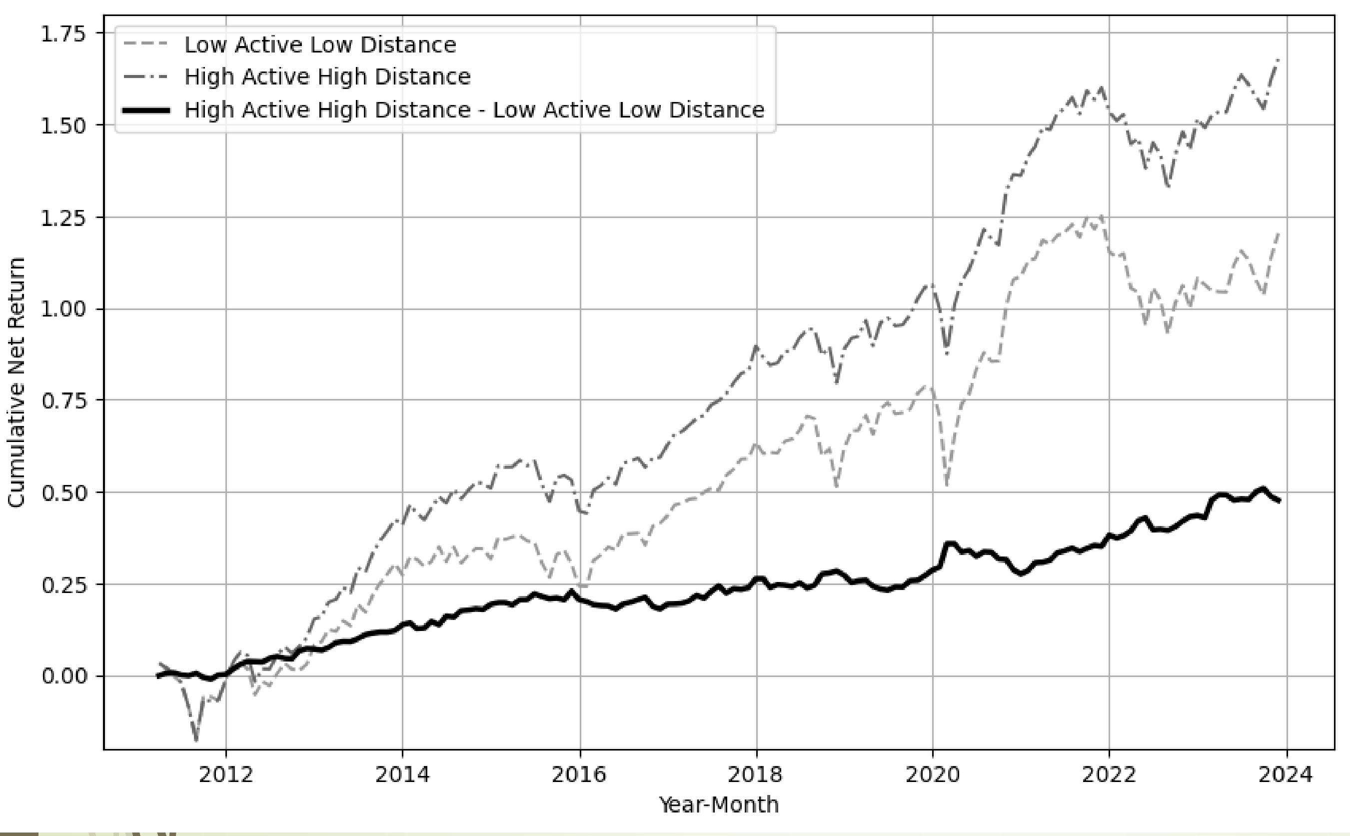
Xiyuan Ma (Singapore Management University)
Matthew Spiegel (Yale University)
Hong Zhang (Singapore Management University)
Yijun Zhou (Florida State University)

Motivation

- In theory, fund skill is critical for **market efficiency** (Gârleanu & Pedersen 2018; Gervais & Strobl 2023)
- Existing skill measures focus on **realized performance** or **portfolio activeness**
  - Carhart (1997), Cremers & Petajisto (2009), Kacperczyk et al. (2005, 2008)

- **Research gap:** Existing measures do not explain **why** funds could realize performance or want to be active in the first place—i.e., how skill works in stock selection.
- We propose a new framework:
  - Skilled managers **actively** apply domain expertise to process **difficult-to-understand information (DUI)**
  - **Fund-specific DUI** = semantic **distance** between fund prospectuses and firm disclosures (10-K)
  - **Skill = Activeness × Distance**; the purpose of activeness is to process information about and then invest in "Distant" firms.

Preview of main analysis



- **Key results: Distant investment**
  - predicts fund performance
  - predicts stock returns
  - Improves price efficiency

Roadmap

- **Fund-Firm Distance and Distant Investments**
- Data and Variables Construction
- Empirical Evidence
  - A Diagnostic Test
  - Distance and Fund Performance
  - Return Predictive Power for Stocks
  - Implications for Market Efficiency
- Additional Tests and Robustness checks
  - Low skilled managers
  - Why distance? Why not just fund or firm information?
  - Robustness checks

One example

Fund Prospectus:
Investing in "*sustainable social, environmental and economic development* *(Impact Companies).*"

Putnam VT Sustainable Future Fund

Putnam INVESTMENTS

This 10-K is ***easy to interpret*** and aligns well with the fund criteria.

But how to quantify?

xylem Let's Solve Water

10-K (Item 1): "*a leading global **water technology** company,*" *which addresses "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*."

One example

Putnam VT Sustainable Future Fund

Putnam INVESTMENTS

xylem Let's Solve Water

Fund-firm distance = 0.6006 (Cosine Distance), 1.096 (L2 Distance), (both around the10% in distribution)

Following Acemoglu, Mühlbach, and Scott (2022), we apply **Sentence-BERT** (Liu et al. 2019; Reimers and Gurevych 2019) to compute embeddings for fund prospectus and firms' *10-K (Item1)*.

Pairwise Fund-firm Distance (between fund f and firm s) is measured as the Euclidean distance between the two embeddings:

$$Pairwise\ Fund\ Firm\ Distance_{f,s} = \| V_f - V_s \| = \sqrt{\sum_{i=1}^{n} \left(V_{f,i} - V_{s,i}\right)^2}. \qquad (1)$$

One example

Putnam VT Sustainable Future Fund

Putnam INVESTMENTS

"***sustainable social, environmental and economic development*** *(Impact Companies).*"

New products are ***technical and hard to comprehend***, semantic distance = ***0.9597 (Cosine)***, ***1.385 (L2)***, both at 99% of the distribution

SeattleGenetics

10-K (Item 1): the company is "*commercializing **ADCETRIS®**, or brentuximab vedotin, for the trbreatment of certain CD30-expressing lymphomas, and **PADCEV™**, or enfortumab vedotin-ejfv, for the treatment of certain metastatic urothelial **cancers**.*

Evaluating SeaGen's investment opportunities requires greater managerial effort and expertise.

One example

Putnam VT Sustainable Future Fund

Putnam INVESTMENTS

xylem Let's Solve Water

SeattleGenetics

- ***Putnam VT*** invested in both firms
- Esp. it started investing in SeaGen after "***PADCEV™***" (one of the new products) first appeared in its 10-K in 2019.
- This investment is highly profitable: *Pfizer* acquired the firm in 2023 for 43 billion USD—nearly twice its April 2020 valuation.

Insights from the example (1):
Distant Investment and two clarifications

- This example illustrates a general idea:
  - **High Semantic Distance** = DUI from the fund's revealed expertise
  - **Distant investment** = Investing in firms with **high semantic distance**
  - It reveals managerial skills to overcome the difficulty of processing DUI.

Xylen: an obvious match, no DUI

- **low distance (no extra performance)**

SeaGen: DUI ➔ uncapitalized opportunity

- **good distance (skill & performance)**

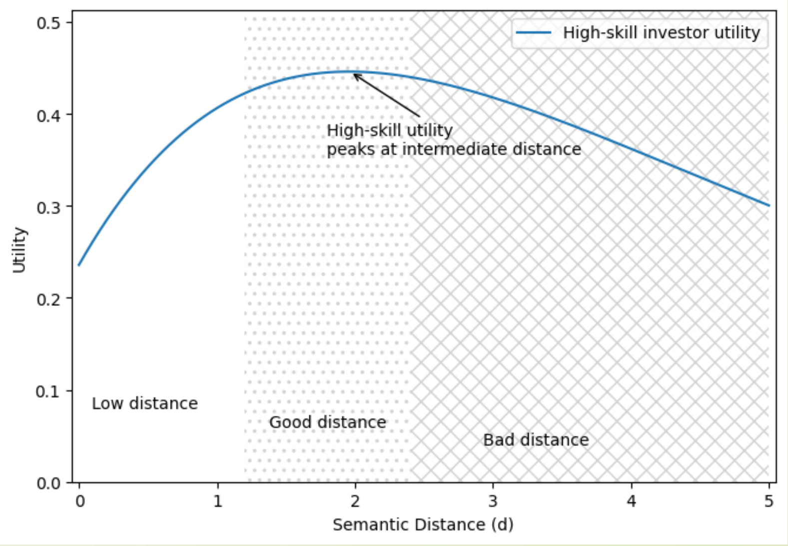
Other unrelated firms

- **bad distance (do not invest)**

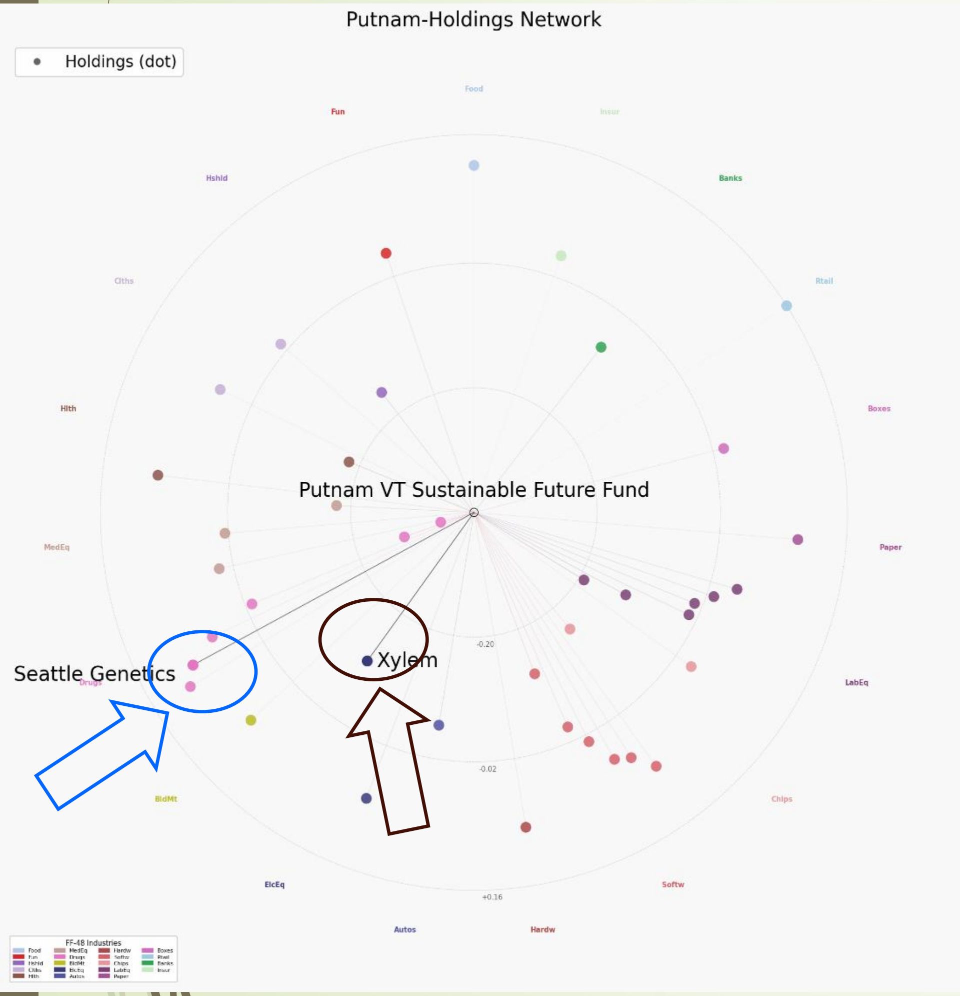
Insights from the example (2):
The economics of "distance"

- In theory, investors **interpret the same news differently** (e.g., Rubinstein 1993), rendering **the value of information** inherently **investor-specific** (Van Nieuwerburgh and Veldkamp 2009; Farboodi et al. 2025).
- **Distance** captures this **investor-specific information value**.
  - **Prospectuses** = Revealed **domain of expertise**
  - **Distance** = **DUI** depends on **who is reading**
  - **Skill** = **Expertise extension**, the ability to apply domain-specific expertise to interpret processable **DUI**.

- A simple model (Kyle 1985) demonstrates that there is an optimal distance (good distance) to apply this skill.
- Here, fund managers extract valuable signals (ρ) from firm disclosure.
  - **Signal precision** $\downarrow$ with **distance** d; **Signal precision** $\uparrow$ with **skill** θ.
  - E.g., $\rho(d, \theta) = e^{-d/\theta}$



Insights from the example (3):
Our Empirical Measure



- **<u>Fund-firm distance</u>**: We focus on firm-specific information and hence use **industry-adjusted distance**.
  - Left: Putnam vs. invested stocks
- **<u>Holding distance of funds</u>**: We then use **active weights** (Doshi, Elkamhi, and Simutin 2015) to aggregate the distance of firms at the fund level.
- **<u>Our empirical measure of skill</u>**: This example and framework leads to a testable measure of **skilled distant investment (SDI)**:

SDI = Activeness (what skilled managers do) **×** ***Distance*** (why)

Roadmap

- Fund-Firm Distance and Distant Investments
- Data and Variables Construction
- **Empirical Evidence**
  - A Diagnostic Test
  - Distance and Fund Performance
  - Return Predictive Power for Stocks
  - Implications for Market Efficiency
- Additional Tests and Robustness checks
  - Why distance? Why not just fund or firm information?
  - Low skilled managers
  - Robustness checks

2. Distance and Fund Performance

- We provide two tests on fund-level performance.
- Double Sorting on
  - ***Activeness*** = average rank of **return gaps** & **active weights**
  - ***Holding Distance* (our measure)**.
  - Fama-French-Carhart (1997) four-factor adjusted **before-fee and after-fee** performance are calculated, along with the performance spread between the highest and lowest quintiles.
- Fama-MacBeth (1973) Regression
  - Fund performance are measured by four-factor adjusted before-fee and after-fee alphas and ***value-added*** (Berk and van Binsbergen 2015).
  - Controlling for fund characteristics such as fund size (TNA), age, turnover, expense, past flow, past returns, and return volatility.

$$Perf_{f,t} = \beta_1 \times Holding\ Dist_{f,t-1} + \beta_2 \times Fund\ Active_{f,t-1} + \beta_3 \times SDI_{f,t-1} + \cdots,$$

2.1 Double Sorting: Activeness predicts performance only among high-distance funds.

- Within the top-distance quintile, the most active funds outperform the most inactive ones by a risk-adjusted before-fee return of **5.4% when annualized**.

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***	-0.09	-0.12**	-0.13	-0.09	0.13	(1.52)
Distance 2	-0.17***	-0.05	-0.15***	-0.05	-0.17*	0.00	(0.04)
Distance 3	-0.13**	-0.09	0.01	-0.01	-0.04	0.09	(0.98)
Distance 4	-0.11	-0.11*	-0.09	0.08	0.00	0.11	(0.83)
High Distance	-0.36***	-0.12*	-0.14**	0.05	0.09	0.45***	(2.90)
HML	-0.14	-0.04	-0.02	0.17*	0.18**	0.32**	
T	(-1.01)	(-0.40)	(-0.32)	(1.70)	(2.58)	(1.99)	

The difference between the HML Active spread within high-Distance funds and that within low-Distance funds. The difference is 0.32% per month (or **3.84% when annualized**), confirming a significant role for distant investments.

2.2 Fama-MacBeth Regression:

SDI significantly enhances predictive power

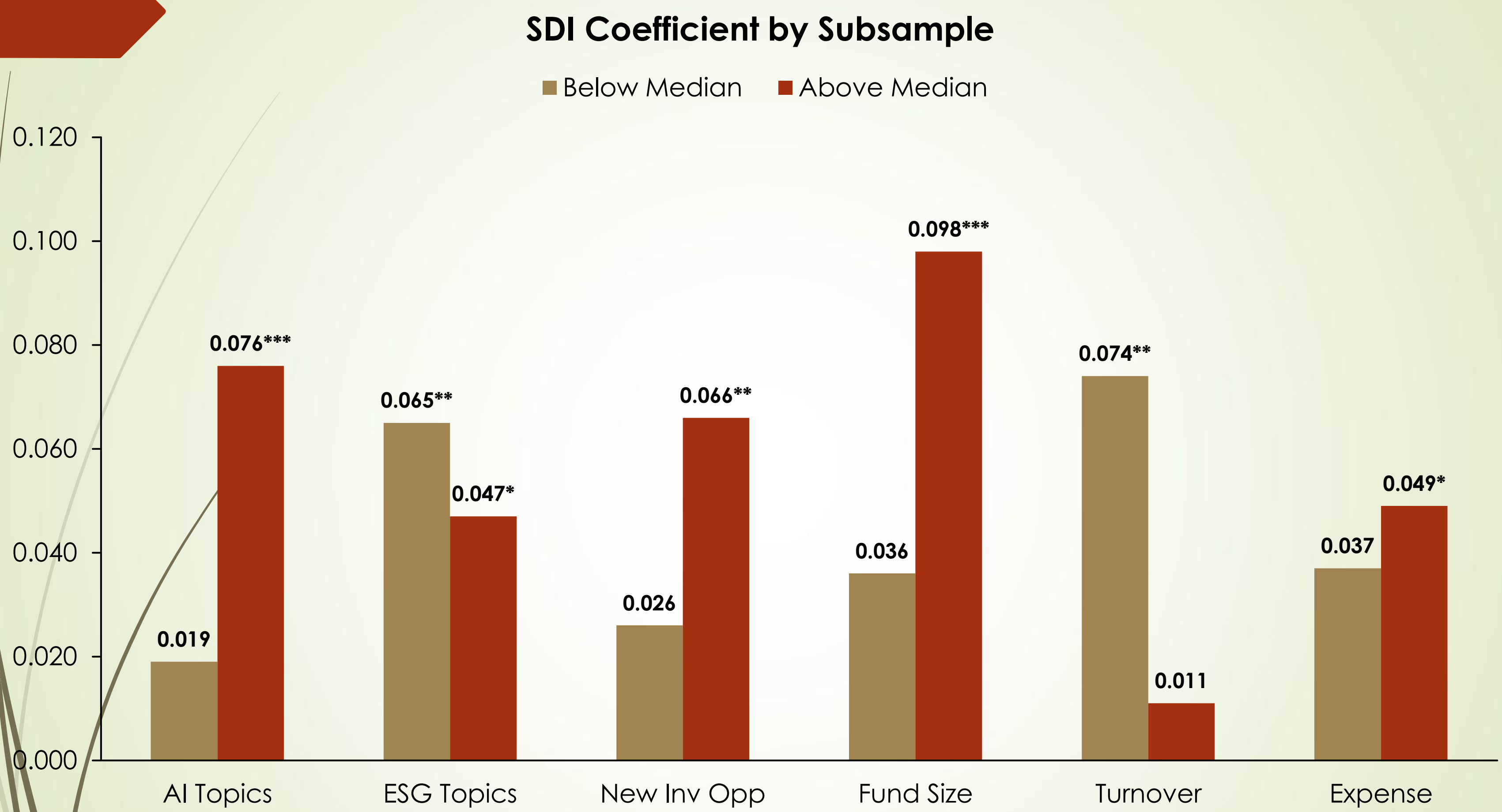
$$Perf_{f,t} = \beta_1 \times Holding\ Dist_{f,t-1} + \beta_2 \times Fund\ Active_{f,t-1} + \beta_3 \times SDI_{f,t-1} + \cdots,$$

	(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 $_{f,t-1}$	0.006		-0.044	0.005		-0.044	-0.124		-1.920
	(0.16)		(-1.03)	(0.14)		(-1.03)	(-0.11)		(-1.31)
Fund Activeness $_{f,t-1}$		0.100**	0.102**		0.098**	0.099**		2.636	3.033
		(2.69)	(2.66)		(2.63)	(2.61)		(1.45)	(1.59)
SDI $_{f,t-1}$			0.049**			0.048**			2.419**
			(2.24)			(2.21)			(2.56)

- **SDI**—the interaction between fund activeness and holding distance—predicts significant out-of-sample performance across all these measures.
- A one-standard-deviation increase in fund activeness is associated with an increase of **3.03** million USD in quarterly value-added.
- **SDI** can enhance this number by **2.42** million USD, representing an **80% relative increase** in value-added.

2.3 Subsample Analyses (Table 5)

SDI coefficient on before-fee Carhart alpha, by subsample



SDI effect is stronger for firms with more emerging opportunities and for larger, higher-turnover, higher-expense funds

- The predictive power concentrates on funds with above-median exposures to **emerging AI** and **comprehensive investment opportunities**.
- The predictive power concentrates on funds with **above-median sizes**, **below-median turnover** ratios, and **above-median fees**.

3. Return Predictive Power for Stocks

- If skilled funds extract value, their trades should predict stock returns.
  - To test, we link **DGTW** adjusted stock returns to **order imbalance** (OI) (Da, Gao, and Jagannathan 2011; Jones et al. 2024).
- We construct a MF OI measure, **skilled distant investment order imbalance (*SDI_OI*)**.

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

  - Specifically, the buy (sell) order — $\boldsymbol{Buy}\ (\boldsymbol{Sell})_{f,s,t}$ — is defined as the percentage increase (decrease) in the invested value of fund in the invested value of fund f in stock s in quarter t.
  - A positive ***SDI_OI*** can be interpreted as a net buy by mutual funds originating from their skilled distant investments.

3. Return Predictive Power for Stocks:

Trading by skilled distant investors predicts returns

$$DGTW_{s,t} = \alpha + \beta \times SDI\_OI_{s,t-1} + \boldsymbol{\Gamma} \times \boldsymbol{Controls}_{s,t-1} + \varepsilon_{s,t}$$

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	***DGTW-adjusted Returns*** $_{s,t}$ ***(%/quarter)***					
SDI_OI $_{s,t-1}$	0.362***		0.395***		0.374***	
	(3.24)		(3.36)		(3.30)	
High SDI_OI $_{s,t-1}$		0.314***		0.306***		0.290***
		(3.35)		(2.93)		(2.83)
AWOI $_{s,t-1}$			-0.087	-0.002	-0.103	-0.022
			(-0.94)	(-0.02)	(-1.13)	(-0.25)

- A one-standard-deviation increase in ***SDI_OI*** is associated with **a 0.362%** higher quarterly DGTW return.

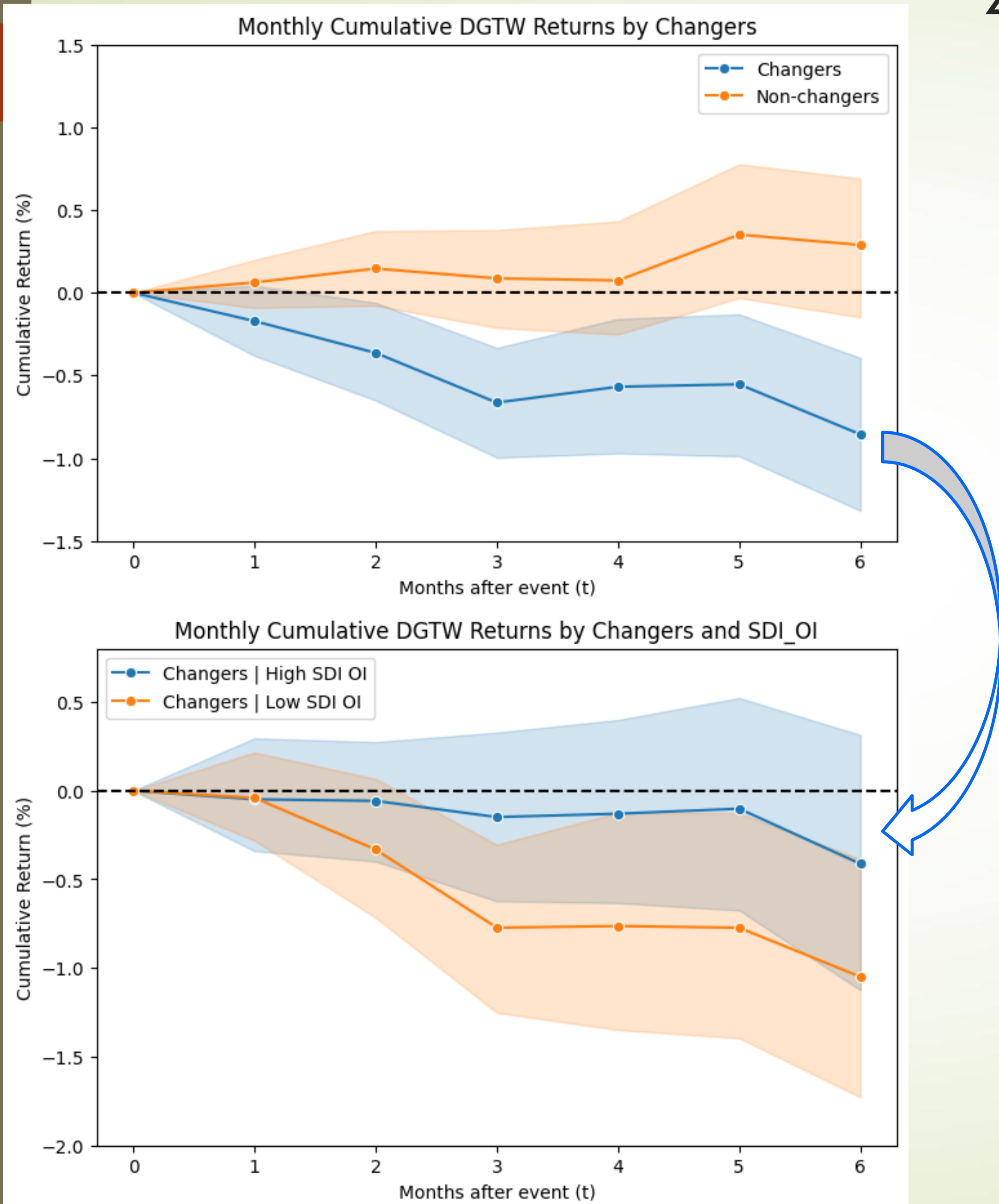
- ***Controls*** include Size, BM, past returns, asset growth (*Investment*), operating profitability, illiquidity, and analyst coverage.

- Adding ***AWOI*** **(activeness implied trading)** or more ***Controls*** such as idiosyncratic volatility, investor attention (captured by Google Search), geographic proximity to funds, and the firm's 10-K does not change our results.

4. Two tests on market efficiency

- Do these trades improve market efficiency?
  - The “lazy price” effect of Cohen, Malloy and Nguyen (2020)
  - Information Shares

4.1 Distant Investment vs. Laze price



- Top Left: lazy price (Cohen, Malloy and Nguyen 2020)
  - Monthly VW cumulative DGTW returns for **"Changers"** and **"Nonchangers"** (above/below median of *Firm Embedding Change)* after the public release of 10-Ks
- Bottom Left
  - **"Changers"** with high skilled mutual fund trading (above median) no longer exhibit lazy price.

Shaded areas present the [5%, 95%] confidence interval.

4.2 Market Efficiency: Information Shares

Table 8: SDI_TI on Return Variance Decomposition (Brogaard et al. 2021)

Key Takeaway: Distant-investing funds **increase firm-specific private information** while **reducing market-wide noise** — consistent with skilled processing of DUI

Roadmap

- Fund-Firm Distance and Distant Investments
- Data and Variables Construction
- Empirical Evidence
  - A Diagnostic Test
  - Distance and Fund Performance
  - Return Predictive Power for Stocks
  - Implications for Market Efficiency
- **Additional Tests and Robustness checks**
  - Low skilled managers
  - Why distance? Why not just fund or firm information?
  - Robustness checks

5.1 Distinctiveness from Text-Based Measures

Table 9: None of the standalone text proxies subsume SDI

Text Proxy	Source	β_3 (Interaction)	Predictsa?
10-K Complexity	Loughran & McDonald (2024)	Insig.	✗ No
10-K Textual Changes	Cohen, Malloy & Nguyen (2020)	Insig.	✗ No
New AI Topics	ChatGPT-identified keywords	Insig.	✗ No
New ESG Topics	ChatGPT-identified keywords	Insig.	✗ No
New ML Inv. Opps.	Basu, Ma & Briscoe-Tran (2022)	Insig.	✗ No
Fund Change	Prospectus textual change	Insig.	✗ No
Prospectus Distinct.	Kostovetsky & Warner (2020)	Insig.	✗ No

SDI = Holding Distance × Fund Activeness $\beta_3 = 0.05^{**}$ ($t = 2.24$) ✓ **Significant**

Insight: Skill requires the **intersection** of fund-specific and firm-specific text — neither alone predicts performance

5. Robustness & Additional Analyses

1 Low-Skill Managers

Flow-driven distant investment by low-activeness funds → buy attention-grabbing stocks → poor performance

⚠

2 Alt. Activeness

Active Share, R^2, tracking error, industry concentration — all produce consistent results

✓

3 Bag-of-Words

Traditional BoW approach confirms results with smaller economic magnitude

✓

4 Full 10-K Item 1 Text

Averaging across all chunks of Item 1 produces consistent but weaker results

✓

5 Placebo: Item 1A

Distance using risk factors (Item 1A) shows insignificant results → confirms strategic specificity

✗

6 Persistence

SDI persists strongly: autocorrelation of 0.85 at 1 quarter, 0.75 at 4 quarters → enduring skill

✓

All tests confirm: SDI captures enduring, economically meaningful managerial skill

Conclusions

- **Main takeaway**

  Skill is not just deviation from benchmarks

- **Skill = the ability to interpret difficult information**
  - Distant investment reveals this ability
  - Improves performance and market efficiency

- We extend textual analysis from "**what is written**" (i.e., traditional ***document-centric measures*** like sentiment) to "**who is reading**" (e.g., Ash, Chen, and Naidu 2026).

Thank you very much!