From Words to Portfolios: Disentangling Narrative and Reality in Mutual Fund Differentiation

Ge Yu[1]

July 31, 2026

Abstract

This paper studies the difference between genuine portfolio innovation and narrative differentiation in mutual funds. Developing a transformer-based portfolio embedding model and combining it with pre-trained large language model embeddings of fund prospectuses, I construct two novel measures of mutual fund differentiation from portfolio holdings and prospectus narratives. Prospectus uniqueness commands higher fees and attracts additional capital when accompanied by an outstanding performance history, while portfolio uniqueness predicts future risk-adjusted performance among skilled managers. The results differ between retail and institutional investors: the flow response to prospectus uniqueness comes primarily from retail investors, while the higher fees it commands are borne mainly by institutional investors. Together, these findings indicate that narrative differentiation in fund disclosures operates primarily as a marketing and pricing tool, whereas real product differentiation through genuinely distinctive portfolios is what generates alpha.

JEL Classification: G11, G23, G14

Keywords: Mutual Fund Differentiation, Deep Learning, Embeddings, Prospectuses

[1]Stockholm School of Economics, Department of Finance. Email: ge.yu@hhs.se. I am grateful to Mariassunta Giannetti, Riccardo Sabbatucci, and Paul Huebner for invaluable guidance and support. I thank Bo Becker, Ramin Baghai, Alvin Chen, Jan Starmans, Gualtiero Azzalini, and Tobias Sichert for helpful comments. I also thank Semyon Malamud for insightful discussions and suggestions, and seminar participants at Stockholm School of Economics for valuable feedback. All errors are my own.

"Such a basis (for product differentiation) may be real or fancied, so long as it is of any importance whatever to buyers, and leads to a preference for one variety of the product over another."

— Edward Chamberlin, *The Theory of Monopolistic Competition* (1933)

1 Introduction

What does it mean for a mutual fund to be unique? The answer depends on how uniqueness is measured, and different measures can point to very different funds. By conventional single-axis metrics, two funds can appear equally differentiated and yet behave nothing alike: one attracts investor flows and charges higher fees without generating alpha, while the other generates alpha without attracting flows or commanding higher fees. This asymmetry is difficult to reconcile with the one-dimensional approach to fund differentiation standard in the literature, such as holdings overlap, active share, tracking error, or prospectus topic distance. The goal of this paper is to separate the ways in which mutual funds differentiate themselves, measure these dimensions independently in a geometry where they can be compared, and document their economic consequences.

This asymmetry is systematically understated in the prior literature because existing measures of fund uniqueness are single-axis by construction. One strand measures uniqueness exclusively from portfolio holdings or returns (Lakonishok, Shleifer and Vishny, 1992; Cremers and Petajisto, 2009; Wahal and Wang, 2011; Sun, Wang and Zheng, 2012). Another measures it exclusively from narrative signals such as fund names or prospectus language (Cooper, Gulen and Rau, 2005; Hoberg and Phillips, 2016). Studies that construct both types of measures treat them as competing proxies in horse-race specifications (Kostovetsky and Warner, 2020) or as robustness checks for one another (Bonelli, Buyalskaya and Yao, 2024; Vozlyublennaia and Wu, 2017). In all three cases, the implicit assumption is that fund uniqueness is a single latent construct: the measures differ only in noise, not in economic

content.

In this paper, I construct two novel measures using deep-learning methods that separate the two margins: *portfolio uniqueness*, capturing a fund's holdings deviation from peers, and *prospectus uniqueness*, capturing the distinctiveness of its Principal Investment Strategy narrative. These measures are weakly correlated (Spearman $\rho = 0.23$). The weak correlation is not a measurement artifact: I show that the two measures predict different outcomes through different economic channels, and that their interaction contains information that neither measure alone can capture. Specifically, prospectus uniqueness commands higher fees but has no predictive content for future risk-adjusted performance; portfolio uniqueness predicts forward alpha among funds with demonstrated managerial skill but has no pricing power; and prospectus uniqueness attracts capital among top-performing funds, while portfolio uniqueness has no flow effect at any performance level.

I construct the two measures using two deep-learning embedding pipelines that translate fund attributes into comparable geometric representations. On the holdings side, I extend the portfolio embedding approach of Gabaix, Koijen, Richmond and Yogo (2025) with a non-linear input structure that retains information about portfolio weights, allowing the model to distinguish between a 0.5% drift and a 3% concentrated bet in the same stock, a distinction central to measuring active management intensity. The retained weight information matters substantially for portfolio representation: trained on CRSP holdings data of all U.S. equity funds from 2011 to 2025, the model achieves 95% top-1 retrieval accuracy on held-out funds and a managed-portfolio benchmark score of 0.713, exceeding the best-performing model in Gabaix et al. (2025). On the text side, I embed each fund's Principal Investment Strategy narrative from its Form N-1A using a pre-trained language embedding model. I then combine the two embedding spaces into cross-modal weighted measures: prospectus uniqueness uses holdings similarity to define each fund's peer group and measures the fund's distance from a holdings-weighted text centroid; portfolio uniqueness reverses the direction. This cross-modal construction makes each measure ask whether a fund is distinctive in one

dimension among funds that are similar in the other dimension. The measures also carry information beyond the full set of existing differentiation proxies: stock-level and industry-level concentration, active share, breadth, tracking error, returns-based selectivity, and the keyword-based measures of Kostovetsky and Warner (2020) jointly explain about a third of the variation in portfolio uniqueness and under 5% of the variation in prospectus uniqueness within category and month, and the fee and flow results strengthen, while the alpha interaction is stable against each proxy individually, when this full proxy set enters the regressions as controls (Appendix G). Figure 1 shows that my measures respond to major economic events: prospectus uniqueness exhibits a secular decline as regulation pushes toward more standardized disclosures, portfolio uniqueness drops sharply during the COVID-19 crash as correlated selling compresses holdings differentiation, and portfolio uniqueness declines again from 2024 as the AI-driven Magnificent Seven concentration pulls portfolios toward overlapping mega-cap positions.

I test the predictions of the two-margin framework using a panel of U.S. equity mutual funds from January 2011 to October 2025 with progressively saturated fixed effects, following Kostovetsky and Warner (2020) and Abis and Lines (2024). For fund fees, a one-standard-deviation increase in prospectus uniqueness raises annualized fees by approximately 1.6 basis points, significant at the 1% level across six specifications, while portfolio uniqueness has no pricing power in any specification. For future alpha, portfolio uniqueness predicts forward risk-adjusted performance only when paired with managerial skill. I interact uniqueness with past-performance terciles: the interaction with the top tercile is positive and significant in both Fama–MacBeth and panel fixed-effects specifications, while the base coefficient on portfolio uniqueness for average managers is small or negative. The interaction captures information beyond performance persistence alone. It remains statistically significant after directly controlling for the continuous level of past alpha and its interactions with tercile indicators. The result is also replicated using an external classification based on Morningstar star ratings. Moreover, the interaction's t-statistic lies at the 99th percentile of the distribution

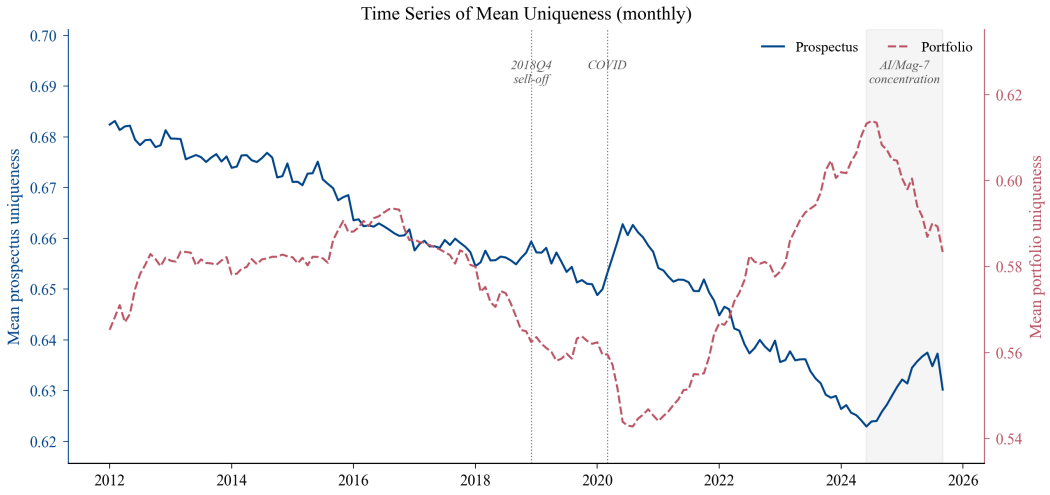


Figure 1
Time Series of Prospectus and Portfolio Uniqueness, 2011–2025

This figure shows monthly cross-sectional means of prospectus uniqueness and portfolio uniqueness, January 2011 through October 2025. Both measures are constructed as cross-modal weighted distances using the embedding pipelines described in Section 3.2. Annotations highlight three episodes discussed in the text: the secular decline in prospectus uniqueness associated with disclosure standardization, the sharp drop in portfolio uniqueness during the COVID-19 crash in early 2020, and the renewed compression in portfolio uniqueness from 2024 onward as the AI-driven mega-cap concentration pulls active portfolios toward common positions.

generated by 200 placebo classifications randomly constructed within category-month cells (Appendix H). This pattern suggests that for skilled managers, strategy uniqueness reflects occupying a differentiated product niche, such as less crowded trades (Hoberg, Kumar and Prabhala, 2018) or lower return correlation with peers (Amihud and Goyenko, 2013), where reduced competition allows skilled managers to earn additional rents. For managers without demonstrated skill, distinctive portfolios may instead reflect costly idiosyncratic bets driven by overconfidence (Barber and Odean, 2001) or excessive trading (Puetz and Ruenzi, 2011) rather than informed deviations from the herd.

Finally, I ask whether and which measure of uniqueness attracts investor capital. The results show that prospectus uniqueness attracts capital exclusively among top performers, while portfolio uniqueness has no flow effect at any performance level. The interaction structure is the inverse of the alpha result: prospectus uniqueness attracts flows where

portfolio uniqueness generates alpha. Together, the three findings establish that prospectus uniqueness operates as a marketing and pricing tool while portfolio uniqueness operates as a skill-dependent performance channel, and the two are largely non-overlapping in their economic roles. The flow interaction remains informative after accounting for the convexity of the flow–performance relation (Sirri and Tufano, 1998; Huang, Wei and Yan, 2007). Funds with unique prospectuses are less volatile and sit no higher within the winner bucket, the interaction survives replacing the performance buckets with within-month rank vigintile dummies, none of 200 random within-month placebo gates reproduces it, and the response comes from retail investors in the steepest segment of the curve (Appendix I).

The results above do not necessarily warrant a causal interpretation: managers choose their portfolios and prospectuses, raising standard reverse-causality and omitted-variable concerns. I address these concerns with two distinct instrumental variable strategies, one for each causal channel. For the prospectus-uniqueness-to-fees channel, I instrument prospectus uniqueness with the 12-month first difference in the fund family's AI workforce share, constructed from Revelio Labs employment data. The intuition follows Babina, Fedyk, He and Hodson (2024): AI adoption standardizes prospectus language and reduces text uniqueness. First-differencing strips out slow-moving family characteristics such as size, sophistication, and technology orientation, and isolates idiosyncratic timing variation in AI hiring. The instrument yields a 2SLS estimate of 33 basis points per standard deviation of prospectus uniqueness, substantially larger than the OLS estimate of 1.6 basis points, consistent with attenuation bias from measurement error and a local average treatment effect among complier funds whose prospectus language is most responsive to AI adoption. Placebo tests show that the instrument does not predict alpha, flows, returns, or fund turnover, and pre-trend tests confirm that it does not capture pre-existing fee dynamics.

For the portfolio-uniqueness-to-alpha channel, I instrument portfolio uniqueness with the similarity-weighted fraction of within-category peers that exited in the prior six months. The intuition is that peer exits reflect liquidity pressures at the exiting funds (Coval and

Stafford, 2007; Lou, 2012) rather than information about the focal fund's future alpha, but they nonetheless mechanically alter the focal fund's position relative to its remaining peers. The instrument is strong and yields a 2SLS estimate of 24 basis points per standard deviation of portfolio uniqueness on next-month forward FFC6 alpha. Tercile-subsample 2SLS preserves the skill-dependent monotonic pattern from the panel regressions: the effect is positive and marginally significant for top-tercile funds, positive for middle-tercile funds, and small and insignificant for bottom-tercile funds. Validity tests show that the peer-exit instrument does not predict key fund characteristics such as Morningstar rating, Sharpe ratio, or management fees, and pre-trend tests confirm that it does not capture an existing trend in fund alpha. Both instruments deliver weak-instrument-robust statistical significance, supporting a causal interpretation.

Decomposing the panel regressions on fees and flows by CRSP fund-type flags, I further find that the higher fees and additional inflows associated with prospectus uniqueness come from different populations of investors. Among institutional funds, a one-standard-deviation increase in prospectus uniqueness raises fees by 2.6 to 2.8 basis points, with t-statistics roughly twice as large as in the retail subsample. Among retail funds, the same effect is approximately 1 basis point and only marginally significant. Portfolio uniqueness remains insignificant in both subsamples. The flow channel runs in the opposite direction. Retail flows respond strongly to prospectus uniqueness among top performers and even penalize poor performers with unusual prospectuses by withdrawing capital. Institutional flows are essentially unresponsive to prospectus uniqueness at any performance level. A distribution-channel split traces the retail side of this asymmetry to its point of sale: both the winner-chasing flows and the loser penalty on prospectus uniqueness concentrate in broker-sold share classes, where an intermediary delivers the strategy text to the investor (Section 5.4). The retail and institutional segments therefore split along orthogonal dimensions: institutional investors pay higher fees for distinctive prospectuses but do not chase them with capital, while retail investors chase distinctive prospectuses with capital but pay

only marginally higher fees. I argue that the higher fees paid by institutional investors are best understood as a rent on the justification value of distinctive language inside consultant-mediated and committee-mediated allocation processes. Crucially, portfolio uniqueness, the channel that actually generates alpha, is insignificant for fees in both subsamples. Whatever institutional investors are paying for, they are not paying for genuine product differentiation.

Overall, this paper contributes to the literature by showing that mutual fund differentiation operates along two distinct margins, portfolio choice and prospectus language, and that these two margins have different implications for fees, flows, and performance. Methodologically, I develop a scalable framework for measuring portfolio and prospectus uniqueness in a comparable setting and extend recent embedding-based approaches to portfolio representation by incorporating economically meaningful variation in portfolio weights. A suite of interpretability analyses demonstrates that the learned representations capture economically meaningful information. Embedding distance predicts realized return comovement among widely held stocks and across funds; out-of-sample probes recover fund characteristics and return-based properties not observed during training; and the resulting embedding space enables direct detection of closet indexing and measurement of style drift (Appendix E). Empirically, prospectus uniqueness is associated with higher fees and performance-contingent flows, whereas portfolio uniqueness predicts alpha only among skilled managers, challenging the common assumption that uniqueness is a single latent characteristic that can be measured interchangeably across different fund attributes. Finally, two separate instrumental-variable strategies provide causal evidence for these channels. These findings speak to several strands of literature, including work on mutual fund competition, the industrial organization of asset management, mandatory fund disclosures, and the use of machine learning in finance.

Related Literature. This paper contributes to four distinct strands of the asset management literature.

Mutual fund competition. A long literature studies how mutual funds compete with

one another and how that competition shapes performance, fees, and flows (e.g., Lakonishok et al., 1992; Wermers, 1999; Berk and Green, 2004; Kacperczyk, Sialm and Zheng, 2005; Cremers and Petajisto, 2009; Wahal and Wang, 2011; Hoberg et al., 2018; Bonelli et al., 2024). The closest precedents are Sun et al. (2012), who construct a Strategy Distinctiveness Index from hedge fund returns and show that distinctive funds earn persistent alpha, and Hoberg et al. (2018), who measure mutual fund competition from prospectus text and find that funds in less competitive niches earn higher alpha. Both papers measure distinctiveness on a single axis. My contribution is to add the missing axis. By measuring holdings and prospectus distinctiveness jointly in a cross-modal geometry, I show that the two dimensions do different economic work: holdings distinctiveness rewards skill but earns no fees, while prospectus distinctiveness earns fees and attracts performance-contingent flows but carries no information about future returns.

Industrial organization of asset management. A second strand studies how heterogeneous investors interact with fund families, how fees are set, and how marketing intermediates investor choice (e.g., Sirri and Tufano, 1998; Hortaçsu and Syverson, 2004; Choi, Laibson and Madrian, 2010; Christoffersen, Evans and Musto, 2013; Del Guercio and Reuter, 2014; Gennaioli, Shleifer and Vishny, 2015; Berk and van Binsbergen, 2015; Roussanov, Ruan and Wei, 2021). The closest precedents are Hortaçsu and Syverson (2004), who show that search frictions sustain fee dispersion uncorrelated with quality even among S&P 500 index funds; Kostovetsky and Warner (2020), who measure fund product differentiation from both holdings and text but treat the two as competing proxies for a single construct; and Bonelli et al. (2024), who formalize the differentiation-and-pricing trade-off on a single broad-vs-niche axis. My contribution is to identify a segment-level asymmetry between who pays for prospectus uniqueness and who chases it: the higher fees are concentrated in the institutional segment, while the corresponding flow response comes primarily from retail investors. This pattern is hard to reconcile with settings in which fund selection and fee incidence are aligned within the same investor type, and it suggests that institutional discipline of fund pricing may be

incomplete when differentiation takes a narrative form.

Mutual fund disclosures. A third strand studies what funds say in their mandatory regulatory disclosures and how those disclosures shape investor behavior (e.g., Cooper et al., 2005; Hillert, Niessen-Ruenzi and Ruenzi, 2025; Cohen, Malloy and Nguyen, 2020; deHaan, Song, Xie and Zhu, 2021; Abis and Lines, 2024). The closest precedent is Abis and Lines (2024), who document systematic style drift between prospectus language and actual portfolio behavior, the most direct evidence that fund disclosures and fund actions can diverge over time. My contribution is to introduce a scalable cross-modal framework that compares disclosure content directly against actual portfolio holdings within the same fund-month, measuring narrative and portfolio distinctiveness in the cross-section rather than studying disclosure in isolation or studying drift over time.

AI and embedding methods in finance. Finally, this paper extends an emerging literature applying deep-learning embeddings and large language models to financial data (e.g., Tetlock, 2007; Loughran and McDonald, 2011; Jegadeesh and Wu, 2013; Loughran and McDonald, 2016; Hoberg and Phillips, 2016; Araci, 2019; Li, Mai, Shen and Yan, 2021; Dolphin, Smyth and Dong, 2022; Gabaix et al., 2025). I extend the portfolio embedding architecture of Gabaix et al. (2025) with cardinal weight preservation via a learned nonlinear projection, and combine it with pre-trained text embeddings into a cross-modal weighted measurement framework. To the best of my knowledge, this is the first such combination in the asset management literature.

The remainder of the paper is organized as follows. Section 2 develops a two-stage model of fund differentiation that yields three predictions linking prospectus uniqueness and portfolio uniqueness to fees, alpha, and segment-specific flows. Section 3 describes the data sources, the deep-learning embedding pipelines for portfolios and prospectuses, and the construction of the cross-modal uniqueness measures. Section 4 presents the panel regression and instrumental variable evidence. Section 5 documents the segment-level asymmetry between retail and institutional investors in who pays the higher fees and who chases the additional

flows. Section 6 concludes. Formal apparatus and proofs for the theoretical framework are collected in Appendix A.

2 Theoretical Framework

In this section, I develop a model of monopolistic competition among delegated portfolio managers who differentiate on two margins: the strategy narrative they disclose in the prospectus and the portfolio they implement. The two margins have different production costs and different observability. Narrative is cheap to produce and visible to every investor at the point of selection. A distinctive portfolio is costly, raises expected return only for skilled managers, and is observed before allocation by institutional investors but not by retail investors. Fund performance is subject to decreasing returns to scale, so the framework respects the equilibrium accounting of delegated asset management (Sharpe, 1991; Berk and Green, 2004; Fama and French, 2010). The model nests a frictionless monopolistic-competition benchmark and adds one friction: institutional capital enforces consistency between what a fund claims and what it holds. This enforcement friction generates the distinctive predictions of the paper.

I use the model to organize the empirical patterns documented in Sections 3 and 4. Three predictions anchor the empirical work. Share-class fees rise with prospectus uniqueness, and the institutional response is attenuated where claims outrun holdings. Gross alpha, conditional on scale, rises with portfolio uniqueness among skilled managers, while prospectus uniqueness does not predict forward gross alpha. Retail and institutional share classes price prospectus uniqueness differently, and the difference carries an enforcement component. I state the model in this section and collect notation, assumptions, the lemma chain, and proof sketches in Appendix A.

2.1 Environment

A continuum of funds is indexed by fundamental skill $\theta_i \in [\underline{\theta}, \bar{\theta}]$, with $\underline{\theta} < 0 < \bar{\theta}$ and continuous density $g(\theta_i)$. The support crosses zero, so a positive mass of funds destroys value before fees, consistent with the value-weighted accounting of Fama and French (2010). Neither the manager nor outside investors observe θ_i. I impose a common posterior $\theta \equiv \mathbb{E}[\theta_i \mid \mathcal{I}]$ held by the manager and institutional investors, formed from a common information set $\mathcal{I}$ (portfolio observability, due-diligence signals, realized performance), with posterior variance $\Sigma \equiv \mathrm{Var}[\theta_i \mid \mathcal{I}]$ that declines with the information history and is therefore smaller for older funds. I write θ for the posterior at the decision time. Retail investors do not share this posterior; their signal is the fund's past performance ϱ, a noisy proxy for skill. The common-posterior assumption shuts down signaling through the narrative and isolates the enforcement channel.

Each fund chooses four actions in two stages. At Stage 1 the manager commits to a prospectus narrative $s_T \in [0, 1]$ and posts retail and institutional share-class fees $p_R, p_I \geq 0$. At Stage 2 the manager implements a portfolio $s_P \in [0, 1]$. Both s_T and s_P are uniqueness indices: 0 is conformity with category peers and 1 is maximal differentiation. Institutional investors observe s_P before allocating; retail investors do not. First-period assets equal realized demand, $a = D_1$, and assets feed back into performance through decreasing returns to scale, so a is a rational-expectations fixed point.

2.2 Fund Performance

Realized risk-adjusted return depends on skill, portfolio differentiation, and scale:

$$r = \theta_i \left(1 + \lambda_\alpha s_P\right) - \Gamma(s_P)\, a + \xi, \qquad \Gamma(s_P) = \Gamma_0 + \Gamma_1 s_P, \qquad \xi \sim \mathcal{N}\big(0,\ \sigma_\xi^2 \left(1 + \rho s_P^2\right)\big). \quad (1)$$

The term $\lambda_\alpha s_P \theta_i$ is a skill-portfolio complementarity: the payoff to differentiation rises with skill, $\lambda_\alpha > 0$. The term $\Gamma(s_P)\, a$ is the decreasing-returns-to-scale drag of Berk and Green

(2004), documented by Chen, Hong, Huang and Kubik (2004) and Pástor, Stambaugh and Taylor (2015). The capacity coefficient is steeper for differentiated portfolios, $\Gamma_1 > 0$, because niche strategies absorb capital with greater price impact, consistent with active share declining in fund size (Cremers and Petajisto, 2009). The term ρs_P^2 makes niche portfolios intrinsically noisier.

Equation (1) is the realized risk-adjusted return. The fund's alpha is its expected value formed on the common posterior, the expected excess return over the category benchmark:

$$\alpha \;\equiv\; \mathbb{E}[r \mid \theta, s_P, a] \;=\; \theta - \Gamma_0 a + \lambda_\alpha \tilde{\theta}\, s_P, \qquad \tilde{\theta} \equiv \theta - \frac{\Gamma_1}{\lambda_\alpha}\, a, \tag{2}$$

with $\tilde{\theta}$ the capacity-adjusted posterior. The marginal alpha from differentiation is $\partial\alpha/\partial s_P = \lambda_\alpha \tilde{\theta}$: skill clears the scale hurdle before differentiation pays, and large funds face a higher hurdle. The realized return's conditional variance, $\mathrm{Var}[r \mid \theta, s_P] = \sigma_\xi^2(1+\rho s_P^2) + (1+\lambda_\alpha s_P)^2\,\Sigma$, combines idiosyncratic noise with posterior skill uncertainty Σ; the second term is largest for young funds. The closed forms below use the seasoned-fund approximation $\Sigma \approx 0$, and I recover its content as an age interaction in Prediction 4.

A normalization fixes the location of the skill distribution. The category benchmark is defined so that size-weighted alpha integrates to zero within each category, $\int a^*(i)\,\alpha_i^*\,dG(i) = 0$. Funds are atomistic and take the benchmark as given, so the normalization pins the location of g without entering any fund's first-order condition. Because the fund sector trades against non-fund investors, the identity holds only approximately within the sector (Pedersen, 2018). However, none of the predictions from the model requires the identity to hold exactly.

2.3 Investor Demand

There are two investor segments, retail (R) and institutional (I), with masses $\mu_R + \mu_I = 1$. A segment-j investor who selects fund j receives utility $u_j = \widehat{v}_j - p_j + \varepsilon_{T,j}(s_T)$, where

$\widehat{v}_j$ is expected fund quality and $\varepsilon_{T,j}$ is an idiosyncratic match shock with distribution $F_{s_T,j}$. Following Bar-Isaac, Caruana and Cuñat (2012), $F_{s_T,j}$ rotates around a segment-specific fixed point as s_T rises: more distinctive narrative moves mass into both tails of the match shock while holding the mean fixed. The institutional match-shock distribution is less dispersed than the retail one.

Retail investors observe the narrative and past performance. Their valuation is

$$\widehat{v}_R = \varrho\,(1 + \varphi_c\, s_T)\,, \tag{3}$$

where ϱ is category-time-demeaned past performance and $\varphi_c > 0$ is a salience amplification through the narrative channel. The salience loads on the distribution channel c: it is stronger where an intermediary pitches the fund, so φ_c is larger in broker-sold than in direct-sold channels (Sirri and Tufano, 1998; Del Guercio and Reuter, 2014). Retail investors do not detect a gap between the prospectus and the portfolio at the point of selection; they learn of it after returns disappoint.

Institutional investors observe the narrative, past performance, the implemented portfolio, the scale of the fund, and the common posterior. Their valuation is

$$\widehat{v}_I = q\,\theta + \omega\, s_T + \mu_\alpha\,\alpha - \Psi_I(s_T, s_P), \tag{4}$$

where $q > 0$ weights a quality signal correlated with skill, $\omega > 0$ is the gross value of a distinctive disclosed strategy, $\mu_\alpha \geq 0$ weights the fund's alpha α from equation (2) (which embeds the capacity drag), and Ψ_I is the enforcement discount defined below. Demand in each segment is $D_j = \mu_j\,[1 - F_{s_T,j}(U - \widehat{v}_j + p_j)]$, where U is the passive outside option, and aggregate first-period assets are $a = D_1 = D_R + D_I$.

2.4 Consistency Enforcement

The distinctive primitive of the model is a friction that penalizes a fund whose prospectus claims more than its portfolio delivers. Institutional capital flows through a delegation chain of consultants, outsourced chief investment officers, investment committees, and boards. Their due-diligence protocols treat consistency between stated strategy and holdings as a contractible, auditable criterion. Capital governed by such protocols cannot be deployed to a fund that fails the criterion, whatever any individual allocator privately believes about skill. The penalty is therefore an agency artifact of verifiability in delegation. It operates even though, under the common posterior, the gap conveys no information that institutions do not already possess. I refer to the premise that the narrative does not enter the return technology (1) as the non-productive-narrative premise.

Let the over-promising gap be $(s_T - s_P)_+$ and the under-promising gap be $(s_P - s_T)_+$. A fraction $q_I \in (0, 1]$ of institutional capital enforces the standard, applying the valuation discount

$$\Psi_I(s_T, s_P) = q_I \left[\psi^+ (s_T - s_P)_+^2 + \psi^- (s_P - s_T)_+^2\right], \qquad \psi^+ > \psi^- \geq 0, \tag{5}$$

and the manager bears a direct legal-and-reputational cost on the over-promising side, $c_L = \kappa_L (s_T - s_P)_+^2$ with $\kappa_L \geq 0$. Enforcement is stronger on the over-promising side, $\psi^+ > \psi^-$ (Condition AE in Appendix A). Two features support the asymmetry. Liability for materially misleading statements and the Names Rule's investment-policy requirement attach to affirmative claims, and there is no comparable action against conservative language paired with distinctive holdings, so the legal cost κ_L exists only on the over-promising side. Style-drift policing operates largely at the benchmark-assignment margin, which the category-by-month empirical design absorbs, so the within-cell under-promising intensity ψ^- is small. Retail enforcement $q_R \geq 0$ is small in the baseline and rises over the sample as third-party drift monitoring and Regulation Best Interest take hold. The baseline compar-

ative statics hold at $q_R = 0$, and the secular rise in retail monitoring gives the segment contrast a testable time trend.

2.5 The Fund's Problem

The manager maximizes a certainty-equivalent objective with risk aversion γ_M over fee revenue,

$$\mathrm{CE} = \mathbb{E}[\pi] - \gamma_M \mathrm{Var}[\pi] - c_T(s_T) - c_P(s_P) - c_L(s_T, s_P), \tag{6}$$

with production costs $c_T = \kappa_T s_T^2$ and $c_P = \kappa_P s_P^2$ and $\kappa_T \ll \kappa_P$. Profit accrues over two periods, $\pi = p_R D_R + p_I D_I + \delta \bar{p}\,(D_1 + \Phi(r) + \varepsilon_A)$, where $\bar{p} = \mu_R p_R + \mu_I p_I$ is the segment-mass-weighted fee and $\delta \in (0,1)$ is the discount factor. Period-2 assets respond to the realized return through the flow schedule Φ, with $\Phi' > 0$ and $\Phi'' \geq 0$. The convexity of Φ is the flow-performance relation of Sirri and Tufano (1998) and Chevalier and Ellison (1997). A second-order expansion around the fund's alpha gives $\mathbb{E}[\Phi(r)] \approx \Phi(\alpha) + \frac{1}{2}\chi_2 \mathrm{Var}[r]$, so the certainty equivalent loads on $\mathrm{Var}[r]$ with coefficient $-\tilde{\gamma}$, where

$$\tilde{\gamma} = \gamma_M \delta^2 \bar{p}^2 \chi^2 - \tfrac{1}{2}\delta \bar{p}\, \chi_2. \tag{7}$$

The manager is risk-averse in fee revenue through γ_M and variance-loving through the convex flow schedule through χ_2. I solve the model on the region where risk aversion dominates, $\tilde{\gamma} > 0$. Where the local convexity is large, empirically in the bottom tail of past performance, the sign reverses and a low-skill manager can choose high portfolio uniqueness as a gamble on variance (Brown, Harlow and Starks, 1996). I treat that region as a risk-shifting regime with its own prediction (Prediction 5) and scope the results below to $\tilde{\gamma} > 0$.

I solve by backward induction. At Stage 2 the manager treats scale as the fixed point a^* and chooses s_P. The first-order condition trades the marginal expected-alpha gain $\lambda_\alpha \tilde{\theta}$ against the marginal variance cost $2\tilde{\gamma}\rho\sigma_\xi^2 s_P$, the marginal production cost $2\kappa_P s_P$, the gap-closing pull $2Q^+ (s_T - s_P)_+$ on the over-promising side, and the drift drag $2Q^- (s_P - s_T)_+$ on

the under-promising side, with $Q^+ = \tau_I \mu_I \bar{f}_I q_I \psi^+ + \kappa_L$ and $Q^- = \tau_I \mu_I \bar{f}_I q_I \psi^-$. The optimum is piecewise linear:

$$s_P^* = \begin{cases} 0 & \tilde{\theta} \leq -(\nu/\zeta) s_T \quad \text{(closet-indexing region)}, \\ \zeta\, \tilde{\theta} + \nu\, s_T & -(\nu/\zeta) s_T < \tilde{\theta} < s_T/\zeta_0 \quad \text{(over-promising region)}, \\ \zeta^- \tilde{\theta} + \nu^- s_T & \tilde{\theta} \geq s_T/\zeta_0 \quad \text{(under-promising region)}, \end{cases} \tag{8}$$

with coefficients given in Appendix A. Four properties drive the empirical work. Portfolio uniqueness rises with the posterior in every interior regime, through $\tilde{\theta}$. It rises with the narrative in the over-promising region, with slope $\nu \in (0, 1)$, because the manager closes the gap to relax the enforcement pull. Funds with sufficiently negative capacity-adjusted skill hold the benchmark portfolio $s_P = 0$, so closet indexing emerges as an equilibrium mass. And in a window-dressing region, $\tilde{\theta} \in (-\nu s_T/\zeta, 0)$, the fund sustains positive portfolio uniqueness because the committed narrative and the enforcement pull demand it, even though differentiation lowers its expected alpha. Larger funds choose less distinctive portfolios, $\partial s_P^*/\partial a^* < 0$.

At Stage 1 the manager sets fees and narrative under rational expectations. Each fee first-order condition is an inverse-hazard pricing rule (Bonelli et al., 2024): the optimal share-class fee is proportional to the inverse hazard rate of the segment's match-shock distribution at the indifference threshold. More distinctive narrative rotates the distribution and lowers the hazard rate, so both fees rise with s_T. Decreasing returns add a capacity-pricing term: because demand erodes alpha, raising the fee protects performance, which raises the optimal markup where $\Gamma(s_P)$ is large. The institutional fee carries the enforcement penalty in the over-promising region, which attenuates its response to s_T relative to the no-enforcement benchmark. The narrative first-order condition balances the rotation and salience gains in demand against the enforcement terms. For a low-skill manager the gap at any high s_T is large and the enforcement term binds, so the manager chooses interior narrative $s_T^* < 1$. This breaks the corner race of a pure rotation model, in which narrative is cheap and every

manager sets $s_T = 1$.

Text production costs are heterogeneous. A fraction π_H of funds carries a high coefficient κ_T^H, capturing boilerplate templating: many families generate prospectus language from shared legal templates, and deviating carries compliance and operational costs unrelated to skill. Templated funds hold the narrative down at any skill level, so the population of funds with distinctive holdings and conventional language is populated by templated skilled funds. This distinction between templating and enforcement becomes a separation test in Prediction 6.

2.6 Equilibrium Predictions and Empirical Hypotheses

The model delivers three predictions that anchor the empirical work, stated here and derived in Appendix A.

Prediction 1 (Fees and prospectus uniqueness)**.** *Both share-class fees rise with prospectus uniqueness through the inverse-hazard rotation channel, so the cross-sectional relation between fees and ProspectusUniq is positive and operates through observability and search frictions (Choi et al., 2010; Hortaçsu and Syverson, 2004; Roussanov et al., 2021). In the region where claims outrun holdings the institutional fee response to ProspectusUniq weakens, and the weakening grows with enforcement intensity.*

Prediction 2 (Alpha and portfolio uniqueness)**.** *Conditional on scale, expected gross alpha rises with portfolio uniqueness among funds with positive capacity-adjusted skill and falls with it in the window-dressing region, so the uniqueness-alpha gradient is $\lambda_\alpha \tilde{\theta}$ and switches sign at $\tilde{\theta} = 0$. Tests use gross, category-adjusted alpha with log fund size and log family size in the conditioning set, and the Berk and van Binsbergen (2015) value-added measure as the robustness skill object. The interaction of PortfolioUniq with the skill posterior is the discriminating specification, because pure decreasing returns predict a uniform uniqueness premium while the model predicts a premium concentrated where the capacity-adjusted poste-*

rior is positive. ProspectusUniq does not predict forward gross alpha after this conditioning set.

Prediction 3 (Segment-level pricing and flows)**.** *The institutional and retail fee responses to ProspectusUniq differ, and the institutional response carries an enforcement component absent from the retail response; the difference is proportional to the enforcement gap* $q_I - q_R$ *and compresses as retail monitoring rises. Retail flows respond to ProspectusUniq in proportion to past performance, so distinctive narrative attracts capital among recent winners and repels it among recent losers, with the amplification stronger in broker-sold channels.*

The framework therefore decomposes fund differentiation into objects with distinct empirical incidence. Prospectus uniqueness maps to fees, salience, and retail flow sensitivity. Section 5 directly tests the distribution-channel component of Prediction 3. Classifying share classes by distribution channel shows that both the winner amplification and the loser penalty associated with *ProspectusUniq* are concentrated in broker-sold retail classes, whereas institutional classes exhibit neither response. This pattern suggests that the amplification arises where an intermediary conveys the narrative. Portfolio uniqueness maps to gross alpha conditional on scale, to risk, and to capacity. The segment split in pricing reflects institutional enforcement of consistency between the two margins. The framework sharpens this interpretation into two further discriminating implications: enforcement intensity rises with institutional oversight of the fund, and templating compresses prospectus uniqueness within fund families. The templating pattern is visible in Appendix F, where a fund's nearest text neighbor belongs to its own family 41% of the time against a 0.6% chance rate and prospectus uniqueness declines with family size.

3 Data and Methodology

The data I use in this paper come from various sources. This section describes the datasets, major cleaning steps, and the methodology to build two novel measures for fund unique-

ness in prospectuses and portfolios based on deep learning algorithms. First, I introduce the main sample of US mutual fund information, prospectuses, and holdings from CRSP Survivor-Bias-Free US Mutual Fund Database, Morningstar Direct, and SEC's Electronic Data Gathering, Analysis, and Retrieval system (EDGAR). Second, I detail the steps to build a deep learning model based on an improved methodology from Gabaix et al. (2025) to generate high-dimensional representations of portfolio strategies. Third, I describe the cross-modal method to calculate the uniqueness measures for prospectus language and portfolio holdings. Last, I provide comprehensive summary statistics for major variables.

3.1 Data Sources

The primary dataset in this paper is holdings data from the CRSP Mutual Fund Holdings database. I restrict the sample by limiting the focus of my analysis to US equity funds, excluding international funds, sector funds, and underlying variable annuities following Abis, Buffa, Javadekar and Lines (2022). I also exclude funds with less than $5 million in total net assets (TNA) (Kacperczyk, Sialm and Zheng, 2008). The sample covers 178 months from January 2011 to October 2025. I map individual securities to firm identifiers using the CRSP link table, summing market values across share classes of the same firm, ensuring that a fund holding both Class A and Class B shares of a company is not counted as two separate positions. Since not all holdings are reported monthly, I forward-fill reported positions to monthly frequency (up to 6 months). For forward-fill months, I hold share counts constant and update market value accordingly. For stability of the portfolio embedding model training, I also apply the following filters: (i) minimum 20 institutional holders per security-month, which ensures the embedding model has sufficient co-occurrence signal for each stock; (ii) minimum 20 holdings per portfolio-month, which ensures enough context for the transformer to form meaningful representations; (iii) maximum 75% single-holding concentration, which removes single-stock vehicles and trivial portfolios. After filtering, my sample covers around 6,000 different portfolios, 7,500 assets, and 880,000 portfolio-month

observations.

I then gather prospectus filings from SEC's Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) for my sample period January 2011 to October 2025. Since each open-end fund is required to file Form N-1A, which includes a "Principal Investment Strategies" section describing the fund's investment approach, asset selection criteria, and risk philosophy, I extract the strategy narrative section for each fund-quarter similar to the work by Abis (2020) and Abis et al. (2022). To improve the quality of the prospectus dataset, I remove texts that contain mostly symbols or scrambled text. I also remove texts that are fewer than 50 words. Where multiple filings exist for the same fund-quarter (amendments, supplements), I retain the most recent. I link SEC filings to CRSP portfolio numbers via a two-step process: (i) map filing CIK codes to fund identifiers using the SEC Company Tickers file, (ii) match to CRSP portfolio numbers via the CRSP fund-share class mapping. This produces approximately 14,000 to 19,000 unique fund-quarter text observations (coverage varies by filing availability and linkage success). Since prospectus filings are updated quarterly, I assign each quarter's text to its three constituent months for alignment with the monthly holdings data.

I complement the dataset using fund summary variables from CRSP Mutual Fund Summary, including monthly net returns, total net assets, turnover ratio, expense ratio, and inception date; and Morningstar Direct, including monthly Morningstar category assignment, star rating, fees, manager start date, investment style consistency, Sharpe ratio, and fund flows.

In addition, I gather fund management company-level data using the Revelio Labs dataset. It provides workforce intelligence by aggregating and standardizing hundreds of millions of public employment records such as LinkedIn profiles, resumes, and job postings to create a universal HR database. This dataset allows me to track the employees' career trajectories (such as in the works by Gao, Ma and Xu, 2023; He, Mostrom and Sufi, 2024) and build an aggregate company-level dataset. I match Revelio Labs data with my major dataset

using fuzzy mapping of company names, cross-validated by company's NAICS industry code and geographical locations. Section 4.4.1 provides further details of instrument variable construction from the dataset.

3.2 Embeddings

My major methodology in building measures for portfolio and prospectus uniqueness relies on embeddings. Embeddings constitute dense vector representations of data entities projected into a continuous, high-dimensional latent space. In this paper, these entities include lexical tokens in prospectuses and financial instruments in portfolios. Such representations enable me to better evaluate the similarities between fund portfolios and prospectuses and therefore construct comparable measures for uniqueness in portfolio holdings and prospectus texts. This subsection introduces how I obtain the embeddings from portfolio holdings data and prospectus textual data and discusses their advantages over traditional methods.

3.2.1 Portfolio Embeddings

Traditional approaches represent a portfolio as a weight vector $\mathbf{w}$ over a fixed asset universe (Markowitz, 1952; Merton, 1972; Black and Litterman, 1992). Yet this representational paradigm suffers from well-documented pathologies: Ledoit and Wolf (2004) and DeMiguel, Garlappi and Uppal (2009) demonstrate that equal-weight allocations frequently dominate optimized weight vectors out of sample, precisely because the weight-vector framework indexes assets as distinct coordinates and captures cross-asset relationships only indirectly through estimated covariance, rather than through a continuous representation of asset similarity. Subsequent remedies, whether shrinkage estimators, hierarchical clustering (López de Prado, 2016), or risk parity formulations (Qian, 2006), ameliorate the estimation problem but do not resolve the deeper representational deficiency: the weight vector encodes *how much* capital is allocated but not *what* the assets fundamentally are. Embedding methods address this deficit by recovering latent factor structures that observable characteristics fail

to capture (Dolphin et al., 2022; Gabaix et al., 2025).

I represent each fund's portfolio as a dense vector in $\mathbb{R}^K$ using a modified BERT-style model (Devlin, Chang, Lee and Toutanova, 2019) trained on the universe of fund holdings. The approach follows Gabaix et al. (2025), who introduce Portfolio-Shares BERT (PS-BERT): by analogy with natural language processing, individual securities are treated as the words and portfolio holdings play the role of sentences. A model trained to predict masked securities from their portfolio context learns which stocks serve similar economic roles across funds.[2] The resulting embeddings encode implied investment strategy reflecting the latent commonality among portfolios that hold similar securities in similar proportions. Two funds holding different mega-cap technology stocks, e.g., one overweighting Apple, the other overweighting Microsoft, receive similar embeddings because these stocks co-occur in thousands of other portfolios as near-substitutes. The trained model captures this relationship clearly. Apple and Microsoft have a cosine similarity of 0.664 in the learned asset space, far above the mean pairwise similarity of 0.062, and Microsoft is Apple's nearest neighbor in the 7,554-asset universe (Table E.1). Conversely, a fund holding the same stocks as peers but with dramatically different weights receives a distinct embedding. I further augment the original PS-BERT framework with an active-weight-aware input layer that preserves cardinal portfolio weight information. Relative to the baseline implementation, my modification places greater emphasis on cardinal active weights, allowing the model to distinguish between a 0.5% drift and a 3% concentrated bet in the same stock, a distinction central to measuring active management intensity.

For the model input, each portfolio-month is encoded as an ordered sequence of up to $M = 62$ holdings, sorted by portfolio weight (descending). For portfolios exceeding 62 holdings, I retain the top 62 by weight. Each holding j in the sequence enters the model

[2] Appendix B describes the self-attention mechanism and full architecture hyperparameters.

through three additive input channels:

$$\mathbf{x}_j = \text{Emb}_{\text{token}}(s_j) + \text{Emb}_{\text{pos}}(j) + f_w(\tilde{w}_j) \tag{9}$$

where $\text{Emb}_{\text{token}}(s_j) \in \mathbb{R}^K$ is a learned vector for the identity of security s_j, drawn from a universe of $V \approx 7{,}500$ unique assets; $\text{Emb}_{\text{pos}}(j) \in \mathbb{R}^K$ is a learned vector for rank position j; $f_w(\tilde{w}_j) \in \mathbb{R}^K$ is a learned weight projection that maps the scalar active weight into embedding space.

Specifically, for each security-month, I compute the portfolio active weight as the deviation from the market-implied weight:

$$\tilde{w}_{ij} = \frac{w_{ij} - w_j^{\text{mkt}}}{\sigma_t^w} \tag{10}$$

where w_{ij} is fund i's portfolio weight in security j, $w_j^{\text{mkt}} = \sum_i \text{mv}_{ij} / \sum_i \sum_k \text{mv}_{ik}$ is the aggregate market weight computed across all institutional holders, and σ_t^w is the cross-sectional standard deviation of active weights in month t. Active weights are winsorized at the 2.5th/97.5th percentiles before standardization.

While I retain the masked language model pre-training objective, vocabulary construction from asset identifiers, mean-pooling extraction, and the overall transformer encoder architecture, my implementation improves the original PS-BERT framework of Gabaix et al. (2025) in two ways:

Cardinal weight injection. The original PS-BERT architecture discards portfolio weight magnitudes in model training, encoding holdings only through their rank ordering. This suffices for portfolio construction and return prediction, where the identity and relative ordering of holdings carry most of the signal. For measuring management *intensity*, however, cardinal weights are essential: a fund overweighting Apple by 0.5% versus 3% reflects a categorically different level of conviction, even if Apple occupies the same rank position. I

therefore add an explicit weight channel $f_w(\tilde{w}_j)$ to the input representation, additive with the token and position embeddings.

Nonlinear weight projection via MLP. I use a two-layer multilayer perceptron (MLP) to map the scalar active weight into embedding space. Appendix C details the mathematical settings. The economic motivation for a nonlinear projection is that different magnitudes of active weight carry qualitatively different information. A small overweight ($\tilde{w} \approx 0.1$) may reflect passive drift or minor rebalancing; a moderate overweight ($\tilde{w} \approx 1$) signals a deliberate active bet; a large concentrated position ($\tilde{w} > 3$) indicates high-conviction investment. A linear projection forces all regimes onto a single axis, preventing the model from distinguishing drift from conviction. The MLP allows each regime to activate different directions in embedding space.

Figure 2 confirms that the trained model learns to differentiate weight regimes. I plot the representation shift, measured as the cosine distance between the model's internal representation at active weight $\tilde{w}$ and its representation at market weight ($\tilde{w} = 0$), computed after layer normalization and averaged across securities. The representation shift increases monotonically with active weight, reaching cosine distance ≈ 0.3–0.5 for overweight positions ($\tilde{w} \in [1, 4]$). Without the weight projection, the Gabaix et al. (2025) model produces zero representation shift, as all positions with the same security and rank are treated identically regardless of the manager's conviction.

I train the model in two stages: masked language model (MLM) style pre-training to learn asset-level representations, followed by contrastive fine-tuning to produce portfolio-level embeddings optimized for cosine similarity comparison. In the first stage, I randomly mask 15% of holdings in each portfolio sequence and train the model to predict the masked securities from the remaining context (Devlin et al., 2019). In the second stage, I construct a positive pair by splitting holdings into two half-portfolios by rank parity (Reimers and Gurevych, 2019; Gabaix et al., 2025). Since each half-portfolio encodes the same underlying

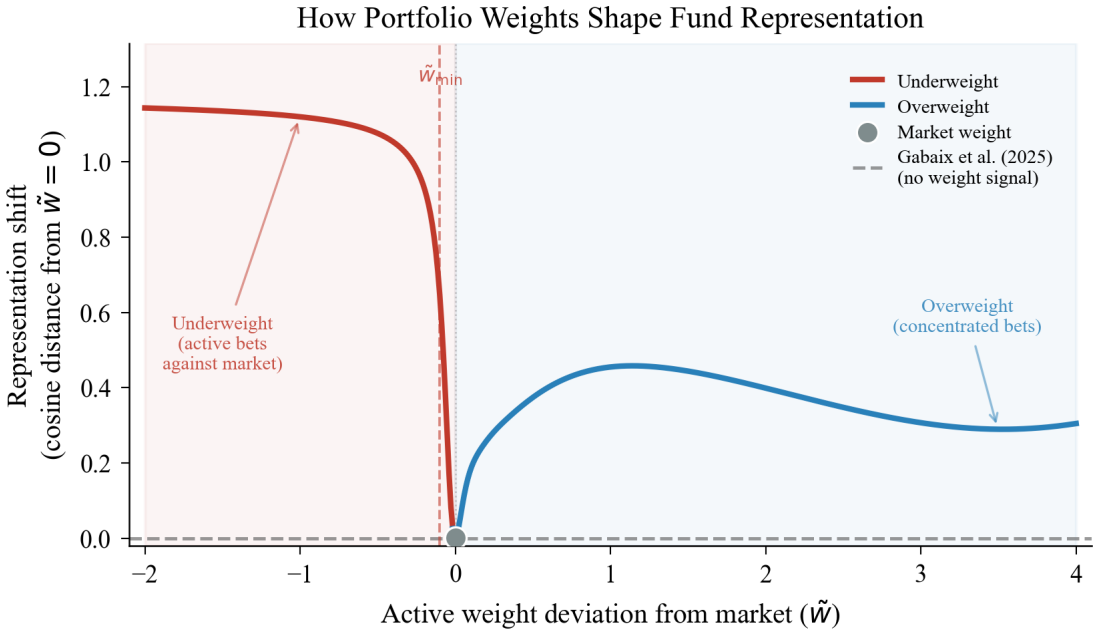


Figure 2
Learned Weight Projection

This figure shows the post-LayerNorm representation shift (cosine distance from the market-weight representation at $\tilde{w} = 0$) as a function of the standardized active weight, computed from the trained model and averaged over 500 random (token, position) samples. The dashed vertical line at $\tilde{w}_{\min} = -0.1$ marks the empirical lower bound of observed active weights; the region to its left is extrapolation. The Gabaix et al. (2025) baseline at zero indicates no weight-dependent signal.

strategy, the model is fine-tuned to match these pairs. I use AdamW with cosine learning rate scheduling and early stopping on validation loss (pre-training) and retrieval accuracy (fine-tuning). Appendix D details the specific technical procedures.

After training, I extract portfolio embeddings for the full sample. For each portfolio-month, the final embedding is the L2-normalized mean of the contextualized asset representations at asset positions:[3]

$$\mathbf{e}_i^h = \frac{\sum_{j=1}^{n_i} \mathbf{h}_j^{(L)}}{\left\| \sum_{j=1}^{n_i} \mathbf{h}_j^{(L)} \right\|_2} \in \mathbb{R}^K \tag{11}$$

where n_i is the number of holdings (up to 62). The final output is approximately 880,000 256-dimensional portfolio-month embedding vectors.

I validate this methodology in various ways. First, using the managed portfolio (MP) benchmark proposed in Gabaix et al. (2025), for each held-out portfolio-month, I mask the second-largest holding and measure the model's ability to predict the correct asset from the full universe V. The MP score, defined as $\text{MP} = 1 + \overline{\ln P_{\text{correct}}} / \ln V$, is 0.713 across 85,190 predictions spanning 178 months. This exceeds the best-performing model in their study, which achieves an MP score of 0.61, indicating a better performance than the models in Gabaix et al. (2025).

Second, I test portfolio retrieval. The finetuned model achieves 92–95% top-1 retrieval accuracy on held-out investors. This means given half of a fund's holdings, the model correctly identifies the fund from among hundreds of candidates. This far exceeds the random baseline of 0.3 to 0.6% and confirms that the 256-dimensional embeddings faithfully encode portfolio identity.

Last, I examine whether the model recovers economic structure without supervision. Figure 3 visualizes the learned asset embedding space via UMAP 2-D projection. Panel A highlights five major Fama–French industries assigned from CRSP SIC codes: Finance concentrates along the rim of the map, Healthcare and Business Equipment form their own

[3]Special tokens are excluded from the mean-pooling computation.

regions, and Energy and Telecom collect in distinct pockets, despite the model receiving no industry labels during training. Panel B colors assets by distinctiveness: peripheral assets held by specialized strategies are more distinctive than ubiquitous mega-caps near the center.

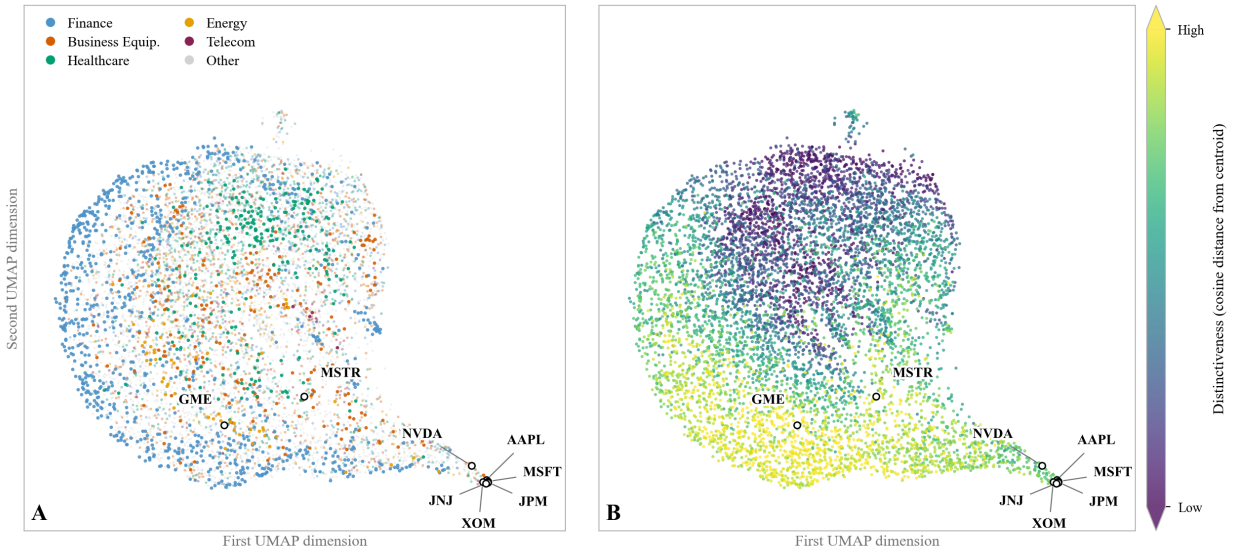


Figure 3
Learned Asset Embedding Space

This figure displays a UMAP projection of the 7,554 asset embeddings ($K = 256$). Panel A highlights assets in five major Fama-French industries (Finance, Business Equipment, Healthcare, Energy, and Telecom), which I assign from the modal CRSP SIC code of the issuing firm; assets in all other industries appear as small gray dots. The strongest color marks regions where an industry concentrates. The model observes only co-holding patterns, and the industry clustering that emerges confirms that the embeddings recover economic structure. Panel B colors the same projection by embedding distinctiveness (cosine distance from the centroid of all assets), with the color scale clipped at the 5th and 95th percentiles and labeled at its low and high endpoints. Selected tickers are labeled for orientation.

Finally, I assess the embedding pipeline for look-ahead bias using a split-sample retraining exercise. I train the holdings encoder exclusively on 2011–2017 data, freeze both the model and its vocabulary, and reconstruct the two uniqueness measures for 2018–2025, thereby preventing any post-2017 information from entering the representation pipeline. The resulting measures closely match the production measures across 473,635 common fund-months, with pooled Spearman rank correlations of 0.850 for *PortfolioUniq* and 0.845 for *ProspectusUniq*. The conditional-alpha finding in Section 4.2 is quantitatively similar over

2018–2021, before substantial vocabulary deterioration. The interaction of *PortfolioUniq* with the top past-alpha tercile is 0.267 ($t = 2.03$) under the frozen model and 0.288 ($t = 1.97$) under the production model. The estimated fee association with *ProspectusUniq* is also stronger under the frozen model, at 0.061 ($t = 2.09$) compared with 0.033 under the production model over the full 2018–2025 window. Differences emerge primarily in 2022–2025, when the out-of-vocabulary portfolio weight rises from 8.5 to 12.3 percent and agreement between the two sets of measures declines accordingly, a pattern consistent with an increasingly stale vocabulary.

3.2.2 Interpreting the Embedding Space

The visual clustering in Figure 3 admits a quantitative upgrade. Assigning each asset its modal SIC code from the CRSP monthly stock file, a leave-one-out nearest-neighbor classifier on the embeddings recovers the Fama–French 48-industry label at 32.4% accuracy against a 16.5% majority baseline, and within-industry pairwise cosine exceeds cross-industry cosine at every granularity I examine (Table E.1, Panel B). Named pairs behave as the co-holding logic predicts: Apple and Microsoft have cosine similarity 0.664 against an all-pairs mean of 0.062, six standard deviations above baseline, and Exxon–Chevron, Coca-Cola–PepsiCo, and JPMorgan–Bank of America are all four to five standard deviations above it (Panel A). The structure is local: nearest neighbors recover industry, while global k-means clusters align with co-holding cohorts instead (adjusted Rand index with industry 0.02), and widely co-held mega-caps across industries score as high as same-industry pairs. The space therefore encodes which stocks play similar roles in portfolios, a relation that overlaps with industry but is measured from investor behavior.

Distance in the space predicts comovement. Among all 1.97 million pairs of the 2,000 most widely held stocks, embedding cosine forecasts the realized 2011–2025 return correlation with a Spearman rank correlation of 0.161, monotonically across all twenty cosine bins, and contributes incremental explanatory power beyond same-industry indicators (incremental R^2

of 0.020 after Fama–French 10 and two-digit SIC dummies; Table E.1, Panel C). The relation is concentrated where co-holding data are rich: it weakens in the 2,001–4,000 holding-breadth band and is absent beyond it. The same holds one level up: across 2.15 million fund pairs from six annual snapshots, portfolio-embedding cosine predicts the funds' return correlation over the following 24 months (pooled Spearman 0.230) and retains incremental information after conditioning jointly on Morningstar category, raw holdings overlap, and prospectus-text similarity (partial Spearman 0.09–0.11; Table E.3). Raw holdings cosine remains the strongest single predictor of future fund comovement; the embedding contributes a compact strategy summary that adds information beyond it.

Linear probes quantify what the 256 dimensions contain. With whole months held out, the embedding recovers portfolio-construction characteristics (top-10 share $R^2 = 0.66$, industry concentration 0.63, active share 0.58–0.60, sector tilts up to 0.54) and, more tellingly, return-based quantities the model never observed in training: idiosyncratic volatility ($R^2 = 0.42$), SMB loading (0.42), and $1 - R^2$ (0.39), all computed from returns while the model sees only holdings (Table E.2). A multinomial probe assigns the correct Morningstar category among 57 at 66.8% accuracy against a 13.8% baseline and detects index funds with an AUC of 0.92. At the same time, thirty observables and category dummies jointly span 8.4% of embedding variance out of sample: the embedding contains the standard characteristics and a large residual component, and the uniqueness measures built from it draw on both.

This interpretation sharpens what *PortfolioUniq* measures. Distinctiveness of the individual assets a fund holds is a separate property from the distinctiveness of the portfolio that combines them: holdings-weighted asset distinctiveness correlates with *PortfolioUniq* at -0.03 pooled ($+0.04$ within category-month), so portfolio uniqueness reflects the combination, the weights and the company each position keeps, in line with the context-dependence of the representation. Appendix E reports the full interpretation battery, including applications of the embedding space to closet-index detection, style-drift measurement, and category crowding, and a forward-pass demonstration of context dependence.

3.2.3 Prospectus Embeddings

I embed each fund's prospectus strategy narrative into a 768-dimensional vector [4] using a pre-trained embedding model: Gemini embedding model (gemini-embedding-001; Lee, Chen, Dua, Cer et al., 2025). Unlike the portfolio embeddings, which are trained from scratch on holdings data, the text embeddings rely on a foundation model pre-trained on a broad text corpus. This approach allows the representation to capture semantic proximity beyond the surface lexical overlap of traditional bag-of-words and topic model approaches. For instance, the two strategy narratives: "growth equity in technology sectors" and "capital appreciation through innovative technology companies" are closer in textual embedding space, while their lexical overlap stays low. As argued in many previous works such as Araci (2019), Li et al. (2021), and Jehnen, Villalba-Diez and Ordieres-Mere (2026), this high-dimensional output can provide a rich representation of strategy narrative and specificity. The model takes as input all the "Principal Investment Strategies" sections extracted from each fund's N-1A filing and records their corresponding vector representations. These vectors are then L2-normalized, aligning with the portfolio embeddings' normalization.

Since most of the strategy sections in prospectuses share substantial boilerplate content, such as risk disclaimers, regulatory disclosures, fee descriptions, or other standard investment process language, the raw embeddings cluster in a narrow positive-similarity cone. Figure 4, Panel A shows the narrow distribution of pairwise cosine similarities among raw embeddings. This narrow distribution makes it hard to distinguish fund narratives. To address this concern, I subtract the monthly cross-sectional mean text embedding, which proxies for the typical prospectus, and re-normalize:

$$\tilde{\mathbf{e}}_i^t = \frac{\mathbf{e}_i^t - \bar{\mathbf{e}}_t^t}{\|\mathbf{e}_i^t - \bar{\mathbf{e}}_t^t\|_2}, \qquad \bar{\mathbf{e}}_t^t = \frac{1}{N_t}\sum_{j=1}^{N_t} \mathbf{e}_j^t \tag{12}$$

Figure 4, Panel B shows the result: pairwise cosine similarities now span the full

[4] `gemini-embedding-001` originally output a 3072-dimensional embedding. It is truncated to 768 dimensions for computational efficiency using the Matryoshka Representation Learning (MRL).

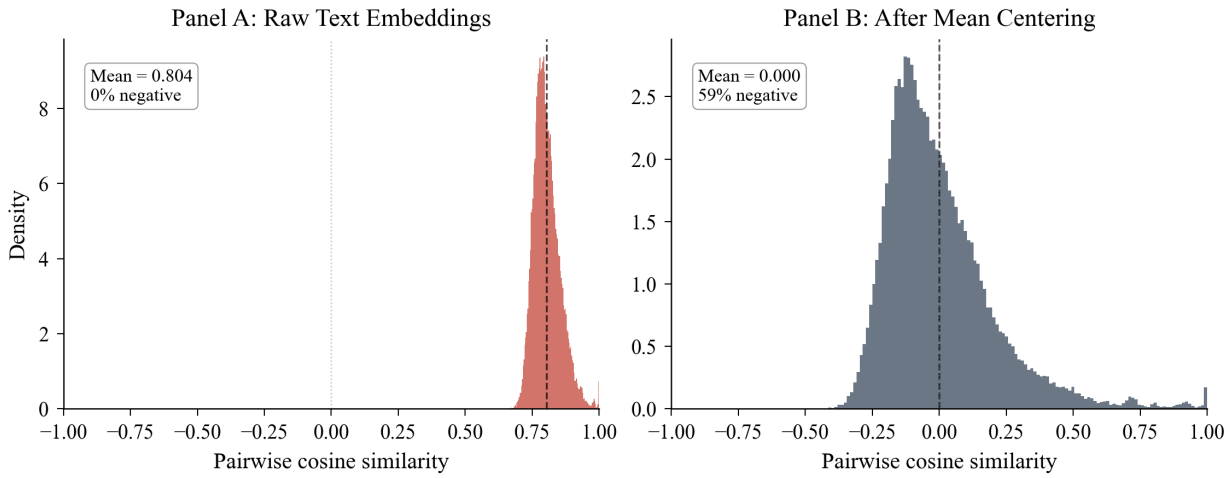


Figure 4
The Cone Problem and Mean Centering Fix

This figure shows the distribution of pairwise cosine similarities among prospectus text embeddings. Panel A: raw embeddings cluster in a narrow positive-similarity cone (mean = 0.804, 0% negative). Panel B: after subtracting the monthly mean embedding and re-normalizing, similarities have mean ≈ 0 with 59% negative values, providing meaningful variation for measuring narrative uniqueness.

$[-0.4, +1.0]$ range. The centered embeddings isolate fund-specific deviations from the typical prospectus, which is exactly the variation relevant for measuring narrative uniqueness.

3.3 Cross-Modal Uniqueness Measures

After obtaining portfolio holdings and prospectus text embeddings, a natural starting point is to measure how unique a fund is within each modality separately, such as in Kostovetsky and Warner (2020). Nevertheless, this single-modal measure misses a key dimension of fund differentiation: the alignment between what a fund says and what it actually does. A fund may be a pure marketing differentiator, holding a standard portfolio while describing an unusual strategy; or it may be a silent innovator, pursuing a distinctive portfolio while describing it in generic terms. These misalignments are economically meaningful (for example, as shown in Brown, Cederburg and Towner, 2024) and invisible to single-modal measures.

I therefore construct cross-modal uniqueness measures that use similarity in one modality to define a fund's peer group and measure distance in the other modality. This

design ensures that each measure captures information that the other modality alone cannot provide. The detailed calculations are as follows.

Prospectus Uniqueness Weighted by Holdings Similarity. For each fund i in month t, I construct a holdings-similarity-weighted centroid in text space and measure the fund's distance from it. The procedure has four steps.

First, compute the pairwise holdings cosine similarity matrix across all N_t funds with both embeddings available in month t:

$$S^h_{ij,t} = \cos(\mathbf{e}^h_i, \mathbf{e}^h_j) \tag{13}$$

Second, apply a rectified linear unit (ReLU) filter, which zeroes out negative similarities, and row-normalize to obtain peer weights:[5]

$$W^h_{ij,t} = \frac{\max(S^h_{ij,t},\, 0)}{\sum_{k \neq i} \max(S^h_{ik,t},\, 0)}, \qquad j \neq i \tag{14}$$

Third, form the holdings-weighted text centroid, which represents the typical narrative of funds with similar portfolios to fund i:

$$\mathbf{c}^{\text{text}}_{i,t} = \sum_{j \neq i} W^h_{ij,t} \cdot \tilde{\mathbf{e}}^t_j \tag{15}$$

Last, measure the distance to the centroid. This distance is then my measure of fund uniqueness in prospectus strategy narratives:

$$\mathit{ProspectusUniq}_{i,t} = 1 - \cos\big(\tilde{\mathbf{e}}^t_i,\ \mathbf{c}^{\text{text}}_{i,t}\big) \tag{16}$$

Figure 5 provides a simple illustration of how peers' narratives impact the uniqueness

[5] Approximately 25% of pairwise holdings similarities are negative after mean centering; these represent funds pursuing sufficiently different strategies that including them in the centroid would dilute the peer signal.

measure. *ProspectusUniq* captures the extent to which a fund's prospectus narrative is distinctive relative to the narratives of funds with similar portfolio compositions. Intuitively, it measures whether a fund describes itself differently from its portfolio peers. A higher *ProspectusUniq* reflects greater divergence in stated strategy despite similarity in underlying holdings, whereas a lower *ProspectusUniq* reflects narrative conformity among funds with comparable portfolios.

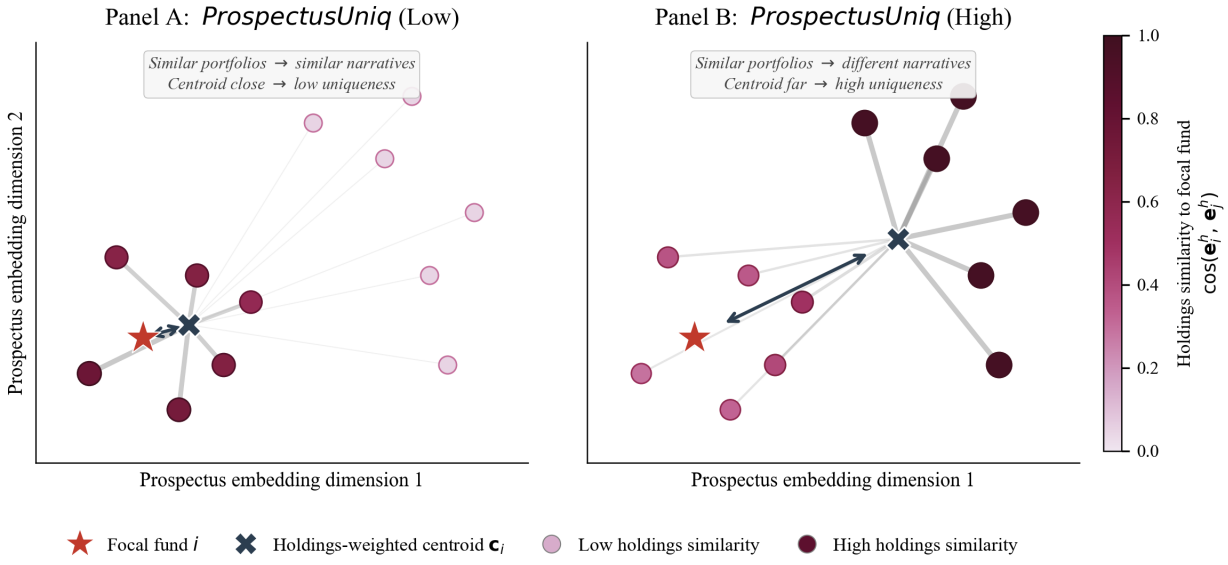


Figure 5
Cross-Modal Uniqueness: Prospectus Uniqueness Illustration

This figure illustrates how holdings-similarity-weighted peer narratives determine *ProspectusUniq*. Funds are positioned in text embedding space; dot color and size reflect holdings similarity to the focal fund (star). The holdings-weighted text centroid ($\times$) and the resulting distance measure *ProspectusUniq* are shown.

Portfolio Uniqueness Weighted by Text Similarity. Symmetrically, I use text similarity to weight peers and measure distance in holdings space. The uniqueness measure for a fund's portfolio is then given by:

$$PortfolioUniq_{i,t} = 1 - \cos\left(\mathbf{e}_i^h, \sum_{j \neq i} W_{ij,t}^t \cdot \mathbf{e}_j^h\right) \tag{17}$$

where

$$W_{ij,t}^{t} = \frac{\max\big(\cos(\tilde{\mathbf{e}}_i^t, \tilde{\mathbf{e}}_j^t),\, 0\big)}{\sum_{k \neq i} \max\big(\cos(\tilde{\mathbf{e}}_i^t, \tilde{\mathbf{e}}_k^t),\, 0\big)}, \qquad j \neq i \tag{18}$$

PortfolioUniq captures the extent to which a fund's portfolio is distinctive relative to the holdings of funds with similar prospectus narratives. Intuitively, it measures whether a fund diverges from what its narrative peers hold.

The Spearman correlation $\rho(\textit{ProspectusUniq}, \textit{PortfolioUniq}) = 0.23$ indicates that the two measures are weakly positively associated, confirming that they capture genuinely different dimensions of fund distinctiveness. Figure 6 displays the joint distribution: funds populate every combination of the two margins, including the marketing differentiators and silent innovators introduced above. This low correlation is the empirical foundation for treating portfolio uniqueness and prospectus uniqueness as separate economic channels throughout the paper. In Appendix F, I re-estimate the main results using single-modal local uniqueness measures and find that the portfolio-uniqueness findings are robust and that the prospectus fee premium reflects the strategy-relative comparison that the cross-modal measures encode.

3.4 Relation to Existing Measures

The literature measures fund differentiation with several established proxies, and a natural question is how much of *PortfolioUniq* and *ProspectusUniq* these proxies already span. I compute the full set on my sample: active share against the market and against the category (Cremers and Petajisto, 2009), portfolio concentration through the Herfindahl index of portfolio weights and the top-ten share, the industry concentration index of Kacperczyk et al. (2005) computed over the Fama–French ten industries against the aggregate institutional industry weights, portfolio breadth through the number of holdings, the holdings-based and text-based uniqueness measures of Kostovetsky and Warner (2020) replicated from raw holdings and from the Principal Investment Strategies text, the returns-based selectivity measure $1 - R^2$ of Amihud and Goyenko (2013) together with tracking error, idiosyncratic volatility,

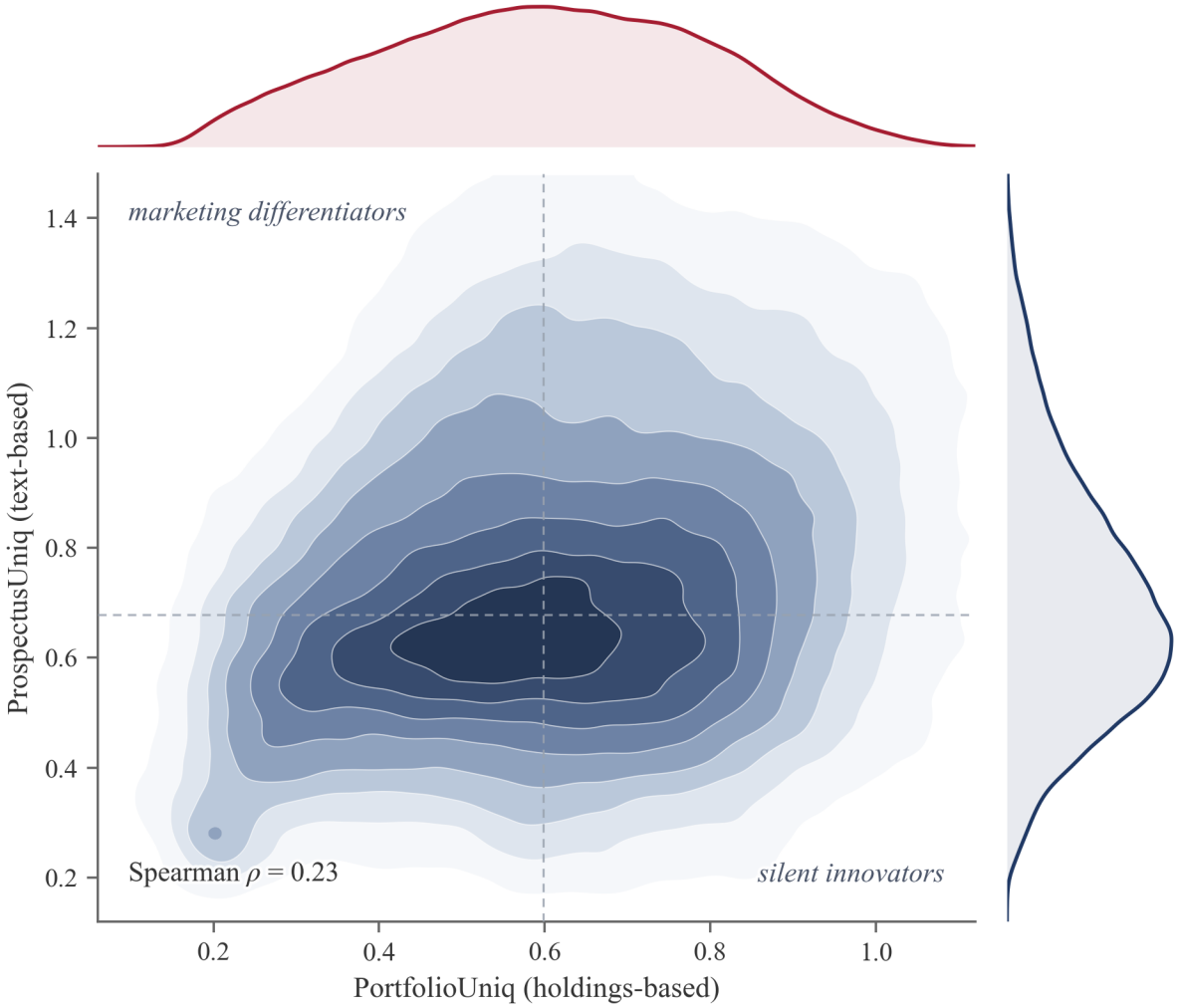


Figure 6
The Joint Distribution of Portfolio and Prospectus Uniqueness

This figure shows the joint distribution of *PortfolioUniq* (horizontal axis) and *ProspectusUniq* (vertical axis) over the full measure universe of $N = 843{,}374$ portfolio-months. The central panel draws the smoothed sample density as filled highest-density contours on a single light-to-dark ramp; from lightest to darkest, the bands enclose 99.8, 97, 90, 78, 62, 45, 28, and 12 percent of the sample mass. Dashed guides mark the two sample medians and split the panel into quadrants. The two off-diagonal quadrants correspond to the fund types described in Section 3.3: funds in the upper-left region (*marketing differentiators*) hold standard portfolios while describing unusual strategies, and funds in the lower-right region (*silent innovators*) pursue distinctive portfolios while describing them in generic terms. The marginal density of *PortfolioUniq* appears above the panel in crimson and the marginal density of *ProspectusUniq* to its right in navy. The two measures share broad joint support with a modest positive association (Spearman $\rho = 0.23$): funds populate every combination of portfolio and prospectus uniqueness, so the two margins measure genuinely different objects.

and peer return uniqueness in the spirit of Sun et al. (2012), prospectus length, readability, and boilerplate share, and the Morningstar style-consistency metric. Table G.1 reports the correlations. The strongest traditional correlates of *PortfolioUniq* are the stock-level concentration measures, with Spearman correlations of 0.34; the industry concentration index follows at 0.24, the returns-based measures at 0.17 to 0.19, and active share and the Kostovetsky and Warner (2020) holdings measure sit at 0.09 and below. For *ProspectusUniq*, boilerplate share is the strongest traditional correlate at −0.21, keyword-based text uniqueness reaches 0.17, and length and readability stay at or below 0.10. The only large correlations, 0.73 and 0.53, belong to the single-modal nearest-neighbor measures of Appendix F, which I build from the same embeddings and which therefore serve as an internal upper benchmark.

Panel C of Table G.1 quantifies the joint span. Regressing each measure on its complete competitor set delivers an R^2 of about 0.60 for *PortfolioUniq*, and 26 to 29 percentage points of that total come from the same-embedding nearest-neighbor sibling alone. The traditional proxies collectively explain 33% of *PortfolioUniq* pooled and 31% within category×month, and 4% and 3% of *ProspectusUniq*; within category and month, the largest single traditional contributor is market active share, with an incremental R^2 of 1.9 percentage points, and the industry concentration index adds 0.7 percentage points pooled and 0.1 within. Roughly two thirds of the variation in *PortfolioUniq*, and over 95% of the variation in *ProspectusUniq*, is information beyond the entire existing toolkit.

A specific concern is that *PortfolioUniq* could proxy portfolio breadth, since funds holding thousands of names sit close to any centroid. The rank correlation with the number of holdings is −0.23, and the relation is weakly monotone: the log holdings count explains 0.4% of the variance of the winsorized measure, with a Pearson correlation of −0.06. Residualizing *PortfolioUniq* on log breadth, the adjustment that Kostovetsky and Warner (2020) apply to their holdings measure, leaves the measure essentially unchanged, and its correlation with their holdings uniqueness stays at 0.073 to the third decimal. *PortfolioUniq* captures

the composition of what a fund holds relative to its narrative peers, and its information content sits almost entirely outside breadth, concentration, and active share. Appendix G carries this conclusion to the paper's headline results: the fee and flow findings strengthen when the complete proxy set enters as controls, the alpha finding is stable against every proxy individually, and both instrumental-variable estimates move by less than 2% under the relevant proxy controls.

3.5 Summary Statistics

After data cleaning and merging different datasets following the steps described in the previous subsections, my estimation sample consists of 666,375 fund-month observations spanning 6,024 unique portfolios and 180 months from January 2011 to December 2025. The cross-modal weighted uniqueness measures *PortfolioUniq* and *ProspectusUniq* are available for 458,190 fund-months calculated from the subset with both holdings and prospectus embeddings. Table 1 reports summary statistics for key variables. All variables are winsorized at the 1st and 99th percentiles.

4 Main Findings

In this section, I examine how the two different kinds of uniqueness affect key mutual fund characteristics. I first study the impact of uniqueness on future fund fees and fund performance, then I study the impact on future fund flows. This section concludes with two instrumental variables to address the endogeneity concerns.

4.1 Uniqueness and Fund Fees

I first ask whether funds with more distinctive prospectus language or more distinctive strategy derived from portfolios charge higher total fees. I estimate the following specification:

$$Fee_{i,t} = \beta_1 \, PortfolioUniq_{i,t-k} + \beta_2 \, ProspectusUniq_{i,t-k} + \gamma' \mathbf{X}_{i,t-k} + \boldsymbol{\delta} + \varepsilon_{i,t}, \tag{19}$$

where $Fee_{i,t}$ is the annualized fee for fund i in month t, winsorized at the 1st and 99th percentiles. *PortfolioUniq* and *ProspectusUniq*, representing the uniqueness in portfolio strategies and uniqueness in prospectus texts, are estimated using the methodology introduced in the last section, and they are lagged by $k \in \{1, 3\}$ months. The control vector $\mathbf{X}$ includes log total net assets, portfolio turnover ratio, fund return, and fund age. $\boldsymbol{\delta}$ denotes fixed effects. Following Kostovetsky and Warner (2020) and Abis and Lines (2024), I progressively saturate: column (1) includes family, category, and month fixed effects; column (2) replaces these with family×month and category fixed effects; and column (3) uses family×month and category×month fixed effects. Columns (4)–(6) repeat the same structure with a three-month lag. I cluster all standard errors at the fund and month levels.

Table 2 reports the results. The coefficient on *ProspectusUniq* is positive and significant at the 1% level across all six specifications, ranging from 0.083 to 0.087. A one-unit increase in prospectus uniqueness raises annualized fees by approximately 8 basis points. This finding is consistent with the work of Vozlyublennaia and Wu (2017). However, their measure is based purely on clustering fund returns. In contrast, the coefficient on *PortfolioUniq* is economically small and statistically insignificant in every specification. The pattern is unchanged between one-month and three-month lags. Appendix G shows that the premium is a property of the narrative content itself: on a constant sample it survives controls for prospectus length, readability, boilerplate share, and the keyword-based text uniqueness of Kostovetsky and Warner (2020) one at a time, more than doubles when all five text controls enter jointly, and holds within filing agents and within family×month cells.

The results show that the text differentiation in prospectuses can command higher fees while strategy differentiation in portfolios cannot. These higher fees from unique prospectus descriptions provide funds with a direct incentive to emphasize distinctiveness in their prospectus language.

4.2 Uniqueness and Future Alpha

I next examine whether portfolio or prospectus uniqueness predicts future risk-adjusted performance. Following previous work, including Hoberg et al. (2018) and Abis and Lines (2024), I construct forward-looking alphas that cleanly separate the estimation of factor loadings from the evaluation period. For each fund-month t, I estimate factor loadings $\hat{\beta}$ by regressing fund excess returns on the Fama–French–Carhart six-factor model (market, SMB, HML, RMW, CMA, and momentum) over the 24-month estimation window $[t-24, t-1]$. Forward alpha at horizon h is then:

$$\alpha^{\text{fwd}}_{i,t+h} = r_{i,t+h} - \hat{\beta}'_{i,t}\,\mathbf{f}_{t+h}, \tag{20}$$

where $r_{i,t+h}$ is the fund's realized excess return and $\mathbf{f}_{t+h}$ is the vector of factor returns, both at the future month $t+h$. Since the factor loadings are estimated entirely from pre-t data, forward alpha provides an out-of-sample measure of abnormal performance.

I allow the effect of uniqueness to vary with managerial skill (Berk and Green, 2004) by interacting uniqueness with past-performance terciles:

$$\alpha^{\text{fwd}}_{i,t+h} = \sum_{r\in\mathcal{R}} \mathbf{1}_{[r]}\big(\beta_{1r}\,PortfolioUniq_{i,t} + \beta_{2r}\,ProspectusUniq_{i,t} + \alpha_r\big) + \gamma'\mathbf{X}_{i,t} + \boldsymbol{\delta} + \varepsilon_{i,t+h}, \tag{21}$$

where $\mathcal{R} = \{\text{bottom } 30\%, \text{middle}, \text{top } 30\%\}$ denotes past-performance terciles assigned within category×month using backward 12-month rolling alpha. The middle tercile serves as the reference group, so $\beta_{1,\text{middle}}$ captures the base effect and the interaction coefficients

$(\beta_{1,\text{top}} - \beta_{1,\text{middle}})$ and $(\beta_{1,\text{bot}} - \beta_{1,\text{middle}})$ capture differential effects by past performance. $\boldsymbol{\delta}$ denotes fixed effects.

I estimate the specifications using standard Fama–MacBeth regression and two-way fixed effects (TWFE) model. The control vector $\mathbf{X}$ includes log total net assets, portfolio turnover ratio, number of managers, and fund age. I cluster all standard errors at the fund and month levels.

Table 3 reports the results for horizons of one and three months. The interaction with the top-tercile dummy is positive and significant across all four specifications ($t = 3.16$ in Fama–MacBeth, $t = 2.24$ in panel FE at one month). The implied total effect for top-performing funds is positive: 0.102 in Fama–MacBeth and 0.117 in panel FE, meaning that portfolio uniqueness predicts superior forward alpha only among managers who have demonstrated recent skill. The bottom-tercile interaction is small and insignificant. In contrast, the coefficient on *ProspectusUniq* and its interactions are uniformly insignificant. Distinctive prospectus language carries no predictive content for future performance regardless of past skill. This finding is consistent with earlier work on fund competitions such as Hoberg et al. (2018): uniqueness in strategies pays off only when paired with demonstrated managerial ability. In Appendix G I race this interaction against eleven portfolio-based differentiation proxies, including active share, concentration, breadth, the industry concentration index of Kacperczyk et al. (2005), and the holdings uniqueness of Kostovetsky and Warner (2020): the coefficient stays between 0.019 and 0.023 per standard deviation in every specification and is significant at the 5% level against ten of the eleven proxies individually, no competitor interaction replicates the skill-gated pattern, and the joint specification with all eleven proxies and their interactions keeps the magnitude while inflating the standard errors among the correlated terms.

In addition to the literature finding that unique funds outperform their herding peers (Wei, Wermers and Yao, 2015; Jiang and Verardo, 2018; Son, 2022), the results also reveal an opposite impact for average-skilled and low-skilled managers: the base coefficient on

PortfolioUniq is negative and marginally insignificant ($t = -1.06$ in Fama–MacBeth, $t = 0.05$ in panel FE for the one-month horizon), suggesting that for average-performing funds, more unique strategies do not outperform. On the contrary, a more unique strategy can even hurt future performance for average-skilled managers. This asymmetry in the impact of portfolio uniqueness demonstrates that strategy differentiation reflects distinct economic mechanisms for differently skilled managers. For skilled managers, strategy uniqueness means occupying a differentiated product niche, such as less crowded trades (Hoberg et al., 2018), less return correlation with peers (Amihud and Goyenko, 2013), or pricing power over a distinct investor clientele. The less competition for the same alpha opportunities helps managers earn additional rents and raises margins. For average managers, strategy uniqueness indicates that managers are competing outside their competence boundary. Their deviations from herding strategies are not driven by actual skills or private information as in skilled managers. Instead, these deviations reflect arbitrary bets, potentially driven by style drift, tournament behavior (Brown et al., 1996), overconfidence (Barber and Odean, 2001), excessive trading (Puetz and Ruenzi, 2011), or simply incompetence as argued in Pástor and Stambaugh (2012). By moving away from the collective intelligence reflecting shared fundamental analysis and informational externality of peer validation, these managers lose the benefits of consensus-correlated positions without gaining an actual competitive edge. Speaking from the perspective of the Hotelling analogy (Hotelling, 1929), differentiation reduces competition and raises margins only if the strategies are located near actual alpha opportunities. Differentiating away from genuine opportunities produces isolation without reward, not actual profitable product innovation.

Because the interaction conditions on past winners, a natural concern is that it repackages performance persistence: funds sorted into the top tercile on past alpha carry higher subsequent alpha for reasons of persistence alone, and any variable correlated with past alpha inside the tercile inherits that gradient. The concern has empirical bite in my sample. Within the top tercile, funds in the highest *PortfolioUniq* quintile have past 12-month alpha

1.32 percentage points above the lowest quintile ($t = 13.8$) and higher idiosyncratic volatility. I therefore control for persistence directly (Appendix H). Adding the continuous backward 12-month alpha and its interactions with both tercile dummies to equation (21), on identical observations, leaves the top-tercile interaction significant: $PortfolioUniq \times I(\text{Top})$ moves from 0.181 ($t = 3.16$) to 0.114 ($t = 2.10$) in Fama–MacBeth and from 0.108 ($t = 2.19$) to 0.094 ($t = 1.92$) in the panel, so the panel estimate holds its magnitude at the 10% level. The persistence control does real work: the past-alpha level predicts forward alpha (0.013 to 0.014, $t \approx 2.4$), yet its own interaction with the top tercile, the slope a persistence story requires, is 0.0001 ($t = 0.04$) in Fama–MacBeth and -0.006 ($t = -2.13$) in the panel. Replacing the tercile intercepts with past-performance decile dummies gives the same answer (0.135, $t = 2.67$ in Fama–MacBeth), and the pattern holds with all regressors lagged one month and with a gross-of-fee dependent variable. The interaction therefore carries information beyond the persistence gradient it conditions on. The threshold structure matters: a fully continuous specification that interacts $PortfolioUniq$ with past alpha instead of the tercile dummies finds no effect ($t \leq 1.4$ in all four estimator-horizon cells), so the premium is a property of being among the winners rather than a smooth gradient in past alpha.

The result also survives gates built from information outside the fund's own recent alpha. Replacing the past-alpha terciles with Morningstar star-rating buckets, an external skill signal whose top group overlaps the baseline winner tercile at a correlation of only 0.25, reproduces the interaction in both estimators (0.123, $t = 2.55$ in Fama–MacBeth; 0.119, $t = 1.96$ in the panel), with and without the continuous rating as a control (Table H.2). A gate built on the 36-month trailing Berk and van Binsbergen (2015) value added reproduces it in Fama–MacBeth (0.135, $t = 2.58$). Gates built on the t-statistic of trailing alpha and on manager tenure produce no interaction, so the effect concentrates among funds whose recent risk-adjusted returns are high, the winner sets that 12-month alpha and star ratings identify, and it does not extend to statistical-precision definitions of skill. Placebo gates sharpen the reading: permuting the tercile assignments at random within category×month

cells 200 times yields a placebo distribution in which the actual panel t-statistic sits at the 99th percentile (permutation $p = 0.030$), and a size gate produces no positive top interaction. The peer-exit instrumental-variable estimate of Section 4.4.2 is insensitive to the persistence control, moving from 1.253 to 1.259 with the first-stage F unchanged at 75.3 (Table H.5).

The interaction is a conditional cross-sectional slope, and I state its scope precisely. In calendar time, an equal-weighted spread portfolio long the most unique and short the least unique winners earns 6.3 basis points per month in net excess returns ($t = 1.18$), matching the 5.5 basis points the regression coefficients imply, and a raw double sort shows the same top-tercile spread of 5.3 basis points ($t = 1.88$); at the observed spread volatility, 179 months detect an effect of this size only at roughly twice its magnitude, and the spread loads on a generic persistence factor (Table H.4). The winners' uniqueness premium is therefore visible in the cross-section with controls, and its economic magnitude sits below what a calendar-time portfolio test can certify at this sample length.

This finding is directly connected to classical work on mutual fund managers' career concerns and herding behavior by Chevalier and Ellison (1997, 1999), where they argued that managers herd to protect their careers and those who deviate without a strong track record face disproportionate career risk if their bets fail. This creates an adverse selection mechanism where average-performing managers who hold unique portfolios are more likely to be those with weaker career discipline or worse risk management, and these managers are closer to those who cannot deliver higher future alpha.

This finding is also conceptually similar to the work of Kacperczyk et al. (2005), where they showed that industry-concentrated funds outperform among skilled managers; Kacperczyk and Seru (2007), where they argued that skilled managers' portfolio deviations from public information optimal allocations earn positive alpha; and Cremers and Petajisto (2009), where they empirically showed that performance persistence is especially strong among funds that have both high active share and strong prior returns. Additionally, Feldman, Saxena and Xu (2020) show that a more concentrated active fund management industry

can lead to improved productivity. My findings are consistent with their theoretical model where if a manager is merely average, the costs of searching for alpha outweigh the benefits. Therefore, herding is the most economically efficient response for average-skilled managers.

The results reveal another asymmetry between prospectus uniqueness and portfolio uniqueness. While the previous subsection establishes that only prospectus uniqueness can justify higher fees, this subsection shows that portfolio uniqueness is the actual measure for genuine product innovation and can attract extra alpha for skilled managers.

4.3 Uniqueness and Fund Flows

After establishing the two asymmetries in uniqueness measures with total fund fees and fund performance, I now ask whether and which measure of uniqueness attracts investor capital. I estimate a similar specification to that in the previous subsection:

$$Flow_{i,t} = \sum_{r \in \mathcal{R}} \mathbf{1}_{[r]} \big(\beta_{1r}\, PortfolioUniq_{i,t-1} + \beta_{2r}\, ProspectusUniq_{i,t-1} + \alpha_r \big) + \gamma' \mathbf{X}_{i,t-1} + \boldsymbol{\delta} + \varepsilon_{i,t}, \tag{22}$$

where $Flow_{i,t}$ is fund net flow as a percentage of fund net assets (winsorized at the 1st and 99th percentiles), and $\mathcal{R} = \{\text{bottom } 30\%, \text{middle}, \text{top } 30\%\}$ denotes past performance rank based on 36-month alpha terciles, computed within each month. The model fully interacts both uniqueness measures with performance rank, allowing the flow sensitivity to uniqueness to vary across the performance distribution. The control vector $\mathbf{X}$ includes log total net assets, portfolio turnover ratio, fund return, and fund age. Column (1) includes family, category, and month fixed effects; column (2) replaces family and month fixed effects with family×month interactions. I cluster all standard errors at the fund and month levels.

Table 4 reports the results. Two patterns stand out. First, *PortfolioUniq* has no significant effect on flows at any performance level: neither the base coefficient nor any interaction term is statistically different from zero. Second, *ProspectusUniq* attracts flows *only* among top performers: the interaction term *ProspectusUniq*×Top 30% is 0.54 ($t = 2.24$)

in column (1) and 0.65 ($t = 2.63$) in column (2), while the base effect and the middle interaction are insignificant. Investors reward distinctive prospectus language with capital, but only when that distinctiveness is accompanied by strong past returns. Appendix G shows that this interaction survives prospectus length, readability, boilerplate, keyword-based, and embedding-based text controls and strengthens by a third when all of them enter jointly; the keyword uniqueness of Kostovetsky and Warner (2020) carries a parallel skill-gated flow effect, and *ProspectusUniq* remains significant on top of it. Section 5.4 traces this winner-gated response to its distribution channel: it concentrates in broker-sold retail share classes, survives controls for paid distribution and fund-name distinctiveness, and leaves no detectable retail footprint around prospectus rewrites themselves.

The results demonstrate that text differentiation in prospectuses operates primarily as a marketing and pricing tool. Unique narratives attract higher inflows but only for top performers. A one-unit increase in prospectus uniqueness raises monthly flows by 0.54–0.65 percentage points for top-performing funds, but has no effect on bottom- or middle-performing ones. In contrast to the findings by Kostovetsky and Warner (2020), where they argued that unique strategy descriptions attract more inflows in the first three years, my results show that investors are indeed rewarding distinctive narratives in prospectus with capital, but only when distinctiveness is backed by strong past return records.

The results also show that while portfolio uniqueness can genuinely affect funds' performance, it has no flow effect at any performance level. This provides evidence of flow misallocation by investors: several previous studies provide evidence that investors are pursuing unique strategies using different proxies for uniqueness (Vozlyublennaia and Wu, 2017; Kostovetsky and Warner, 2020; Bonelli et al., 2024), my results, nevertheless, find that investors are more misguided by their perceived uniqueness from prospectuses, and the uniqueness-pursuing capital fails to flow into actual innovative strategies.

Because the interaction conditions on a past-performance bucket, a natural concern is that it repackages the convexity of the flow-performance relation (Sirri and Tufano, 1998;

Huang et al., 2007): the 30/40/30 gate discretizes a convex curve, so any variable correlated with rank position inside the top bucket, or with the volatility that pushes funds to extreme ranks, inherits the steep upper end of the curve. The premise of this concern is reversed in my sample (Appendix I). Funds with unique prospectuses are less volatile, with Spearman correlations of -0.11 to -0.13 between lagged *ProspectusUniq* and trailing volatility and maximum monthly returns, and within the top bucket they sit at the same rank position as other winners (rank correlation -0.000; the top-to-bottom uniqueness quintile difference in rank is -0.001 with $t = -0.30$). I nevertheless control for the curve directly. Adding the continuous performance rank moves the interaction to 0.59 ($t = 2.49$); replacing the linear lagged return with the piecewise rank slopes of Sirri and Tufano (1998) gives 0.62 ($t = 2.62$); and the sharpest specification replaces the gate's level dummies with nineteen within-month rank vigintile dummies, so that identification comes only from variation within twentieths of the rank distribution, and the interaction is 0.66 ($t = 2.77$) with family, category, and month fixed effects and 0.76 ($t = 3.07$) with family×month fixed effects (Table I.2). The result also survives rebuilding the gate itself: ranking the 36-month alpha within category×month instead of within month gives a baseline interaction of 0.68 ($t = 2.91$) and a vigintile-controlled interaction of 0.53 ($t = 2.34$), and a gate built from the trailing-twelve-month compounded return rank, the exact Sirri and Tufano (1998) sort variable, strengthens the interaction to 0.96 ($t = 4.16$).

Volatility works against the result rather than for it (Table I.3). Trailing volatility enters flows negatively in both the level and the top-bucket interaction (-0.36 and -0.13 in the full specification), a continuous *ProspectusUniq*×volatility term is zero ($t = -0.50$), and a specification combining the vigintile dummies with volatility and maximum-return levels and their top-bucket interactions leaves the interaction at 0.61 ($t = 2.53$). A double sort makes the same point without regression structure: sorting winners into performance-rank bins and *ProspectusUniq* terciles within month, mean rank is essentially identical across terciles inside every bin (maximum gap 0.004 rank points), yet the high-minus-low uniqueness flow

spread is 0.33 to 0.37 percentage points per month in the top two rank bins ($t = 2.87$ and 3.28). Permuting the gate labels within month two hundred times produces no placebo draw whose t-statistic reaches the actual one (empirical $p \approx 0.005$), and replacing the performance gate with 30/40/30 gates built on size, age, or volatility produces no positive interaction. Aggregating flows to the portfolio level preserves the result (0.70, $t = 2.57$), so it is not a share-class weighting artifact.

Estimating the flow-performance curve separately by uniqueness tercile locates the effect precisely: prospectus uniqueness steepens the top segment of the curve. Under the within-month gate, the slope on the top piece of the rank rises from 5.69 in the lowest uniqueness tercile to 8.20 in the highest, a steepening of 2.50 ($t = 2.48$), while the low and middle segments and the top-bucket intercept are unchanged (Table I.4); ranking within category×month, the steepening is positive at 1.03 but imprecise ($t = 1.03$), so I read the curve exhibit as descriptive of where the interaction operates. The incidence is retail. The retail-flow interaction is 0.84 ($t = 3.07$) and holds under the vigintile and volatility controls (0.90 and 0.89, $t > 3$), while the institutional interaction is null in every specification ($t \leq 1.12$; Table I.5). This pattern sharpens the contrast with Kostovetsky and Warner (2020): in their data, distinctive strategy text flattens the flow-performance relation for young funds, shielding them from the punishment for weak performance, whereas in mine, distinctive prospectus text amplifies the reward at the top of the realized performance distribution among established winners, and its level effect among poor performers is negative for retail investors. Under rational updating with decreasing returns to scale (Berk and Green, 2004), distinctive text would matter most where information about skill is scarcest; its concentration in the steepest segment of the curve, among retail investors, conditional on rank, volatility, size, and age, points instead to marketing amplification of the search among recent winners, the region where participation costs bind least (Huang et al., 2007).

Combining with the two previous subsections, my results find that narrative differentiation in fund disclosures operates primarily as a marketing and pricing tool, whereas real

product differentiation implemented through genuinely distinctive strategies derived from portfolio holdings contributes to reduced competition when coupled with managerial skill.

4.4 Identification

To mitigate the concern that unobserved fund attributes could be correlated with both uniqueness measures and key dependent variables, in this section I pursue causal identification for the two central findings: prospectus uniqueness raises fees, and portfolio uniqueness raises forward alpha by employing two novel instrumental variable strategies.

4.4.1 Instrumental Variable: AI Adoption Instrument for Prospectus Uniqueness

Inspired by the AI and product differentiation work by Babina et al. (2024), the key underlying idea is that AI adoption standardizes prospectus language, reducing prospectus uniqueness. This can occur for two complementary reasons: AI-assisted drafting leads prospectuses to rely more heavily on standardized language, resulting in lower textual uniqueness in prospectus; and with the increasing popularity of generative AI tools, AI-assisted drafting naturally leads to analogous outcomes. To exploit this effect for identification, I then require a measurement of AI adoption independent from key characteristics driving fee-setting decisions.

Following this logic, I utilize the family-level AI workforce share from Revelio Labs data and strip out the slow-moving family characteristics by first differencing, isolating idiosyncratic timing variation in AI hiring that is unlikely to be driven by fee-setting decisions. The AI workforce share is defined as follows. First, for each mutual fund management company, I find all employees' information using their employment histories. This generates the total number of employees and their user IDs reported in Revelio Labs. Next, using their user IDs, I search for their education records, especially their degree field. Then, using

a keyword-matching method,[6] I define an employee as an AI worker if they have at least one bachelor's degree majoring in fields related to AI, machine learning, natural language processing, and other keywords closely related to generative AI. Last, for each month, I can then compute the fund company's AI employee share by dividing the total number of AI workers by the total number of workers.

I then instrument prospectus uniqueness with the timing of family-level AI workforce adoption. The instrument is the 12-month first difference of the fund family's AI employee share, constructed from Revelio Labs workforce data using the aforementioned methodology:

$$Z_{i,t} = AI_User_Pct_{f(i),t} - AI_User_Pct_{f(i),t-12}, \tag{23}$$

where $AI_User_Pct_{f(i),t}$ is the percentage of employees in AI-related roles at fund i's parent family f. First differencing is critical: the raw level of AI employee share captures broad family transformation such as technology sophistication, scale, corporate strategy and predicts many non-fee outcomes. The 12-month change strips these slow-moving family characteristics and isolates idiosyncratic *timing* variation in AI hiring potentially driven by HR processes, candidate availability, and internal project timelines rather than fee-setting decisions. I estimate the following using a two-stage least squares (2SLS) procedure:

$$\text{First stage:} \quad ProspectusUniq_{i,t} = \pi\, Z_{i,t} + \gamma' \mathbf{X}_{i,t} + \mu_i + \delta_{c,t} + \nu_{i,t}, \tag{24}$$

$$\text{Second stage:} \quad Fee_{i,t} = \beta\, \widehat{ProspectusUniq}_{i,t} + \gamma' \mathbf{X}_{i,t} + \mu_i + \delta_{c,t} + \varepsilon_{i,t}, \tag{25}$$

where μ_i denotes fund fixed effects and $\delta_{c,t}$ denotes category×time fixed effects. The first-stage sign is negative ($\hat{\pi} < 0$): AI adoption standardizes prospectus language, reducing text uniqueness. Controls include log TNA, monthly return, and fund flow scaled by TNA. Standard errors are double-clustered by fund and month.

Table 5 reports the results. The first-stage coefficient on ΔAI User Pct is negative and

[6] See Appendix J for the complete list of AI-related keywords used to classify employees.

significant under both fixed-effect specifications (columns 1 and 4), with a Kleibergen–Paap F-statistic of 11.2 in the more demanding and preferred fixed effects specification including category×time, above the threshold of 10 for rejecting weak instruments (Stock and Yogo, 2005). This confirms the hypothesis that higher AI adoption at the company level makes mutual funds under management produce more generic prospectus language, potentially because of heavier reliance on standardized language.

The second-stage 2SLS estimate (column 5) shows that a one-unit increase in instrumented prospectus uniqueness raises annualized fees by 1.75 percentage points ($t = 3.14$), with an Anderson–Rubin p-value (Anderson and Rubin, 1949) below 0.001 confirming significance under weak-instrument-robust inference. The IV estimate is substantially larger than the OLS estimate, consistent with attenuation from measurement error in the text uniqueness proxy and a local average treatment effect among complier funds whose prospectus language is most responsive to AI adoption.

To address the concerns that AI adoption affects the total fees through some unobservable channels, I run a series of placebo and pre-trends tests. First, I test whether the instrument predicts Fama–French five-factor alpha, fund flows, monthly returns, or portfolio turnover. If the instrument satisfies the exclusion restriction, it should not predict non-fee outcomes after controlling for fund and category×month fixed effects. Second, I test whether the instrument predicts past fee changes at lags of 3, 6, 12, and 18 months. If the instrument is simply capturing a pre-trend of fee dynamics, it should generate significant coefficients at some lagged periods.

Table 6 presents the instrument validity tests. Columns (1)–(4) report placebo regressions: the instrument does not predict Fama–French five-factor alpha, fund flows, monthly returns, or portfolio turnover, confirming that it is orthogonal to non-fee outcomes after conditioning on fund and category×time fixed effects. Columns (5)–(8) test for pre-trends by regressing *past* fee changes (ΔFee_{t-k}) on the instrument at lags of 3, 6, 12, and 18 months. All eight pre-trend coefficients are statistically insignificant, ruling out the possibility that

the instrument captures pre-existing fee dynamics.

4.4.2 Instrumental Variable: Peer Exit for Portfolio Uniqueness

The key underlying idea of this instrumental variable strategy is that peer exits are likely to reflect fund-level liquidity and redemption pressures at the exiting fund, rather than information about the focal fund's future alpha. This argument is consistent with the fire-sale and flow-induced trading literature, which shows that investor flows can generate non-fundamental trading pressure and price effects through forced sales and correlated trading needs (Coval and Stafford, 2007; Lou, 2012; Greenwood and Thesmar, 2011). Accordingly, peer exits can be viewed as shocks to the competitive environment within a category: they alter the set of surviving peers and hence the category composition against which the focal fund is evaluated. As a result, even holding the focal fund's own portfolio fixed, peer exits can shift the fund's measured uniqueness, especially after controlling for category-level shocks.

Based on this idea, I instrument portfolio uniqueness with the similarity-weighted peer fund exit rate. When a peer fund exits a category, the competitive landscape shifts, mechanically altering the focal fund's position relative to its remaining peers, and these landscape shifts are weighted by how close the peer is to the focal fund. The instrument is therefore defined as:

$$Z_{i,t}^{peer} = \frac{\sum_{j \neq i,\, j \in c} \text{sim}(i,j)_{t-6} \cdot \mathbf{1}[j \text{ exited by } t]}{\sum_{j \neq i,\, j \in c} \text{sim}(i,j)_{t-6}}, \tag{26}$$

where $\text{sim}(i,j)_{t-6}$ is the portfolio holdings-based cosine similarity between funds i and j measured six months prior to avoid mechanical contamination. The instrument captures differential exposure to peer exits: a fund whose closest competitors exit experiences a larger mechanical shift in uniqueness than a fund whose distant competitors exit, even within the

same category and month. I estimate:

$$\text{First stage:} \quad PortfolioUniq_{i,t} = \pi\, Z^{peer}_{i,t} + \gamma' \mathbf{X}_{i,t} + \mu_i + \delta_{c,t} + \nu_{i,t}, \tag{27}$$

$$\text{Second stage:} \quad \alpha^{\text{fwd}}_{i,t+1} = \beta\, \widehat{PortfolioUniq}_{i,t} + \gamma' \mathbf{X}_{i,t} + \mu_i + \delta_{c,t} + \varepsilon_{i,t}, \tag{28}$$

where $\alpha^{\text{fwd}}_{i,t+1}$ is the forward FFC6 alpha at $t{+}1$ as defined in equation (20), μ_i denotes fund fixed effects, and $\delta_{c,t}$ denotes category×time fixed effects. Category×time fixed effects are essential: they absorb all category-level time-varying shocks such as sector downturns, style rotations, and category-wide exit waves that could simultaneously drive peer exits and the focal fund's performance. After demeaning, identification comes from *differential* exposure to exiting peers within the same category and month. Performance tercile dummies (I(Top), I(Bottom)) are included as exogenous controls but are not interacted with the instrument, because terciles based on backward alpha are endogenous: interacting Z^{peer} with an endogenous grouping variable would violate the exclusion restriction even if Z^{peer} itself is valid. Controls include log TNA, lagged return, flow/TNA, turnover ratio, and expense ratio. Standard errors are double-clustered by fund and month.

Table 7 reports the results. The first stage (column 1) confirms that peer exits are a strong predictor of portfolio uniqueness, with a first-stage coefficient of -0.270 ($t = -8.18$) and a Kleibergen–Paap F-statistic of 66.9, well above the Stock and Yogo (2005) 10% critical value of 16.38 for one instrument. The negative sign, higher similarity-weighted peer exit reduces the fund's uniqueness measure, reflects the compression of competitive niches. Since portfolio uniqueness is a cross-modal measure (text-similarity-weighted holdings centroid distance), the effect operates through category composition. Exiting funds are disproportionately niche or extreme, and funds with high Z^{peer} (holdings-similar to exiters) occupy the same fragile niches. When these niche peers exit, the remaining category becomes more tightly clustered around the mainstream: the text-weighted holdings centroid shifts toward a more central position, mechanically reducing the distance for funds in the depopulated

periphery. The 6-month lag between similarity measurement and outcome also allows for behavioral convergence: with fewer close competitors, funds face less pressure to differentiate and may reposition toward more standard strategies.

The second-stage 2SLS estimate (column 2) shows that a one-unit increase in instrumented portfolio uniqueness raises next-month forward alpha by 1.22 percentage points ($t = 2.19$, $p = 0.028$). The Anderson–Rubin p-value of 0.006 confirms significance under weak-instrument-robust inference. Column (3) reports the reduced-form regression: peer exits directly predict lower forward alpha ($t = -2.75$), consistent with the causal chain. Both instrumental-variable estimates are insensitive to the competitor proxies of Appendix G: adding active share, $1 - R^2$, log breadth, and industry concentration moves the peer-exit 2SLS coefficient by less than 1%, and adding length and boilerplate controls moves the AI-adoption 2SLS coefficient by about 2% while raising the first-stage F-statistic.

To examine whether the causal effect varies with managerial skill, I estimate separate 2SLS regressions within each past-performance tercile subsample. This approach avoids the econometric pitfall of interacting the instrument with endogenous tercile dummies, which would violate the exclusion restriction even if the base instrument is valid. Table 8 reports the results. The first-stage F-statistics are strong in all three subsamples (54.9, 36.6, and 29.3), confirming that the instrument has power across the performance distribution. The 2SLS estimates reveal a striking monotonic pattern: the effect of portfolio uniqueness on forward alpha is negative and insignificant for bottom-tercile funds (-0.45, $t = -0.53$), positive and marginally significant for middle-tercile funds (1.55, $t = 1.75$), and positive and marginally significant for top-tercile funds (2.11, $t = 1.71$). This monotonicity, consistent with the panel interaction regression results in Table 3, confirms that portfolio uniqueness generates alpha only when paired with managerial skill. For unskilled managers, uniqueness may reflect idiosyncratic bets rather than informed positions as previously argued.

To address the concerns that peer exits affect future performance through some unobservable channels, I run a series of placebo and pre-trends tests. First, I test whether

the instrument is related to past risk-adjusted performance. If there exists a pre-trend or unobservable persistent fund characteristics driving future alpha, I should also be able to find clear relationships of the instrument and past performance. For robustness, I check for Fama–French five-factor alpha and Carhart four-factor alpha, both calculated from past 12 months and 36 months returns. Next, I run a series of placebo tests, asking whether the instrument is related to key fund characteristics, including fees, turnover ratio, fund age, manager tenure length, Morningstar ratings, fund monthly returns, fund Sharpe ratio and fund style consistency. If the instrument impacts future performance through some other mechanisms, I can find its significant relationship with some of the key characteristics.

Table 9 presents instrument validity tests. Columns (1)–(4) test for pre-trends by regressing backward-looking alpha on Z^{peer} in reduced-form regressions. As previously argued, if the instrument captured persistent fund characteristics rather than exogenous competitive shocks, it would predict past alpha as well as future alpha. None of the four backward alpha measures is significant at the 5% level ($|t| < 1.33$ across all specifications). Columns (5)–(12) test whether the instrument predicts eight placebo outcomes spanning fund fees, trading behavior, structural characteristics, and backward risk-adjusted performance. All eight placebo coefficients are statistically insignificant, covering management fees ($t = -0.24$), turnover ($t = 0.36$), fund age ($t = 1.76$), manager tenure ($t = 0.23$), star rating ($t = 0.08$), monthly return ($t = -0.15$), Sharpe ratio ($t = 0.45$), and style consistency ($t = -0.43$). Zero out of twelve tests significant at 5%. This comprehensive pass rate provides strong support for the instrument's exclusion restriction.

5 Retail and Institutional Investors

Who is buying strategy narratives in mutual fund prospectuses? This section asks two specific questions: who pays the higher fees documented in Section 4.1, and who chases the additional flows among top-performing funds documented in Section 4.3. To answer them, I

decompose the panel regressions in Tables 2 and 4 by CRSP fund-type variables, following prior work (Chen, Goldstein and Jiang, 2010; Evans and Fahlenbrach, 2012; Christoffersen et al., 2013). I cross-validate and complement the retail and institutional flags using CRSP fund names following the methodology in Gil-Bazo and Ruiz-Verdú (2009). All regression specifications are identical to those in Section 4.

5.1 Fees by Fund Type

I first examine who pays the higher fees by re-estimating equation (19), splitting the sample by CRSP fund-type flags and CRSP fund names. Because the flags are reported quarterly, I forward-fill them to monthly frequency. The results are reported in Table 10.

For retail funds, the relationship between prospectus uniqueness and total fees is marginal and economically small, around 5 basis points. The magnitude and significance are substantially stronger for institutional funds: the coefficient ranges from 14 to 15 basis points, with t-statistics roughly twice as large. Consistent with the pooled results in Table 2, portfolio uniqueness remains insignificant in both subsamples, indicating once again that differentiation in portfolio holdings does not command higher fees.

This finding is puzzling. Institutional investors are conventionally treated as more sophisticated, yet they appear to pay more for distinctive prospectus language than retail investors do. If they pay more for distinctive prospectuses, do they also chase them with capital?

5.2 Flows by Fund Type

I answer this question by re-estimating equation (22) using the same sample split. I proxy retail and institutional flows using the share-class net flows as a percentage of share-class TNA. The results are reported in Table 11.

The results reveal a sharp behavioral contrast between the two investor groups. Retail investors chase narrative uniqueness with additional capital among funds with superior past

performance, and they appear to punish poor-performing funds with unusual prospectuses by withdrawing capital. Institutional flows, by contrast, are essentially unresponsive to distinctive narratives. There is weak evidence that distinctive portfolios may hurt institutional flows among poor performers, suggesting that institutional investors may pay closer attention to actual holdings; this evidence, however, is not robust across specifications. Combined with the fee result, this finding creates a second asymmetry: institutional investors are willing to pay more for distinctive prospectuses but do not allocate more capital to them. Section 5.4 takes this contrast one step further on the retail side by examining how the narrative is distributed to investors.

5.3 The Retail-Institutional Asymmetry

That retail investors are more responsive to narrative content is broadly consistent with prior evidence. Cooper et al. (2005) show that mutual funds adopting hot-style names experience large positive abnormal flows in the year after the name change, even when their holdings do not align with the new label. Solomon, Soltes and Sosyura (2014) show that media coverage of fund holdings drives subsequent flows through investor attention rather than through information content. Roussanov et al. (2021) estimate a structural model in which marketing is nearly as important as performance and fees in shaping fund size, and the marketing channel operates primarily through retail investors. Del Guercio and Tkac (2008) show that discrete Morningstar rating changes drive substantial retail flows independent of underlying performance changes. Against this background, the puzzle in my data is not that retail investors react to narratives, but that institutional investors are paying more for narratives without flowing into them. As Section 4.2 establishes, the distinctive portfolios of skilled managers do generate additional alpha, yet these funds are not the ones collecting the higher fees. Price is therefore not a clean reflection of fund quality, even in the institutional market.

Two channels can potentially explain the puzzle. The first is that distinctive prospec-

tuses correlate with specialized institutional mandates such as sector-specific or ESG strategies, where competitive pressure is lower and managers can charge higher fees. To the extent that these funds diverge from their stated strategies and instead hold generic portfolios, this channel also explains why the data does not support a positive portfolio uniqueness effect, even in the institutional sample.

The second channel is a principal-agent friction in institutional allocation. Institutional fund selection is typically mediated by consultants and investment committees who must justify their selections to plan sponsors. Distinctive narratives provide post-hoc justification capital, so the higher fees become a rent on the justification value of distinctive language rather than on its informational content.

Taken together, these results suggest that the higher fees paid by institutional investors for prospectus uniqueness are best understood as a rent on narrative production. The segment-level asymmetry between who pays and who chases is difficult to reconcile with a single rational-agent benchmark in which the same investor both selects funds and bears their fees. The findings do not imply that institutional investors are inattentive. Rather, they imply that the institutional channel separates the consumer of the narrative, namely the consultant or committee using it as justification, from the bearer of its cost, namely the underlying plan beneficiary. This decoupling weakens the standard intuition that institutional markets discipline mutual fund pricing more effectively than retail markets, and it reframes prospectus distinctiveness as a product attribute that commands higher fees not from what it reveals about portfolios, but from the role it plays inside institutional decision processes.

5.4 The Distribution Channel of Narrative Flows

The retail flow evidence invites a natural objection: retail investors rarely read statutory prospectuses, and experimental evidence shows that fee and disclosure information moves allocation choices far less than a frictionless benchmark predicts (Choi et al., 2010). The

objection targets the document. The measure, however, captures text that reaches investors through the sales interface. Since 2009, Rule 498 lets funds satisfy their delivery obligation with a summary prospectus whose first substantive item is the principal investment strategies section, the same section my text embeddings represent, and the same strategy language populates fund profile pages on brokerage platforms and the materials intermediaries draw on when they present a fund to a client. Prospectus uniqueness therefore measures how differentiated the fund's displayed strategy pitch is from its category peers at the point of sale. On this reading, the winner-flow gate of Table 4 is a prediction about distribution: a distinctive pitch helps most where an intermediary actively sells the fund and past performance makes the pitch credible (Sirri and Tufano, 1998; Roussanov et al., 2021).

The distribution channel is observable at the share-class level. I classify share classes as broker-sold or direct-sold from three CRSP signals: the class letter in the fund name, the presence of a 12b-1 plan, and front or deferred load schedules. Table 12 re-estimates the flow specification within each channel. The winner-flow response to prospectus uniqueness concentrates in broker-sold retail classes, at 0.66 to 0.80 percentage points per unit of uniqueness across both fixed-effects schemes, larger than the pooled estimate. The base effect sharpens in the same channel: broker-sold retail classes punish poorly performing funds with distinctive prospectuses at -0.67 to -0.70, which accounts for the retail-subsample punishment in Table 11. Institutional classes show a well-powered null on both margins, with over 92,000 observations. Direct-sold retail estimates are positive and imprecise; the cell holds under 8% of the broker-sold observations, so the split localizes the effect where I can measure it and leaves the direct-sold channel undetermined. Both the winner gate and the loser punishment on prospectus uniqueness live where an intermediary carries the strategy text to the investor, the segment Del Guercio and Reuter (2014) show is served by advice-seeking investors who delegate fund selection. A within-fund comparison of broker and direct classes of the same portfolio would sharpen this further; my estimation panel retains roughly one share class per portfolio, so the dual-class universe contains six to nine funds (Table 12, Panel B) and

supports no reliable inference.

Two artifacts could still generate the channel pattern, and Table 15 rejects the first. If the flow response merely priced paid shelf space, it would sit inside 12b-1 share classes. The winner-flow coefficient survives at full size in classes with no 12b-1 plan, and the pooled triple interaction with 12b-1 status is small and insignificant in every specification, so the response is statistically indistinguishable across paid-distribution status. A suggestive cut in the same table locates the response among funds rated four or five stars, the funds platforms and brokers showcase, consistent with the salience of discrete ratings for retail money (Del Guercio and Tkac, 2008); star coverage is selected and correlated with the performance gate, so I read this split as descriptive. The second candidate artifact is the fund's name, the one piece of text every investor sees. Name distinctiveness, measured against category-month peers, correlates with prospectus uniqueness at 0.21 and with portfolio uniqueness at 0.06: distinctive narratives come wrapped in distinctive names, and name changes coincide with large prospectus rewrites at 2.6 to 2.9 times the base rate (Table 14). The flow gate survives the name surface intact: controlling for name distinctiveness and its performance interaction leaves the prospectus-uniqueness interaction significant in every specification, at 0.50 to 0.75 across both gates and both fixed-effects schemes, while the name interaction itself is unstable in sign across gate definitions. Matched flow comparisons around name changes show funds bleeding relative to peers in the change year and no flow gain in the year after, so at annual resolution renaming looks like a response to distress; the post-change flow surge that Cooper et al. (2005) document at monthly frequency around hot-style renames operates within a window my annual name data cannot resolve.

The event-study evidence bounds the mechanism. Table 13 stacks flows around 1,884 large prospectus rewrites, using near-verbatim refilings to net out the filing cycle. Retail flows show no detectable response to the rewrite itself: the cumulative twelve-month estimate is 0.4 percentage points of TNA with a confidence interval of ± 3.7, and the sharpest cell, rewrites unaccompanied by any holdings change, is also null. The one response the design detects

is institutional: a cumulative 8.2 percentage points of TNA over twelve months, building from the third month after the filing, with clean pre-trends. The cross-sectional gate and the time-series null fit together under the sales-interface reading. The gate rewards the level of differentiation of the displayed strategy text among funds whose performance makes them saleable; rewriting the document, holding the fund's competitive position otherwise fixed, moves no retail money, which matches the experimental evidence on disclosure redesign (Choi et al., 2010). Retail narrative-chasing is a distribution phenomenon: the text does its work in the hands of the intermediary, at the point of sale, when performance cooperates. Institutional capital responds to the filing event itself, with a delay consistent with due-diligence cycles, the same monitoring behavior that generates the enforcement discount in Section 2.

6 Conclusion

Mutual funds can differentiate themselves along two distinct dimensions: their actual investment strategies, reflected in portfolio holdings, and their strategy descriptions, reflected in prospectus narratives. While the majority of the literature approaches mutual fund differentiation from a single-dimensional perspective, this paper shows that uniqueness in actions and uniqueness in disclosures are systematically different and have different economic consequences.

Using deep-learning embedding techniques applied to mutual fund holdings and prospectus filings, I construct two novel cross-modal measures of fund uniqueness in portfolios and prospectuses. I show that the two forms of uniqueness, in prospectus narratives and in portfolio holdings, have distinct economic roles. Prospectus uniqueness raises fees by approximately 1.6 basis points per standard deviation but has no predictive content for future risk-adjusted performance. Portfolio uniqueness predicts forward alpha among funds with demonstrated managerial skill but has no pricing power. Prospectus uniqueness attracts

capital among top-performing funds, while portfolio uniqueness has no flow effect at any performance level. To alleviate endogeneity concerns, I instrument portfolio uniqueness by peer exits and prospectus uniqueness by AI workforce adoption, and find consistent results. The effects also differ between retail and institutional investors: the flow response to prospectus uniqueness comes primarily from retail investors, while the higher fees it commands are borne mainly by institutional investors.

These findings suggest that distinctive prospectuses signal marketing effort, not necessarily actual portfolio innovation. The flow-alpha mismatch implies that capital chases narrative distinctiveness in a market where alpha is generated by holdings distinctiveness. Investors who select funds on the basis of prospectus language are systematically paying for the wrong kind of differentiation. Moreover, the divergent economic roles of the two measures suggest that current N-1A disclosure requirements do not effectively convey how a fund's actual strategy compares to those of its peers. Enhanced disclosure formats that link narrative claims more directly to observable portfolio behavior could reduce the scope for marketing differentiation to substitute for genuine product innovation.

Table 1
Summary Statistics

This table reports summary statistics for the estimation sample. The sample consists of 666,375 fund-month observations for 6,024 unique portfolios across 180 months (January 2011 to December 2025). *ProspectusUniq* and *PortfolioUniq* are the cross-modal weighted uniqueness measures defined in Section 3.3. Annualized fee is computed from the gross-net monthly return gap. Forward alpha is the realized abnormal return at horizon h, computed using FFC6 factor loadings estimated over months $[t-24, t-1]$. All continuous variables are winsorized at the 1st and 99th percentiles within each month. N varies across variables due to data availability in CRSP, Morningstar, and SEC EDGAR sources.

	N	Mean	Std	P5	P25	Median	P75	P95
Panel A: Uniqueness Measures								
ProspectusUniq	458,190	0.658	0.189	0.368	0.524	0.643	0.779	0.994
PortfolioUniq	458,190	0.576	0.193	0.261	0.431	0.575	0.720	0.894
Panel B: Dependent Variables								
Annualized fee (%)	663,804	1.086	0.467	0.190	0.843	1.099	1.334	1.916
Fund flow (% of TNA)	656,504	−0.093	1.967	−1.069	−0.043	−0.006	0.017	0.640
Forward alpha, FFC6 t+1 (%)	608,774	−0.272	2.310	−4.206	−1.365	−0.190	0.866	3.382
Forward alpha, FFC6 t+3 (%)	596,429	−0.301	2.402	−4.421	−1.431	−0.205	0.885	3.465
Forward alpha, FFC6 avg 12M (%)	543,459	−0.250	0.744	−1.625	−0.535	−0.177	0.058	0.871
Panel C: Fund Characteristics								
TNA ($M)	657,604	1,190	8,634	0.5	13.4	87.2	479.3	4,497
Log TNA	657,604	4.279	2.680	−0.693	2.595	4.468	6.172	8.411
Fund age (years)	666,375	13.5	11.3	1.0	5.1	11.3	19.2	31.5
Turnover ratio	611,324	0.647	0.767	0.060	0.230	0.440	0.770	1.840
Monthly return (%)	661,258	0.81	4.59	−7.47	−1.83	1.08	3.55	8.10
Expense ratio (%)	612,451	1.11	0.47	0.20	0.86	1.12	1.35	1.95
Panel D: Performance & Risk								
Alpha, C4 12M (%)	640,555	−2.287	9.472	−18.928	−6.786	−1.782	2.286	13.227
Alpha, FF5 36M (%)	600,876	−2.361	4.755	−10.914	−4.651	−1.885	0.192	4.850
Sharpe ratio (12M)	631,517	1.006	1.179	−0.730	0.184	0.884	1.708	3.209
Star rating (1–5)	562,181	3.01	1.05	1.0	2.0	3.0	4.0	5.0
Style consistency	431,337	14.0	8.6	4.7	8.1	11.9	17.4	31.2
Panel E: Manager Variables								
N managers	665,426	2.97	2.44	1.0	2.0	2.0	3.0	7.0
Manager tenure (months)	480,721	70.0	75.5	0.0	6.0	48.0	107.0	222.0
Manager retention 1Y (%)	579,494	93.3	6.7	80.0	90.9	94.3	99.2	100.0

Table 2
Uniqueness and Fund Fees

This table reports the effect of lagged uniqueness measures on annualized fund fees, derived from the gap between gross and net monthly returns. The dependent variable is annualized fee (in percentage points), winsorized at the 1st and 99th percentiles. Columns (1)–(3) lag uniqueness by one month; columns (4)–(6) lag by three months. Controls include log TNA, turnover ratio, fund return, and fund age (all lagged). Standard errors are double-clustered by fund and month. t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	(1)	(2)	(3)	(4)	(5)	(6)
	1-Month Lag			3-Month Lag		
$PortfolioUniq_{t-k}$	−0.0150	−0.0112	−0.0125	−0.0150	−0.0116	−0.0132
	(−0.78)	(−0.54)	(−0.59)	(−0.78)	(−0.56)	(−0.62)
$ProspectusUniq_{t-k}$	0.0837***	0.0867***	0.0832***	0.0832***	0.0861***	0.0828***
	(3.39)	(3.18)	(2.99)	(3.37)	(3.16)	(2.98)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Family FE	Yes			Yes		
Category FE	Yes	Yes		Yes	Yes	
Month FE	Yes			Yes		
Family×Month FE		Yes	Yes		Yes	Yes
Category×Month FE			Yes			Yes
R^2	0.749	0.751	0.755	0.750	0.751	0.755
Observations	425,354	382,291	381,336	421,304	378,570	377,651

Table 3
Portfolio Uniqueness, Past Performance, and Forward Alpha

This table reports the effect of uniqueness measures and their interactions with past performance terciles on forward FFC6 alpha (monthly %, not annualized). Forward alpha at month t uses betas estimated from $t-24$ through $t-1$ and computes realized abnormal return at $t+1$ (columns 1–2) or $t+3$ (columns 3–4). Performance terciles are assigned within category×month using backward 12-month rolling alpha. Middle tercile is the reference group. Columns (1) and (3) report Fama–MacBeth estimates with Newey–West t-statistics. Columns (2) and (4) include portfolio and month fixed effects with double-clustered standard errors (portfolio × month). Controls include log TNA, fund age, number of managers, and turnover ratio. t-statistics in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

	(1)	(2)	(3)	(4)
	α^{FFC6}_{t+1}		α^{FFC6}_{t+3}	
	FM	Panel	FM	Panel
$PortfolioUniq_t$	−0.0785	0.0045	−0.1110	−0.0308
	(−1.06)	(0.05)	(−1.42)	(−0.37)
$ProspectusUniq_t$	−0.0663	−0.1102	−0.0563	−0.0057
	(−0.85)	(−1.24)	(−0.66)	(−0.06)
Performance Interactions (reference: Middle tercile)				
$PortfolioUniq \times I(\text{Top})$	0.1805***	0.1123**	0.1390***	0.0900*
	(3.16)	(2.24)	(2.57)	(1.90)
$PortfolioUniq \times I(\text{Bottom})$	0.0501	0.0144	0.1045	0.0574
	(0.79)	(0.28)	(1.46)	(1.19)
$ProspectusUniq \times I(\text{Top})$	−0.0172	0.0236	−0.0050	0.0033
	(−0.35)	(0.42)	(−0.07)	(0.06)
$ProspectusUniq \times I(\text{Bottom})$	0.0612	0.0495	0.0292	−0.0245
	(1.07)	(0.97)	(0.50)	(−0.56)
$I(\text{Top})$	−0.0274	−0.0535	−0.0301	−0.0587
	(−0.72)	(−1.35)	(−0.78)	(−1.45)
$I(\text{Bottom})$	−0.1236***	−0.0587	−0.1015**	−0.0104
	(−2.75)	(−1.50)	(−2.02)	(−0.29)
Controls	Yes	Yes	Yes	Yes
Portfolio FE	—	Yes	—	Yes
Month FE	—	Yes	—	Yes
Observations	3,416/mo	398,241	3,395/mo	394,944
R^2 (within)	—	0.001	—	0.001

Table 4
Uniqueness, Past Performance, and Fund Flows

This table reports the effect of lagged uniqueness measures on fund net flows (as a percentage of fund net assets, winsorized at 1st/99th percentiles), interacted with past performance rank. Performance rank is based on 36-month Fama–French five-factor alpha terciles (bottom 30%, middle, top 30%) computed within each month. Panel A reports base effects (for bottom-30% performers). Panel B reports performance rank intercepts. Panels C and D report the differential effect of uniqueness for middle and top performers relative to the bottom group. Controls include log TNA, turnover ratio, fund return, and fund age (all lagged). Standard errors are double-clustered by fund and month. t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	(1)	(2)
Panel A: Base effects (bottom 30% performers)		
$PortfolioUniq_{t-1}$	0.0146	0.0867
	(0.08)	(0.50)
$ProspectusUniq_{t-1}$	−0.1750	−0.3039
	(−0.96)	(−1.64)
Panel B: Past performance rank		
Middle	0.4710***	0.3635**
	(3.28)	(2.48)
Top 30%	1.0661***	0.9275***
	(5.55)	(4.83)
Panel C: Portfolio Uniqueness × Performance		
$PortfolioUniq_{t-1}$ × Middle	0.1886	0.0590
	(1.00)	(0.30)
$PortfolioUniq_{t-1}$ × Top 30%	0.1918	0.0261
	(0.87)	(0.12)
Panel D: Prospectus Uniqueness × Performance		
$ProspectusUniq_{t-1}$ × Middle	0.0781	0.2288
	(0.40)	(1.11)
$ProspectusUniq_{t-1}$ × Top 30%	0.5426**	0.6526***
	(2.24)	(2.63)
Controls	Yes	Yes
Family FE	Yes	
Category FE	Yes	Yes
Month FE	Yes	
Family×Month FE		Yes
R^2	0.051	0.218
Observations	396,453	356,530

Table 5
IV Estimates: Effect of Prospectus Uniqueness on Fees

This table reports 2SLS estimates of the effect of prospectus uniqueness on annualized management fees. The instrument is the 12-month first difference of the fund family's AI employee share (ΔAI_User_Pct$_{12M}$), constructed from Revelio Labs workforce data. First differencing removes slow-moving family characteristics (size, sophistication, technology orientation) that contaminate the level-based instrument. Columns (1)–(3) include fund and time fixed effects; columns (4)–(6) replace time FE with category×month FE. Controls include log total net assets, monthly return, and fund flow scaled by TNA. KP Wald F is the Kleibergen–Paap rank Wald F-statistic for weak identification (Stock–Yogo 15% critical value: 8.96). AR p is the Anderson–Rubin weak-instrument-robust test. Standard errors are double-clustered by fund and month. t/z-statistics in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

	Model I			Model II		
	(1)	(2)	(3)	(4)	(5)	(6)
	First Stage	2SLS	Red. Form	First Stage	2SLS	Red. Form
Dep. var.:	*ProspUniq*	Fee	Fee	*ProspUniq*	Fee	Fee
ΔAI User Pct$_{12M}$	−0.021***		−0.034***	−0.019***		−0.033***
	(−3.37)		(−4.98)	(−3.34)		(−5.17)
ProspectusUniq		1.617***			1.745***	
		(3.20)			(3.14)	
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fund FE	Yes	Yes	Yes	Yes	Yes	Yes
Category×Month FE				Yes	Yes	Yes
KP Wald F	11.4	11.4	—	11.2	11.2	—
AR p	—	<0.001	—	—	<0.001	—
Observations	184,220	184,220	184,220	183,276	183,276	183,276
Funds	2,034	2,034	2,034	2,032	2,032	2,032

Table 6
IV Validity Tests: Prospectus Uniqueness Instrument

This table reports instrument validity tests for ΔAI_User_Pct$_{12M}$, the 12-month first difference of the fund family's AI employee share. *Columns (1)–(4):* Placebo outcomes. If the instrument satisfies the exclusion restriction, it should not predict non-fee outcomes after controlling for fund and category×month fixed effects. α^{FF5} is the 36-month rolling Fama–French five-factor alpha; Flow is net fund flow scaled by TNA; Return is the monthly fund return; Turnover is the turnover ratio. None is significant at the 5% level. *Columns (5)–(8):* Fee change pre-trends. I test whether the instrument predicts *past* 6-month fee changes at lags of 3, 6, 12, and 18 months. All four pre-trend coefficients are insignificant, confirming the instrument is not picking up pre-existing fee dynamics. All specifications include fund and category×month fixed effects. Controls match the main specification. Standard errors are double-clustered by fund and month. t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Placebo Outcomes				Fee Change Pre-Trends			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var.:	α^{FF5}	Flow	Return	Turnover	ΔFee$_{t-3}$	ΔFee$_{t-6}$	ΔFee$_{t-12}$	ΔFee$_{t-18}$
ΔAI User Pct$_{12M}$	0.070	−0.024	−0.017	−0.022	−0.002	0.001	0.000	0.002
	(1.19)	(−0.58)	(−0.91)	(−1.90)	(−0.82)	(0.52)	(0.12)	(0.87)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fund FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Category×Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Significant at 5%?	No	No	No	No	No	No	No	No

Table 7
IV Estimates: Effect of Portfolio Uniqueness on Forward Alpha

This table reports instrumental variable estimates of the effect of portfolio uniqueness on forward FFC6 alpha (monthly %, not annualized). Forward alpha at month t uses factor loadings estimated from $t-24$ through $t-1$ and computes the realized abnormal return at $t+1$. Column (1) reports the first-stage regression of *PortfolioUniq* on the excluded instrument Z^{peer} (similarity-weighted fraction of within-category peers that exited in the prior six months). Column (2) reports the second-stage 2SLS estimate. Column (3) reports the reduced-form regression of forward alpha directly on Z^{peer}. Performance tercile dummies (I(Top), I(Bottom)) are included as exogenous controls but not interacted with Z^{peer}, because terciles based on backward alpha are endogenous. All specifications include portfolio and category×month fixed effects. The Kleibergen–Paap (KP) Wald F-statistic tests for weak instruments (Stock–Yogo 10% critical value: 16.38 for one instrument). The Anderson–Rubin (AR) p-value tests the null of zero causal effect and is robust to weak instruments. Standard errors are two-way clustered by portfolio and month. t-statistics in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

	(1) First Stage	(2) 2SLS	(3) Reduced Form
Dep. var.:	*PortfolioUniq*	α^{FFC6}_{t+1}	α^{FFC6}_{t+1}
Z^{peer}	−0.2700***		−0.3738***
	(−8.18)		(−2.75)
PortfolioUniq		1.2217**	
		(2.19)	
Log TNA	−0.0001	−0.0357***	−0.0358***
	(−0.09)	(−8.87)	(−12.19)
Monthly Return	0.0363	1.2165	1.1190***
	(1.63)	(0.94)	(3.37)
Flow / TNA	0.0002	0.0024	0.0034**
	(0.51)	(1.35)	(2.32)
Turnover	0.0015	−0.0465***	−0.0447***
	(0.58)	(−3.91)	(−3.76)
I(Top)	Yes	Yes	Yes
I(Bottom)	Yes	Yes	Yes
Portfolio FE	Yes	Yes	Yes
Category×Month FE	Yes	Yes	Yes
KP Wald F	66.9	66.9	—
AR p-value	—	0.006	—
Observations	379,794	379,794	395,725
Portfolios	3,820	3,820	3,923

Table 8
IV Estimates by Past Performance Tercile

This table reports separate 2SLS estimates within each past-performance tercile. The dependent variable is forward FFC6 alpha at $t+1$ (monthly %). Each column restricts the sample to funds in the bottom, middle, or top tercile of within-category-month backward 12-month alpha. The instrument is Z^{peer} (similarity-weighted peer exit fraction). The subsample approach allows the local average treatment effect (LATE) to vary across performance groups without imposing invalid instrument interactions. All specifications include portfolio and category×month fixed effects. Controls include log TNA, monthly return, flow/TNA, turnover, and expense ratio. Standard errors are two-way clustered by portfolio and month. t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Bottom	Middle	Top
	(1)	(2)	(3)
Panel A: First Stage ($Z^{peer} \rightarrow PortfolioUniq$)			
Z^{peer}	−0.2975***	−0.2327***	−0.2315***
	(−7.41)	(−6.05)	(−5.41)
KP Wald F	54.9	36.6	29.3
Panel B: 2SLS			
PortfolioUniq	−0.4483	1.5528*	2.1103*
	(−0.53)	(1.75)	(1.71)
Controls	Yes	Yes	Yes
Portfolio FE	Yes	Yes	Yes
Category×Month FE	Yes	Yes	Yes
Observations	130,175	136,045	113,211
Portfolios	3,580	3,566	3,362

Table 9
IV Validity: Pre-Trends and Placebo Outcomes

All specifications include controls, tercile dummies (I(Top), I(Bottom)), portfolio FE, and category×month FE, matching Table 7. Panel A: the instrument should not predict backward-looking alpha; all four pre-trend coefficients are insignificant. Panel B: the instrument should not predict fund characteristics unaffected by the portfolio uniqueness channel. None of the 12 tests is significant at the 5% level. t-statistics in parentheses; SEs clustered by portfolio. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Panel A: Pre-Trends (Backward Alpha)

Dep. var.:	(1) α^{FF5}_{12M}	(2) α^{FF5}_{36M}	(3) α^{C4}_{12M}	(4) α^{C4}_{36M}
Z^{peer}	−1.097	−0.509	−0.706	−0.776
	(−1.04)	(−0.80)	(−0.81)	(−1.33)
Sig. at 5%?	No	No	No	No
Observations	411,675	393,065	411,675	393,065

Panel B: Placebo Outcomes

Dep. var.:	(5) Fee	(6) Turnover	(7) Age	(8) Tenure	(9) Stars	(10) Return	(11) Sharpe	(12) Style
Z^{peer}	−0.000	0.030	1.453	2.122	0.016	−0.000	0.024	−0.710
	(−0.24)	(0.36)	(1.76)	(0.23)	(0.08)	(−0.15)	(0.45)	(−0.43)
Sig. at 5%?	No	No	No	No	No	No	No	No
Observations	411,675	411,675	411,675	315,099	371,774	411,675	411,675	287,991

Table 10
Uniqueness and Fund Fees by Investor Type

This table reports the effect of lagged uniqueness measures on annualized fund fees, derived from the gap between gross and net monthly returns, separately for funds classified as retail and institutional by CRSP. The dependent variable is annualized fee (in percentage points), winsorized at the 1st and 99th percentiles. Columns (1)–(2) and (5)–(6) lag uniqueness by one month; columns (3)–(4) and (7)–(8) lag by three months. Columns (1)–(4) restrict to funds classified as retail by CRSP; columns (5)–(8) restrict to funds classified as institutional. Controls include log TNA, turnover ratio, fund return, and fund age (all lagged). Standard errors are double-clustered by fund and month. t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Retail Funds				Institutional Funds			
	1-Month Lag		3-Month Lag		1-Month Lag		3-Month Lag	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$PortfolioUniq_{t-k}$	−0.0259	−0.0277	−0.0264	−0.0281	−0.0183	0.0016	−0.0177	0.0017
	(−1.26)	(−1.24)	(−1.28)	(−1.26)	(−0.65)	(0.05)	(−0.62)	(0.06)
$ProspectusUniq_{t-k}$	0.0509*	0.0527*	0.0501*	0.0525*	0.1392***	0.1458***	0.1410***	0.1474***
	(1.91)	(1.77)	(1.88)	(1.76)	(4.25)	(3.75)	(4.28)	(3.77)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Family FE	Yes		Yes		Yes		Yes	
Category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes		Yes		Yes		Yes	
Family×Month FE		Yes		Yes		Yes		Yes
R^2	0.085	0.093	0.084	0.091	0.042	0.045	0.041	0.045
Observations	296,360	263,294	291,696	259,021	126,396	106,180	124,244	104,322

Table 11
Uniqueness, Past Performance, and Fund Flows by Investor Type

This table reports the effect of lagged uniqueness measures on fund net flows (as a percentage of fund net assets, winsorized at 1st/99th percentiles), interacted with past performance rank, separately for funds classified as retail and institutional by CRSP. Performance rank is based on 36-month Fama–French five-factor alpha terciles (bottom 30%, middle, top 30%) computed within each month. Panel A reports base effects (for bottom-30% performers). Panel B reports performance rank intercepts. Panels C and D report the differential effect of uniqueness for middle and top performers relative to the bottom group. Columns (1)–(2) restrict to funds classified as retail by CRSP; columns (3)–(4) restrict to funds classified as institutional. Controls include log TNA, turnover ratio, fund return, and fund age (all lagged). Standard errors are double-clustered by fund and month. t-statistics in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

	Retail Funds		Institutional Funds	
	(1)	(2)	(3)	(4)
Panel A: Base effects (bottom 30% performers)				
$PortfolioUniq_{t-1}$	0.1767	0.2036	−0.6990**	−0.1724
	(0.90)	(1.02)	(−1.99)	(−0.47)
$ProspectusUniq_{t-1}$	−0.5043***	−0.5116**	0.5757	−0.0592
	(−2.62)	(−2.45)	(1.50)	(−0.15)
Panel B: Past performance rank				
Middle	0.2246	0.2212	0.8652***	0.7893**
	(1.44)	(1.32)	(2.67)	(2.23)
Top 30%	0.9440***	0.9172***	1.2948***	1.1431***
	(4.63)	(4.31)	(3.30)	(2.75)
Panel C: Portfolio Uniqueness × Performance				
$PortfolioUniq_{t-1}$ × Middle	0.1260	0.0357	0.5518	−0.0435
	(0.61)	(0.17)	(1.38)	(−0.10)
$PortfolioUniq_{t-1}$ × Top 30%	0.0756	0.0227	0.7032	−0.0322
	(0.31)	(0.09)	(1.54)	(−0.07)
Panel D: Prospectus Uniqueness × Performance				
$ProspectusUniq_{t-1}$ × Middle	0.3641*	0.3767	−0.4488	−0.0397
	(1.71)	(1.58)	(−1.10)	(−0.10)
$ProspectusUniq_{t-1}$ × Top 30%	0.6633***	0.5996**	0.1790	0.6462
	(2.59)	(2.16)	(0.35)	(1.25)
Controls	Yes	Yes	Yes	Yes
Family FE	Yes		Yes	
Category FE	Yes	Yes	Yes	Yes
Month FE	Yes		Yes	
Family×Month FE		Yes		Yes
R^2	0.020	0.021	0.017	0.016
Observations	281,931	250,703	111,954	93,865

Table 12
Uniqueness, Past Performance, and Fund Flows by Distribution Channel

This table splits the share-class flow specification of Table 4 by distribution channel. The dependent variable is share-class net flow as a percentage of share-class TNA, winsorized at the 1st/99th percentiles. I classify share classes from three CRSP signals: the class-letter suffix of the CRSP fund name (broker letters: A, B, C, T, M, Advisor; direct letters: Institutional, I, Y, Z, R5/R6, Investor, Admiral, Service), the presence of a 12b-1 plan, and front/CDSC load schedules. Broker-sold combines a broker letter with, for unsuffixed classes, a 12b-1 plan; direct-sold requires a direct letter and no 12b-1 plan. Broker-sold retail and direct-sold retail intersect the channel flag with the CRSP retail-fund flag; the institutional column uses the CRSP institutional-fund flag. Performance rank is based on 36-month five-factor alpha terciles computed each month. Panel A reports the key coefficients per subsample; full regressor sets are in the source file. A letters-only channel flag gives nearly identical results: the broker-sold retail interaction ranges from 0.6492 to 0.8123 across the two fixed-effects schemes, and a within-category-month performance gate gives 0.6036 to 0.7385 (broker-sold), all significant at the 10% level or better. Panel B reports the within-fund dual-class design: share-class flow on a broker-class dummy and its interactions with the portfolio-level top-performance dummy and the top-interacted prospectus uniqueness, absorbing portfolio×month fixed effects. The estimation panel retains roughly one share class per portfolio, so the strict broker-versus-direct universe contains only 6 to 7 portfolios and the loose broker-versus-nonbroker universe 9; with so few entity clusters the reported t-statistics rest on asymptotics that do not apply, and I report Panel B for completeness only. Columns (1)–(2) of the pooled block reproduce the Table 4 specification on the current data vintage. Controls include log TNA, turnover ratio, fund return, and fund age (all lagged). Standard errors are double-clustered by portfolio and month. t-statistics in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

	Pooled		Broker-sold retail		Direct-sold retail		Institutional	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Channel subsamples								
$ProspectusUniq_{t-1}$ × Top 30%	0.5513**	0.6784***	0.7986***	0.6641**	0.9185	0.2944	0.2347	0.7133
	(2.31)	(2.75)	(2.59)	(1.98)	(0.61)	(0.13)	(0.46)	(1.39)
$ProspectusUniq_{t-1}$	−0.1877	−0.3397*	−0.6661***	−0.7026***	−0.5410	−0.2123	0.5443	−0.0992
	(−1.05)	(−1.85)	(−2.92)	(−2.87)	(−0.38)	(−0.10)	(1.43)	(−0.26)
$PortfolioUniq_{t-1}$ × Top 30%	0.1889	0.0384	0.2279	0.1449	1.4090	1.8929	0.6617	−0.0505
	(0.85)	(0.17)	(0.71)	(0.47)	(0.97)	(0.95)	(1.46)	(−0.11)
$PortfolioUniq_{t-1}$	0.0072	0.0862	0.2355	0.3379	−1.1765	−1.2592	−0.7015**	−0.1531
	(0.04)	(0.50)	(0.99)	(1.48)	(−1.08)	(−0.80)	(−2.05)	(−0.42)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Family, Category, Month FE	Yes		Yes		Yes		Yes	
Family×Month, Category FE		Yes		Yes		Yes		Yes
Observations	391,416	352,009	199,711	182,418	14,310	10,385	110,581	92,688
Portfolios	3,901	3,461	2,039	1,840	232	164	1,463	1,233

	Strict: broker vs. direct classes				Loose: broker vs. non-broker classes			
	Composite flag		Letters flag		Composite flag		Letters flag	
Gate:	Month	Cat.×mo.	Month	Cat.×mo.	Month	Cat.×mo.	Month	Cat.×mo.
Panel B: Within-fund dual-class design (portfolio×month FE; 6 to 9 portfolios; completeness only)								
Broker class	0.2398	0.4268	0.8678***	1.1171***	−0.2188	−0.0776	0.1401	0.2964
	(0.37)	(0.58)	(3.45)	(4.52)	(−0.59)	(−0.22)	(0.36)	(0.78)
Broker × Top 30%	3.3321***	4.4976	2.9027***	3.0485**	1.3311	1.3822	2.1644***	3.0467*
	(3.01)	(1.33)	(5.74)	(1.97)	(1.05)	(0.52)	(3.05)	(1.74)
Broker × $ProspectusUniq$ × Top	−8.9678***	−12.8095**	−9.1656***	−11.1119***	−1.8029	−2.5990	−2.8609**	−4.8331**
	(−4.99)	(−2.47)	(−4.86)	(−3.24)	(−0.90)	(−0.70)	(−2.35)	(−1.96)
Observations	971	971	824	824	1,790	1,790	1,831	1,831
Portfolio-months	458	458	412	412	718	718	754	754
Portfolios / classes	7 / 16	7 / 16	6 / 12	6 / 12	9 / 24	9 / 24	9 / 23	9 / 23

Table 13
Prospectus Rewrites and Fund Flows: Stacked Event-Study Estimates

This table reports stacked difference-in-differences estimates of monthly fund flows around prospectus rewrites. A treated event is a portfolio-month in which a filing's text embedding moves by more than 0.05 in cosine distance from the fund's previous filing; controls are near-verbatim refilings (distance below 0.01, annual-update forms only), which net out the filing cycle. Event windows run from $\tau = -6$ to $\tau = +12$ months around each portfolio's own filing month, with $\tau = -1$ as the reference period. The dependent variable is retail or institutional flow as a percentage of TNA, winsorized at the 1st/99th percentiles within month, on the portfolio-level panel. Regressions include treated and event-time dummies, treated×event-time interactions, lagged performance rank, lagged return, and the change in the fee gap, with portfolio and calendar-month fixed effects; standard errors are double-clustered by filer CIK and calendar month. "Cumulative" sums the treated×τ coefficients over $\tau = 1$ to 12, in percentage points of TNA. "Pure text" restricts to events whose portfolio embedding moved less than the gap-matched median (words changed, holdings did not); "joint" is the complement. The performance-top row fails the joint pre-trend test ($p = 0.027$) and its cumulative estimate continues a pre-existing negative differential trend; I report it without a causal interpretation. Minimum detectable cumulative effects at 95% confidence are approximately 3.7 (main retail), 5.9 (pure text retail), and 7.4 (performance top) percentage points of TNA, against a control-group retail flow of −0.43 percent per month (s.d. 6.30). Panel B reports the treated×τ profiles for the three headline specifications. t-statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Specification	Flow DV	Events	$\beta_{\tau=0}$	Cumulative $\tau = 1..12$	95% CI	Pre-trend p
Panel A: Cumulative abnormal flows						
Main (all big rewrites)	Retail	1,884	−0.128 (−0.61)	0.408 (0.21)	[−3.34, 4.15]	0.280
Main (all big rewrites)	Institutional	1,971	0.024 (0.07)	8.231** (1.99)	[0.11, 16.35]	0.525
Pure text (holdings unchanged)	Retail	559	−0.072 (−0.23)	−4.334 (−1.45)	[−10.20, 1.53]	0.300
Pure text (holdings unchanged)	Institutional	575	0.317 (0.76)	6.041 (1.47)	[−1.99, 14.07]	0.054
Joint text and holdings change	Retail	797	−0.408 (−1.49)	−0.541 (−0.19)	[−6.03, 4.95]	0.479
Uniqueness-raising rewrites	Retail	726	−0.348 (−1.27)	−1.344 (−0.49)	[−6.72, 4.03]	0.677
Uniqueness-lowering rewrites	Retail	639	−0.205 (−0.58)	−3.519 (−1.02)	[−10.25, 3.22]	0.204
Performance top 30% at $\tau = -1$	Retail	466	−0.645 (−1.86)	−9.816*** (−2.59)	[−17.23, −2.40]	0.027
Performance bottom 70%	Retail	1,227	0.201 (0.85)	3.347 (1.43)	[−1.22, 7.92]	0.874
Huge rewrites (distance > 0.10)	Retail	983	0.132 (0.51)	3.519 (1.62)	[−0.74, 7.78]	0.284
Excluding 2020Q1–Q2	Retail	1,773	−0.155 (−0.72)	−0.132 (−0.07)	[−3.98, 3.72]	0.471
Annual updates (485BPOS) only	Retail	1,135	−0.324 (−1.14)	−2.351 (−0.83)	[−7.87, 3.17]	0.555
Supplements (497) only	Retail	749	0.172 (0.65)	4.972* (1.68)	[−0.81, 10.76]	0.777
Clean controls (no big event ±12m)	Retail	1,884	−0.143 (−0.67)	0.371 (0.19)	[−3.44, 4.18]	0.139

	Main, Retail	Main, Institutional	Pure text, Retail
Panel B: Treated × event-time profiles (reference $\tau = -1$)			
$\tau = -6$	−0.350 (−1.45)	0.453 (1.16)	−0.429 (−0.97)
$\tau = -5$	−0.140 (−0.67)	0.314 (0.90)	−0.346 (−0.94)
$\tau = -4$	−0.314 (−1.34)	0.394 (1.17)	−0.594* (−1.71)
$\tau = -3$	−0.288 (−1.51)	0.006 (0.02)	−0.670* (−1.91)
$\tau = -2$	−0.398* (−1.84)	−0.017 (−0.05)	−0.694** (−2.14)
$\tau = 0$	−0.128 (−0.61)	0.024 (0.07)	−0.072 (−0.23)
$\tau = +1$	−0.180 (−1.29)	0.376 (0.95)	−0.178 (−0.61)
$\tau = +2$	−0.096 (−0.43)	0.409 (0.91)	0.051 (0.17)
$\tau = +3$	0.079 (0.34)	0.829** (2.37)	−0.151 (−0.42)
$\tau = +4$	−0.274 (−1.07)	0.685* (1.79)	−0.900* (−1.93)
$\tau = +5$	0.254 (1.05)	0.790** (2.25)	−0.300 (−0.75)
$\tau = +6$	−0.087 (−0.28)	0.553 (1.42)	−0.276 (−0.70)
$\tau = +7$	0.535** (2.19)	0.820** (2.21)	0.508 (1.14)
$\tau = +8$	0.159 (0.80)	0.590 (1.60)	−0.347 (−1.15)
$\tau = +9$	−0.235 (−1.03)	0.871** (2.25)	−1.236*** (−3.12)
$\tau = +10$	0.055 (0.30)	0.860** (2.04)	−0.524 (−1.58)
$\tau = +11$	0.137 (0.56)	0.691** (1.99)	−0.497 (−1.47)
$\tau = +12$	0.062 (0.22)	0.758* (1.71)	−0.485 (−1.23)
Observations	538,726	568,787	516,606

Table 14
Fund Name Distinctiveness and the Narrative Flow Channel

This table asks whether the flow response to prospectus uniqueness proxies for the fund's marketing name. I measure name distinctiveness as one minus the mean TF-IDF cosine similarity of the fund's marketing name (the CRSP fund name stripped of trust prefix and class suffix) to its category-month peers, in a word-token and a character 3-gram variant. Name distinctiveness correlates with prospectus uniqueness (Spearman 0.21 in the paper sample; 0.14 pooled within category-month) and barely with portfolio uniqueness (0.06). Panel A adds the lagged name measure (winsorized 1st/99th within month) and its interaction with the top-performance dummy to the flow specification of Table 4, re-estimating the baseline on the identical fixed sample. The month gate ranks 36-month five-factor alpha within month (the Table 4 convention); the category gate ranks within category-month. Panel B compares the rate of big prospectus rewrites (filing-to-filing text-embedding cosine distance above 0.05) between fund-years with and without a marketing-name change, on the at-risk universe of panel fund-years with any filing within one year; the z-statistics treat fund-years as independent and overstate precision. Panel C reports matched differences in annualized flows around name changes: each changer is matched to its nearest category-year neighbor on size and performance ranks among funds with no change within one year; flow detection is at annual resolution, so within-year timing is unobserved. Standard errors in Panel A are double-clustered by portfolio and month; Panel C reports paired t-statistics. t-statistics in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

	Month gate		Category×month gate	
	(1)	(2)	(3)	(4)
Panel A: Does the name surface absorb the flow gate?				
Baseline, identical sample:				
$ProspectusUniq_{t-1}$ × Top 30%	0.5986**	0.7320***	0.6884***	0.6673***
	(2.48)	(2.95)	(2.94)	(2.88)
Adding token name distinctiveness:				
$ProspectusUniq_{t-1}$ × Top 30%	0.5519**	0.6933***	0.7502***	0.7182***
	(2.27)	(2.77)	(3.18)	(3.08)
$NameUniq_{t-1}$ × Top 30%	1.2626*	1.1754	−1.4628*	−1.2488
	(1.67)	(1.49)	(−1.91)	(−1.61)
$NameUniq_{t-1}$	−0.5586	−0.4244	0.0725	0.1554
	(−1.01)	(−0.76)	(0.14)	(0.29)
Adding 3-gram name distinctiveness:				
$ProspectusUniq_{t-1}$ × Top 30%	0.5310**	0.6792***	0.7205***	0.6992***
	(2.19)	(2.72)	(3.07)	(3.02)
$NameUniq_{t-1}$ × Top 30%	1.8804**	1.6623**	−0.8368	−0.8855
	(2.54)	(2.10)	(−1.14)	(−1.18)
Controls	Yes	Yes	Yes	Yes
Family, Category, Month FE	Yes		Yes	
Family×Month, Category FE		Yes		Yes
Observations	388,367	349,171	387,417	348,282

	Window	Changers	Non-changers	Ratio	Two-prop. z
Panel B: Big-rewrite rates around name changes (fund-years)					
Marketing-name changes	±1 year	0.268	0.092	2.91	29.77
Marketing-name changes	Same year	0.172	0.034	5.05	35.31
All genuine changes	±1 year	0.235	0.092	2.56	26.98
All genuine changes	Same year	0.146	0.034	4.30	31.83

	Outcome year	Flow DV	Pairs	Mean diff. (t)
Panel C: Matched annualized flows around name changes (pp of TNA)				
Marketing-name changes	$y+1$	Retail flow	2,005	0.06 (0.06)
Marketing-name changes	$y+1$	Share-class flow	2,625	−1.23 (−1.37)
Marketing-name changes	y (change year)	Retail flow	2,104	−3.07** (−2.44)
Marketing-name changes	y (change year)	Share-class flow	2,749	−4.24*** (−3.99)
All genuine changes	$y+1$	Retail flow	2,346	−0.29 (−0.31)
All genuine changes	$y+1$	Share-class flow	3,253	−0.53 (−0.66)
All genuine changes	y (change year)	Retail flow	2,502	−2.97** (−2.48)
All genuine changes	y (change year)	Share-class flow	3,446	−3.79*** (−3.99)

Table 15
Paid Distribution and Visibility: 12b-1 Status and Morningstar Ratings

This table asks whether the winner-flow response to prospectus uniqueness is an artifact of paid distribution. A share class has a 12b-1 plan when CRSP reports an actual or maximum 12b-1 fee; 56.5% of the estimation sample carries a plan. Panel A splits the flow specification of Table 4 by 12b-1 status under both performance-gate definitions (the within-month gate of Table 4 and a within-category-month variant). Panel B reports the pooled specification with the 12b-1 dummy fully interacted with the measure and gate block; the triple interaction tests the difference between the two subsamples at full sample power. Panel C splits the rated subsample by Morningstar star level. I read Panel C as suggestive: 94.8% of the gated sample is rated, star coverage is highly selected, and the star level is itself correlated with the performance gate. In the fully interacted rating-coverage specification the implied rated-fund slope sits at the pooled baseline (0.48 to 0.60) and the unrated remnant is a thin, unusual 5% of the sample, so I draw no conclusion from the coverage margin. Top denotes the top 30% performance rank under the stated gate. Uniqueness measures enter with a one-month lag. Controls include log TNA, turnover ratio, fund return, and fund age (all lagged). FE(1) denotes family, category, and month fixed effects; FE(2) denotes family×month and category fixed effects. Standard errors are double-clustered by portfolio and month. t-statistics in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

	Month gate				Category×month gate			
	No 12b-1 plan		Has 12b-1 plan		No 12b-1 plan		Has 12b-1 plan	
	FE(1)	FE(2)	FE(1)	FE(2)	FE(1)	FE(2)	FE(1)	FE(2)
Panel A: 12b-1 subsamples								
ProspectusUniq × Top	0.5259	1.0619***	0.5703**	0.4392	0.6006	0.7139*	0.6430**	0.5187*
	(1.34)	(2.65)	(1.99)	(1.42)	(1.57)	(1.74)	(2.20)	(1.86)
ProspectusUniq	0.1670	−0.4634	−0.5285**	−0.4487**	0.0904	−0.1254	−0.4315**	−0.3306*
	(0.53)	(−1.47)	(−2.52)	(−1.99)	(0.29)	(−0.37)	(−2.21)	(−1.70)
PortfolioUniq × Top	−0.0299	−0.1348	0.2853	0.2771	−0.0404	−0.1838	0.4786	0.5620*
	(−0.08)	(−0.36)	(0.99)	(0.97)	(−0.12)	(−0.53)	(1.61)	(1.87)
Observations	170,159	141,527	221,257	200,324	168,833	140,462	218,633	197,883

	Month gate		Category×month gate	
	FE(1)	FE(2)	FE(1)	FE(2)
Panel B: Pooled, 12b-1 fully interacted				
ProspectusUniq × Top	0.3741	0.9458**	0.6062	0.8035**
	(0.95)	(2.39)	(1.59)	(1.98)
ProspectusUniq × Top × 12b-1	0.2373	−0.4486	0.0452	−0.2702
	(0.50)	(−0.91)	(0.10)	(−0.55)
Observations	391,416	352,009	387,466	348,328

	Month gate				Category×month gate			
	Rated 4–5 stars		Rated 1–3 stars		Rated 4–5 stars		Rated 1–3 stars	
	FE(1)	FE(2)	FE(1)	FE(2)	FE(1)	FE(2)	FE(1)	FE(2)
Panel C: Star-level split (suggestive)								
ProspectusUniq × Top	0.8773**	1.0086**	−0.2160	−0.0565	0.7551	0.5930	0.1781	0.2752
	(1.99)	(2.10)	(−0.85)	(−0.19)	(1.59)	(1.14)	(0.69)	(1.01)
PortfolioUniq × Top	−0.9401**	−0.5844	0.1863	−0.0241	−0.3371	−0.4964	0.1866	0.2122
	(−2.18)	(−1.21)	(0.81)	(−0.10)	(−0.76)	(−1.01)	(0.77)	(0.84)
Observations	119,940	98,987	251,107	217,257	118,994	98,138	248,721	215,168

References

Abis, Simona, "Man vs. Machine: Quantitative and Discretionary Equity Management," 2020. Working paper.

_ **and Anton Lines**, "Broken Promises, Competition, and Capital Allocation in the Mutual Fund Industry," *Journal of Financial Economics*, 2024, *162*, 103948.

_ **, Andrea M. Buffa, Apoorva Javadekar, and Anton Lines**, "Learning from Prospectuses," 2022. Working paper.

Amihud, Yakov and Ruslan Goyenko, "Mutual Fund's R^2 as Predictor of Performance," *Review of Financial Studies*, 2013, *26* (3), 667–694.

Anderson, T. W. and Herman Rubin, "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 1949, *20* (1), 46–63.

Araci, Dogu, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," 2019. arXiv preprint arXiv:1908.10063.

Babina, Tania, Anastassia Fedyk, Alex Xi He, and James Hodson, "Artificial Intelligence, Firm Growth, and Product Innovation," *Journal of Financial Economics*, 2024, *151*, 103745.

Bar-Isaac, Heski, Guillermo Caruana, and Vicente Cu nat, "Search, Design, and Market Structure," *American Economic Review*, 2012, *102* (2), 1140–1160.

Barber, Brad M. and Terrance Odean, "Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment," *Quarterly Journal of Economics*, 2001, *116* (1), 261–292.

Berk, Jonathan B. and Jules H. van Binsbergen, "Measuring Skill in the Mutual Fund Industry," *Journal of Financial Economics*, 2015, *118* (1), 1–20.

_ **and Richard C. Green**, "Mutual Fund Flows and Performance in Rational Markets," *Journal of Political Economy*, 2004, *112* (6), 1269–1295.

Black, Fischer and Robert Litterman, "Global Portfolio Optimization," *Financial Analysts Journal*, 1992, *48* (5), 28–43.

Bonelli, Maxime, Anastasia Buyalskaya, and Tianhao Yao, "Financial Product Incentives to Differentiate: Evidence from Mutual Funds," 2024. Working paper.

Brown, David C., Scott Cederburg, and Mitch Towner, "Dominated ETFs," 2024. Working paper, forthcoming in *Critical Finance Review*.

Brown, Keith C., W. V. Harlow, and Laura T. Starks, "Of Tournaments and Temptations: An Analysis of Managerial Incentives in the Mutual Fund Industry," *Journal of Finance*, 1996, *51* (1), 85–110.

Chen, Joseph, Harrison Hong, Ming Huang, and Jeffrey D. Kubik, "Does Fund Size Erode Mutual Fund Performance? The Role of Liquidity and Organization," *American Economic Review*, 2004, *94* (5), 1276–1302.

Chen, Qi, Itay Goldstein, and Wei Jiang, "Payoff Complementarities and Financial Fragility: Evidence from Mutual Fund Outflows," *Journal of Financial Economics*, 2010, *97* (2), 239–262.

Chevalier, Judith and Glenn Ellison, "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, 1997, *105* (6), 1167–1200.

_ **and** _ , "Career Concerns of Mutual Fund Managers," *Quarterly Journal of Economics*, 1999, *114* (2), 389–432.

Choi, James J., David Laibson, and Brigitte C. Madrian, "Why Does the Law of One Price Fail? An Experiment on Index Mutual Funds," *Review of Financial Studies*, 2010, *23* (4), 1405–1432.

Christoffersen, Susan E. K., Richard B. Evans, and David K. Musto, "What Do Consumers' Fund Flows Maximize? Evidence from Their Brokers' Incentives," *Journal of Finance*, 2013, *68* (1), 201–235.

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

Cooper, Michael J., Huseyin Gulen, and P. Raghavendra Rau, "Changing Names with Style: Mutual Fund Name Changes and Their Effects on Fund Flows," *Journal of Finance*, 2005, *60* (6), 2825–2858.

Coval, Joshua and Erik Stafford, "Asset Fire Sales (and Purchases) in Equity Markets," *Journal of Financial Economics*, 2007, *86* (2), 479–512.

Cremers, K. J. Martijn and Antti Petajisto, "How Active Is Your Fund Manager? A New Measure That Predicts Performance," *Review of Financial Studies*, 2009, *22* (9), 3329–3365.

de Prado, Marcos López, "Building Diversified Portfolios That Outperform Out of Sample," *Journal of Portfolio Management*, 2016, *42* (4), 59–69.

deHaan, Ed, Yang Song, Chloe Xie, and Christina Zhu, "Obfuscation in Mutual Funds," *Journal of Accounting and Economics*, 2021, *72* (2–3), 101429.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal, "Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio?," *Review of Financial Studies*, 2009, *22* (5), 1915–1953.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in "Proceedings of NAACL-HLT" 2019, pp. 4171–4186.

Dolphin, Rian, Barry Smyth, and Ruihai Dong, "Stock Embeddings: Learning Distributed Representations for Financial Assets," 2022. arXiv preprint arXiv:2202.08968.

Evans, Richard B. and Rüdiger Fahlenbrach, "Institutional Investors and Mutual Fund Governance: Evidence from Retail-Institutional Fund Twins," *Review of Financial Studies*, 2012, *25* (12), 3530–3571.

Fama, Eugene F. and Kenneth R. French, "Luck versus Skill in the Cross-Section of Mutual Fund Returns," *Journal of Finance*, 2010, *65* (5), 1915–1947.

Feldman, David, Konark Saxena, and Jingrui Xu, "Is the Active Fund Management Industry Concentrated Enough?," *Journal of Financial Economics*, 2020, *136* (1), 23–43.

Gabaix, Xavier, Ralph S. J. Koijen, Robert Richmond, and Motohiro Yogo, "Asset Embeddings," 2025. NBER Working Paper No. 33651.

Gao, Janet, Wenting Ma, and Qiping Xu, "Access to Financing and Racial Pay Gap Inside Firms," 2023. U.S. Census Bureau Working Paper CES-23-36.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny, "Money Doctors," *Journal of Finance*, 2015, *70* (1), 91–114.

Gil-Bazo, Javier and Pablo Ruiz-Verdú, "The Relation between Price and Performance in the Mutual Fund Industry," *Journal of Finance*, 2009, *64* (5), 2153–2183.

Greenwood, Robin and David Thesmar, "Stock Price Fragility," *Journal of Financial Economics*, 2011, *102* (3), 471–490.

Guercio, Diane Del and Jonathan Reuter, "Mutual Fund Performance and the Incentive to Generate Alpha," *Journal of Finance*, 2014, *69* (4), 1673–1704.

_ **and Paula A. Tkac**, "Star Power: The Effect of Morningstar Ratings on Mutual Fund Flow," *Journal of Financial and Quantitative Analysis*, 2008, *43* (4), 907–936.

He, Bianca, Lauren I. Mostrom, and Amir Sufi, "Investing in Customer Capital," 2024. NBER Working Paper No. 33171.

Henderson, Matthew, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil, "Efficient Natural Language Response Suggestion for Smart Reply," 2017. arXiv preprint arXiv:1705.00652.

Hillert, Alexander, Alexandra Niessen-Ruenzi, and Stefan Ruenzi, "Mutual Fund Shareholder Letters: Flows, Performance, and Managerial Behavior," *Management Science*, 2025, *71* (5), 4453–4473.

Hoberg, Gerard and Gordon Phillips, "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, 2016, *124* (5), 1423–1465.

_ , Nitin Kumar, and Nagpurnanand Prabhala, "Mutual Fund Competition, Managerial Skill, and Alpha Persistence," *Review of Financial Studies*, 2018, *31* (5), 1896–1929.

Hortaçsu, Ali and Chad Syverson, "Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds," *Quarterly Journal of Economics*, 2004, *119* (2), 403–456.

Hotelling, Harold, "Stability in Competition," *Economic Journal*, 1929, *39* (153), 41–57.

Huang, Jennifer, Kelsey D. Wei, and Hong Yan, "Participation Costs and the Sensitivity of Fund Flows to Past Performance," *Journal of Finance*, 2007, *62* (3), 1273–1311.

Jegadeesh, Narasimhan and Di Wu, "Word Power: A New Approach for Content Analysis," *Journal of Financial Economics*, 2013, *110* (3), 712–729.

Jehnen, Simon, Javier Villalba-Diez, and Joaquin Ordieres-Mere, "FinTextSim: A Domain-Specific Sentence-Transformer for Extracting Predictive Latent Topics from Financial Disclosures," *Frontiers in Artificial Intelligence*, 2026, *9*, 1752103.

Jiang, Hao and Michela Verardo, "Does Herding Behavior Reveal Skill? An Analysis of Mutual Fund Performance," *Journal of Finance*, 2018, *73* (5), 2229–2269.

Kacperczyk, Marcin and Amit Seru, "Fund Manager Use of Public Information: New Evidence on Managerial Skills," *Journal of Finance*, 2007, *62* (2), 485–528.

_ , Clemens Sialm, and Lu Zheng, "On the Industry Concentration of Actively Managed Equity Mutual Funds," *Journal of Finance*, 2005, *60* (4), 1983–2011.

_ , _ , and _ , "Unobserved Actions of Mutual Funds," *Review of Financial Studies*, 2008, *21* (6), 2379–2416.

Kostovetsky, Leonard and Jerold B. Warner, "Measuring Innovation and Product Differentiation: Evidence from Mutual Funds," *Journal of Finance*, 2020, *75* (2), 779–823.

Lakonishok, Josef, Andrei Shleifer, and Robert W. Vishny, "The Impact of Institutional Trading on Stock Prices," *Journal of Financial Economics*, 1992, *32* (1), 23–43.

Ledoit, Olivier and Michael Wolf, "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices," *Journal of Multivariate Analysis*, 2004, *88* (2), 365–411.

Lee, Jinhyuk, Feiyang Chen, Sahil Dua, Daniel Cer et al., "Gemini Embedding: Generalizable Embeddings from Gemini," 2025. arXiv preprint arXiv:2503.07891.

Li, Kai, Feng Mai, Rui Shen, and Xinyan Yan, "Measuring Corporate Culture Using Machine Learning," *Review of Financial Studies*, 2021, *34* (7), 3265–3315.

Lou, Dong, "A Flow-Based Explanation for Return Predictability," *Review of Financial Studies*, 2012, *25* (12), 3457–3489.

Loughran, Tim and Bill McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance*, 2011, *66* (1), 35–65.

_ **and** _ , "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research*, 2016, *54* (4), 1187–1230.

Markowitz, Harry, "Portfolio Selection," *Journal of Finance*, 1952, *7* (1), 77–91.

Merton, Robert C., "An Analytic Derivation of the Efficient Portfolio Frontier," *Journal of Financial and Quantitative Analysis*, 1972, *7* (4), 1851–1872.

Pedersen, Lasse Heje, "Sharpening the Arithmetic of Active Management," *Financial Analysts Journal*, 2018, *74* (1), 21–36.

Puetz, Alexander and Stefan Ruenzi, "Overconfidence Among Professional Investors: Evidence from Mutual Fund Managers," *Journal of Business Finance & Accounting*, 2011, *38* (5–6), 684–712.

Qian, Edward, "On the Financial Interpretation of Risk Contribution: Risk Budgets Do Add Up," *Journal of Investment Management*, 2006, *4* (4), 1–11.

Reimers, Nils and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in "Proceedings of EMNLP-IJCNLP" 2019, pp. 3982–3992.

Roussanov, Nikolai, Hongxun Ruan, and Yanhao Wei, "Marketing Mutual Funds," *Review of Financial Studies*, 2021, *34* (6), 3045–3094.

Sharpe, William F., "The Arithmetic of Active Management," *Financial Analysts Journal*, 1991, *47* (1), 7–9.

Sirri, Erik R. and Peter Tufano, "Costly Search and Mutual Fund Flows," *Journal of Finance*, 1998, *53* (5), 1589–1622.

Solomon, David H., Eugene Soltes, and Denis Sosyura, "Winners in the Spotlight: Media Coverage of Fund Holdings as a Driver of Flows," *Journal of Financial Economics*, 2014, *113* (1), 53–72.

Son, Matthew G., "It is Better to be Yourself: A Network Analysis on Mutual Fund Distinctiveness and Performance," 2022. Working paper, SSRN 4199469.

Stock, James H. and Motohiro Yogo, "Testing for Weak Instruments in Linear IV Regression," in Donald W. K. Andrews and James H. Stock, eds., *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Cambridge University Press, 2005, pp. 80–108.

Sun, Zheng, Ashley Wang, and Lu Zheng, "The Road Less Traveled: Strategy Distinctiveness and Hedge Fund Performance," *Review of Financial Studies*, 2012, *25* (1), 96–143.

Tetlock, Paul C., "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *Journal of Finance*, 2007, *62* (3), 1139–1168.

Ľuboš Pástor and Robert F. Stambaugh, "On the Size of the Active Management Industry," *Journal of Political Economy*, 2012, *120* (4), 740–781.

Ľuboš Pástor, Robert F. Stambaugh, and Lucian A. Taylor, "Scale and Skill in Active Management," *Journal of Financial Economics*, 2015, *116* (1), 23–45.

Vozlyublennaia, Nadia and Youchang Wu, "Mutual Funds Apart from the Crowd," 2017. Working paper, SSRN 2769161.

Wahal, Sunil and Albert (Yan) Wang, "Competition among Mutual Funds," *Journal of Financial Economics*, 2011, *99* (1), 40–59.

Wei, Kelsey D., Russ Wermers, and Tong Yao, "Uncommon Value: The Characteristics and Investment Performance of Contrarian Funds," *Management Science*, 2015, *61* (10), 2394–2414.

Wermers, Russ, "Mutual Fund Herding and the Impact on Stock Prices," *Journal of Finance*, 1999, *54* (2), 581–622.

Appendix A Formal Apparatus and Proof Sketches

This appendix collects notation, the assumptions, the lemma chain with proof sketches, and the formal statements of the predictions in Section 2.6 with proof sketches. Full derivations follow the standard implicit-function and contraction-mapping arguments and are available on request.

A.1 Notation and Assumptions

Table A.1 summarizes notation. I maintain the following assumptions. **(A1) Match-shock rotation.** For each segment $F_{s_T,j}$ is twice differentiable, log-concave, and mean-preserving-rotating in s_T around a fixed point, with $F_{s_T,I}$ second-order stochastically dominating $F_{s_T,R}$. The exogenous bounds on fees and valuations confine every attainable indifference threshold to an operating range above the rotation point, on which rotation lowers the hazard rate uniformly. **(A2) Parameter bounds.** Fees lie in $[p^{\text{lo}}, p^{\text{hi}}]$, densities and their derivatives are bounded, CE is strictly concave in each fee, and the unconstrained s_P^* lies below one. **(A3) Enforcement.** Institutions observe s_P, a fraction q_I enforces the two-sided discount (5), the manager bears c_L, and Condition AE holds, $\psi^+ > \psi^- \geq 0$. **(A4) Risk-aversion dominance.** $\tilde{\gamma} > 0$ on the maintained region. **(A5) Scale.** $\mu_I \bar{f} \mu_\alpha (\Gamma_0 + \Gamma_1) < 1$, and the capacity-ordering bound makes $\tilde{\theta}$ strictly increasing in θ. **(A6) Text-cost heterogeneity.** $\kappa_{T,i} \in \{\kappa_T^L, \kappa_T^H\}$ with mass π_H on the high type and a templating bound under which a templated fund's narrative never exceeds the low cutoff ℓ_T used in Proposition 1.

Table A.1
Notation summary.

Symbol	Definition
$s_T, s_P \in [0,1]$	Prospectus and portfolio uniqueness; Stage-1 and Stage-2 choices.
$\theta_i \in [\underline{\theta}, \bar{\theta}]$	Fundamental skill, $\underline{\theta} < 0 < \bar{\theta}$ (fixed, unobserved).

Symbol	Definition
$\theta = \mathbb{E}[\theta_i \mid \mathcal{I}]$	Common posterior on skill at the decision time.
$\Sigma = \mathrm{Var}[\theta_i \mid \mathcal{I}]$	Posterior skill variance; declines with the information history.
$r;\ \alpha$	Realized risk-adjusted return, eq. (1); alpha $\alpha = \mathbb{E}[r \mid \theta, s_P, a]$, the expected excess return, eq. (2).
$\tilde{\theta} = \theta - (\Gamma_1/\lambda_\alpha)a^*$	Capacity-adjusted posterior.
a, a^*	First-period assets and their fixed point, $a = D_1 = D_R + D_I$.
$p_R, p_I;\ \bar{p}$	Share-class fees; segment-mass-weighted fee.
$\tau_j = p_j + \delta\bar{p}$	Segment-j effective revenue weight.
$\lambda_\alpha; \Gamma(s_P) = \Gamma_0 + \Gamma_1 s_P; \rho$	Skill complementarity; capacity drag; variance amplification.
$\gamma_M; \Phi, \chi, \chi_2; \tilde{\gamma}$	Risk aversion; flow schedule and its slope and convexity; net variance coefficient (7).
$\kappa_P; \kappa_T^L, \kappa_T^H, \pi_H; \kappa_L$	Portfolio cost; text costs and templating mass; legal cost.
$F_{s_T,j}, f_{s_T,j}, h_{s_T,j}$	Match-shock CDF, density, hazard for segment j.
$\varrho; \varphi_c; \omega; q; \mu_\alpha$	Past performance; channel salience; narrative value; quality weight; alpha weight.
$q_I, q_R; \psi^+, \psi^-$	Enforcement intensities; over-promising and under-promising penalties.
$Q^+ = \tau_I\mu_I\bar{f}_I q_I\psi^+ + \kappa_L;$ $Q^- = \tau_I\mu_I\bar{f}_I q_I\psi^-$	Composite over-promising and under-promising pulls.
$D_0 = \kappa_P + \tilde{\gamma}\rho\sigma_\xi^2;$ $D = D_0 + Q^+;$ $D^- = D_0 + Q^-$	Stage-2 denominators.
$\bar{\eta}_\alpha = \delta\bar{p}\chi + \tau_I\mu_I\mu_\alpha\bar{f}_I$	Alpha-pricing pass-through.
$\zeta_0 = \bar{\eta}_\alpha\lambda_\alpha/2D_0;$ $\zeta = \bar{\eta}_\alpha\lambda_\alpha/2D;\ \nu = Q^+/D$	No-penalty, over-promising skill and gap-closing coefficients.

Symbol	Definition
$\zeta^- = \bar{\eta}_\alpha \lambda_\alpha / 2D^-$; $\nu^- = Q^-/D^-$	Under-promising coefficients, with $\nu > \nu^-$ under AE.

A.2 Scale and Stage-2

Lemma 1 (Scale fixed point)**.** *Under (A1)–(A5) the map $a \mapsto D_R(a) + D_I(a)$ is a continuous, strictly decreasing self-map of $[0,1]$, so a^* exists and is unique; it is differentiable with $\partial a^*/\partial p_j < 0$ and $da^*/d\theta > 0$ bounded under (A5).*

Sketch. Institutional valuation falls in a through the capacity term $-\mu_\alpha \Gamma(s_P) a$, so demand is strictly decreasing in a; a unique fixed point follows from monotonicity, and the comparative statics from the implicit-function theorem with the contraction bound in the denominator. □

Lemma 2 (Stage-2 policy)**.** *Under (A1)–(A6) and a joint dominance condition bounding the demand-side curvature below the production-and-variance curvature $2D_0$, the Stage-2 optimum is unique and given by (8), with $D_0 = \kappa_P + \tilde{\gamma}\rho\sigma_\xi^2$, $D = D_0 + Q^+$, $D^- = D_0 + Q^-$, and the coefficients in Table A.1. The corner $s_P^* = 0$ obtains exactly when $\bar{\eta}_\alpha \lambda_\alpha \tilde{\theta} + 2Q^+ s_T \leq 0$. The policy is continuous: each branch equals s_T at the kink $\tilde{\theta} = s_T/\zeta_0$ because $D = D_0 + Q^+$ and $D^- = D_0 + Q^-$, and the over-promising branch reaches zero at the corner threshold.*

Derivation. Fix s_T, the fees, and the scale a^*, and collect the s_P-dependent terms of the certainty equivalent (6). Expected alpha enters the objective through two channels, the flow term $\delta\bar{p}\chi\,\mathbb{E}[\alpha]$ in expected profit and the institutional valuation $\mu_\alpha \mathbb{E}[\alpha]$ that raises demand D_I; together they give the alpha-pricing pass-through $\bar{\eta}_\alpha = \delta\bar{p}\chi + \tau_I \mu_I \mu_\alpha \bar{f}_I$, and since $\partial \mathbb{E}[\alpha]/\partial s_P = \lambda_\alpha \tilde{\theta}$ the marginal alpha benefit of differentiation is $\bar{\eta}_\alpha \lambda_\alpha \tilde{\theta}$. The production cost $\kappa_P s_P^2$ and the variance penalty $\tilde{\gamma}\rho\sigma_\xi^2 s_P^2$ (from $-\tilde{\gamma}\mathrm{Var}[\alpha]$ with $\mathrm{Var}[\alpha] = \sigma_\xi^2(1 + \rho s_P^2)$ under the seasoned-fund approximation $\Sigma \approx 0$) give the joint marginal cost $2D_0 s_P$ with

$D_0 = \kappa_P + \tilde{\gamma}\rho\sigma_\xi^2$. The enforcement discount Ψ_I (5) and the legal cost c_L supply a two-sided, piecewise term. Collecting these, the Stage-2 first-order condition is

$$\frac{\partial\,\mathrm{CE}}{\partial s_P} = \bar{\eta}_\alpha\lambda_\alpha\tilde{\theta} \;-\; 2D_0 s_P \;+\; 2Q^+\,(s_T - s_P)_+ \;-\; 2Q^-\,(s_P - s_T)_+ \;=\; 0, \tag{29}$$

with $Q^+ = \tau_I\mu_I\bar{f}_I q_I\psi^+ + \kappa_L$ (demand discount plus the direct legal cost) and $Q^- = \tau_I\mu_I\bar{f}_I q_I\psi^-$. Under (A4) ($\tilde{\gamma} > 0$) and the joint-dominance condition, the objective is strictly concave on each side of the kink at $s_P = s_T$, where the quadratic penalties have zero slope, so $\partial\mathrm{CE}/\partial s_P$ is continuous and (29) has a unique root. Because the positive-part terms switch off piecewise, the condition is linear in s_P on each branch and solves in closed form:

(i) *Over-promising* ($s_P < s_T$, so $(s_P - s_T)_+ = 0$): (29) reads $\bar{\eta}_\alpha\lambda_\alpha\tilde{\theta} + 2Q^+ s_T = 2(D_0 + Q^+)s_P$, hence $s_P^* = \zeta\,\tilde{\theta} + \nu\,s_T$ with $\zeta = \bar{\eta}_\alpha\lambda_\alpha/(2D)$, $\nu = Q^+/D$, $D = D_0 + Q^+$.

(ii) *Under-promising* ($s_P > s_T$): the drift term gives $\bar{\eta}_\alpha\lambda_\alpha\tilde{\theta} + 2Q^- s_T = 2(D_0 + Q^-)s_P$, hence $s_P^* = \zeta^-\tilde{\theta} + \nu^- s_T$ with $\zeta^- = \bar{\eta}_\alpha\lambda_\alpha/(2D^-)$, $\nu^- = Q^-/D^-$, $D^- = D_0 + Q^-$.

(iii) *Kink* ($s_P = s_T$): both penalties vanish, leaving $\bar{\eta}_\alpha\lambda_\alpha\tilde{\theta} = 2D_0 s_P$, i.e. $s_P^* = \zeta_0\tilde{\theta}$ with $\zeta_0 = \bar{\eta}_\alpha\lambda_\alpha/(2D_0)$; this branch is selected where $\zeta_0\tilde{\theta} = s_T$, i.e. $\tilde{\theta} = s_T/\zeta_0$, the boundary between (i) and (ii).

(iv) *Corner*: branch (i) reaches zero at $\zeta\tilde{\theta} + \nu s_T = 0$, i.e. $\tilde{\theta} = -(\nu/\zeta)s_T$; below this the unconstrained root is negative and the constrained optimum is $s_P^* = 0$, equivalently $\partial\mathrm{CE}/\partial s_P\big|_{s_P=0} = \bar{\eta}_\alpha\lambda_\alpha\tilde{\theta} + 2Q^+ s_T \le 0$.

At the kink the over-promising branch gives $\zeta(s_T/\zeta_0) + \nu s_T = [(D_0 + Q^+)/D]\,s_T = s_T$ and the under-promising branch gives $[(D_0 + Q^-)/D^-]\,s_T = s_T$ (using $\zeta/\zeta_0 = D_0/D$ and $\zeta^-/\zeta_0 = D_0/D^-$), so the three branches join continuously and (8) follows. Setting $\psi^- = \kappa_L = \Gamma_1 = 0$ and $\underline{\theta} = 0$ gives $Q^- = 0$, $\zeta^- = \zeta_0$, $\nu^- = 0$, and $\tilde{\theta} = \theta$, collapsing (8) to the two-regime frictionless policy. □

A.3 Stage-1 and the Fee FOCs

Lemma 3 (Fee responses and capacity pricing)**.** *Under (A1)–(A5) each fee FOC has a unique interior solution admitting the inverse-hazard analogy* $p_j^* \sim 1/h_{s_T,j}$*. Rotation signs* $\partial p_j^*/\partial s_T > 0$ *on the operating range. The marginal revenue of a fee increase is augmented by a capacity-pricing term proportional to* $\Gamma(s_P)\,|\partial a^*/\partial p_j|$*, raising the optimal markup where* Γ *is large. The institutional FOC carries the enforcement penalty* $-2q_I\psi^+\left(s_T - s_P^*\right)_+ \tau_I \mu_I f_{s_T,I}$ *in the over-promising region; the legal cost* κ_L *does not enter any fee FOC.*

Sketch. Differentiate CE in p_j, apply the implicit-function theorem under a diagonal-dominance condition on the two-fee Jacobian whose off-diagonal entries each carry a factor $\delta\mu_R\mu_I$, $\gamma_M\delta^2\mu_R\mu_I$, or $\mu_R\mu_I\bar{f}\mu_\alpha\Gamma$, and read the signs from the scalar approximation. □

Lemma 4 (Corner suppression)**.** *Under (A1)–(A6) and an enforcement-dominance condition under which the over-promising pull at the corner* $s_T = 1$ *exceeds the rotation and salience gains, there is* $\tilde{\theta}_\ell > 0$ *such that* $s_T^* < 1$ *for all funds with* $\tilde{\theta} < \tilde{\theta}_\ell$ *of either text-cost type.*

Sketch. At $s_T = 1$ a low-skill manager has a large gap, so the enforcement term in the narrative FOC dominates the bounded rotation and salience terms; continuity in θ extends the strict inequality to a neighborhood. □

A.4 Propositions and Proof Sketches

The predictions in Section 2.6 follow from the propositions below.

Proposition 1 (Relative-mass asymmetry)**.** *Under (A1)–(A6), Condition AE, and* $\pi_H > 0$*, there is a non-degenerate set of cutoffs on which the joint mass with high prospectus uniqueness and low portfolio uniqueness is strictly below the joint mass with high portfolio uniqueness and low prospectus uniqueness. Under symmetric enforcement with homogeneous text costs the construction delivers no ordering, so the observed asymmetry is evidence on Condition AE.*

Sketch. By Lemma 4 and the templating bound, no fund clears a sufficiently high narrative cutoff while falling below a low portfolio cutoff, so the high-narrative low-portfolio cell is empty at extreme cutoffs. Templated skilled funds clear a high portfolio cutoff while holding the narrative below the low cutoff, so the high-portfolio low-narrative cell has positive mass. Under $\psi^+ = \psi^-$ and $\kappa_L = 0$ the under-promising pull raises narrative toward holdings just as the over-promising pull raises holdings toward narrative, and the two cells are symmetric. □

Proposition 2 (Skill complementarity in gross alpha)**.** *Conditional on scale, the fund's alpha at the Stage-2 optimum is* $\alpha^*(\theta, s_T, a^*) = \theta - \Gamma_0 a^* + \lambda_\alpha \tilde{\theta}\, s_P^*$*, which is strictly convex in* θ *within each interior branch and has uniqueness gradient* $\partial\alpha/\partial s_P = \lambda_\alpha \tilde{\theta}$*, positive for* $\tilde{\theta} > 0$ *and negative on the window-dressing region.*

Sketch. Substitute (8) into α in equation (2) and use $\partial\tilde{\theta}/\partial\theta = 1$ at fixed a^*; the second derivative in θ is $2\lambda_\alpha\zeta$ or $2\lambda_\alpha\zeta^-$, both positive. □

Proposition 3 (Fee responses and segment contrast)**.** $\partial p_R^*/\partial s_T > 0$ *and* $\partial p_I^*/\partial s_T > 0$ *in the under-promising region. In the over-promising region the enforcement penalty attenuates the institutional response,* $\partial p_I^*/\partial s_T < \partial p_I^*/\partial s_T|_{q_I\psi^+=0}$*. The cross-partial* $\partial^2 p_I^*/\partial s_T \partial s_P$ *carries a strictly positive enforcement component* $2q_I\psi^+\tau_I\mu_I f_{s_T,I}/|G_{I,p_I}|$ *in the over-promising region; the retail cross-partial carries the analogous component scaled by* q_R*, which vanishes at the baseline.*

Sketch. Apply Lemma 3. The enforcement penalty enters G_{I,s_T} with a negative sign in the over-promising region and zero elsewhere, which signs the attenuation; differentiating that penalty in s_P gives the positive enforcement component of the cross-partial. □

Proposition 4 (Retail flow amplification)**.** *The cross-partial of expected retail demand in* s_T *and* ϱ *has a strictly positive direct salience component* $\mu_R f_{s_T,R}\varphi_c$*. Under a salience-dominance condition the total cross-partial is positive and the sign of the retail flow response*

to s_T equals the sign of ϱ. The amplification increases in φ_c and is stronger in broker-sold channels.

Sketch. Differentiate $\mathbb{E}[D_R] = \mu_R[1 - F_{s_T,R}(U - \widehat{v}_R + p_R^*)]$ in s_T and then in ϱ, using $\partial\widehat{v}_R/\partial s_T = \varphi_c\varrho$; the residual price-response and density terms are bounded below the direct component under the condition. □

A.5 Auxiliary Predictions

Prediction 4 (Age interaction in niche risk)**.** *PortfolioUniq predicts tracking error and idiosyncratic volatility, and the slope is steeper for young funds because the posterior-uncertainty term $(1 + \lambda_\alpha s_P)^2\Sigma$ in the realized-return variance declines with the information history.*

Prediction 5 (Risk shifting in the bottom tail)**.** *Where $\tilde{\gamma} \leq 0$, empirically the bottom tercile of past performance, high PortfolioUniq among low-posterior funds arises as a variance gamble and predicts elevated closure, manager turnover, and outflows; the relation is weak or absent in the top terciles.*

Prediction 6 (Enforcement versus templating)**.** *Suppression of the high-narrative low-portfolio cell loads on enforcement intensity, institutional ownership, and consultant coverage, and not on templating proxies. The mass of the high-portfolio low-narrative cell loads on templating proxies, family size and similarity of a fund's text to the family modal prospectus, and not on enforcement intensity.*

Appendix B Portfolio Embedding Architecture Details

The combined input vectors $\{\mathbf{x}_j\}_{j=0}^{M+1}$ pass through $L = 6$ identical transformer encoder layers. Each layer contains a multi-head self-attention sublayer followed by a position-wise feed-forward sublayer, with residual connections and layer normalization after each.

In self-attention, every holding attends to every other holding in the portfolio, allowing the model to learn contextual relationships: e.g., that a large AAPL position in the presence

of NVDA and AVGO signals a technology-focused strategy, while AAPL alongside JNJ and PG signals a diversified blue-chip approach. A forward-pass experiment verifies this mechanism directly: holding the anchor position's identity, weight, and sequence position fixed, switching the surrounding names from a technology cohort to a defensive cohort moves the anchor's contextualized representation by cosine distance 0.29 to 0.46, ten to twenty times the distance between two independent random-context draws (Appendix E). The multi-head structure ($H = 8$ heads, each operating on $K/H = 32$ dimensions) allows parallel attention to different types of inter-holding relationships. Formally, for attention head h in layer ℓ:

$$\text{Attn}^{(\ell,h)}(\mathbf{X}) = \text{softmax}\left(\frac{\mathbf{Q}^{(\ell,h)}(\mathbf{K}^{(\ell,h)})^\top}{\sqrt{d_h}}\right)\mathbf{V}^{(\ell,h)} \tag{30}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V} \in \mathbb{R}^{(M+2)\times d_h}$ are learned linear projections of the input and $d_h = K/H = 32$. The outputs of all heads are concatenated and projected back to $\mathbb{R}^K$.

Each layer also applies a two-layer feed-forward network with intermediate expansion to $4K = 1{,}024$ dimensions and GELU activation, providing nonlinear transformation capacity. A padding mask ensures the model ignores empty positions in portfolios with fewer than 62 holdings.

Table B.1
Architecture Hyperparameters

Parameter	Value	Description
K	256	Embedding dimension
L	6	Transformer layers
H	8	Attention heads
d_h	32	Per-head dimension (K/H)
d_{ff}	1,024	Feed-forward intermediate dim ($4K$)
M	62	Maximum holdings per sequence
Dropout	0.1	Applied after embeddings and within layers
V	∼7,500	Vocabulary size (unique firms + special tokens)

Appendix C Weight Projection MLP Details

The two-layer MLP used for weight projection is set as follows:

$$f_w(\tilde{w}) = \mathbf{W}_2 \cdot \mathrm{GELU}(\mathbf{W}_1 \tilde{w} + \mathbf{b}_1) + \mathbf{b}_2 \tag{31}$$

where $\mathrm{GELU}(x) = x\,\Phi(x)$ and $\Phi(x)$ is the cumulative distribution function of the standard Gaussian distribution. Here $\tilde{w} \in \mathbb{R}$ is the standardized active weight of a single holding, $\mathbf{W}_1 \in \mathbb{R}^{K/4 \times 1}$ and $\mathbf{b}_1 \in \mathbb{R}^{K/4}$ are the first-layer parameters that expand the scalar into a $K/4 = 64$-dimensional intermediate representation, and $\mathbf{W}_2 \in \mathbb{R}^{K \times K/4}$ and $\mathbf{b}_2 \in \mathbb{R}^K$ project the result into the full $K = 256$-dimensional embedding space. The output $f_w(\tilde{w}) \in \mathbb{R}^K$ is added element-wise to the token and position embeddings before entering the transformer.

Appendix D Two-Stage Training Procedure

I train the model in two stages: masked language model (MLM) pre-training to learn asset-level representations, followed by contrastive fine-tuning to produce portfolio-level embeddings optimized for cosine similarity comparison.

Stage 1: Masked Language Model Pre-training. Following Devlin et al. (2019), I randomly mask 15% of holdings in each portfolio sequence and train the model to predict the masked securities from the remaining context. Of the masked positions, 80% are replaced with a special [MASK] token, 10% are replaced with a random security from the vocabulary, and 10% are left unchanged. The training objective is cross-entropy over the security vocabulary, computed only at masked positions:

$$\mathcal{L}_{\mathrm{MLM}} = -\sum_{j \in \mathcal{M}} \log P(s_j \mid \mathbf{x}_{\setminus \mathcal{M}}, \{\tilde{w}_k\}_{k=1}^{n_i}) \tag{32}$$

A critical design choice: the active weight $\tilde{w}_j$ is *preserved* even when the security identity at position j is masked. The model thus observes the weight signal and must infer which security occupies that "slot". This forces the model to learn associations between weight magnitudes and security types (e.g., a large positive active weight at the top rank position is more likely to be a high-conviction stock than a market-weight mega-cap).

Stage 2: Contrastive Fine-Tuning. MLM pre-training produces contextualized per-asset representations but does not directly optimize portfolio-level embeddings for similarity comparison. To bridge this gap, I fine-tune using a contrastive learning objective adapted from sentence embedding methods in NLP (Reimers and Gurevych, 2019).

Pair construction. For each portfolio-month with ≥ 4 holdings, I construct a positive pair by splitting holdings into two half-portfolios by rank parity: even-ranked positions (1st, 3rd, 5th, ...) form half A, odd-ranked positions (2nd, 4th, 6th, ...) form half B. Each half is a valid sub-portfolio that should encode the same underlying strategy.

Portfolio embedding. For each half-portfolio, I compute the embedding as the L2-normalized mean of the contextualized asset representations output by the final transformer layer (excluding special tokens):

$$\mathbf{e} = \frac{\sum_{j=1}^{n} \mathbf{h}_j^{(L)}}{\left\| \sum_{j=1}^{n} \mathbf{h}_j^{(L)} \right\|_2} \tag{33}$$

Loss function. The Multiple Negatives Ranking Loss (MNRL; Henderson, Al-Rfou, Strope, Sung, Lukács, Guo, Kumar, Miklos and Kurzweil, 2017) trains the model so that $\mathbf{e}_i^A$ is most similar to $\mathbf{e}_i^B$ (its own other half) rather than to any $\mathbf{e}_j^B$ for $j \neq i$:

$$\mathcal{L}_{\text{MNRL}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\big(\cos(\mathbf{e}_i^A, \mathbf{e}_i^B)/\tau\big)}{\sum_{j=1}^{B} \exp\big(\cos(\mathbf{e}_i^A, \mathbf{e}_j^B)/\tau\big)} \tag{34}$$

where $B = 512$ is the batch size and $\tau = 0.1$ is the temperature parameter. The large batch size is essential: each anchor portfolio must be distinguished from $B - 1 = 511$ negatives, providing a challenging discrimination task. The model achieves a validation

retrieval accuracy of 49.8% (255× above the random baseline of $1/B \approx 0.2\%$), confirming that the embeddings encode meaningful portfolio-level information.

The train/validation split is by portfolio identity (90/10), not by month, ensuring no fund appears in both sets.

Table D.1
Training Hyperparameters

Parameter	Value	Description
Masking rate	15%	Fraction of holdings masked during MLM
B (pre-train)	256	Batch size for MLM
B (fine-tune)	512	Batch size for contrastive learning
τ	0.1	Temperature for MNRL loss
LR (pre-train)	5×10^{-4}	AdamW learning rate
LR (fine-tune)	5×10^{-5}	AdamW learning rate

Appendix E Interpretation and Applications of the Embedding Space

This appendix asks what the learned embeddings mean in economic terms and answers in three parts. First, a portfolio embedding is a position in product space: it identifies the fund's strategy type from holdings alone. Second, distance in the space approximates substitutability, and I validate this against realized return comovement at both the asset and the fund level. Third, the embedding decomposes into a component spanned by standard observable characteristics and a large latent residual, and I measure both. Throughout, I report the validation failures alongside the successes.

E.1 Asset-Level Geometry

Table E.1 collects the asset-level evidence. Panel A shows that every named same-industry pair sits above the all-pairs cosine baseline of 0.062, with the classic pairs four to six stan-

dard deviations above it (Apple–Microsoft 0.664, Exxon–Chevron 0.584, Coca-Cola–PepsiCo 0.498, Pfizer–Merck 0.474, JPMorgan–Bank of America 0.441). Panel B replaces the name-inferred sector labels of Figure 3 with each asset's modal SIC code from the CRSP monthly stock file: a leave-one-out kNN-10 classifier recovers the Fama–French 48 industry at 32.4% accuracy against a 16.5% baseline and the 10-industry label at 48.1% against 37.5%, and within-industry cosine exceeds cross-industry cosine under both schemes.

Two features of the geometry deserve equal prominence. The organizing principle is co-holding breadth: widely co-held mega-caps across industries (JPMorgan–Exxon 0.463, Apple–Johnson & Johnson 0.442) score as high as same-industry bank or tech pairs, NVIDIA–AMD reaches only 0.176, and NVIDIA's nearest neighbors are its growth-portfolio cohort (Tesla, Broadcom, Meta, Netflix) with AMD absent from the top ten. And the industry signal is local: global k-means clusters of the embeddings align with co-holding cohorts (adjusted Rand index with industry 0.02, adjusted mutual information 0.05–0.09), so industry recovery operates through nearest neighborhoods. I accordingly describe the space as organizing assets by the economic role they play in portfolios, with industry as one recovered dimension.

Panel C tests whether distance carries forward-looking content. Over all 1,969,439 pairs of the 2,000 most widely held stocks, embedding cosine predicts the realized 2011–2025 pairwise return correlation with Spearman 0.161, strictly monotonically across twenty cosine bins (mean realized correlation 0.233 in the lowest bin to 0.343 in the highest), within industries (0.202) and across them (0.136), and with incremental R^2 of 0.020 beyond same-industry dummies. The prediction is conditional on data richness: over a 1M random sample of the full universe the rank correlation is -0.038, driven by thinly held stocks (rank beyond 4,000 by number of holders: -0.047) for which the model observes few co-holding signals. Embedding distance measures substitutability where the holdings data identify it.

E.2 What the Portfolio Embedding Encodes

Table E.2 and Figure E.2 report linear probes on 293,833 reported portfolio-months with whole months held out. The embedding recovers construction characteristics (top-10 share $R^2 = 0.66$, KSZ industry concentration 0.63, log holdings count 0.61, active shares 0.58–0.60, health and technology sector weights 0.52–0.54) and return-based quantities external to the training data (idiosyncratic volatility 0.42, SMB beta 0.42, $1 - R^2$ 0.39, HML beta 0.20). Classification probes identify the Morningstar category at 66.8% (top-3: 89.5%) against a 13.8% baseline and flag index funds with AUC 0.92. Figure E.1 visualizes the implied map: category groups occupy coherent regions, the index-fund centroid sits on a detached island, and the three case funds are individually locatable.

Two boundaries on this evidence. Holdings-derived targets share raw inputs with the embedding, so their high R^2 is partly mechanical; the return-based probes carry the external-validation weight. And because the cross-validation blocks months while the same fund appears at different dates, persistence inflates the figures relative to a fund-blocked design. Conversely, the variance budget bounds the observable content from above: the thirty probe targets jointly explain 5.9% of embedding variance out of sample, 8.4% with category dummies added. The measures built from the embedding therefore use standard characteristics plus a large latent component; the probes document that the latent component exists and that part of it is economically meaningful (it predicts comovement and return-based fund properties), while its remainder is not separable from noise by these methods.

E.3 Fund-Pair Validation: Comovement and Flows

Table E.3 and Figure E.3 evaluate the substitutability interpretation at the fund level on 2,146,274 fund pairs from six annual snapshots. Embedding cosine predicts the pair's return correlation over the following 24 months (pooled Spearman 0.230; monotone binscatter both overall and within category) and adds information beyond category, raw holdings overlap, and prospectus-text similarity jointly (partial Spearman 0.087–0.106). A fund's ten embed-

ding nearest neighbors overlap its ten strongest future-comovement partners at 10.2%, above the 8.5% category-random baseline and the 0.5% unconditional baseline. Raw holdings cosine is the stronger raw predictor throughout (pooled Spearman 0.671; neighbor overlap 24.5%), and equal-weighted category peer averages track funds' future returns better than embedding-kNN peer averages (mean R^2 0.920 versus 0.884). The embedding's value at the pair level is incremental information in a 256-dimensional summary that requires no pairwise holdings computation; for raw comovement prediction, holdings overlap remains the benchmark.

Panel D tests substitutability in quantities: if nearby funds are substitutes, a peer's good performance should divert flows from the focal fund, and more so for closer peers. On 12.6 million directed pair-months across eleven snapshots, the similarity-interacted cross-sensitivities are small, positive, and insignificant, with a minimum detectable effect of 15% of the own-performance flow slope. The data rule out flow substitution along embedding distance at any economically substantial fraction of the own-performance effect. The substitutability interpretation of embedding distance rests on return comovement, where the evidence is strong, and receives no support from flows.

E.4 Applications

Table E.4 reports three applications. First, a closet-index detector: a fund's distance to the month's index-fund centroid correlates with active share (Spearman 0.43) and holdings uniqueness (0.59) while disagreeing with active share in informative directions, flagging factor-tilted quasi-index products (DFA US Vector, volatility-weighted 500 funds) as far from the index centroid despite bottom-2% active share, and placing small-cap active funds near a centroid that itself contains small-cap index funds. Index-hugging funds charge 22 bp/yr less on average, and the detector isolates roughly 141 expensive index-huggers per month, quintile-1 distance funds charging above-median fees.

Second, style drift: month-over-month embedding displacement validates against two

external benchmarks in the predicted direction, correlating at Spearman +0.40 with Morningstar's style-consistency metric across 3,743 funds and rising 1.35× in months when a fund switches Morningstar category (Welch $t = 7.2$). Figure E.4 traces the twelve-month cumulative displacement of three flagship funds against the cross-sectional distribution; Fidelity Contrafund combines small monthly steps with large persistent drift (mean twelve-month displacement 0.60 against a cross-sectional median near 0.30).

Third, category crowding: the exact mean pairwise cosine among category members. From its July 2024 trough to September 2025 the aggregate index rises 19% and every large category rises monotonically (Large Growth +44%, Large Value +33%, Large Blend +31%), consistent with the concentration of active portfolios in a narrow set of mega-cap names over this period (Figure E.5). Two qualifications accompany this series: the 2024–2025 level remains below its 2015–2023 mean, so the episode is a rebound within the historical range; and crowding is uncorrelated with subsequent category-average alpha (pooled Pearson −0.012), so I use the index descriptively. A fourth exercise, selecting maximally distant funds for a fund-of-funds, reduces realized pairwise correlation relative to random selection (0.746 versus 0.790) and is dominated by naive one-per-category selection (0.678); embedding distance identifies outlier funds, and spreading across categories diversifies better than maximizing spread in the space.

E.5 Context Dependence of the Representation

The architecture in Appendix B claims that the meaning of a position depends on the portfolio around it. A forward-pass experiment isolates this claim. I construct synthetic 21-name portfolios that hold an anchor stock (Apple, NVIDIA, JPMorgan, Exxon, or Johnson & Johnson) at a fixed weight and sequence position while the twenty surrounding names switch between a technology cohort, a defensive blue-chip cohort, and independent random draws, with weights standardized exactly as in the training pipeline. Switching the context from technology to defensive moves the anchor's final-layer representation by cosine distance 0.29–

0.46, ten to twenty times the 0.01–0.10 distance between two independent random-context draws. The same Apple position means something different next to NVIDIA and Broadcom than next to Johnson & Johnson and Procter & Gamble, and the static asset embedding is context-free by construction (its lookup row matches the published asset-embedding file at cosine distance zero).

The experiment also bounds the weight channel. Raising the anchor's raw active weight from 0.5% to 10% within a fixed context moves the representation by 0.005–0.126, an order of magnitude less than changing the context, and the training pipeline's within-month winsorization caps raw active weights near 3%, so a 3% and a 10% position map to nearly identical inputs. The model distinguishes passive drift from a deliberate overweight up to the winsorization cap, and context identity dominates own-position weight in the learned representation.

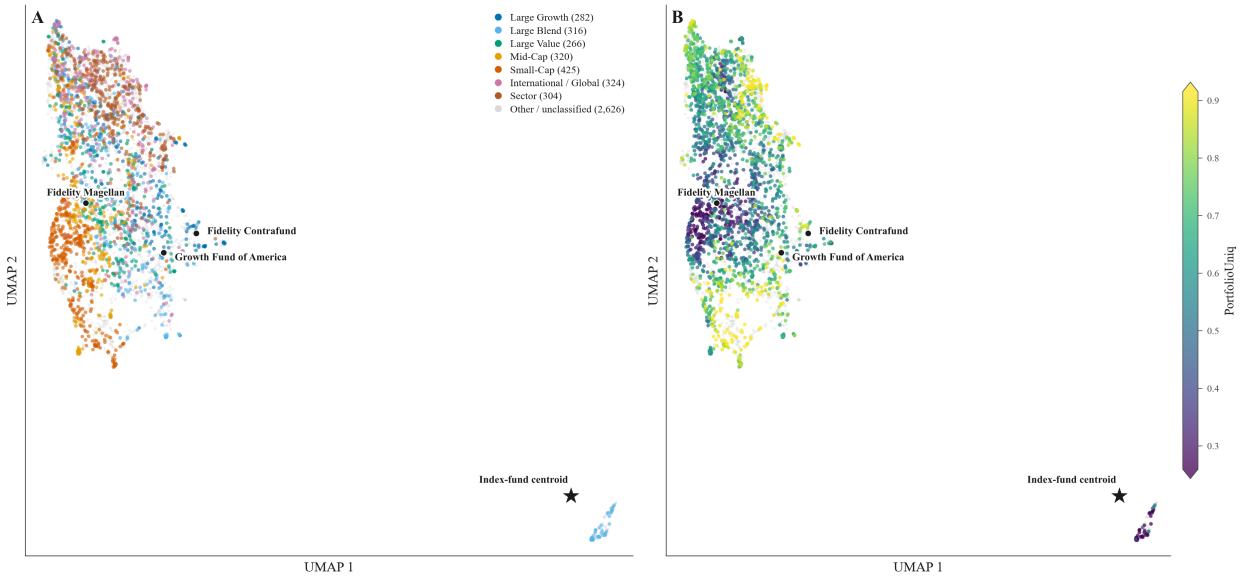


Figure E.1
Portfolio Embedding Map

This figure shows a UMAP projection of 4,863 reported portfolio embeddings in June 2023. Panel A colors funds by Morningstar category group; Panel B colors funds by *PortfolioUniq*. Annotations locate Fidelity Contrafund, Fidelity Magellan, Growth Fund of America, and the centroid of pure index funds, which projects onto a detached island of index trackers away from the active-fund cloud.

Table E.1
Asset Embedding Geometry: Named Pairs, Industry Recovery, and Return Comovement

This table validates the learned asset embedding space ($V = 7{,}554$ assets, $K = 256$). Panel A reports cosine similarities for named stock pairs against the all-pairs baseline (mean 0.062, standard deviation 0.097); z is the distance from the all-pairs mean in standard deviations. Panel B reports industry recovery using each asset's modal SIC code from the CRSP monthly stock file mapped into Fama–French 10 and 48 industry schemes: leave-one-out kNN-10 classification accuracy on L2-normalized embeddings versus the majority-class baseline, and the mean within-industry versus cross-industry pairwise cosine. k-means clustering of the same embeddings does not align with industry (ARI 0.02, AMI 0.05–0.09), so the recovery is a local nearest-neighbor property. Panel C reports the Spearman rank correlation between embedding cosine and realized pairwise return correlation over 2011–2025 (pairwise-complete, at least 60 common months of 180), for all pairs of the 2,000 most widely held stocks and for a 1M random sample of the full universe, with subsets by Fama–French 10 industry and by holding breadth. Mean realized correlation by within-set cosine quintile is shown below each set. In a rank regression on the top-2000 set, cosine adds incremental R^2 of 0.020 beyond same-FF10 and same-SIC2 dummies (rank-cosine coefficient 0.143); in the random sample the cosine slope is negative (-0.038), so predictive power is concentrated where co-holding data are rich.

Panel A: Named asset pairs (cosine similarity)

Same industry (FF48)				Cross industry			
Pair	Industry	Cosine	z	Pair	Industries	Cosine	z
AAPL–MSFT	Comps/BusSv	0.664	6.24	JPM–XOM	Banks/Oil	0.463	4.16
AAPL–GOOG	Comps/BusSv	0.626	5.85	AAPL–JNJ	Comps/Drugs	0.442	3.94
XOM–CVX	Oil	0.584	5.41	AAPL–XOM	Comps/Oil	0.440	3.92
KO–PEP	Soda	0.498	4.52	PFE–JPM	Drugs/Banks	0.400	3.50
PFE–MRK	Drugs	0.474	4.27	WMT–CVX	Rtail/Oil	0.322	2.70
JPM–BAC	Banks	0.441	3.93	NVDA–JNJ	Chips/Drugs	0.166	1.08
T–VZ	Telcm	0.356	3.05	KO–NVDA	Soda/Chips	0.091	0.30
GS–MS	Fin	0.300	2.47				
UNP–CSX	Trans	0.263	2.09				
HD–LOW	Rtail	0.250	1.95				
WMT–TGT	Rtail	0.229	1.73				
DIS–NFLX	Fun	0.215	1.58				
NVDA–AMD	Chips	0.176	1.18				

Panel B: Industry recovery from real SIC labels

Scheme	kNN-10 acc.	Baseline	Ratio	Within cos.	Cross cos.
FF10	0.481	0.375	1.28×	0.075	0.058
FF48	0.324	0.165	1.96×	0.090	0.060
FF10 excl. Other	0.451	0.291	1.55×	0.081	0.056

Panel C: Cosine vs. realized return correlation, 2011–2025

Set	Subset	Pairs	Spearman
Top-2000 most held, all pairs	All	1,969,439	0.161
	Within FF10	379,200	0.202
	Cross FF10	1,590,239	0.136
	Excl. multi-PERMNO trusts	1,877,351	0.161
Random 1M, full universe	All	1,000,000	−0.038
	Both in top 2000 by holders	162,396	0.161
	Both rank 2001–4000	350,530	0.067
	Both rank >4000	487,074	−0.047

Mean realized correlation by cosine quintile (Q1 → Q5):
Top-2000: 0.243, 0.258, 0.269, 0.283, 0.314 (strictly monotone)
Random 1M: 0.189, 0.191, 0.190, 0.189, 0.177

Table E.2
What the Portfolio Embedding Encodes: Out-of-Sample Probes and Variance Budget

This table reports linear probes of the 256-dimensional portfolio embeddings on the reported-status portfolio panel (293,833 portfolio-months, 4,356 portfolios, 2011–2025). Panel A reports pooled out-of-fold R^2 from ridge regressions of each target on the embedding, with five contiguous month blocks as outer cross-validation folds (whole months held out) and the ridge penalty chosen by month-blocked inner cross-validation. Sector weights and factor betas are computed from holdings and from 24-month FFC6 return regressions respectively; the model is trained on holdings only, so the return-based targets are external to the training data. Panel B reports classification probes on 30 sampled months. Panel C reports the share of embedding variance jointly explained by the 30 observables of Panel A (plus 50 Morningstar category dummies), computed across the top-50 embedding principal components (65.4% of total variance) with components 51–256 counted as unexplained. The same fund appears in training and held-out months at different dates, and embeddings are persistent, so Panel A figures are upper bounds relative to a fund-blocked design; holdings-derived targets also share raw inputs with the embedding, so the return-based probes carry the external-validation weight.

Panel A: Ridge probes, pooled out-of-fold R^2

Portfolio construction	R^2	Sector weights and betas	R^2
Top-10 share	0.656	Health weight	0.540
Industry concentration (KSZ)	0.635	HiTec weight	0.522
Log # holdings	0.615	Utilities weight	0.422
Active share (category)	0.604	Energy weight	0.336
Active share (market)	0.577	Manufacturing weight	0.303
Idiosyncratic vol. (FFC6, 24m)	0.420	Shops weight	0.298
$1 - R^2$ (FFC6, 24m)	0.387	SMB beta	0.417
HHI	0.331	HML beta	0.204
Expense ratio	0.243	Market beta	0.086
Tracking error (category, 24m)	0.178	RMW / CMA / MOM beta	0.03–0.05
Log TNA	0.105	Return uniqueness (24m)	0.067
Turnover ratio	0.052	Fund age	0.058
		Other / NoDur / Telcm / Durbl weight	0.50 / 0.20 / 0.14 / 0.08

Panel B: Classification probes (30 sampled months, N = 47,969)

Probe	Metric	Value	Baseline
Morningstar category (57 classes)	Accuracy	0.668	0.138
Morningstar category	Top-3 accuracy	0.895	0.138
Index-fund flag	AUC	0.922	0.500

Panel C: Variance budget (N = 245,206 complete-case)

Feature set	Share of total variance	Latent share
30 observables jointly	5.9%	94.1%
30 observables + category dummies	8.4%	91.6%

Table E.3
Fund-Pair Validation: Embedding Distance, Future Comovement, and Flows

This table asks whether distance in the portfolio embedding space measures economic substitutability between funds. The sample pools six June snapshots (2013–2023): all within-category fund pairs plus an equal-size cross-category sample, 2,146,274 pairs (2,013,990 with prospectus-text coverage). The outcome is the pairwise correlation of the two funds' returns over the following 24 months. Panel A reports pooled Spearman rank correlations of each predictor with the outcome; hold_cos is the raw holdings-weight cosine, text_cos the mean-centered prospectus-embedding cosine, emb_cos the portfolio-embedding cosine. Panel B reports R^2 from regressions of the outcome's within-snapshot percentile rank on predictor ranks with snapshot fixed effects, and the incremental R^2 and partial Spearman of emb_cos over the listed baselines. Pair observations share funds, so conventional standard errors overstate precision and I report magnitudes only. Panel C reports the pooled share of a fund's ten embedding nearest neighbors that fall in its top ten future-comovement neighbors, against analytic baselines, and the mean R^2 of each fund's future returns on equal-weighted peer averages. Panel D reports the flow-substitutability regression on 12,615,288 directed pair-months from eleven June snapshots: fund i's next-month flow (% of TNA) on peer j's current-month return interacted with embedding-similarity terciles, own performance as a control, fund and month fixed effects, standard errors clustered by fund. The substitution prediction is a more negative cross-sensitivity in tercile 3; the estimated interactions are small, positive, and insignificant, and the minimum detectable effect (2.8× SE) on the tercile-3 interaction is 15% of the own-performance flow slope, so the data rule out cross-flow sensitivities above roughly one-seventh of the own effect. t-statistics in parentheses.

Panel A: Spearman with future 24m correlation

Predictor	Pooled	Within-cat.
Holdings cosine	0.671	0.493
Same-category dummy	0.605	–
Text cosine	0.401	0.155
Embedding cosine	0.230	0.153

Panel B: Rank regressions

Specification	R^2
Embedding only	0.054
Holdings only	0.473
Category only	0.387
Category + holdings	0.531
+ embedding	0.536
Category + holdings + text	0.544
+ embedding	0.547

Incremental value of embedding cosine: $\Delta R^2 = 0.0053$ over category + holdings (partial Spearman 0.106); $\Delta R^2 = 0.0034$ over category + holdings + text (partial 0.087).

Panel C: Neighbor overlap and peer-group quality (pooled, six snapshots)

	Embedding	Holdings	Category-random	Any-random
NN-overlap@10 with future-comovement top-10	0.102	0.245	0.085	0.005
	Category peers	kNN-20 peers		
Mean R^2 on EW peer average (future 24m)	0.920	0.884		
Median	0.950	0.929		

Panel D: Flow substitutability (N = 12,615,288; 3,630 funds, 132 months)

	Coefficient	MDE (2.8× SE)
Peer performance$_j$	−0.0030	
	(−2.63)	
Peer performance$_j$ × similarity T2	0.0006	0.0016
	(1.14)	
Peer performance$_j$ × similarity T3	0.0008	0.0020
	(1.16)	
Own performance$_i$	0.0134	
	(5.60)	

Table E.4
Applications of the Embedding Space: Closet Indexing, Style Drift, and Crowding

This table reports three applications built directly from the portfolio embeddings. Panel A defines a fund's index distance as one minus the cosine between its embedding and the month's centroid of pure index funds (CRSP index flag D; 342 index portfolios, mean 234 per month) and evaluates it on 409,152 active portfolio-months; fee quintiles are formed within month. Index-huggers are quintile-1 fund-months with an above-median fee. Panel B validates month-over-month embedding displacement (reported-status months only) against Morningstar's style-consistency metric, whose sign convention I verify in the data (a higher metric indicates lower consistency, so a positive rank correlation is the predicted direction), and against category-switch months. Panel C reports the category-crowding index, the exact mean pairwise cosine among category members, from its July 2024 trough to September 2025, the last month passing a pre-stated sample-reliability rule that removes filing-lag artifacts at the series ends; the 2024–2025 reliable-month mean (0.184) remains below the 2015–2023 mean (0.202), so the rise is a rebound. Crowding is uncorrelated with next-period category-average alpha (pooled Pearson −0.012). Panel D reports an illustrative fund-of-funds exercise: ten funds selected per method at each June snapshot 2013–2022, equal-weighted over the next 24 months, 200 draws per method; embedding-based greedy max-min selection lowers the average pairwise correlation relative to random selection and one-per-category selection lowers it further.

Panel A: Closet-index detector (active funds, N = 409,152)

Correlation of index distance with: active share (market) Spearman 0.427, Pearson 0.526; holdings uniqueness Spearman 0.592.

Fee by distance quintile (%/yr)	Q1	Q2	Q3	Q4	Q5	Q5−Q1
Mean	0.980	1.120	1.168	1.193	1.197	+0.217
Median	0.98	1.10	1.15	1.17	1.16	

Expensive index-huggers: 24,911 fund-months, 1,342 portfolios, mean 141 funds/month.

Panel B: Style-drift validation

Test	Statistic	p	N
Mean displacement vs. MS style consistency (Spearman)	0.399	4.4×10^{-143}	3,743
12m-cumulative version (Spearman)	0.122	5.3×10^{-13}	3,470
Displacement in category-switch months	0.127		1,421
Displacement in non-switch months	0.094		295,318
Switch vs. non-switch (Welch t; ratio 1.35×)	7.23	7.6×10^{-13}	296,739

Panel C: Category crowding, 2024 trough to September 2025

Series	Trough (month)	2025-09	Change
Aggregate (member-weighted)	0.170 (2024-07)	0.203	+19.4%
US Fund Large Growth	0.081 (2024-07)	0.116	+43.8%
US Fund Large Value	0.200 (2024-04)	0.266	+33.1%
US Fund Large Blend	0.117 (2024-07)	0.153	+30.8%
US Fund Small Growth	0.251 (2024-06)	0.315	+25.4%
US Fund Small Blend	0.143 (2024-06)	0.157	+10.4%
US Fund Mid-Cap Growth	0.289 (2024-09)	0.317	+9.6%

Panel D: Fund-of-funds selection demo (illustrative)

Method	Avg. pairwise corr.	Ann. vol.	Ann. Sharpe
Uniform random	0.790	0.153	0.70
One per random category	0.678	0.142	0.55
Greedy max-min embedding distance	0.746	0.157	0.66

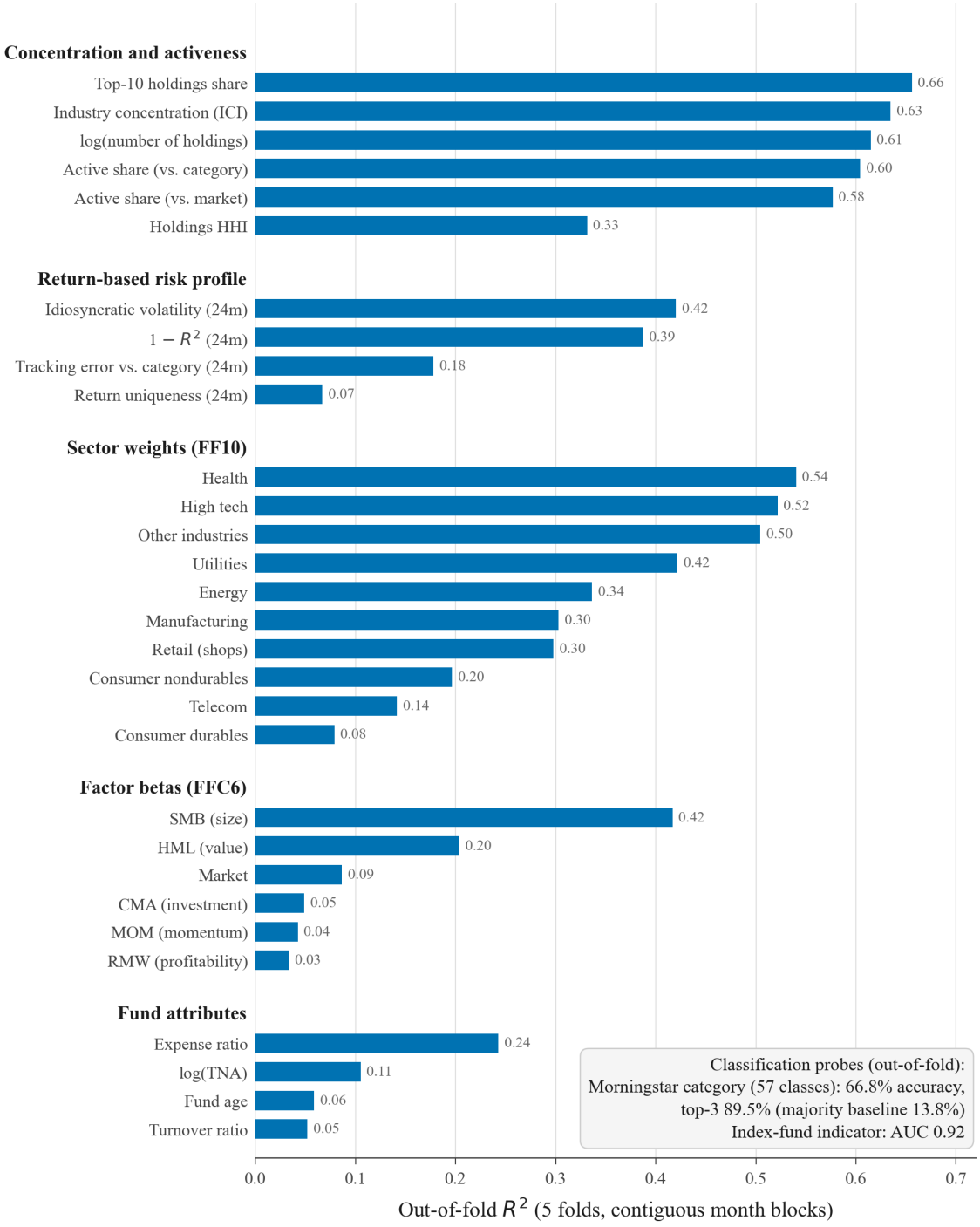


Figure E.2
Out-of-Sample Probes of the Portfolio Embedding

This figure plots pooled out-of-fold R^2 from ridge probes of the 256-dimensional portfolio embedding on portfolio characteristics, sector weights, and return-based quantities, with whole months held out (Table E.2, Panel A). Return-based targets are external to the holdings-only training data.

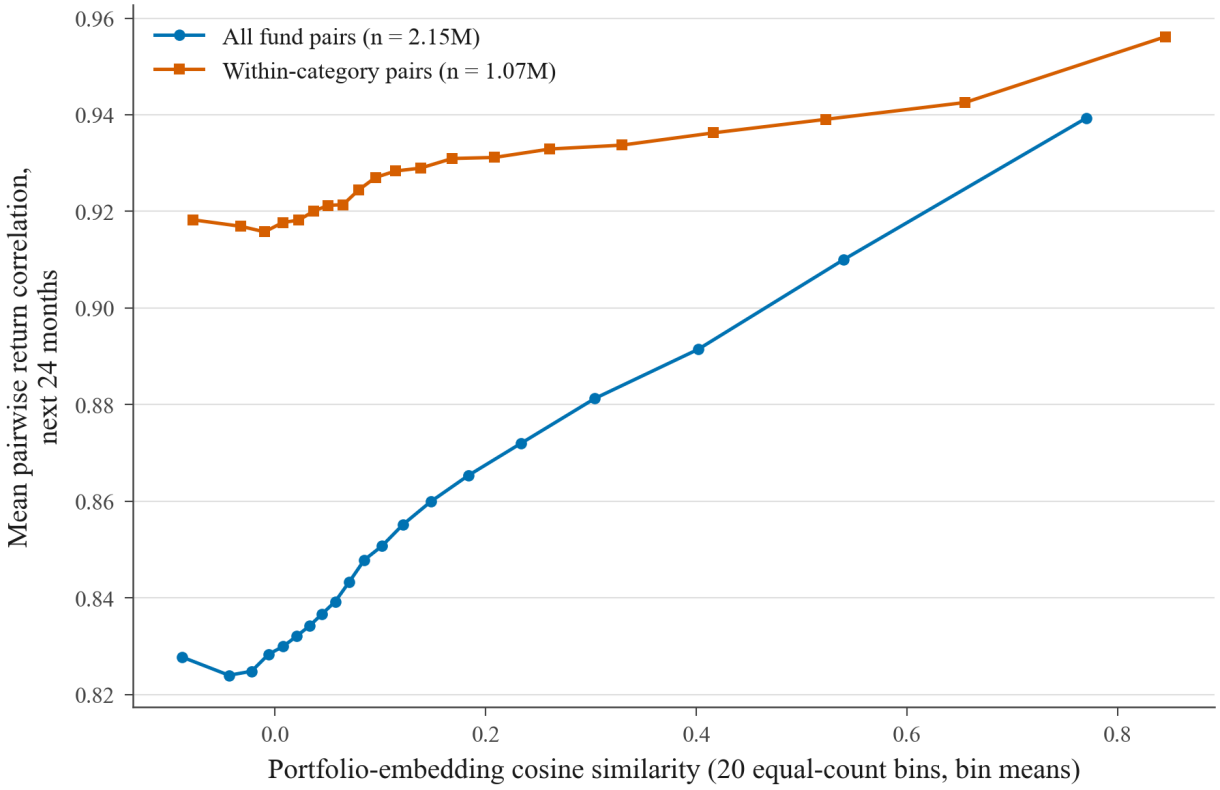


Figure E.3
Embedding Distance and Future Fund Comovement

This figure bins 2.15 million fund pairs from six annual snapshots into twenty groups by portfolio-embedding cosine and plots the mean pairwise return correlation over the following 24 months, for all pairs and for within-category pairs. Both profiles rise monotonically.

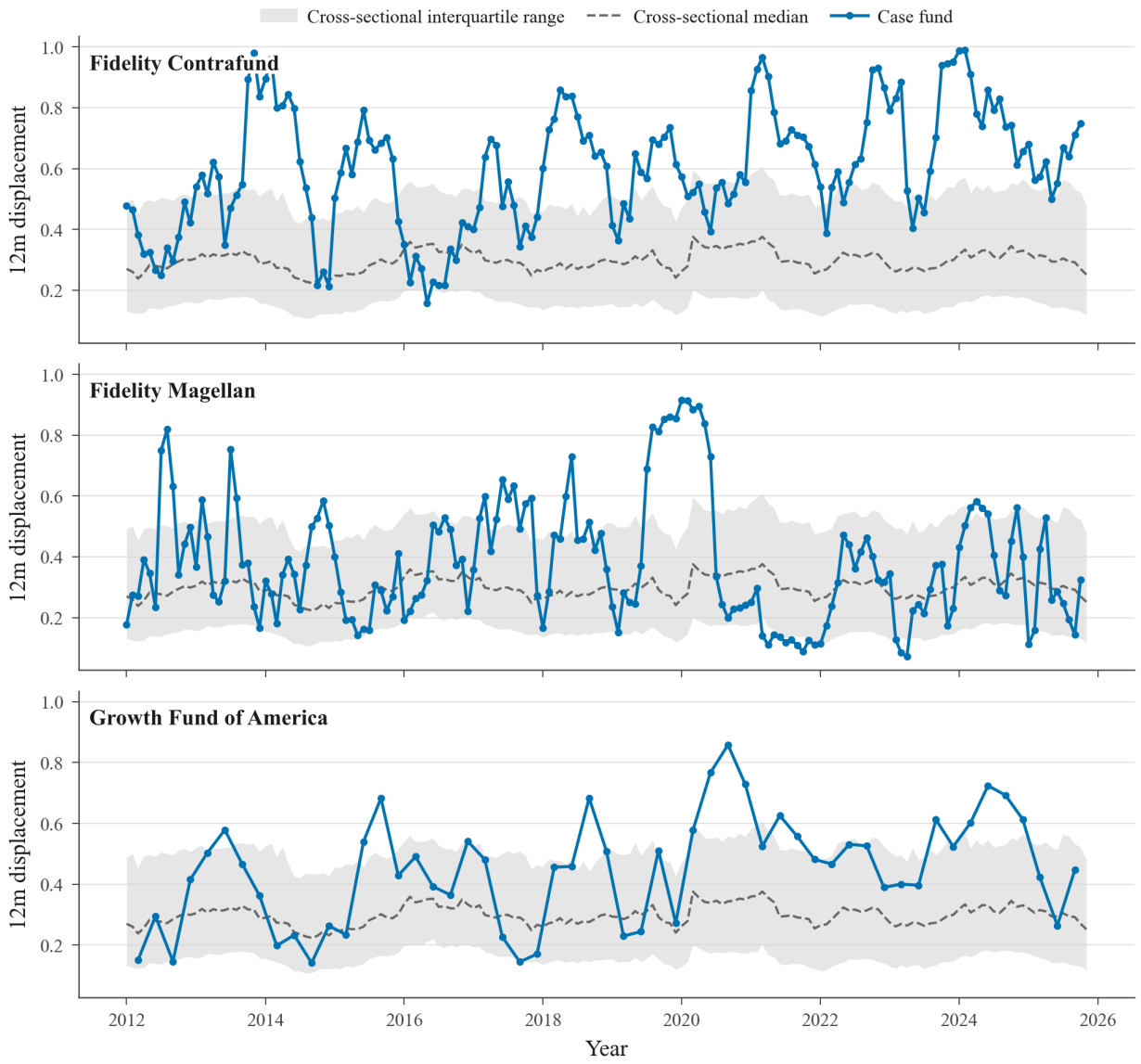


Figure E.4
Style-Drift Trajectories

This figure plots twelve-month cumulative embedding displacement for Fidelity Contrafund, Fidelity Magellan, and Growth Fund of America against the cross-sectional median and interquartile band (511,132 fund-month observations, reported holdings only). Fund-level mean displacement correlates with Morningstar's style-consistency metric at Spearman +0.40 and rises 1.35× in category-switch months.

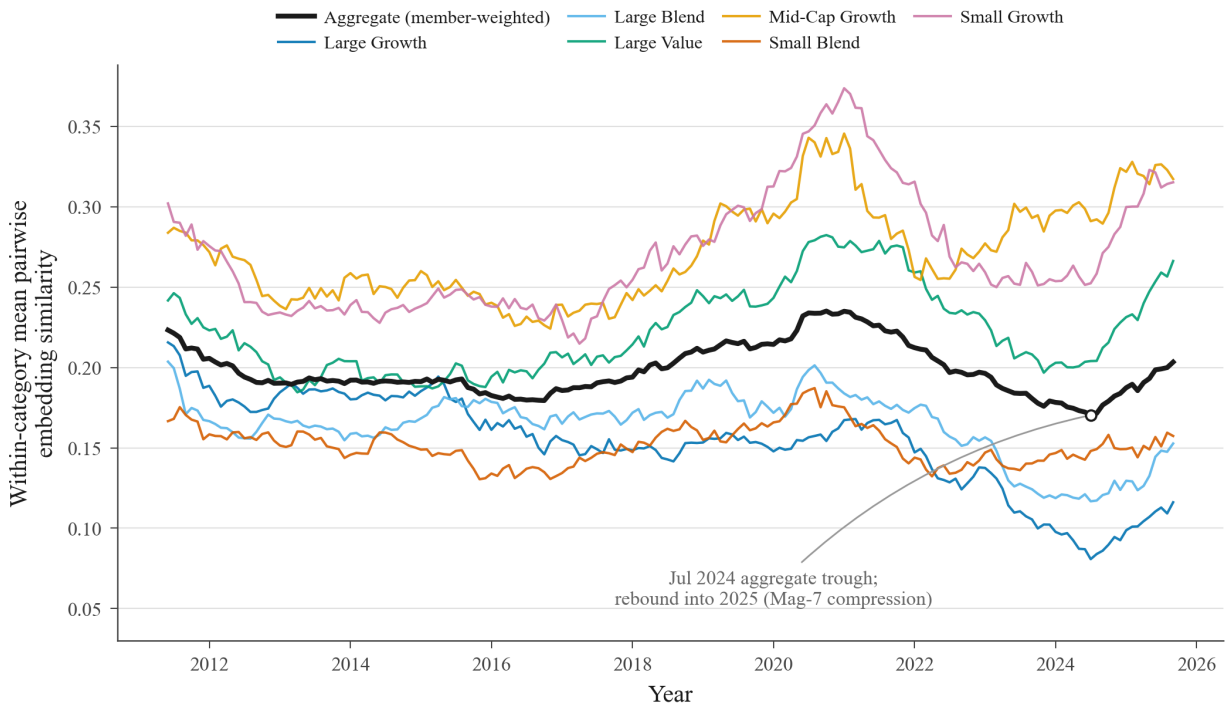


Figure E.5
Category Crowding Index

This figure plots the category crowding index, the exact mean pairwise embedding cosine among category members, for the aggregate and the six largest categories over reliable sample months. Crowding falls to a trough in mid-2024 and rises monotonically through September 2025 in every large category, within its historical range.

Appendix F Single-Modal Local Uniqueness Measures

The cross-modal measures in Section 3.3 use similarity in one modality to define a fund's peer group and distance in the other modality to measure uniqueness. This appendix re-estimates the main specifications under an alternative family of measures that defines the peer group and measures distance within a single modality. The portfolio-uniqueness findings hold under these measures, and the prospectus fee premium operates through the strategy-relative comparison that the cross-modal design encodes.

Construction. For each fund i in month t, I rank all funds with an available embedding by cosine similarity within a modality and collect the twenty nearest neighbors $\mathcal{N}_{20}(i,t)$. The single-modal local uniqueness is the average cosine distance to these neighbors,

$$PortfolioUniq_{i,t}^{\text{kNN}} \;=\; 1 - \frac{1}{20} \sum_{j \in \mathcal{N}_{20}^{h}(i,t)} \cos\!\left(\mathbf{e}_i^{h},\, \mathbf{e}_j^{h}\right), \tag{35}$$

and symmetrically $ProspectusUniq_{i,t}^{\text{kNN}}$ in text space using the mean-centered prospectus embeddings $\tilde{\mathbf{e}}^{t}$.[7] Each measure uses a single modality to define both the peer set and the distance, so it captures within-modality distinctiveness alone. The measures are available for the same 458,190 fund-months used in the main analysis. They are related to their cross-modal counterparts while remaining far from collinear: the Spearman correlation is 0.73 for portfolio uniqueness and 0.51 for prospectus uniqueness.

The main results under single-modal measures. Table F.1 re-estimates the fee, forward-alpha, and flow specifications with the single-modal local measures in place of their cross-modal counterparts, holding the sample and all controls fixed. The portfolio-uniqueness results are robust. Portfolio uniqueness continues to predict forward risk-adjusted perfor-

[7] I set the neighborhood size to twenty; the measures are robust to ten and fifty neighbors, with cross-K rank correlations above 0.95. Mean centering of the text embeddings follows the procedure in Section 3.2.3 and is what gives the single-modal prospectus measure a usable spread.

mance among past winners, and the interaction with the top performance tercile strengthens from 0.181 to 0.283 ($t = 2.52$). Portfolio uniqueness also commands higher fees under the single-modal measure, with a coefficient of 0.190 ($t = 6.04$): a one-standard-deviation increase is associated with a fee that is higher by approximately 3 basis points, which is consistent with skilled managers in less crowded niches pricing that differentiation. The prospectus flow result is likewise robust. Distinctive narrative attracts capital among top performers, and the interaction with the top tercile is 1.236 ($t = 3.15$), broadly consistent with the cross-modal estimate of 0.551. The prospectus fee premium is a property of the cross-modal measure, and Table F.2 shows why.

The prospectus fee premium and the peer group. Panel A of Table F.2 traces the single-modal prospectus fee coefficient as fixed effects are progressively added. The single-modal measure is positively associated with fees across fund families: the coefficient is 0.157 ($t = 2.74$) with no fixed effects and 0.275 ($t = 4.82$) within categories. This association lives between families. Within a fund family, prospectus language is largely templated by a central legal function, so a fund's nearest text neighbors are disproportionately its own family. A fund's single nearest neighbor in text space belongs to its own family 41% of the time, against a 0.6% chance benchmark, and single-modal prospectus uniqueness declines with family size (Spearman -0.34). Narrative differentiation is therefore a family-level positioning choice. Once family fixed effects absorb the between-family variation, the single-modal coefficient is -0.149 ($t = -2.64$), while the cross-modal coefficient remains positive at 0.083 ($t = 3.34$). Deduplicating the neighbor set to exclude same-family funds removes the cross-sectional association entirely, with a coefficient of -0.03 and no fixed effects, and it leaves the within-family coefficient negative, as the family-deduplicated row of Panel B reports.

Panel B of Table F.2 enters the cross-modal and single-modal prospectus measures together with portfolio uniqueness within fund families. The cross-modal measure carries the fee premium, with a coefficient of 0.110 ($t = 4.35$), and portfolio uniqueness is independently

priced at 0.173 ($t = 5.61$). The single-modal prospectus measure enters with a negative loading under both its mean-centered and family-deduplicated definitions, at -0.256 and -0.221. The premium attaches to narrative that is distinctive relative to a fund's portfolio peers, which is the comparison that *ProspectusUniq* encodes. Consistent with Section 5, the premium is borne by institutional investors: the cross-modal fee response is 0.136 ($t = 4.24$) for institutional share classes and 0.046 ($t = 1.70$) for retail.

Taken together, these results support the cross-modal design of Section 3.3. The portfolio-uniqueness findings hold under a single-modal local measure, and the prospectus fee premium reflects the strategy-relative narrative distinctiveness that the cross-modal construction is built to capture.

Appendix G Horse Races Against Existing Uniqueness Proxies

Section 3.4 shows that existing differentiation proxies span a modest share of the cross-modal measures. This appendix asks the sharper question: do the paper's three headline results survive when those proxies enter the regressions directly as competitors? I race *ProspectusUniq* against five text-based controls in the fee and flow specifications, and *PortfolioUniq* against eleven portfolio-based proxies in the forward-alpha specification, one at a time and jointly. I then re-estimate each baseline with fully orthogonalized measures and re-run both instrumental-variable designs with the relevant proxies as controls.

Design. Each race holds the sample fixed across all specifications: the constant sample keeps fund-months for which every competitor proxy is available, and the uniqueness measures and all proxies are standardized, so coefficients are per standard deviation. The specifications mirror the corresponding main-text tables, and on the full paper sample my estimation pipeline reproduces the published baselines exactly (a *ProspectusUniq* fee coefficient

Table F.1
Robustness of the Main Results to Single-Modal Local Uniqueness Measures

This table re-estimates the main fee, forward-alpha, and fund-flow specifications, replacing the cross-modal uniqueness measures with their single-modal local twenty-nearest-neighbor analogues defined in equation (35). The sample is the 458,190 fund-months for which both measure families are available, and it is held fixed across columns. Column (1) reproduces the baseline cross-modal estimates from Tables 2, 3, and 4; column (2) uses the single-modal local measures, with the mean-centered embeddings for the prospectus measure. Panel A reports the annualized fee specification in equation (19) with family, category, and month fixed effects and a one-month lag (approximately 420,000 observations). Panel B reports Fama–MacBeth estimates of the forward FFC6 alpha at the one-month horizon, with on average 3,400 funds per monthly cross-section and the middle performance tercile as the reference group. Panel C reports the fund-flow specification with family, category, and month fixed effects (approximately 391,000 observations). Controls in each panel match the corresponding main-text specification. Standard errors are double-clustered by fund and month in Panels A and C and computed with Newey–West in Panel B; t-statistics are in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

	Cross-modal (baseline)	Single-modal local (kNN)
Panel A: Annualized fee		
$PortfolioUniq_{t-1}$	−0.016	0.190***
	(−0.83)	(6.04)
$ProspectusUniq_{t-1}$	0.085***	−0.173***
	(3.44)	(−3.13)
Panel B: Forward FFC6 alpha at t+1 (Fama–MacBeth)		
$PortfolioUniq$	−0.079	−0.269*
	(−1.06)	(−1.74)
$PortfolioUniq \times I(\text{Top})$	0.181***	0.283**
	(3.16)	(2.52)
Panel C: Fund flows (% of TNA)		
$ProspectusUniq \times$ Top 30%	0.551**	1.236***
	(2.31)	(3.15)

Table F.2
The Prospectus Fee Premium and the Peer Group

This table relates the prospectus fee premium to the choice of peer group. The dependent variable is the annualized fee, and the sample is the common set of approximately 420,000 fund-months with a one-month lag and the baseline controls. Panel A reports the coefficient on the prospectus uniqueness measure as fixed effects are progressively added, separately for the single-modal local measure (column 1) and the cross-modal measure (column 2). Panel B reports joint specifications that include a single-modal prospectus measure, the cross-modal prospectus measure, and the single-modal portfolio measure together, with family, category, and month fixed effects. The mean-centered and family-deduplicated single-modal prospectus measures, where the latter excludes same-family funds from each fund's neighbor set, are entered in separate regressions because they are nearly collinear; the cross-modal and portfolio coefficients are reported from the mean-centered specification and are essentially identical in the family-deduplicated specification, at 0.107 and 0.173. Each coefficient is placed in the column corresponding to its measure family. Standard errors are double-clustered by fund and month; t-statistics are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Annualized fee	
	Single-modal (kNN)	Cross-modal
Panel A: Prospectus fee coefficient by fixed-effect saturation		
No fixed effects	0.157***	0.084***
	(2.74)	(2.59)
Category, month	0.275***	0.116***
	(4.82)	(3.74)
Family, category, month	−0.149***	0.083***
	(−2.64)	(3.34)
Panel B: Joint specifications, family, category, and month fixed effects		
ProspectusUniq$^{\text{kNN}}$, mean-centered	−0.256***	
	(−4.47)	
ProspectusUniq$^{\text{kNN}}$, family-deduplicated	−0.221***	
	(−4.05)	
ProspectusUniq (cross-modal)		0.110***
		(4.35)
PortfolioUniq$^{\text{kNN}}$	0.173***	
	(5.61)	

of 0.085 with $t = 3.44$, a *PortfolioUniq* × Top alpha interaction of 0.108 with $t = 2.19$, and a *ProspectusUniq* × Top flow interaction of 0.551 with $t = 2.31$), so differences inside the races are attributable to the added controls and the sample definition. One sample-composition fact matters for the alpha race: requiring Morningstar style consistency, a variable that never enters any race, removes a third of the paper sample and takes the alpha baseline below significance on its own. I therefore report the alpha race both on the full constant sample and on the constant sample without the style-consistency requirement, where the baseline matches the paper.

The fee premium. Table G.2 races the fee result against prospectus length, readability, boilerplate share, the keyword-based text uniqueness of Kostovetsky and Warner (2020), and the single-modal kNN text measure built from the same embeddings. *ProspectusUniq* keeps a positive coefficient significant at the 5% level on four of the five one-at-a-time rungs; against keyword uniqueness alone it retains significance at the 10% level (0.0115, $t = 1.74$). In the joint specification the coefficient more than doubles relative to the baseline, to 0.0407 per standard deviation ($t = 5.75$), because the single-modal kNN control enters negative, and it survives fixed effects for 243 EDGAR filing agents (0.0296, $t = 4.53$) and family×month fixed effects (0.0385, $t = 6.18$). Filing-agent fixed effects fully absorb the length control, which confirms that document length is an agent-level templating attribute; the premium on *ProspectusUniq* is a fund-level attribute that survives inside filing agents and inside family-months. The fee premium therefore reflects distinctive narrative content itself, beyond length, readability, boilerplate, keyword uniqueness, filing-agent templating, and family-level pricing.

The skill-gated alpha result. Table G.3 races the forward-alpha interaction against eleven portfolio-based proxies, each entering in levels and interacted with the past-performance terciles. On the informative sample the *PortfolioUniq* × Top coefficient stays within a tight band of 0.019 to 0.023 per standard deviation across every specification and is significant

at the 5% level against ten of the eleven proxies one at a time. Two specifications keep the coefficient while widening the confidence interval, and I state them plainly. Adding the same-embedding kNN sibling, whose correlation with *PortfolioUniq* is 0.73, raises the coefficient to 0.0230 and moves the t-statistic to 1.52; the joint specification with all eleven proxies and their tercile interactions, twenty-three correlated terms in total, keeps the coefficient at 0.0191 and moves the t-statistic to 1.30. The magnitude is stable in every specification, and the loss of precision in the joint specification comes from standard-error inflation among correlated interactions. The converse results sharpen the interpretation: on the informative sample not a single competitor×Top interaction is significant in any specification, one at a time or joint, so the skill-gated pattern is specific to *PortfolioUniq* among all measured dimensions of portfolio differentiation. The industry concentration index of Kacperczyk et al. (2005) illustrates the point: it leaves the interaction at 0.0217 ($t = 2.11$), and it enters with a positive and insignificant level effect and a negative and insignificant Top interaction, so within-fund industry concentration carries no skill-gated alpha signal of its own in this design.

The skill-gated flow result. Table G.4 races the flow interaction against the five text-based controls. *ProspectusUniq*×Top survives every rung, with the weakest estimate against keyword uniqueness (0.0931, $t = 1.87$), and strengthens by a third in the joint specification (0.1424, $t = 2.59$). One competitor shares the pattern: the keyword uniqueness interaction is independently significant in the joint specification (0.1812, $t = 3.78$), so keyword-based narrative distinctiveness carries a parallel skill-gated flow effect. *ProspectusUniq* remains significant on top of it, which establishes the two measures as complementary descriptions of how distinctive language attracts performance-contingent capital.

Orthogonalized measures. Panel A of Table G.5 re-estimates each baseline after orthogonalizing the measures within category×month on their full competitor sets. The fee and flow results strengthen: the residualized *ProspectusUniq* fee coefficient rises from 0.0170 to

0.0294 ($t = 5.05$) and the flow interaction from 0.1056 to 0.1341 ($t = 2.63$), so the component of *ProspectusUniq* correlated with the competitors was diluting the estimates. The alpha result is mixed across samples: the residualized interaction is significant on the full constant sample (0.0253, $t = 2.30$), where the raw baseline is insignificant, and attenuates to the 10% level on the larger sample (0.0163, $t = 1.81$).

Instrumental variables. Panels B and C of Table G.5 add the relevant proxies as controls to both 2SLS designs. The peer-exit estimate for *PortfolioUniq* is 1.331 with the industry concentration index as the sole proxy control and 1.339 ($t = 2.22$, first-stage Kleibergen–Paap F above 73) when market active share, $1 - R^2$, and log breadth enter alongside it. The AI-adoption estimate for *ProspectusUniq* moves from 1.479 to 1.447 ($t = 2.94$) when length and boilerplate share enter, and the added controls raise the first-stage F. Both causal estimates move by less than 2% under the proxy controls.

Taken together, the races establish that the fee and flow results hold and strengthen against every measured competitor, that the alpha interaction is stable in magnitude against all of them and statistically robust to each traditional proxy individually, and that the identification results are insensitive to the proxy controls. The cross-modal measures do economic work that the existing toolkit leaves undone.

Appendix H The Winner-Gated Alpha Result and Performance Persistence

The forward-alpha interaction of Section 4.2 conditions on past winners, so the natural objection is that it repackages performance persistence: any variable positively correlated with past alpha inside the top tercile inherits the persistence gradient. This appendix measures that channel directly. The premise of the objection holds in my sample, and I lead with it: Panel A of Table H.3 shows that within the winner tercile, funds in the top

Table G.1
Relation of the Cross-Modal Uniqueness Measures to Existing Proxies

This table relates *PortfolioUniq* and *ProspectusUniq* to existing measures of fund differentiation on the 458,190 fund-month paper sample, pairwise complete. Panel A reports correlations with *PortfolioUniq*; Panel B with *ProspectusUniq*. Spearman correlations use raw values; Pearson correlations use values winsorized at the 1st and 99th percentiles within month. Active share follows Cremers and Petajisto (2009), computed against the aggregate institutional market portfolio and against the equal-weighted category portfolio; HHI is the sum of squared portfolio weights and top-10 share the sum of the ten largest weights; the industry concentration index follows Kacperczyk et al. (2005), computed over the Fama–French ten industries as the sum of squared deviations of fund industry weights from the aggregate institutional industry weights, with SIC ranges from French's data library, and the industry Herfindahl is the sum of squared industry weights; $1 - R^2$ and idiosyncratic volatility come from 24-month FFC6 regressions following Amihud and Goyenko (2013); tracking error is the 24-month volatility of the return gap to the category mean; return uniqueness is the sign-flipped average pairwise 24-month return correlation with category peers, in the spirit of Sun et al. (2012); the holdings and text uniqueness replications follow Kostovetsky and Warner (2020); boilerplate share is the fraction of a filing's distinct 5-grams appearing in more than 5% of contemporaneous filings; style consistency is the Morningstar metric. *PortfolioUniq*$^{\text{kNN}}$ and *ProspectusUniq*$^{\text{kNN}}$ are the single-modal twenty-nearest-neighbor measures built from the same embeddings as the paper measures and serve as an upper benchmark rather than an outside competitor. The sign flip between the Spearman and Pearson correlations of the raw holdings count reflects the heavy right tail of the count distribution, so the rank statistic is the reliable one. Panel C regresses each winsorized measure on its full competitor set on complete cases and reports pooled R^2, within category×month R^2 (iterative demeaning), and both statistics with the same-embedding kNN sibling excluded; the *PortfolioUniq* competitor sets include the industry concentration index and exclude the redundant sign-flipped average similarity duplicate, the industry Herfindahl, and the raw holdings count, which enters as a log.

	Spearman	Pearson (wins.)	N
Panel A: Correlates of PortfolioUniq			
PortfolioUniq$^{\text{kNN}}$ (same embeddings)	0.727	0.727	458,190
Top-10 share	0.341	0.347	458,190
HHI	0.340	0.391	458,190
Industry concentration index (FF10)	0.237	0.129	458,190
Industry Herfindahl (FF10)	0.222	0.140	458,190
Tracking error vs. category (24M)	0.190	0.197	405,001
$1 - R^2$, FFC6 (24M)	0.174	0.142	400,138
Return uniqueness (24M)	0.172	0.160	404,734
Idiosyncratic volatility (24M)	0.165	0.166	400,138
Style consistency (Morningstar)	0.092	0.100	313,440
Active share, market	0.089	0.144	458,190
KW holdings avg. similarity (sign-flipped)	0.085	0.083	457,259
KW holdings uniqueness	0.084	0.073	457,259
Active share, category	0.035	0.048	457,259
Log number of holdings	−0.230	−0.063	458,190
Number of holdings (raw count)	−0.230	0.146	458,190
Panel B: Correlates of ProspectusUniq			
ProspectusUniq$^{\text{kNN}}$ (same embeddings)	0.533	0.523	458,190
KW text uniqueness	0.170	0.132	441,875
Log word count	0.097	0.083	443,106
Gunning fog index	0.051	0.066	443,106
Flesch–Kincaid grade	0.040	0.057	443,106
Boilerplate share	−0.206	−0.167	443,106

Panel C: Joint explanatory power of the competitor set

		Full set		Excl. kNN sibling	
Dependent variable	N	Pooled	Within	Pooled	Within
PortfolioUniq (incl. style consistency)	278,727	0.603	0.601	0.332	0.310
PortfolioUniq (excl. style consistency)	398,278	0.600	0.604	0.340	0.325
ProspectusUniq	441,875	0.290	0.273	0.041	0.031

Table G.2
The Prospectus Fee Premium Against Text-Based Competitors

This table re-estimates the fee regression of equation (19) with category×month fixed effects while adding text-based competitor controls, on a constant sample of 257,936 fund-months for which every proxy is available. The dependent variable is the annualized fee (in percentage points), winsorized at the 1st and 99th percentiles. The uniqueness measures and all competitor controls are lagged one month and standardized, so coefficients are per standard deviation; competitor controls are winsorized at the 1st and 99th percentiles within month before standardizing. Panel A adds one control at a time: log word count of the Principal Investment Strategies section, Flesch–Kincaid grade, boilerplate share (the fraction of a filing's distinct 5-grams appearing in more than 5% of contemporaneous filings), the text uniqueness replication of Kostovetsky and Warner (2020), and the single-modal kNN text measure $ProspectusUniq^{\mathrm{kNN}}$ built from the same embeddings. Panel B enters all five controls jointly; column (2) adds fixed effects for 243 EDGAR filing agents and column (3) adds family×month fixed effects. Controls include log TNA, turnover ratio, fund return, and fund age (all lagged). Standard errors are double-clustered by fund and month. t-statistics in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Panel A: One competitor at a time

	(1)	(2)	(3)	(4)	(5)	(6)
Added control	None	Length	FK grade	Boilerplate	KW text	$ProspectusUniq^{\mathrm{kNN}}$
$ProspectusUniq_{t-1}$	0.0170***	0.0128**	0.0152**	0.0193***	0.0115*	0.0340***
	(2.59)	(1.97)	(2.34)	(2.95)	(1.74)	(4.58)
$PortfolioUniq_{t-1}$	−0.0144**	−0.0116*	−0.0122**	−0.0127**	−0.0131**	−0.0137**
	(−2.38)	(−1.95)	(−2.06)	(−2.09)	(−2.19)	(−2.27)
Added control$_{t-1}$		0.0645***	0.0436***	0.0186***	0.0463***	−0.0310***
		(8.68)	(7.17)	(3.22)	(6.77)	(−4.20)

Panel B: All five competitors jointly

	(1)	(2)	(3)
$ProspectusUniq_{t-1}$	0.0407***	0.0296***	0.0385***
	(5.75)	(4.53)	(6.18)
$PortfolioUniq_{t-1}$	−0.0057	−0.0056	−0.0034
	(−1.00)	(−1.06)	(−0.78)
Log word count	0.0651***	−0.0002	0.0013
	(5.67)	(−0.02)	(0.12)
Flesch–Kincaid grade	0.0340***	0.0351***	0.0263***
	(5.49)	(6.19)	(4.98)
Boilerplate share	0.0284***	0.0134**	0.0024
	(4.95)	(2.27)	(0.37)
KW text uniqueness	0.0074	0.0140	−0.0156*
	(0.69)	(1.58)	(−1.69)
$ProspectusUniq^{\mathrm{kNN}}$	−0.0492***	−0.0571***	−0.0698***
	(−6.68)	(−8.04)	(−9.17)
Category×Month FE	Yes	Yes	Yes
Filing-agent FE		Yes	
Family×Month FE			Yes
Observations	257,936	257,936	257,936

Table G.3
The Skill-Gated Alpha Result Against Portfolio-Based Competitors

This table re-estimates the forward-alpha regression of equation (21) at the one-month horizon with portfolio and month fixed effects while adding portfolio-based competitor proxies, each entering in levels and interacted with the past-performance terciles. The dependent variable is forward FFC6 alpha (monthly %, winsorized at the 1st and 99th percentiles). The uniqueness measures and all proxies are standardized, so coefficients are per standard deviation; proxies are winsorized at the 1st and 99th percentiles within month before standardizing. Two constant samples are used: the full constant sample requiring every proxy including Morningstar style consistency ($N = 253{,}749$), on which the baseline interaction is itself insignificant because of the conditioning on style-consistency availability, and the constant sample excluding the style-consistency requirement ($N = 357{,}996$), on which the baseline matches the paper. Competitor proxies are active share against the market and the category (Cremers and Petajisto, 2009), the holdings uniqueness replication of Kostovetsky and Warner (2020), log number of holdings, HHI, the industry concentration index of Kacperczyk et al. (2005) computed over the Fama–French ten industries, $1 - R^2$ and idiosyncratic volatility from 24-month FFC6 regressions (Amihud and Goyenko, 2013), 24-month tracking error against the category, return uniqueness in the spirit of Sun et al. (2012), and the single-modal kNN measure $PortfolioUniq^{kNN}$ built from the same embeddings. Panel A adds one proxy at a time and reports the interaction of each variable with the top past-performance tercile; the joint row enters all eleven proxies and all their tercile interactions simultaneously (no proxy dropped; all pairwise correlations below 0.85). Panel B reports every competitor×Top coefficient inside the joint specification. Controls include log TNA, number of managers, and turnover ratio. Standard errors are double-clustered by fund and month; t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Constant sample		Constant sample, ex style	
	PortfolioUniq × Top	Proxy × Top	*PortfolioUniq* × Top	Proxy × Top
Panel A: One proxy at a time				
Baseline (no proxy)	0.0135 (1.20)		0.0207** (2.10)	
+ Active share, market	0.0129 (1.14)	0.0036 (0.28)	0.0199** (1.99)	0.0048 (0.41)
+ Active share, category	0.0129 (1.15)	0.0197 (1.32)	0.0197** (1.99)	0.0202 (1.50)
+ KW holdings uniqueness	0.0132 (1.17)	0.0091 (0.57)	0.0200** (2.01)	0.0135 (0.96)
+ Log number of holdings	0.0129 (1.15)	−0.0156 (−1.33)	0.0196** (2.00)	−0.0095 (−0.88)
+ HHI	0.0150 (1.25)	−0.0058 (−0.32)	0.0208** (2.02)	−0.0012 (−0.07)
+ Industry concentration (KSZ)	0.0132 (1.11)	−0.0037 (−0.19)	0.0217** (2.11)	−0.0128 (−0.74)
+ $1 - R^2$ (24M)	0.0132 (1.15)	−0.0020 (−0.09)	0.0217** (2.14)	−0.0127 (−0.68)
+ Tracking error (24M)	0.0109 (0.97)	0.0047 (0.21)	0.0193** (2.00)	−0.0019 (−0.09)
+ Return uniqueness (24M)	0.0125 (1.11)	0.0033 (0.20)	0.0200** (2.04)	−0.0032 (−0.21)
+ Idiosyncratic vol. (24M)	0.0149 (1.29)	−0.0202 (−0.73)	0.0221** (2.25)	−0.0173 (−0.70)
+ $PortfolioUniq^{kNN}$ (same embeddings)	0.0256 (1.49)	−0.0171 (−0.81)	0.0230 (1.52)	−0.0036 (−0.20)
Joint (all eleven + all × Top)	0.0209 (1.26)		0.0191 (1.30)	
Panel B: Competitor × Top coefficients in the joint specification				
Active share, market		−0.0153 (−0.54)		−0.0068 (−0.28)
Active share, category		0.0026 (0.05)		0.0125 (0.28)
KW holdings uniqueness		0.0096 (0.24)		−0.0036 (−0.10)
Log number of holdings		−0.0269 (−0.77)		−0.0156 (−0.54)
HHI		−0.0190 (−0.59)		−0.0183 (−0.67)
Industry concentration (KSZ)		0.0159 (0.56)		0.0045 (0.18)
$1 - R^2$ (24M)		0.0514 (1.12)		0.0156 (0.38)
Tracking error (24M)		0.0765** (2.10)		0.0362 (1.08)
Return uniqueness (24M)		−0.0400 (−1.44)		−0.0144 (−0.56)
Idiosyncratic vol. (24M)		−0.0646 (−1.06)		−0.0211 (−0.42)
$PortfolioUniq^{kNN}$ (same embeddings)		−0.0078 (−0.37)		0.0061 (0.35)
Portfolio + Month FE		Yes		Yes
Observations		253,749		357,996

Table G.4
The Skill-Gated Flow Result Against Text-Based Competitors

This table re-estimates the fund-flow regression of equation (22) with family, category, and month fixed effects while adding text-based competitor controls, on a constant sample of 253,630 fund-months for which every proxy is available. The dependent variable is the share-class net flow as a percentage of TNA, winsorized at the 1st and 99th percentiles. The uniqueness measures and all competitor controls are lagged one month and standardized, so coefficients are per standard deviation; each rung of Panel A adds one control in levels together with its interaction with the top 36-month-alpha tercile. The controls are log word count, Flesch–Kincaid grade, boilerplate share, the text uniqueness replication of Kostovetsky and Warner (2020), and the single-modal kNN text measure $ProspectusUniq^{\mathrm{kNN}}$ built from the same embeddings. Panel B reports the control coefficients inside the joint specification, which enters all five controls and their top interactions simultaneously. Controls include log TNA, turnover ratio, fund return, and fund age (all lagged). Standard errors are double-clustered by fund and month; t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	ProspectusUniq × Top 30%	*ProspectusUniq* (base)	Control × Top 30%	Control (level)
Panel A: One control at a time				
Baseline (no control)	0.1071** (2.18)	−0.0395 (−1.00)		
+ Log word count	0.1050** (2.13)	−0.0400 (−1.01)	0.0194 (0.54)	0.0949** (2.44)
+ Flesch–Kincaid grade	0.1073** (2.18)	−0.0370 (−0.93)	−0.0293 (−0.83)	−0.0182 (−0.76)
+ Boilerplate share	0.1118** (2.27)	−0.0415 (−1.04)	0.0423 (1.12)	−0.0125 (−0.38)
+ KW text uniqueness	0.0931* (1.87)	−0.0403 (−1.01)	0.0887*** (2.71)	0.0308 (0.88)
+ $ProspectusUniq^{\mathrm{kNN}}$ (same embeddings)	0.1309** (2.38)	−0.0781* (−1.86)	−0.0486 (−1.19)	0.1016*** (2.80)
Joint (all five)	0.1424*** (2.59)	−0.0797* (−1.90)		
Panel B: Control coefficients in the joint specification				
Log word count			−0.0861* (−1.65)	0.1463*** (2.84)
Flesch–Kincaid grade			−0.0368 (−1.04)	−0.0317 (−1.35)
Boilerplate share			0.0438 (1.16)	0.0278 (0.76)
KW text uniqueness			0.1812*** (3.78)	−0.0455 (−0.94)
$ProspectusUniq^{\mathrm{kNN}}$ (same embeddings)			−0.0972** (−2.26)	0.1166*** (3.10)
Family, Category, Month FE		Yes		
Observations		253,630		

Table G.5
Residualized Measures and Instrumental-Variable Robustness to Competitor Controls

Panel A orthogonalizes each uniqueness measure within category×month on its full competitor set (*PortfolioUniq* on all eleven portfolio-based proxies of Table G.3 in the alpha rows and on the same set excluding the industry concentration index in the fee row; *ProspectusUniq* on the five text-based controls of Table G.2), re-standardizes the residual, and re-estimates each headline baseline on the identical fixed sample; the fee and alpha rows use category×month and portfolio+month fixed effects respectively, and the flow row uses family, category, and month fixed effects. Panel B re-estimates the peer-exit 2SLS regression of equation (27) (portfolio and category×month fixed effects; endogenous measure in raw units); both specifications include the industry concentration index of Kacperczyk et al. (2005) as a control, and the proxy-controls specification adds market active share, $1 - R^2$, and log number of holdings, each winsorized within month. Panel C re-estimates the AI-adoption 2SLS regression of equation (24) with the 12-month instrument, adding log word count and boilerplate share as controls. KP-F is the Kleibergen–Paap first-stage F-statistic. The constant sample restricts to fund-months with every competitor proxy available. Standard errors are double-clustered by fund and month; t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Panel A: Headline baselines with fully orthogonalized measures

Baseline	Variable	Original	Residualized	N
Fees	$ProspectusUniq_{t-1}$	0.0170*** (2.58)	0.0294*** (5.05)	255,803
Fees	$PortfolioUniq_{t-1}$	−0.0144** (−2.37)	−0.0153*** (−3.03)	255,803
Alpha, constant	$PortfolioUniq \times I(\text{Top})$	0.0135 (1.20)	0.0253** (2.30)	253,749
Alpha, ex style	$PortfolioUniq \times I(\text{Top})$	0.0207** (2.10)	0.0163* (1.81)	357,996
Flows	$ProspectusUniq \times$ Top 30%	0.1056** (2.12)	0.1341*** (2.63)	252,387

Panel B: Peer-exit IV for PortfolioUniq → forward alpha

Sample	Spec	First stage Z	KP-F	2SLS β	N
Full paper	Baseline controls	−0.2561*** (−8.56)	73.2	1.3314** (2.22)	356,188
Full paper	+ Proxy controls	−0.2555*** (−8.55)	73.0	1.3389** (2.22)	356,188
Constant	Baseline controls	−0.2772*** (−8.41)	70.7	1.6693*** (2.65)	247,820
Constant	+ Proxy controls	−0.2747*** (−8.43)	71.1	1.6859*** (2.65)	247,820

Panel C: AI-adoption IV for ProspectusUniq → fees

Sample	FE	Spec	KP-F	2SLS β	N
Full paper	Fund + Month	Baseline controls	12.9	1.356*** (2.95)	245,102
Full paper	Fund + Month	+ Text controls	13.7	1.339*** (3.03)	245,102
Full paper	Fund + Cat×Month	Baseline controls	11.9	1.479*** (2.85)	245,102
Full paper	Fund + Cat×Month	+ Text controls	12.8	1.447*** (2.94)	245,102
Constant	Fund + Month	Baseline controls	13.6	0.903*** (2.91)	162,929
Constant	Fund + Month	+ Text controls	14.4	0.901*** (2.95)	162,929
Constant	Fund + Cat×Month	Baseline controls	12.7	1.096*** (2.93)	162,929
Constant	Fund + Cat×Month	+ Text controls	14.0	1.076*** (3.02)	162,929

PortfolioUniq quintile have past 12-month alpha 1.32 percentage points above the bottom quintile ($t = 13.8$), more idiosyncratic volatility (+0.28, $t = 11.8$), and more holdings, with category×month fixed effects. The question is whether the interaction survives once that gradient enters the regression.

Direct persistence controls. Table H.1 adds the continuous backward 12-month alpha to the specification of equation (21) on identical observations ($N = 392{,}249$), first as a level and then interacted with both tercile dummies. The level control is the sharper test than it may appear, because persistence is present: past alpha predicts one-month forward alpha at 0.012 to 0.014 (t between 2.0 and 2.5) across specifications. The *PortfolioUniq*$\times I(\text{Top})$ coefficient attenuates by roughly a third in Fama–MacBeth and remains significant at the 5% level in every tercile-based specification: 0.119 ($t = 2.19$) with the level, 0.114 ($t = 2.10$) with the level and its tercile interactions, and 0.135 ($t = 2.67$) when past-performance decile dummies replace the tercile intercepts. The panel estimate holds its magnitude at 0.083 to 0.094 and moves to the 10% level (t between 1.86 and 1.92). The decisive coefficients are the persistence slopes themselves: past alpha $\times I(\text{Top})$, the term a persistence-in-disguise story requires to be positive, is 0.0001 ($t = 0.04$) in Fama–MacBeth and -0.006 ($t = -2.13$) in the panel. Holding the alpha level fixed also turns the $I(\text{Top})$ intercept significantly negative (about -0.10, $t \approx -2.3$ to -2.8): winners with common portfolios underperform the middle group, so the winner premium in this sample lives in the uniqueness interaction. The pattern holds with every regressor lagged one month (0.122, $t = 2.29$ in Fama–MacBeth with lagged alpha interactions) and with a gross-of-fee dependent variable (0.178, $t = 3.04$ at baseline). Two boundaries emerge and I state them plainly. At the three-month horizon the Fama–MacBeth interaction sits at the 10% level under the sharp control (0.097, $t = 1.94$) and the panel interaction is insignificant already at baseline. And a fully continuous parameterization, *PortfolioUniq*× past alpha with no terciles, finds nothing ($t \leq 1.4$ in all four cells): the premium is a top-tercile threshold effect, and the tercile structure is where all of the signal

lives.

Alternative skill gates. Table H.2 rebuilds the interaction with six alternative gates whose top dummies correlate only 0.18 to 0.30 with the baseline winner tercile (manager tenure: 0.00), so each is a genuinely different sort. The Morningstar star-rating gate, an external signal, reproduces the result in all eight cells across estimators, bucket definitions, and gate-level controls (0.105 to 0.123, t between 1.93 and 2.55). The 36-month Berk and van Binsbergen (2015) value-added gate reproduces it in Fama–MacBeth (0.135, $t = 2.58$; 0.106, $t = 2.22$ with the gate level) and attenuates to insignificance in the panel (0.083, $t = 1.45$). Gates built on the t-statistic of trailing FFC6 alpha over 24- or 36-month windows produce no interaction ($t \leq 1.1$ throughout), the 36-month five-factor alpha gate is marginal in Fama–MacBeth only, and the tenure gate, a noisy skill proxy on which survival and family practices load, produces a null. The effect therefore concentrates among funds whose recent risk-adjusted returns are high, the sets that 12-month alpha and star ratings identify, and it does not extend to statistical-precision definitions of long-horizon skill.

Inside the winner tercile. Panels B and C of Table H.3 complete the microscope. Restricted to winner fund-months alone, the continuous *PortfolioUniq* slope with the past-alpha control is positive and insignificant in both fixed-effects schemes (0.011 with category×month effects, 0.077 within portfolio), so the identification comes from the top-versus-middle slope contrast rather than from continuous variation inside the winner group. The raw double sort requires no regression machinery: mean forward alpha improves across *PortfolioUniq* quintiles only in the top past-performance row, and the top-row extreme-quintile spread of 5.3 basis points per month ($t = 1.88$) sits next to the 5.5 basis points the regression coefficients imply.

Placebos, non-skill gates, and the instrumented estimate. Table H.5 closes the identification loop. Permuting the top and bottom tercile assignments jointly at random

within category×month cells 200 times yields a placebo distribution centered near zero (mean $t = -0.31$, SD 0.98, share of $|t| \geq 1.96$ equal to 6.5%, close to nominal); the actual panel t-statistic of 2.19 sits at the 99th percentile, the permutation p-value is 0.030, and one placebo draw in 200 reaches the actual coefficient. A size gate produces no positive top interaction; an age gate produces one marginal panel cell (0.112, $t = 1.88$) with no Fama–MacBeth counterpart, and I report it as it comes out. The peer-exit 2SLS estimate is insensitive to the persistence control on identical rows: 1.253 to 1.259 ($t = 2.24$ to 2.26), with the first-stage Kleibergen–Paap F at 75.3 in both variants.

Calendar time. Table H.4 translates the interaction into portfolio form and reports the power arithmetic alongside the estimates. Among winners, the primary construction (within-month *PortfolioUniq* quintiles, next-month net excess returns) earns 6.3 basis points per month ($t = 1.18$) with an FFC6 alpha of 7.3 basis points ($t = 1.40$), magnitudes that match the regression-implied 5.5 basis points; the realized-abnormal-return construction and the category-month quintile variant show no spread; and the winners' spread loads significantly on a generic persistence factor (β between 0.17 and 0.33) with all spanning alphas at or below 4 basis points. Monthly information coefficients tell the same story: the winners' Spearman IC is +0.004 raw and +0.001 after stripping past alpha, against a regression-implied +0.010 and a series standard error near 0.008. At the observed spread volatility, 179 months certify an effect of this kind only at roughly twice the implied magnitude. I therefore present the winners' uniqueness premium as a conditional cross-sectional slope whose implied economic magnitude is consistent with, and below the detection threshold of, calendar-time portfolio tests at this sample length.

Taken together, the battery confirms the objection's premise, measures the channel it implies, and finds the interaction intact where the theory places it: the persistence slopes are flat or negative where the objection needs them positive, an external skill gate reproduces the pattern, random gates reach the observed coefficient once in 200 draws, and the causal

estimate is unmoved.

Appendix I The Winner-Gated Flow Result and Flow-Performance Convexity

The flow interaction of Section 4.3 conditions on a past-performance bucket, so the natural objection is that it repackages the convexity of the flow-performance relation (Sirri and Tufano, 1998; Huang et al., 2007): the 30/40/30 gate discretizes a convex curve, and any variable correlated either with rank position inside the top bucket or with the volatility that carries funds to extreme ranks inherits the steep upper end of the curve. This appendix tests the premise of that objection, controls for the curve directly, and locates the effect within it.

The premise. Panel A of Table I.1 shows the objection's premise is reversed for *ProspectusUniq*. Lagged prospectus uniqueness correlates negatively with trailing 12- and 24-month volatility (Spearman -0.11 and -0.13) and with the maximum monthly return (-0.11), is essentially uncorrelated with idiosyncratic volatility ($+0.01$), and correlates only weakly with category tracking error ($+0.09$). Within the top bucket, its correlation with rank position is -0.000 under the within-month gate. The premise does hold, modestly, for *PortfolioUniq* ($+0.16$ with idiosyncratic volatility, $+0.11$ with within-top rank), the measure that shows no flow interaction anywhere. Panel B confirms the pattern in a balance table: within winners, funds in the top uniqueness quintile sit at the same rank as the bottom quintile (-0.001, $t = -0.30$), with lower volatility (-0.36, $t = -6.4$) and lower maximum returns (-0.64, $t = -5.5$). They are also smaller (-0.61 log points of TNA) and younger (-1.6 years); size and age are lagged controls in every specification, and the double-sort and gate tests below hold rank, volatility, size, and age structure fixed, but I flag fund size and age as the alternative channel these designs bound rather than eliminate.

Table H.1
The Skill-Gated Alpha Result Under Direct Persistence Controls

This table re-estimates the forward-alpha interaction specification of equation (21) while controlling directly for performance persistence. The dependent variable in Panel A is forward FFC6 alpha at t+1 (monthly %). Base is the published specification of Table 3. +Past α adds the continuous backward 12-month rolling alpha (the same winsorized variable that generates the terciles). +Past α×Terciles additionally interacts it with the top and bottom tercile dummies, so the uniqueness interactions survive only if they carry information beyond a within-tercile linear persistence gradient. Columns (1) to (3) report Fama–MacBeth estimates with Newey–West t-statistics (177 monthly cross-sections, on average 2,216 funds each); columns (4) to (6) include portfolio and month fixed effects with standard errors double-clustered by portfolio and month (N = 392,249; identical rows across columns). Panel B reports the *PortfolioUniq* × I(Top) coefficient (or its continuous analogue *PortfolioUniq*×Past α) under alternative parameterizations: within-category-month past-performance decile dummies replacing the tercile intercepts (reference decile 5), a fully continuous specification with no terciles, all right-hand variables lagged one month within portfolio, a gross dependent variable that adds back the t+1 expense accrual, and the three-month horizon. Controls include log TNA, fund age (Fama–MacBeth only), number of managers, and turnover ratio. t-statistics in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

	(1)	(2)	(3)	(4)	(5)	(6)
	Fama–MacBeth			Panel (portfolio + month FE)		
	Base	+Past α	+Past α ×Terciles	Base	+Past α	+Past α ×Terciles
Panel A: Forward FFC6 alpha at t+1						
PortfolioUniq × I(Top)	0.1805***	0.1194**	0.1135**	0.1084**	0.0905*	0.0941*
	(3.16)	(2.19)	(2.10)	(2.19)	(1.86)	(1.92)
PortfolioUniq × I(Bot.)	0.0501	0.0638	0.0502	0.0256	0.0536	0.0467
	(0.79)	(1.04)	(0.78)	(0.49)	(1.08)	(0.94)
PortfolioUniq	−0.0785	−0.0507	−0.0474	0.0012	−0.0144	−0.0144
	(−1.06)	(−0.73)	(−0.70)	(0.01)	(−0.17)	(−0.17)
ProspectusUniq × I(Top)	−0.0172	0.0139	0.0381	0.0003	0.0046	0.0129
	(−0.35)	(0.31)	(0.86)	(0.01)	(0.09)	(0.24)
ProspectusUniq	−0.0663	−0.1408*	−0.1605**	−0.1104	−0.1065	−0.1102
	(−0.85)	(−1.87)	(−2.21)	(−1.25)	(−1.22)	(−1.27)
I(Top)	−0.0274	−0.0664	−0.1137***	−0.0432	−0.0874**	−0.0997**
	(−0.72)	(−1.64)	(−2.77)	(−1.10)	(−2.14)	(−2.33)
I(Bot.)	−0.1236***	−0.0598	−0.1317***	−0.0701*	−0.0193	−0.0473
	(−2.74)	(−1.22)	(−3.07)	(−1.74)	(−0.47)	(−1.16)
Past α (12M)		0.0116**	0.0135**		0.0100**	0.0139**
		(2.46)	(2.37)		(2.02)	(2.49)
Past α × I(Top)			0.0001			−0.0059**
			(0.04)			(−2.13)
Past α × I(Bottom)			−0.0050			−0.0056*
			(−1.48)			(−1.81)
Panel B: PortfolioUniq × I(*Top*) *under alternative parameterizations*	Fama–MacBeth			Panel		
Decile dummies (ref. D5), t+1	0.1353*** (2.67)			0.0834* (1.87)		
All regressors lagged one month, t+1	0.1619*** (2.74)			0.0954* (1.76)		
Lagged + α interactions, t+1	0.1224** (2.29)			0.0903* (1.67)		
Gross DV, base, t+1	0.1781*** (3.04)			0.1025** (2.00)		
Gross DV + α interactions, t+1	0.1082* (1.86)			0.0881* (1.73)		
Base, t+3	0.1390** (2.57)			0.0584 (1.22)		
+Past α, t+3	0.1023** (1.98)			0.0549 (1.16)		
+Past α×Terciles, t+3	0.0974* (1.94)			0.0603 (1.26)		
Decile dummies, t+3	0.0882* (1.74)			0.0358 (0.84)		
Continuous interaction, no terciles, t+1	0.0008 (0.12)			0.0066 (1.09)		
Continuous interaction, no terciles, t+3	0.0102 (1.37)			0.0005 (0.08)		
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Portfolio FE	—	—	—	Yes	Yes	Yes
Month FE	—	—	—	Yes	Yes	Yes
Observations (Panel A)	2,216/mo	2,216/mo	2,216/mo	392,249	392,249	392,249

Table H.2
Alternative Skill Gates

This table re-estimates the forward-alpha interaction specification of equation (21) at the one-month horizon after replacing the backward 12-month alpha terciles with alternative skill gates. Each gate is assigned within category×month (cells with at least ten non-missing funds, top above the 2/3 rank, bottom at or below the 1/3 rank), except the star-rating gate, which uses the Morningstar buckets High ($\geq$ 4 stars), Low ($\leq$ 2), and middle (= 3) as reference; a rank-tercile variant of the star gate is also reported. The gates are the Morningstar star rating, the t-statistic of the trailing FFC6 alpha over 36- and 24-month windows, the 36-month trailing mean of Berk–van Binsbergen value added, the longest manager tenure, and the 36-month five-factor alpha. V1 keeps the published regressor pattern with the new gate; V2 adds the continuous gate variable and its interactions with the top and bottom gate dummies. Panel A reports the $PortfolioUniq \times I(\text{Top gate})$ coefficient; the first column reports the correlation of each gate's top dummy with the baseline 12-month-alpha top dummy on the paper sample. Panel B reports the V2 gate-level terms. Fama–MacBeth columns use Newey–West t-statistics; Panel columns include portfolio and month fixed effects with standard errors double-clustered by portfolio and month. In the star-gate V2 panel column the gate-level block is nearly collinear with the High dummy (correlation 0.992) and its clustered variance matrix is not positive semi-definite, so those two t-statistics are unavailable (n.a.); the uniqueness interaction is unaffected. Tenure is a noisy skill proxy (survival, family practices, and team composition all load on it), so its null is only weakly informative. Controls match Table 3. t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Corr. with	Fama–MacBeth		Panel		
Gate	baseline Top	V1	V2	V1	V2	N
Panel A: PortfolioUniq × I(Top gate)						
Star rating (High/Low buckets)	0.246	0.1227**	0.1138**	0.1192**	0.1200**	371,683
		(2.55)	(2.42)	(1.96)	(1.99)	
Star rating (rank terciles)	0.247	0.1183**	0.1050**	0.1193*	0.1232**	367,809
		(2.26)	(2.01)	(1.93)	(2.00)	
Alpha t-statistic, 36M	0.259	0.0186	0.0273	0.0050	−0.0004	348,194
		(0.31)	(0.48)	(0.08)	(−0.01)	
Alpha t-statistic, 24M	0.302	0.0564	0.0573	0.0080	0.0011	369,617
		(1.08)	(1.09)	(0.16)	(0.02)	
Value added, 36M	0.178	0.1348***	0.1063**	0.0825	0.0799	320,825
		(2.58)	(2.22)	(1.45)	(1.43)	
Manager tenure	−0.002	−0.0425	−0.0601	0.0601	0.0637	360,702
		(−0.87)	(−1.17)	(0.94)	(0.99)	
Alpha FF5, 36M (cat.×month)	0.269	0.0853*	0.0447	0.0373	0.0439	388,302
		(1.69)	(0.87)	(0.67)	(0.77)	
Panel B: Gate-level terms in V2						
		Fama–MacBeth		Panel		
		Gate G	G×Top	Gate G	G×Top	
Star rating (High/Low buckets)		0.0246	0.0283	−0.0256	0.0094	
		(1.25)	(1.24)	(n.a.)	(n.a.)	
Star rating (rank terciles)		0.3009**	−0.2177*	0.2025*	−0.1974*	
		(2.25)	(−1.66)	(1.71)	(−1.79)	
Alpha t-statistic, 36M		0.1164*	0.0350	−0.0406	0.0294	
		(1.77)	(1.30)	(−0.60)	(1.07)	
Alpha t-statistic, 24M		0.0344	0.0611*	−0.0876	0.0468	
		(0.52)	(1.87)	(−1.10)	(1.41)	
Value added, 36M		0.1249**	−0.1127**	0.0054	−0.0047	
		(2.38)	(−2.33)	(0.16)	(−0.14)	
Manager tenure		0.0465***	−0.0429***	−0.0144	0.0089	
		(2.88)	(−2.73)	(−1.33)	(0.85)	
Alpha FF5, 36M (cat.×month)		0.0331**	0.0003	−0.0106	0.0056	
		(2.23)	(0.04)	(−0.59)	(0.57)	

Table H.3
Inside the Winner Tercile

This table examines the composition of the top past-performance tercile across *PortfolioUniq* quintiles assigned within category×month ($N = 131{,}591$ top-tercile fund-months). Panel A reports mean fund characteristics by quintile; the Q5−Q1 column reports the difference from a regression of each characteristic on a Q5 dummy with category×month fixed effects on the Q1 and Q5 subsample, with standard errors double-clustered by portfolio and month. Panel B restricts the forward-alpha regression to top-tercile fund-months only, with the continuous backward 12-month alpha as a control. Panel C reports mean forward FFC6 alpha at t+1 (monthly %) in the 3×5 double sort of past-performance terciles and *PortfolioUniq* quintiles (cell counts range from 23,176 to 29,002); the spread row reports the mean of monthly top-row Q5−Q1 spreads with a Newey–West t-statistic (177 months, lag 4). t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Panel A: Balance across PortfolioUniq quintiles within the winner tercile

	Q1	Q2	Q3	Q4	Q5	Q5−Q1 (FE)	t
Past alpha, 12M (%)	1.958	2.353	2.642	2.947	3.314	1.318***	(13.82)
Performance rank	0.813	0.826	0.833	0.842	0.849	0.036***	(14.64)
TNA ($M)	3,136	1,824	1,496	1,677	1,940	−1,030	(−0.89)
Log TNA	4.981	4.862	4.819	4.782	4.919	−0.038	(−0.36)
Fund age (years)	16.60	15.69	14.82	14.56	15.04	−1.567***	(−2.86)
Turnover ratio	0.550	0.604	0.594	0.576	0.580	0.012	(0.57)
Expense ratio	0.0100	0.0103	0.0103	0.0104	0.0104	0.0003	(1.52)
Idiosyncratic vol., 24M (%)	1.268	1.394	1.447	1.505	1.664	0.277***	(11.78)
Number of holdings	114.1	111.4	132.2	152.2	197.4	98.9***	(6.15)

Panel B: Forward alpha at t+1, top-tercile fund-months only

	Category×month FE	Portfolio + month FE
PortfolioUniq	0.0108 (0.28)	0.0770 (0.75)
ProspectusUniq	−0.0117 (−0.35)	−0.0004 (−0.00)
Past alpha, 12M	0.0126*** (3.06)	0.0077 (1.43)
Observations	131,591	131,591

Panel C: Mean forward FFC6 alpha at t+1 by past-performance tercile and PortfolioUniq quintile

	Q1	Q2	Q3	Q4	Q5	Q5−Q1 spread (NW t)
Bottom tercile	−0.248	−0.264	−0.253	−0.304	−0.270	
Middle tercile	−0.183	−0.230	−0.192	−0.227	−0.224	
Top tercile	−0.183	−0.148	−0.158	−0.124	−0.132	0.0533* (1.88)

Table H.4
Calendar-Time Portfolios and Information Coefficients

This table translates the forward-alpha interaction into portfolio-level evidence. Panel A forms *PortfolioUniq* quintiles at formation month t (within month, or within category×month) and reports the equal-weighted Q5−Q1 spread realized at t+1 (percent per month, 179 formation months, 2011–2025), separately for top-tercile past winners and the full sample, and for two return constructions: next-month net excess return, and the realized FFC6 abnormal return trimmed at the pooled 1st/99th percentiles. Legs require at least five funds. FFC6 alpha is the intercept of the spread on the six factors; the spanning alpha adds a persistence factor (equal-weighted top-tercile minus bottom-tercile return, same construction and month) whose loading is reported in the last column. All t-statistics are Newey–West with lag 6. For scale, the Fama–MacBeth estimates of Table 3 imply a top-tercile spread of about 0.055% per month for the observed Q5−Q1 *PortfolioUniq* gap, while the monthly spread volatility of about 0.6% means the design reaches $t \geq 2$ only for true spreads of roughly 0.11% per month. Panel B reports the mean monthly Spearman rank correlation between *PortfolioUniq* and the next-month FFC6 abnormal return; the partial version first residualizes the abnormal return on the backward 12-month alpha within month. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Panel A: Equal-weighted Q5−Q1 PortfolioUniq spreads (% per month)

Sample	Quintiles / return	Months	Funds/leg	Mean	FFC6 α	Spanning α	Pers. β
Top tercile	month / net excess	179	162/162	0.063	0.073	0.040	0.204**
				(1.18)	(1.40)	(0.74)	(2.06)
Top tercile	month / FFC6 abnormal	179	151/158	−0.004	—	−0.032	0.328***
				(−0.08)		(−0.87)	(2.89)
Top tercile	cat.×month / net excess	178	151/154	−0.003	0.009	−0.019	0.175***
				(−0.12)	(0.38)	(−0.81)	(3.89)
Top tercile	cat.×month / FFC6 abnormal	178	142/149	−0.007	—	−0.019	0.207***
				(−0.37)		(−1.06)	(3.97)
Full sample	month / net excess	179	506/505	−0.020	−0.003	−0.025	0.137*
				(−0.43)	(−0.07)	(−0.57)	(1.84)
Full sample	month / FFC6 abnormal	179	457/486	−0.033	—	−0.039	0.182*
				(−0.79)		(−1.12)	(1.72)
Full sample	cat.×month / net excess	179	501/506	−0.007	−0.002	−0.010	0.054
				(−0.34)	(−0.10)	(−0.56)	(1.57)
Full sample	cat.×month / FFC6 abnormal	179	454/480	−0.017	—	−0.020	0.062
				(−1.04)		(−1.35)	(1.44)

Panel B: Monthly Spearman information coefficients, PortfolioUniq vs. next-month FFC6 abnormal return

Sample	IC type	Months	Avg. N	Mean IC	NW t	Share months IC > 0
Top tercile	raw	179	782	0.0038	(0.46)	50.3%
Top tercile	partial (past α removed)	179	782	0.0007	(0.09)	49.7%
Full sample	raw	179	2,388	−0.0047	(−0.57)	43.6%
Full sample	partial (past α removed)	179	2,364	−0.0033	(−0.45)	44.7%

Table H.5
Placebo Terciles, Non-Skill Gates, and the Instrumented Estimate

Panel A permutes the pair of top and bottom tercile assignments jointly at random within each category×month cell (200 draws, seed 42) and re-estimates the panel specification of Table 3 at the one-month horizon on each draw; it reports the distribution of the placebo *PortfolioUniq* × I(Top) coefficients and t-statistics, the percentile of the actual estimate in that distribution, and the permutation p-value ($|t_{\text{placebo}}| \geq |t_{\text{actual}}|$). Panel B replaces the past-performance terciles with terciles of fund size (log TNA) and fund age assigned within category×month. Panel C re-estimates the pooled peer-exit 2SLS regression of forward FFC6 alpha on instrumented *PortfolioUniq* (portfolio and category×month fixed effects; controls include the performance tercile dummies, log TNA, fund return, flows, turnover, and expense ratio) with and without the continuous backward 12-month alpha as an included exogenous control, on identical rows. Standard errors are double-clustered by portfolio and month throughout. t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Panel A: Placebo terciles permuted within category×month (200 draws)

	Mean	SD	Min	Max
Placebo coefficient	−0.0120	0.0395	−0.1072	0.1133
Placebo t-statistic	−0.31	0.98	−2.61	2.80
Share of draws with $\lvert t\rvert \geq 1.96$	6.5% (13/200)			
Actual estimate (coefficient, t)	0.1084 (2.19)			
Percentile of actual t among placebos	99.0			
Draws with coefficient ≥ actual	1/200			
Permutation p-value	0.030			

Panel B: Non-skill gates (terciles of size and age within category×month)

	Size gate (N = 387,683)		Age gate (N = 387,753)	
	FM	Panel	FM	Panel
PortfolioUniq	0.0510	0.1124	−0.0291	−0.0126
	(0.64)	(1.21)	(−0.37)	(−0.15)
PortfolioUniq × I(Top gate)	−0.0249	−0.0890	0.0497	0.1122*
	(−0.63)	(−1.39)	(1.05)	(1.88)
PortfolioUniq × I(Bottom gate)	−0.1313***	−0.0879	0.0113	0.0929
	(−2.62)	(−1.28)	(0.24)	(1.53)

Panel C: Peer-exit 2SLS with and without the persistence control (N = 376,866)

	2SLS β	KP F	First stage Z	Reduced form Z	OLS
Without past α (12M)	1.2532**	75.3	−0.2654***	−0.3326**	0.0103
	(2.24)		(−8.68)	(−2.41)	(0.39)
With past α (12M)	1.2590**	75.3	−0.2655***	−0.3343**	0.0097
	(2.26)		(−8.68)	(−2.43)	(0.37)

Rank-position controls. Table I.2, Panel A, adds progressively finer functions of the performance rank. The continuous rank level moves the interaction from 0.55 ($t = 2.31$) to 0.59 ($t = 2.49$); the Sirri and Tufano (1998) piecewise rank slopes give 0.62 ($t = 2.62$); piecewise slopes built from the trailing-twelve-month compounded return rank, with the published gate retained, give 0.64 ($t = 2.66$). The sharp test replaces the gate's level dummies with nineteen within-month rank vigintile dummies, so identification comes only from variation within twentieths of the rank distribution: the interaction is 0.66 ($t = 2.77$) under family, category, and month fixed effects and 0.76 ($t = 3.07$) under family×month fixed effects, and the estimated vigintile pattern is itself cleanly convex, so the control absorbs exactly the curve the objection describes. All four specifications also hold at the 5% level when the gate is rebuilt by ranking within category×month (0.47 to 0.68). Rebuilding the gate from the raw trailing-return rank, the exact Sirri and Tufano (1998) sort, strengthens the interaction to 0.96 ($t = 4.16$); under that gate the *PortfolioUniq* interaction turns significant as well (0.57, $t = 2.65$), so I use the raw-return gate as a robustness memo rather than a primary specification. Panel B shows the within-winner level regression: the *ProspectusUniq* level is positive in all four cells and significant in one, a design with less identifying variation that I report for completeness. Panel C is the nonparametric version: double-sorting winners into rank bins and uniqueness terciles within month, mean rank is essentially flat across terciles inside every bin (maximum gap 0.004), yet the uniqueness flow spread is positive in all six cells and significant in five, including all three bins under the category×month gate.

Volatility and extremeness. Table I.3 adds the second channel's variables directly. Volatility enters flows negatively in the level and in the top-bucket interaction, so the low volatility of unique-prospectus funds works against the measured interaction; maximum return and skewness carry positive levels but no top-bucket interactions; and a continuous *ProspectusUniq* × volatility term is zero ($t = -0.50$), against the masquerade story. The interaction ranges from 0.51 to 0.68 across all of these against same-sample baselines of 0.56

and 0.66. The kitchen sink, vigintile dummies plus volatility and maximum-return levels and top interactions, gives 0.61 ($t = 2.53$) and 0.67 ($t = 2.67$) under the within-month gate; under the category×month gate it attenuates to 0.46 ($t = 2.01$) and 0.39 ($t = 1.73$), which I state as the weakest cells in the battery. Panel D adds 12b-1 fees on the subsample where they exist (54% of observations): the 12b-1 terms themselves move the interaction by less than 0.02 in every column, while the sample restriction takes the family×month and category×month-gate cells below conventional significance, and the *ProspectusUniq* level turns significantly negative in this subsample. The marketing-fee channel therefore neither explains nor absorbs the interaction, and the 12b-1 subsample is the honest boundary of where the result is measured precisely.

The curve by uniqueness. Table I.4 estimates the piecewise flow-performance curve separately by *ProspectusUniq* tercile. Under the within-month gate, prospectus uniqueness steepens the top segment: the high-piece slope rises from 5.69 in the lowest tercile to 8.20 in the highest (+2.50, $t = 2.48$), while the low and middle segments and the top-bucket intercept are unchanged. Ranking within category×month, the steepening is positive but imprecise (+1.03, $t = 1.03$), so the curve exhibit is descriptive of where the interaction operates rather than an independent result. The continuous version implies a flow derivative with respect to prospectus uniqueness of −0.29 at rank 0.05, −0.15 at the median, and +1.04 at rank 0.95; conditional on the same curve, the *PortfolioUniq* high-segment interaction is negative (−5.26, $t = -2.40$), a further contrast between text and holdings at the top of the curve. This locates my result relative to Kostovetsky and Warner (2020): their distinctive strategy text flattens flow-performance sensitivity for young funds, shielding them at the bottom of the curve, whereas prospectus uniqueness in my sample amplifies the top segment among established winners. Rational updating (Berk and Green, 2004) predicts text matters most where information about skill is scarcest; the concentration of the response in the steepest segment of realized performance, among retail investors, points to marketing amplification

where search among winners is most intense (Huang et al., 2007).

Placebos, non-performance gates, investor split, and aggregation. Table I.5 closes the loop. Permuting the gate labels jointly within month 200 times yields a placebo t distribution with mean 0.13 and standard deviation 0.82; 3.5% of draws exceed $|t| = 1.96$, close to nominal size, and no draw reaches the actual t-statistic (empirical $p \approx 0.005$; the closest draw comes within 0.0005). Replacing the performance gate with 30/40/30 gates on size, age, or volatility produces no positive interaction: the volatility gate, the objection's own channel, is null in both directions; the age gate is significantly negative in the replacement design and vanishes alongside the performance gate; the size gate is null in the replacement design, and when added alongside the performance gate its interaction is 0.70 ($t = 2.38$) while the performance interaction sits at 0.47 ($t = 1.93$) on that sample, which I report as it comes out. The investor split shows the incidence: the retail interaction is 0.84 ($t = 3.07$) and survives the corrected gate, the vigintile control, and the volatility interactions, while the institutional interaction is null in every specification ($t \leq 1.12$). Aggregating the dependent variable to the portfolio level preserves the result (0.70, $t = 2.57$; 0.82, $t = 3.07$ with vigintile controls), so the share-class structure does not drive it.

Taken together, the battery reverses the objection's premise for the measure that carries the effect, absorbs the flow-performance curve nonparametrically without losing the interaction, shows volatility pushes against rather than for the result, finds no placebo or non-performance gate that reproduces it, and locates the effect where a marketing channel operates: retail money, at the top of the curve.

Appendix J AI Keyword List

This appendix lists the complete set of keywords used to classify an employee as AI-educated in Section 4.4.1. Classification is applied to the raw degree-field string (`field_raw`) from Revelio Labs' education records. An employee is flagged as AI-educated if at least one

Table I.1
The Flow-Convexity Confound Premise: Volatility, Extremeness, and Rank Position

This table tests the premise of the Sirri–Tufano convexity objection to the flow result of Table 4: that funds with unique prospectuses are more volatile or sit at more extreme performance ranks, so that the *ProspectusUniq* × Top interaction picks up flow convexity (Sirri and Tufano, 1998; Huang et al., 2007) rather than a uniqueness effect. Panel A reports correlations between the lagged uniqueness measures and return-based volatility and extremeness moments on the flow estimation sample: trailing 12- and 24-month return volatility, 24-month FFC6 idiosyncratic volatility, 24-month category tracking error, 12-month return skewness, the maximum monthly return over the trailing 12 months, and, within the top-30% bucket only, the continuous performance rank. Spearman correlations use raw values; Pearson correlations use values winsorized at the 1st/99th percentiles within month (ranks are not winsorized). Panel B sorts the top-30% (winner) bucket into *ProspectusUniq* quintiles within month and reports mean characteristics for the bottom and top quintiles; the Q5−Q1 difference is the coefficient from a regression of the characteristic on the Q5 dummy with month fixed effects on the Q1∪Q5 sample, with standard errors two-way clustered by portfolio and month. The G-month gate ranks 36-month FF5 alpha within month (the published gate); the G-cat gate ranks it within category×month. t-statistics in parentheses. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Panel A: Correlations with lagged uniqueness (flow estimation sample)

	$ProspectusUniq_{t-1}$		$PortfolioUniq_{t-1}$		
	Spearman	Pearson (w)	Spearman	Pearson (w)	N
Volatility, 12M	−0.110	−0.092	0.018	0.023	384,099
Volatility, 24M	−0.132	−0.114	0.033	0.042	365,033
Idiosyncratic vol. (FFC6, 24M)	0.009	0.026	0.161	0.162	367,040
Category tracking error, 24M	0.086	0.112	0.188	0.194	366,277
Return skewness, 12M	−0.024	−0.025	0.034	0.037	384,099
Max monthly return, 12M	−0.106	−0.087	0.016	0.023	384,099
Rank within Top 30% (G-month)	−0.000	0.007	0.109	0.113	132,882
Rank within Top 30% (G-cat)	0.066	0.069	0.091	0.094	119,407

Panel B: Balance within the winner bucket, ProspectusUniq quintiles within month

	G-month (N = 132,882)				G-cat (N = 119,407)			
	Q1	Q5	Q5−Q1	t	Q1	Q5	Q5−Q1	t
Perf. rank (36M FF5)	0.849	0.848	−0.001	(−0.30)	0.841	0.857	0.016***	(5.20)
Volatility, 12M	4.832	4.467	−0.364***	(−6.42)	4.834	4.360	−0.473***	(−8.35)
Max monthly return, 12M	9.183	8.545	−0.639***	(−5.51)	9.134	8.314	−0.818***	(−7.16)
Return skewness, 12M	−0.099	−0.135	−0.036**	(−2.19)	−0.096	−0.134	−0.038**	(−2.22)
Lagged monthly return	0.012	0.011	−0.000	(−0.44)	0.011	0.010	−0.000	(−0.60)
36M FF5 alpha (level)	2.163	2.108	−0.055	(−0.50)	1.532	1.082	−0.451***	(−2.95)
Log TNA	5.313	4.702	−0.611***	(−4.78)	5.302	4.733	−0.569***	(−4.23)
Fund age (years)	16.117	14.547	−1.569**	(−2.50)	15.784	13.348	−2.436***	(−3.91)

Table I.2
The Flow Result Under Flexible Performance-Rank Controls

This table re-estimates the flow specification of equation (22) while controlling for the position in the performance-rank distribution, the first channel of the convexity objection: within a coarse 30/40/30 bucket, any variable correlated with rank position inherits the convex flow-performance relation (Sirri and Tufano, 1998). Panel A reports the $ProspectusUniq_{t-1} \times$ Top coefficient when the baseline adds the continuous performance rank, when Sirri–Tufano piecewise rank slopes (Low $= \min(R, 0.2)$, Mid $= \text{clip}(R - 0.2, 0, 0.6)$, High $= \max(R - 0.8, 0)$) replace the linear lagged return, when the piecewise slopes are built from the trailing-12-month compounded return rank (Sirri and Tufano, 1998) with the published gate retained, and when 19 within-month rank vigintile dummies replace the Top/Middle level dummies entirely (the sharp test: identification comes from within-vigintile variation). The G-month gate ranks 36-month FF5 alpha within month (the published gate); the G-cat gate ranks it within category×month. The memo row reports the plain baseline with the gate itself rebuilt from the trailing-12-month return rank (G-ret, $N = 406{,}117$ FE(1), 365,041 FE(2)). Panel B restricts the sample to the top-30% bucket and regresses flow on the uniqueness levels, the continuous rank, volatility, maximum return, and the flow controls. Panel C double-sorts the winner bucket into performance-rank bins and within-month *ProspectusUniq* terciles and reports mean flow spreads (T3−T1) per bin with Newey–West t-statistics on the monthly spread series (lag 4); mean rank within a bin is essentially identical across terciles (maximum gap 0.004). Standard errors in Panels A and B are two-way clustered by portfolio and month. Fixed effects: FE(1) family, category, and month; FE(2) family×month and category. t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Panel A: $ProspectusUniq_{t-1}\times$ Top under rank controls

	G-month		G-cat	
	FE(1)	FE(2)	FE(1)	FE(2)
Baseline (30/40/30 gate)	0.5513**	0.6784***	0.6813***	0.6623***
	(2.31)	(2.75)	(2.91)	(2.85)
+ Continuous rank level	0.5924**	0.7009***	0.5764**	0.5513**
	(2.49)	(2.83)	(2.50)	(2.41)
+ Sirri–Tufano piecewise rank slopes	0.6241***	0.7196***	0.4959**	0.4688**
	(2.62)	(2.90)	(2.16)	(2.04)
+ ST pieces from 12M return rank	0.6393***	0.7183***		
	(2.66)	(2.86)		
Vigintile dummies replace gate levels (sharp)	0.6605***	0.7606***	0.5280**	0.5049**
	(2.77)	(3.07)	(2.34)	(2.25)
Memo: baseline, G-ret gate (12M return rank)	0.9618***	0.8742***		
	(4.16)	(3.63)		

Panel B: Top-bucket-only level regressions (winners, $I(Top) = 1$)

	G-month		G-cat	
	FE(1)	FE(2)	FE(1)	FE(2)
$ProspectusUniq_{t-1}$	0.3018	0.1593	0.4217**	0.2647
	(1.62)	(0.81)	(2.13)	(1.20)
$PortfolioUniq_{t-1}$	0.0622	0.1846	0.1501	0.1778
	(0.38)	(1.06)	(0.85)	(0.93)
Performance rank (continuous)	5.7294***	5.3703***	4.9347***	4.8408***
	(15.70)	(13.61)	(14.98)	(13.22)
Volatility (12M)	−0.4996***	−0.4878***	−0.5629***	−0.5277***
	(−11.37)	(−9.68)	(−10.99)	(−8.97)
Max return (12M)	0.1870***	0.1862***	0.2011***	0.1823***
	(9.82)	(9.00)	(9.31)	(7.72)
Observations	130,838	107,088	117,225	94,158

Panel C: Double sort within winners, rank bins × ProspectusUniq terciles

	G-month ($N = 132{,}882$)				G-cat ($N = 119{,}407$)			
Rank bin	T1 flow	T3 flow	T3−T1	NW t	T1 flow	T3 flow	T3−T1	NW t
0.7–0.8	−0.238	−0.118	0.102	(1.50)	−0.186	0.037	0.201***	(2.84)
0.8–0.9	−0.028	0.204	0.325***	(2.87)	0.149	0.381	0.259***	(3.58)
0.9–1.0	0.594	0.984	0.371***	(3.28)	0.884	1.162	0.251**	(2.42)

Table I.3
The Flow Result Under Volatility and Extremeness Controls

This table re-estimates the flow specification of equation (22) while controlling for return volatility and extremeness, the second channel of the convexity objection (Sirri and Tufano, 1998; Huang et al., 2007). The dependent variable is share-class net flow (% of TNA, winsorized at 1st/99th percentiles within month). All specifications keep the full baseline (uniqueness levels, gate dummies, all four uniqueness×gate interactions, and lagged controls). Panel A adds, on a common sample, trailing 12-month return volatility with its Top and Middle interactions (column 2), the maximum monthly return over the trailing 12 months with its Top interaction (column 3), 12-month return skewness with its Top interaction (column 4), and a continuous *ProspectusUniq* × Volatility term (column 5); column 1 is the baseline on the identical sample. The memo row reports *ProspectusUniq*×Top from the same specification under family×month and category fixed effects. Panel B substitutes FFC6 idiosyncratic volatility and category tracking error (24M) for total volatility, each with its own same-sample baseline. Panel C is the kitchen sink: 19 within-month performance-rank vigintile dummies replace the Top/Middle level dummies, and volatility and maximum-return levels and their Top interactions enter jointly, under both the published within-month gate (G-month) and the corrected within-category×month gate (G-cat). Panel D adds the lagged 12b-1 fee (in %), its Top interaction, and its interaction with the Sirri–Tufano high-rank piece on the subsample with non-missing 12b-1 data; companion columns re-estimate Panel C on that subsample so sample and control effects are separable. New continuous variables are winsorized at 1st/99th percentiles within month. Fixed effects: FE(1) family, category, and month; FE(2) family×month and category. Standard errors are two-way clustered by portfolio and month. *PU* abbreviates *ProspectusUniq*. *t*-statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Panel A: Volatility, maximum return, and skewness, G-month gate, FE(1), N = 384,099

	(1) Base	(2) +Vol	(3) +Max	(4) +Skew	(5) +*PU*×Vol
ProspectusUniq$_{t-1}$× Top	0.5574**	0.5116**	0.5924**	0.5572**	0.5355**
	(2.31)	(2.12)	(2.46)	(2.30)	(2.23)
Volatility (12M)		−0.1454***			−0.1497***
		(−6.33)			(−4.38)
Volatility × Top		−0.0336			
		(−1.43)			
Volatility × Middle		−0.0211			
		(−1.29)			
Max return (12M)			0.0313***		
			(3.76)		
Max return × Top			0.0164		
			(1.51)		
Skewness (12M)				0.1480***	
				(3.68)	
Skewness × Top				0.0263	
				(0.52)	
ProspectusUniq$_{t-1}$× Volatility					−0.0220
					(−0.50)
Memo: *PU*×Top under FE(2)	0.6580***	0.5929**	0.6849***	0.6593***	0.6376**
(N = 345,392)	(2.66)	(2.38)	(2.74)	(2.65)	(2.58)

Table I.3
The Flow Result Under Volatility and Extremeness Controls (continued)

Continued from Table I.3.

Panel B: Idiosyncratic volatility and tracking error substitutes, FE(1), own samples				
	Base	+Ivol (FFC6, 24M)	Base	+Track. error (24M)
$ProspectusUniq_{t-1}\times$ Top	0.5959**	0.5591**	0.6281**	0.5354**
	(2.41)	(2.25)	(2.54)	(2.19)
Substitute level		0.0081		−0.1922***
		(0.16)		(−3.89)
Substitute × Top		−0.0858		0.2748***
		(−1.63)		(4.12)
Substitute × Middle		−0.1762***		0.0373
		(−3.42)		(0.72)
Observations	367,040	367,040	366,277	366,277

Panel C: Kitchen sink (vigintile dummies + volatility and max levels and Top interactions)				
	G-month		G-cat	
	FE(1)	FE(2)	FE(1)	FE(2)
$ProspectusUniq_{t-1}\times$ Top	0.6094**	0.6713***	0.4586**	0.3941*
	(2.53)	(2.67)	(2.01)	(1.73)
$PortfolioUniq_{t-1}\times$ Top	−0.0338	−0.1340	−0.0228	0.0753
	(−0.15)	(−0.59)	(−0.10)	(0.34)
Volatility (12M)	−0.3561***	−0.3045***	−0.3661***	−0.3161***
	(−11.27)	(−9.01)	(−12.22)	(−10.04)
Volatility × Top	−0.1297***	−0.1423***	−0.1544***	−0.1544***
	(−4.54)	(−4.30)	(−4.73)	(−4.54)
Max return (12M)	0.1258***	0.1044***	0.1254***	0.1045***
	(9.94)	(7.59)	(10.23)	(7.86)
Max return × Top	0.0492***	0.0516***	0.0667***	0.0633***
	(3.15)	(3.04)	(3.61)	(3.23)
Observations	384,099	345,392	380,293	341,846
Memo: same-gate plain baseline	0.5574**	0.6580***	0.6813***	0.6623***
	(2.31)	(2.66)	(2.91)	(2.85)

Panel D: 12b-1 fees (marketing intensity), non-missing-12b-1 subsample								
	G-month				G-cat			
	Sink		Sink + 12b-1		Sink		Sink + 12b-1	
	FE(1)	FE(2)	FE(1)	FE(2)	FE(1)	FE(2)	FE(1)	FE(2)
$PU_{t-1}\times$ Top	0.5955**	0.4300	0.5844**	0.4110	0.4476	0.2904	0.4319	0.3061
	(2.01)	(1.35)	(1.98)	(1.29)	(1.50)	(1.02)	(1.46)	(1.09)
PU_{t-1} (level)	−0.5117**	−0.4244*	−0.5090**	−0.4221*	−0.3503*	−0.2246	−0.3999**	−0.2974
	(−2.35)	(−1.85)	(−2.36)	(−1.88)	(−1.73)	(−1.13)	(−2.07)	(−1.60)
Volatility × Top	−0.2131***	−0.2055***	−0.2028***	−0.1959***	−0.2383***	−0.2216***	−0.2312***	−0.2157***
	(−5.86)	(−5.08)	(−5.68)	(−4.91)	(−6.15)	(−5.44)	(−6.01)	(−5.34)
Max return × Top	0.0724***	0.0644***	0.0706***	0.0636***	0.1000***	0.0849***	0.0985***	0.0832***
	(3.59)	(3.03)	(3.53)	(3.02)	(4.55)	(3.57)	(4.50)	(3.52)
12b-1 fee_{t-1}			−4.1532***	−4.1216***			−4.2307***	−4.1787***
			(−9.79)	(−9.47)			(−10.59)	(−10.38)
12b-1 $fee_{t-1}\times$ Top			−0.6714*	−0.4112			0.1656	0.1037
			(−1.73)	(−1.30)			(0.37)	(0.26)
12b-1 $fee_{t-1}\times$ High piece			8.3682	4.5281			9.1192**	10.7597**
			(1.61)	(0.99)			(2.06)	(2.33)
Observations	208,508	189,457	208,508	189,457	205,980	187,123	205,980	187,123

Table I.4
The Flow-Performance Curve by Prospectus Uniqueness

This table estimates the Sirri–Tufano flow-performance curve separately by prospectus uniqueness and asks where in the curve uniqueness operates. Panels A and B fully interact the piecewise rank slopes (Low = $\min(R, 0.2)$, Mid = $\text{clip}(R - 0.2, 0, 0.6)$, High = $\max(R - 0.8, 0)$) with within-month $ProspectusUniq_{t-1}$ terciles, with family, category, and month fixed effects and the flow controls; the T3−T1 column tests the difference in each segment slope between the highest and lowest uniqueness terciles. The Top-step row adds a level shift $I(\text{Top})\times$ tercile to test whether the bucket intercept, rather than the slope, varies with uniqueness. Panel C reports the continuous version: each uniqueness measure interacted with the three rank pieces, plus piece levels, measure levels, and controls. Implied $\partial Flow/\partial ProspectusUniq$ rows evaluate the linear combination of the level and interaction coefficients at rank 0.05, 0.50, and 0.95 (no standard errors; linear combinations of Panel C coefficients). The G-month gate ranks 36-month FF5 alpha within month; the G-cat gate ranks it within category×month. Standard errors are two-way clustered by portfolio and month. t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Panel A: Segment slopes by ProspectusUniq tercile, G-month (N = 391,416)

Segment	T1 (low)	T2	T3 (high)	T3−T1
Low (rank < 0.2)	1.172	1.978	2.041	0.869
	(1.44)	(2.91)	(2.58)	(0.81)
Mid (0.2–0.8)	1.907	1.863	2.003	0.096
	(12.70)	(13.20)	(12.71)	(0.48)
High (rank > 0.8)	5.691	5.727	8.195	2.504**
	(8.37)	(8.95)	(9.67)	(2.48)
Top-step (I(Top)× tercile)	0.010	0.011	−0.066	−0.077
	(0.16)	(0.17)	(−0.91)	(−0.78)

Panel B: Segment slopes by ProspectusUniq tercile, G-cat (N = 387,466)

Segment	T1 (low)	T2	T3 (high)	T3−T1
Low (rank < 0.2)	3.632	3.312	4.246	0.614
	(4.70)	(6.00)	(6.72)	(0.63)
Mid (0.2–0.8)	1.524	1.621	1.552	0.028
	(10.68)	(12.89)	(10.99)	(0.15)
High (rank > 0.8)	6.927	6.314	7.959	1.032
	(9.67)	(10.09)	(10.09)	(1.03)
Top-step (I(Top)× tercile)	−0.040	−0.003	0.054	0.094
	(−0.61)	(−0.04)	(0.73)	(0.99)

Panel C: Continuous measure × segment interactions

	G-month		G-cat	
	FE(1)	FE(2)	FE(1)	FE(2)
ProspectusUniq× High	7.2586***	4.0275	4.3995*	3.2175
	(3.04)	(1.63)	(1.93)	(1.31)
ProspectusUniq× Mid	0.3296	0.7236	0.4159	0.3009
	(0.71)	(1.46)	(0.97)	(0.65)
ProspectusUniq× Low	0.3144	0.2318	−0.7139	1.4418
	(0.13)	(0.09)	(−0.34)	(0.68)
PortfolioUniq× High	−5.2619**	−4.2818*	−3.6970*	−3.7949*
	(−2.40)	(−1.93)	(−1.76)	(−1.75)
Implied $\partial Flow/\partial PU$ at $R = 0.05$	−0.291	−0.374	−0.154	−0.411
Implied $\partial Flow/\partial PU$ at $R = 0.50$	−0.145	−0.122	−0.137	−0.104
Implied $\partial Flow/\partial PU$ at $R = 0.95$	1.043	0.699	0.648	0.468
Observations	391,416	352,009	387,466	348,328

Table I.5
Placebo Gates, Non-Performance Gates, and the Investor-Type Split

This table reports falsification and decomposition tests for the *ProspectusUniq* × Top flow result. Panel A permutes the (Top, Middle) gate labels jointly within month, preserving bucket counts, 200 times (seed 42) and re-estimates the baseline FE(1) specification on each draw; it reports the placebo t-statistic distribution and the position of the actual estimate. Panel B replaces the performance gate with 30/40/30 within-month gates built on fund size, fund age, and return volatility (column a), and adds the alternative gate terms alongside the performance gate (column b); replacement samples require no performance-rank data and are larger than the baseline sample. Panel C splits flows by investor type using the Morningstar fund-level retail and institutional net-flow aggregates (one row per portfolio-month, the level at which these variables vary; dependent variables winsorized at 1st/99th percentiles within month) and reports $ProspectusUniq_{t-1}$ × Top under the baseline gate specification, the corrected G-cat gate, the vigintile rank control, and the volatility-interaction specification, plus a share-class-level sensitivity. Panel D aggregates the dependent variable to the portfolio level (sum of class flows over sum of lagged class TNA) and re-estimates the baseline and the vigintile specification. All regressions include the lagged controls and FE(1) family, category, and month fixed effects, with standard errors two-way clustered by portfolio and month. *PU* abbreviates *ProspectusUniq*. t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Panel A: Gate-label permutation placebo (200 draws, baseline FE(1) sample)

Placebo t: mean / SD	0.125 / 0.818
Placebo t: 2.5th / 97.5th percentile	−1.42 / 1.89
Placebo t: min / max	−2.13 / 2.3055
Fraction of draws with $\lvert t\rvert \geq 1.96$	0.035 (7/200)
Actual t-statistic	2.3060
Signed percentile of actual t	100.0 (200/200 draws below)
Fraction of draws with $\lvert t\rvert \geq \lvert$actual$\rvert$	0.000 (closest draw 2.3055)

Panel B: Non-performance 30/40/30 gates (size, age, volatility)

	(a) Gate replaces perf. gate		(b) Baseline + gate terms		
Gate variable	*PU* × TopG	*N*	*PU* × Top(perf)	*PU* × TopG	*N*
Size (log TNA)	−0.1731	419,432	0.4663*	0.7006**	390,788
	(−0.53)		(1.93)	(2.38)	
Age (years)	−1.1985***	420,129	0.5566**	0.0251	391,416
	(−3.71)		(2.36)	(0.08)	
Volatility (12M)	−0.0034	406,117	0.5478**	−0.0512	384,099
	(−0.01)		(2.28)	(−0.23)	

Panel C: Retail versus institutional flows, $ProspectusUniq_{t-1}$ × Top

	Retail	*N*	Institutional	*N*
Baseline gate, G-month	0.8401***	260,387	0.2539	267,314
	(3.07)		(0.64)	
Baseline gate, G-cat	0.6527**	257,932	0.3121	265,267
	(2.30)		(0.86)	
Vigintile rank control	0.8984***	260,387	0.4419	267,314
	(3.29)		(1.12)	
Volatility interactions	0.8852***	255,439	0.1939	262,646
	(3.20)		(0.49)	
Share-class sensitivity, baseline	0.8715***	262,054	0.3043	268,745
	(3.20)		(0.78)	

Panel D: Portfolio-level flow (aggregated dependent variable)

	PU×Top	*PortfolioUniq*×Top	*N*
Baseline gate, G-month	0.6956***	0.2693	388,267
	(2.57)	(1.13)	
Baseline gate, G-cat	0.6634**	0.3541	384,317
	(2.52)	(1.49)	
Vigintile rank control	0.8247***	−0.0498	388,267
	(3.07)	(−0.21)	

of their degree records contains any of the keywords below. Matching for full phrases is performed as a case-insensitive substring search; short acronyms are matched with word boundaries to avoid spurious hits (e.g., `ai` must not match inside "maintain"). If any single education record for a user matches, that user is classified as AI-educated.

Substring keywords (case-insensitive). The following phrases trigger a match whenever they appear anywhere inside the degree-field string. Keywords are listed by the source language. Non-Latin-script entries are given as English glosses to keep the appendix portable across typesetting systems; the raw Unicode strings are preserved in the replication code.

- *English:* artificial intelligence; machine learning; deep learning; neural network; natural language processing; computer vision; data science; robotics; large language model; reinforcement learning; statistical learning; predictive analytics; big data; computational intelligence; autonomous systems; cognitive computing; speech recognition; image recognition; generative ai; conversational ai.
- *Chinese (Simplified):* artificial intelligence; machine learning; deep learning; natural language processing; computer vision; data science; neural network.
- *Japanese:* artificial intelligence; machine learning; deep learning.
- *Korean:* artificial intelligence; machine learning; deep learning.
- *French:* intelligence artificielle; apprentissage automatique; apprentissage profond; science des données.
- *German:* künstliche intelligenz; maschinelles lernen.
- *Spanish:* inteligencia artificial; aprendizaje automático; ciencia de datos.
- *Portuguese:* inteligência artificial; aprendizado de máquina; ciência de dados.
- *Italian:* intelligenza artificiale.
- *Dutch:* kunstmatige intelligentie; machinaal leren.

- *Russian:* artificial intelligence; machine learning.
- *Hindi:* artificial intelligence.
- *Polish:* robotyka.
- *Turkish:* yapay zeka.
- *Arabic:* artificial intelligence.

Acronym patterns (word-boundary, case-insensitive). The following short forms are matched only as standalone tokens, delimited by word boundaries: `ai`, `ml`, `llm`, `nlp`, `aiml`, `ai/ml`, `ml/ai`.