Penalty-Heavy Contracts in Asset Management: China's 2025 Fund Fee Reform

Albert Bo Zhao[1]

School of Finance, Nankai University

September 18, 2026

Preliminary. Comments are welcome.

Abstract

In May 2025, China's fund regulator mandated a performance-fee schedule for newly issued active equity funds with an unusual geometry: the penalty for trailing the benchmark by three percentage points is twice the reward for beating it by six, and the fee settles lot by lot over each investor's own holding window. The design moves past the symmetric fulcrum fee, the only performance-fee contract U.S. law has permitted since 1970, to the penalty side. We model such penalty-heavy, vintage-contingent contracts. The schedule taxes tracking variance: managers below a skill threshold optimally closet-index, and the within-year risk response is non-monotone—gambling zones near the downgrade and upgrade lines, an incentive dead zone in deeper losses. Lot-by-lot settlement makes money age a state variable: young lots face a penalty-tilted lottery, while window averaging launders luck out of maturing lots. The mandate requires 60% of new active equity issues at leading companies to adopt the contract, with fixed-fee products continuing alongside; under this partial mandate, fund companies assign floating contracts to their strongest managers, an endorsement reversal visible in the first cohort's appointment data. The welfare account is correspondingly asymmetric: the reform's robust effect is protection, not stimulation—investors of mediocre funds are refunded the noise tax as their managers closet, while the incentive effect on genuinely skilled funds is limited and preference-dependent. First-year evidence from the inaugural cohort is suggestive.

Keywords: performance fees; fulcrum fees; mutual funds; delegated portfolio management; risk-taking incentives; contract design; vintage-contingent settlement; China mutual fund fee reform

JEL classification: G23, G28, D82, D86

[1]Email: zhaoboab@gmail.com. I acknowledge financial support from the Tianjin Philosophy and Social Science Planning Project (Grant No. TJYJ25-02).

1 Introduction

In May 2025, the China Securities Regulatory Commission made a fee schedule mandatory (China Securities Regulatory Commission (中国证券监督管理委员会), 2025). Newly issued active equity funds must charge a management fee that floats against a benchmark. Trailing the benchmark by three percentage points cuts the fee from 1.2% to 0.6% per year; beating it by six points raises the fee to 1.5%. The penalty is twice the reward and sits at half the distance. The fee is settled separately for every lot of shares over that lot's own holding period, so two investors holding the same fund on the same day can owe different rates. And the obligation is partial but binding: leading fund companies must issue at least sixty percent of new active equity products on these terms.

The economist's standard antidote to agency-induced risk-taking is symmetry. When the United States Congress amended the Investment Advisers Act of 1940 in 1970 to curb the risk-shifting of the go-go years, it banned asymmetric performance fees for registered investment companies and left standing only the symmetric fulcrum fee, which adjusts the fee around a benchmark by equal amounts at equal distances (United States Congress, 1970). The fulcrum fee has been little used since: symmetry adds variance to the manager's revenue without adding to its expectation, and investors, who do not shop on management fees, never asked for it. The 2025 schedule does the opposite: it forces the fee past symmetry, to the penalty side. What does such a contract do, and why would a regulator want it?

Three more specific questions organize the analysis, one for each party the contract touches. First, the manager: what does the schedule do to portfolio choice—who remains active under it, how much risk does he take, and when? Second, the fund company and the industry: when the mandate binds only part of new issuance, who receives these contracts, and where do activity and fees settle in equilibrium? Third, the regulator: what institutional concerns make this geometry understandable, and at what cost? We build the model these questions require, solve it at the calibration of the 2025 contract, and confront the predictions with the first year of data from the mandated cohort.

Two incentives shape the manager's problem. A manager of skill μ chooses how far to deviate from the benchmark; deviation buys expected outperformance and buys tracking risk. The fee schedule prices this trade one way: it is a three-tier step, bounded and penalty-heavy, a concave contract that pays for alpha, charges for variance, and stops paying at the floor. The market prices it the opposite way: the convex flow–performance relation (Chevalier and Ellison, 1997; Sirri and Tufano, 1998) rewards variance wherever the ranking is still up for grabs, and in hot markets the flow incentive outweighs the fee.

The manager's objective contains both, minus his own risk and research costs. The contract's incidence is on the manager because it is a supply-side instrument: two voluntary waves of floating fees failed commercially in 2019 and 2023, and the settlement arrives at redemption, after every decision the investor could have made with it. The first set of results is therefore about portfolio choice.

In a normal year, the contract is a screening device. Because the schedule charges for variance and pays for alpha, deviation is a negative-expectation gamble for a manager whose skill cannot cover the noise deviation creates. There is a skill threshold below which the optimal deviation is exactly zero: closet indexing emerges as an equilibrium consequence of the contract's geometry, with no appeal to career concerns (Chevalier and Ellison, 1999), herding (Scharfstein and Stein, 1990), or ranking pressure (Chevalier and Ellison, 1997; Brown et al., 1996) (Proposition 1). At the calibration, a manager must expect about 2.15% of annual alpha per 5% of tracking error before the first step away from the benchmark is worth taking. The level consequences are equally direct. There is a break-even alpha below which a manager's expected fee lies below the base tier (2.35% per year at 4% tracking error; Lemma 2), and the active industry's typical delivered alpha sits well below it: the U.S. average is estimated near zero (Berk and van Binsbergen, 2015; Carhart, 1997), and in China individual outperformance shows no persistence (Lin et al., 2009; Zhuang and Tang, 2004). The median manager's expected fee under the new schedule therefore sits below the 1.2% he used to collect for certain. The end of the rain-or-shine fee has a direction: it ends the guaranteed fee specifically for managers whose deviation produces noise, and refunds the difference to their investors. Among managers who remain active, the contract compresses the low-skilled and, within the mean–variance benchmark, amplifies the high-skilled, so activity is redistributed toward skill (Proposition 1(b)). The two halves differ in robustness: the compression half survives every preference specification examined; the amplification half is mean–variance-specific, and the exact constant-absolute-risk-aversion (CARA) problem of Appendix B shows why: the fee schedule's own lottery variance is a contract-specific risk that the reduced form absorbs.

Within the evaluation year, the contract generates a risk profile that the tournament literature's schedules do not produce. The step schedule's floor and thresholds make the marginal incentive state-dependent, and the profile is non-monotone. A manager one standard deviation below the downgrade line raises risk (the **catch-up gambling zone**), and a manager just below the upgrade line raises risk as well, but a manager deeper than about -7% does not respond at all: once the return distribution sits far enough below the threshold, performance-linked marginal incentives decay toward zero and optimal risk rests at its lower bound. We call this last region the **incentive dead zone**. The dead zone separates the contractual mechanism from behavioral accounts of gambling in losses: those accounts predict a risk response that deepens monotonically as losses accumulate, while the contract predicts an inverted-U with a deep-tail collapse. The profile is also asymmetric in a specific direction: the catch-up peak exceeds the reward peak, where the U.S. tournament evidence finds the stronger response on the reward side (Brown et al., 1996; Kempf and Ruenzi, 2008). What kills the marginal incentive in the deep tail is the floor, not the fact that the fee moves with performance at all: a continuous performance fee keeps its marginal incentive alive at any depth of loss, and the step-plus-floor geometry removes it.

The fee schedule has a boundary: its marginal bite is bounded, and the flow incentive is not. In hot-money episodes the market's convex pricing overwhelms the regulator's concave contract (*flow swamping*; Proposition 2), and the within-year profile flattens as the fee schedule's bite is diluted away. What takes over is a feature that, to our knowledge, no performance fee in the literature carries: the lot-by-lot settlement. Because every lot is evaluated over its own window, a young lot's fee tier is a lottery, penalty-tilted in expectation and a *noise tax* on early money, while *window averaging* launders luck out of a maturing lot's fee at an exponential rate, so the realized charge converges to a step function of the fund's true skill (Proposition 4). Only money sitting exactly on a threshold keeps its marginal incentive forever (Proposition 6), which makes the catch-up zone a young-money phenomenon. And because money gathered at a peak is repriced year after year if the fund then trails its benchmark, a bust now carries a recurring charge against the company's fee revenue: three to four tenths of a bust fund's lifetime fee revenue at the institutional parameters (Proposition 5). The **vintage-contingent fee** turns the company's own appetite for hot money into a regulatory instrument that is active exactly when the fee schedule is not.

At the industry level, the mandate binds, and the company's own revenue logic decides where it lands. Left to fee-revenue motives, the industry would float about one new product in seven; the quota requires three in five. Facing the quota, a company's revenue calculus assigns the floating contracts top-down, to its strongest managers: for a high-skill manager the contract raises revenue, because he becomes more active and draws smart money at almost no fee concession; for a low-skill manager it cuts revenue twice over, through compressed activity and a below-base expected fee. A floating appointment is therefore an endorsement, and the reserved fixed-fee segment shelters the weaker managers rather than crowning the stars (Proposition 7). This *endorsement reversal* runs against the scarcity intuition that the exemption from performance fees would be reserved for stars. The aggregates settle accordingly: total activity declines slightly while its center of gravity moves to the top of the skill distribution, and the middle band is pushed toward the benchmark.

The shape itself is the last question, and the answer is interpretive. Three institutional concerns, each documented in the record, point in the same broad direction: persistent failure should carry a visible cost (the political concern), reward-induced gambling should be limited (the risk concern), and the active industry should remain viable (the viability concern). The observed penalty-heavy geometry can be read as a compromise among these concerns, though the model does not identify the exact regulatory parameters (Section 5). The comparators locate the trade-offs: the symmetric step halves the political force; the bonus-only schedule removes the penalty side and keeps the reward peak's gambling undiminished; and the linear fulcrum, the contract the U.S. Congress chose, creates no recurring charge after a bust, since a floorless linear schedule cannot punish a persistent loser. The welfare decomposition reads by skill band (Proposition 8): the reform's gains concentrate in the lower-middle of the distribution, where marginal managers become closet indexers and their investors are refunded the noise tax, and its costs concentrate at the top. It is a protection device aimed at the investors of mediocre funds, not an incentive device aimed at stars. And the quota's arithmetic is plain: thin in total, negative at the margin: the case for any particular coverage must be carried by values outside the first-round welfare account.

The first cohort's data, one bull-market year after the mandate, trace the predicted trajectory where one year can trace it. Managers appointed to floating-fee products carried pre-appointment excess returns averaging $+1.2\%$, against -6.4% for managers appointed to fixed-fee products: the endorsement reversal, visible in a single cohort of appointment decisions ($p = 0.006$). The young windows behave like the lotteries the theory says they should be: realized first-year tiers pile up at both corners, 46% at the upgrade tier and 38% at the downgrade against a zero-skill lottery's 7% and 23% ($p = 3.1\text{e}{-8}$), and tiers flip about every eleven trading days where a mature window's fee should sit flat. The catch-up zone cannot be found (a single event in its critical bin), and the theory itself predicts that the zone should be invisible in this sample, for two reasons: hot markets dilute the fee schedule's bite (Proposition 2), and the zone lives only in windows both young and near the threshold (Proposition 6), a state a bull market removes. And the flows are consistent with the vintage structure's prediction: the best-performing new products of a bull year were redeemed at a rate of 27% per quarter, with the outflow following the contract's own settlement calendar. This is suggestive evidence from the first cohort: twenty-six funds and one market state. The full tests require the second and third settlement windows, in a market that contains the losing funds the theory is about; those tests become possible as later settlement windows arrive.

Four literatures frame the contribution. The theory of delegated portfolio management studies the fee schedule's incentive content: symmetric fulcrum fees (Starks, 1987; Admati and Pfleiderer, 1997; Ou-Yang, 2003; Dybvig et al., 2010), bonus-only incentive fees and their general-equilibrium price effects (Cuoco and Kaniel, 2011), fee structures as signals of manager type (Das and Sundaram, 2002), and explicit and implicit incentives together (Buffa et al., 2022). The schedules studied in this line are symmetric, reward-only, or flat; to our knowledge, none combines a penalty-heavy shape with lot-by-lot settlement under a mandate. The mandated part matters for how fees can be read: a schedule a manager chooses for himself can signal his type, while a schedule the regulator chooses for him cannot. A second literature prices the optionality inside performance fees and the risk-shifting it invites (Grinblatt and Titman, 1989; Goetzmann et al., 2003; Carpenter, 2000; Panageas and Westerfield, 2009; Massa and Patgiri, 2009; Huang et al., 2011), with Golec and Starks (2004) using the 1970 regulatory shock itself as the experiment. Our kink geometry differs from this line's in one respect that matters for incentives: floored, penalty-heavy steps produce a non-monotone risk response with a dead zone, which a single kink or a continuous schedule does not generate. Third, the flow and tournament literature supplies the implicit incentive our explicit contract must counteract (Chevalier and Ellison, 1997; Sirri and Tufano, 1998; Brown et al., 1996; Berk and Green, 2004; Kempf and Ruenzi, 2008) and the objective in which the two incentives meet (Basak et al., 2007; Dangl et al., 2008); what is new here is the contract on the other side of that opposition, and the frequency: the tournament literature's risk-shifting is annual, while our profile lives inside the evaluation year, on the contract's own settlement calendar. Fourth, fee structures have long been studied as devices that sort investors (Chordia, 1996; Nanda et al., 2000; Barber et al., 2005; Nanda et al., 2009); the 2025 contract sorts managers, and the demand side's documented fee-blindness is precisely why it can.

The nearest neighbor is the Chinese performance-based-fee literature, and the key difference is the floor. Sheng et al. (2014) model continuous performance-fee schedules for Chinese funds and show, statically, that a contract asymmetry can offset the flow asymmetry. A continuous schedule keeps a nonzero marginal fee incentive at every performance level, so a deeply underwater manager always retains a reason to try; the 2025 contract's step-plus-floor geometry kills the marginal incentive in the deep tail, which is what produces the dead zone, the catch-up zone above it, and the state-dependent empirical content they make possible. The vintage settlement, the screening margin, the industry equilibrium under a partial mandate, and the regulator's design problem do not appear in that line.

In sum, the paper provides the first model of a penalty-heavy step contract and of vintage-contingent settlement, characterizes the industry equilibrium the partial mandate creates, and reads the two mandates, symmetry in 1970 and reverse asymmetry in 2025, as directionally consistent with two different constraint sets, illustrating how a regulator may trade off different failures in choosing a contract shape.

Section 2 documents the contract and its history, including the two voluntary failures that locate the contract's incidence on the supply side. Section 3 writes down the manager's problem and solves for the shape of his incentives. Section 4 follows the vintage structure in time and puts the industry back together under the quota. Section 5 reads the mandate's incidence across the skill distribution and interprets the observed geometry against the institutional constraints. Section 6 reports the first cohort's evidence and the dated agenda for the tests that await more data. Section 7 concludes.

2 Institutional Background

2.1 The 2025 reform

On May 7, 2025, the China Securities Regulatory Commission (CSRC) issued the *Action Plan for Promoting High-Quality Development of Publicly Offered Funds* (CSRC Document [2025] No. 21) (China Securities Regulatory Commission (中国证券监督管理委员会), 2025). One of its measures is contractual: newly established actively managed equity funds must charge a management fee that floats with performance relative to the fund's performance benchmark. The mandate came with a quota and a deadline: leading fund-management companies must make such products no less than 60% of their new active equity issuance within one year, in a one-year pilot to be evaluated and then extended gradually. The first batch of twenty-six products was approved on May 23, 2025, sixteen days after the plan.

The fee mandate did not arrive alone; two companion rules define the environment the contract lives in. First, the benchmark became a hard constraint. The *Guidelines on Performance Benchmarks of Publicly Offered Securities Investment Funds* (China Securities Regulatory Commission (中国证券监督管理委员会), 2026), released in January 2026 and effective March 1, 2026, require that a fund's benchmark match its stated investment positioning, not be changed at will, and be enforced through disclosure and correction duties. A floating fee is only as meaningful as its anchor; the guidelines supply the anchor. Second, the manager's own evaluation and pay were rewired to the same object. Under the fund-industry association's assessment rules issued in April 2026 (Asset Management Association of China (中国证券投资基金业协会), 2026), product performance must carry at least 80% of the weight in a manager's evaluation, at least 80% of that on horizons of three years or longer, and a manager who trails the benchmark by more than ten percentage points over three years takes a marked cut in performance pay. Fee schedule, benchmark discipline, and pay rule all pull on the same variable: performance relative to the benchmark.

The regulator said plainly what the reform is for; three stated motives recur in the official documents. The first is to end the "rain-or-shine fee," the fixed percentage of assets collected in good years and bad alike. The second is to turn the industry "from scale-first to return-first," that is, away from maximizing assets under management and toward what those assets earn. The third is to curb "style drift": funds marketing against one benchmark and investing against another. And the plan is explicit that the goal is not to shrink active management: it speaks in the same breath of raising equity investment and strengthening active-management capability.

One structural feature frames the industry analysis later: the obligation is partial. It covers 60% of new active equity issues at leading companies; existing funds are untouched, and the remaining 40% of new issuance may carry conventional fixed fees. Old and new contracts will coexist in the same companies, on the same distribution channels, for years; Sections 4 and 5 study the industry equilibrium this coexistence creates.

2.2 The contract as written

The prospectus of a first-batch product offers the investor not a performance fee in the hedge-fund sense: no share of profits, no high-water mark, no carry. It offers a management fee with three possible annual levels, a rule for which level applies to her money, and a settlement procedure that resolves the rule when she leaves.

The three levels are 1.5%, 1.2%, and 0.6% per year. Which one applies is determined at redemption, separately for every lot of shares, over that lot's own holding period. An investor who has held for at least one year pays the top rate of 1.5% if the fund's annualized excess return over the benchmark, net of the fee premium itself, exceeds 6% *and* her total holding-period return is positive. She pays the floor rate of 0.6% if the annualized excess return is -3% or worse. In all remaining cases she pays 1.2%. Money redeemed within one year pays 1.2% regardless of performance: the floating schedule does not apply to short money. Operationally, the fee accrues daily in the net asset value at the middle rate; at redemption the applicable rate is recomputed over the investor's actual holding window and the difference settled: over-accruals refunded, under-accruals collected. No floating revenue is final until the money leaves.

Two features of the shape organize everything that follows. The first is the tilt. The downgrade costs 60 basis points against the base tier; the upgrade pays 30. The penalty zone begins 3% behind the benchmark, the reward zone 6% ahead: the penalty is twice as heavy and starts half as far away. We call a schedule with this double tilt *penalty-heavy*. The second feature is the unit of account. The fee is not computed on the fund's aggregate performance over a calendar period, as every performance fee in the literature is; it is computed lot by lot, each lot over its own window from entry to exit. Two investors in the same fund on the same day can owe different rates. Borrowing the loan literature's term, we call this a *vintage-contingent fee*. Section 3.2 gives both terms their formal definitions; here we record what the prospectuses say.

Two smaller facts complete the picture. First-batch benchmarks are mostly broad-based indices (the China Securities Index (CSI) 300, CSI A500, CSI 500, and CSI 800 families, with a few products adding Hong Kong or bond components), so the fee is anchored, in most cases, to the market portfolio itself, not to a narrow style index a manager might quietly drift away from. And the upgrade clause's extra condition, a positive total return over and above beating the benchmark, means the reward tier recedes in falling markets while the penalty tier does not; we defer this procyclical asymmetry to the model's extensions.

2.3 What history says

This is not the first time Chinese fund companies could sell a floating management fee. Twice before, the regulator permitted it voluntarily. Twice, the market declined. The two failures, and the reasons behind them, are the best evidence about who this contract is actually for.

In December 2019, the first six voluntary floating-fee funds were approved, on hedge-fund lines: a base fee of 0.8% plus 20% of returns above an 8% absolute hurdle, reward-only, with no benchmark anchor and no penalty side. In the fourth quarter of 2023, a second wave of twenty voluntary "fee-concession" funds linked management fees variously to fund scale, to performance, or to holding period; the performance-linked variant was a three-year closed product charging a tiered base of 0.5% to 1.0% plus a 20% carry. Both waves failed commercially. An industry post-mortem put it in one sentence: the funds are small, fund companies have little incentive to market them, and investor enthusiasm is low. Neither side of the market wanted the contract when the contract was optional.

The demand side's indifference is one of the more robust facts in the fund literature, not a Chinese curiosity. Barber et al. (2005) show that mutual fund investors respond to salient fees (front-end loads, which arrive as a visible deduction) but largely ignore the ongoing management fee, invisible inside the net asset value (NAV). The 2025 contract's fee swing is 90 basis points between floor and cap, arriving as an adjustment at redemption, years after the purchase decision, against the background of equity returns whose annual standard deviation is an order of magnitude larger.

The supply side's history is the reform's immediate context. In 2020 and early 2021, star-manager "blockbuster" funds raised enormous sums at the market's peak, through mobile platform distribution that concentrates retail attention on recent winners, the convex flow–performance relation of Chevalier and Ellison (1997) and Sirri and Tufano (1998). The subsequent crash in the crowded "core assets" left a generation of retail investors locked into underwater positions. Through the whole cycle, the flat-fee revenue model was immune: assets gathered at the peak kept paying their fixed percentage through the bust. This is the concrete content of the "rain-or-shine fee" the action plan promises to end, and of the trust rupture that made ending it politically necessary. The contract arrived in this cycle's aftermath; its vintage-contingent settlement, repricing money gathered at a peak year after year if the fund then trails its benchmark, is the reform's direct answer to that cycle.

2.4 The American mirror

Fifty-five years before the CSRC's action plan, the U.S. Congress legislated the same contract, in the opposite direction. The 1970 amendments to the Investment Advisers Act of 1940 (United States Congress, 1970) effectively banned asymmetric performance fees for registered investment companies. What survived was the *fulcrum fee*: a fee that adjusts symmetrically around a benchmark, rising for outperformance and falling for underperformance by the same amounts, at the same distances.

The 1970 rule answered a specific failure. The late-1960s "go-go years" had produced performance-fee funds whose reward-only schedules gave managers option-like payoffs: heads the manager wins, tails the investor loses. The fear was of incentives too strong:

fees that paid for risk as if it were skill. Symmetry was the cure: a linear schedule centered on the benchmark pays for performance and nothing else, leaving its manager no reason to manufacture variance. In practice the permitted contract barely spread: fulcrum fees remained rare and, where adopted, shallow: adjustments of roughly 0.1 to 0.3 percentage points around the base fee, typically capped. For researchers, the episode's main yield has been as an experiment: Golec and Starks (2004) use the 1970 regulatory shock to identify how performance-fee incentives affect risk shifting.

In 1970, American regulators took a contract that could reward without penalizing and forced it to symmetry. In 2025, Chinese regulators took a contract that had been flat for a quarter century and forced it past symmetry to the other side, to a schedule whose penalty is twice the reward, at half the distance, applied lot by lot. The world's two largest fund markets have now run both experiments: symmetry by mandate, and reverse asymmetry by mandate. The two mandates give opposite answers because they answer different questions; Section 2.5 develops this claim.

2.5 Different failures, different shapes

We take the second reading. Both mandates were deliberate, and each was a rational answer to the constraints its author faced. Four differences between the two episodes organize the comparison; none is cosmetic, and together they explain why the two shapes point in opposite directions without either being mistaken.

The failure being fixed. The 1970 amendments emerged from the go-go years, where the visible abuse was an incentive too strong: bonus-only fees inviting risk shifting at investor expense. Symmetry removed the option. The 2025 reform emerged from the 2020–21 blockbuster cycle, where the visible failure was an incentive too blunt: guaranteed fees, closet indexing, style drift, and an industry that gathered assets at peaks without consequence. A penalty side was the point. Where Congress feared gambling, the CSRC feared inaction.

The investor base. American fund investors of the era were reached largely through advisor-mediated channels; the fee was part of a professional conversation. China's fund investors are mass retail: hundreds of millions of accounts, buying on mobile platforms, chasing lagged performance, holding for short horizons, and, as Section 2.3 documented, inattentive to management fees. A contract that works through investor fee-sensitivity cannot work in this market.

The regulator's objective. Congress and the SEC pursued a single goal: protect investors from incentive-induced risk taking. The CSRC's objective is compound. It must punish the rain-or-shine fee, a political and investor-protection constraint. It must not introduce new gambling incentives, the lesson of the 1970 episode. And it must keep the active equity industry alive and growing; the same action plan calls for raising equity investment and strengthening active management. A symmetric fulcrum satisfies the second constraint but lacks the first's political force; a bonus-only schedule satisfies part of the first but violates the second. A penalty-heavy schedule satisfies the first two at a cost to the third. Whether that trade is sensible, what it costs, and on whom the cost lands, is the question Section 5 takes up.

The market structure. In the U.S. case, the concern was portfolio risk inside a fund. In the Chinese case, the flow–performance relation is convex and runs through platform distribution that amplifies it: money pours into recent winners at market peaks, and money-gathering itself becomes a behavior to be governed. The 2025 contract's vintage-contingent settlement speaks to this margin directly: it determines what the money a company has already gathered is worth to the company—a margin no symmetric fulcrum fee was ever asked to govern.

The two fee shapes are not rival answers to one question; they are two constrained answers to two different failures. The question is therefore not which shape is right, but what each shape does, and what problem a regulator choosing a shape is actually solving.

2.6 Why this is a manager-side device

One question remains before any modeling can start: whose behavior is this contract meant to change? Our reading is that the contract operates on the supply side, on the fund manager's portfolio choice and the fund company's appetite for flows, not on the demand side, as a price instrument aimed at investors. Four facts already on the table make the case.

First, revealed preference: floating management fees were available voluntarily for six years before the mandate, in designs more generous to investors than the 2025 version, and the verdict was commercial failure: if the contract were a demand-side selling point, at least one of the two voluntary waves should have found a clientele. Second, salience: as Section 2.3 documented, investors do not select funds on management fees (Barber et al., 2005), and the 90-basis-point swing at stake is second-order against an equity fund's return risk. Third, the design exempts the investor by construction: money held less than a year pays the base rate no matter what, and the floating adjustment arrives at redemption, after every decision the investor could have made with it; whatever the contract disciplines, it is not the purchase. Fourth, incidence lands squarely on the manager and the company:

the fee schedule reprices their revenue against the benchmark; the companion assessment rules reprice the manager's pay against the same benchmark, with a pay-cut clause at minus ten points over three years; and the reform's three stated targets (guaranteed fees, scale-first growth, style drift) are all supply-side behaviors.

This reading determines what the model must contain. If the contract is a constraint device on managers, the right primitive is the manager's portfolio problem under this fee schedule, the problem Section 3 writes down, with the investor's flow response entering as the market's own incentive system rather than as a fee-elastic demand curve. The demand side is not absent from what follows; it is located where the evidence locates it, in flows that chase performance and ignore fees. What the repricing does to portfolios, to the industry, and to welfare, and why the regulator chose this particular repricing, is the subject of the rest of the paper.

3 The Manager's Problem

A fund promises to beat a public benchmark, and its manager controls one margin that matters: how far the portfolio departs from that benchmark. Departing is the only way to beat the benchmark, and it is also the only way to lose to it; a deviation buys two things at once, expected outperformance and tracking risk. The 2025 reform reprices exactly these two goods, and its price schedule pulls against the market's: the fee schedule, penalty-heavy and bounded, pays for alpha and taxes variance, while the market's flow–performance relation is convex and rewards variance wherever the ranking is still up for grabs. This section builds the smallest environment in which the two forces meet, one fund, one benchmark, one representative holding period, states the manager's problem and its calibration, and traces their imprint on behavior: the skill threshold below which the optimal deviation is exactly zero, the redistribution of activity toward skill among those who remain active, the flow strength at which the market drowns the contract, and the within-year risk profile that the contract's steps and floor generate.

3.1 Environment

Time is a single holding period, normalized to one year. The fund declares a public benchmark B, whose return r_B is exogenous to the manager; we normalize the benchmark's drift to zero and work throughout with the fund's excess return over the benchmark,

$$e = \alpha(x;\mu) + \sigma(x)\, z, \qquad z \sim N(0,1).$$

The manager has skill $\mu \geq 0$ and chooses a deviation $x \in [0, \bar{x}]$ from the benchmark, a monotone transform of the distance between portfolio and benchmark, with active share or tracking error serving equally well, and $x = 0$ corresponding to exact replication, the *closet indexing* of Cremers and Petajisto (2009). Two technologies map deviation into returns:

- **Alpha technology.** $\alpha(x; \mu) = \mu\, a(x)$, with $a(0) = 0$, $a'(x) \geq 0$, $a''(x) \leq 0$: deviation gives skill room to operate, at decreasing marginal product. The baseline is linear, $a(x) = x$, so μ reads directly as expected annual alpha per unit of deviation.

- **Tracking-error technology.** $\sigma(x)$, with $\sigma'(x) > 0$, $\sigma''(x) \geq 0$, and $\sigma(0) = \sigma_0 > 0$. The baseline is linear: $\sigma(x) = \sigma_0 + s\, x$.

The manager's problem is thus a location–scale choice: a deviation of x moves the distribution of e to $N(\mu\, a(x),\, \sigma^2(x))$, picking a mean and a variance together, and the technology forbids buying one without the other. Every incentive system in the model (the fee schedule, the flow relation, the risk penalty) evaluates (α, σ) differently, and the manager cannot unbundle them; this jointness is the primitive tension of the paper. The bound $\bar{x}$ collects the practical limits on deviation (mandate constraints, position limits, research capacity); the baseline $\bar{x} = 2$ corresponds to roughly 11% annual tracking error, the far tail of sustainable deviation against a declared benchmark.

This is a parsimonious location–scale technology, sufficient to study the contract. Because skill multiplies deviation, a more skilled manager buys the same expected alpha with a smaller deviation, and therefore with less tracking risk, than a less skilled one; this heterogeneity in the price of alpha is what the sorting results of the paper are built on: the activity thresholds and cross-contract ordering of Proposition 1, and the industry-level endorsement reversal of Section 4.6 tested in Section 6.2.

Assumption A1. *($\sigma_0 > 0$).* Even a perfect replicator exhibits positive realized tracking error.

Cash drag, sampled replication, subscription and redemption shocks, and lags in benchmark rebalancing push closet indexers' realized tracking error to roughly 1%–3% per year, far above the tens of basis points of explicit index funds. The assumption is empirical, not technical, and it carries one boundary: as $\sigma_0 \to 0$ the contract's thresholds sit infinitely many standard deviations away and the contract loses all bite near $x = 0$ (Appendix D).

Two remarks complete the environment. First, μ is the manager's *perceived skill* throughout: contracts settle on realized returns, but choices are made under beliefs, and the skill-belief extension of Appendix A shows the model with skill beliefs is isomorphic to the baseline with adjusted moments, so nothing below leans on self-knowledge. Second, the baseline treats a single representative holding window whose investor clears the one-year gate defined below; investors who entered at different dates sit in different evaluation windows, and that *vintage* structure is the subject of Section 4.

3.2 The contract

The contract we model is the fee schedule that the China Securities Regulatory Commission required of newly issued active equity funds in May 2025, as written in the prospectuses of the first batch of twenty-six products. Its shape has three features that matter, and each survives formalization exactly: the fee floats with performance relative to the benchmark; the floating is asymmetric: the penalty for falling behind is twice the reward for pulling ahead, and it kicks in at half the distance; and the fee is computed separately for every lot of shares over that lot's own holding period, not for the fund as a whole. The first two features define a one-period pricing rule $f(\cdot)$; the third defines how the rule is applied across investors and time.

The fee schedule. An investor who has held her shares for at least one year is charged an annual management fee that depends on her annualized excess return e over the entire holding period:

$$f(e) = \begin{cases} f_H, & e \geq h \quad \text{(upgrade; requires the total holding-period return to be positive)} \\ f_M, & -\ell < e < h \quad \text{(base tier)} \\ f_L, & e \leq -\ell \quad \text{(downgrade)} \end{cases}$$

The institutional calibration is $f_L = 0.6\%$, $f_M = 1.2\%$, $f_H = 1.5\%$, with thresholds $\ell = 3\%$ and $h = 6\%$. An investor who redeems before one year is charged the base tier f_M regardless of performance: the gate exempts short money from the floating schedule and guarantees that every exposed lot spans a full evaluation window, long enough for the realized excess return to carry signal about skill rather than a fortnight's noise. Two parameters describe the schedule's tilt: the *reward jump* $u \equiv f_H - f_M > 0$ and the *penalty jump* $d \equiv f_M - f_L > 0$, institutionally $u = 30\,bp$ and $d = 60\,bp$. The asymmetry comes in two pieces that do different work:

- **Jump asymmetry**: $d > u$. The downgrade costs the manager $60\,bp$; the upgrade pays only $30\,bp$.

- **Threshold asymmetry**: $\ell < h$. The penalty zone begins 3% behind the benchmark; the reward zone begins 6% ahead.

Definition. *(penalty-heavy asymmetry).* A fee schedule is *penalty-heavy* if both asymmetries tilt the same way: $d > u$ and $\ell < h$.

The institutional contract has $d = 2u$ and $h = 2\ell$: the penalty side is twice as steep and twice as close. The two asymmetries are not redundant: the jump asymmetry governs how much a marginal unit of tracking risk costs near the origin, the threshold asymmetry governs where in the performance distribution the contract starts to bite. A symmetric fulcrum fee ($d = u$, $\ell = h$) and a bonus-only schedule ($d = 0$) serve as the two natural benchmarks, and we use both. The baseline omits the upgrade clause (a positive total holding-period return, over and above beating the benchmark), since the benchmark's drift is normalized to zero and the clause binds only when the benchmark falls; its procyclical content is deferred to the extensions.

Vintage-contingent billing. The third feature is administrative and turns out to be theoretically deep. The fee accrues daily in the NAV at the base rate f_M and is *settled* lot by lot: when an investor redeems, her applicable rate is recomputed as $f(e_{t,\tau})$, where $e_{t,\tau}$ is the annualized excess return over her own holding window $[t, \tau]$, and the difference between what she accrued and what she owes is refunded or collected at exit. We call this a **vintage-contingent fee**: each vintage of shares, indexed by its entry date t, carries its own evaluation window and its own realized fee, and the fund company never books the accrual as final revenue; it books a stream of contingent claims, one per lot, that resolve at each lot's exit. The design removes the free-rider problem of pooled performance fees, and it gives the regulator an instrument that acts on the fund company's *flow* incentive rather than on the manager's portfolio choice, since new money enters at f_M and only later discovers its tier. For the rest of this section we study one representative lot whose holding period clears the gate, writing e for its excess return; the aggregation over vintages is Section 4.

3.3 The manager's objective

The manager faces two incentive systems, not one. The fee schedule is the explicit one, written by the regulator; the implicit one is written by the market: investors chase performance, so this year's return determines next year's assets, and the flow–performance relation is famously convex (Chevalier and Ellison, 1997; Sirri and Tufano, 1998). The fee schedule, being penalty-heavy, dislikes variance; the convex flow relation loves it. The manager's payoff is explicit fee revenue, plus the discounted value of the flows that performance attracts, minus the cost of bearing return risk, minus the cost of deviating from the benchmark:

$$\max_{x \in [0,\bar{x}]} U(x;\mu) \;\equiv\; E[f(e)] \;+\; \Lambda\, E[Q(e)] \;-\; \frac{\rho}{2}\,\sigma^2(x) \;-\; C(x).$$

The first term, $E[f(e)]$, is the expected per-asset fee under the schedule of Section 3.2: because e is normal and f is piecewise constant, it is the step schedule convolved with a Gaussian kernel, whose closed form is derived below; all we need here is that $E[f]$ increases in expected alpha and, under the penalty-heavy calibration, decreases in tracking error.

The second term prices the flow incentive. $Q(e)$ is next period's assets under management (AUM) multiplier and $\Lambda\, E[Q(e)]$ the discounted present value of the future fee income this performance vintage attracts. The convexity of the flow–performance relation is one of the most replicated facts in the fund literature (Chevalier and Ellison, 1997; Sirri and Tufano, 1998); its source is retail attention: inflows concentrate on the top of the ranking table, outflows respond only weakly to poor performance. The same estimates discipline the functional form: the relation is estimated in ranks, ranks are bounded, and the convexity cannot extrapolate past the top of the table; the participation-cost model of Huang et al. (2007) supplies the microfoundation, a finite pool of potential investors that exhausts as performance improves. We adopt the logit saturation form

$$Q(e) = \frac{e^{ke}}{e^{ke} + c},$$

with k the flow semi-elasticity and c the saturation constant, both calibrated in Section 3.5. The form is convex at low performance (the Chevalier–Ellison region where inflows chase winners) and concave near saturation, where the market runs out of new money to chase with. The weight $\Lambda \equiv \lambda f_M Q_0$ converts attracted money into fee income: Q_0 the normalized base scale of the incremental assets a vintage attracts, f_M the base fee they pay, λ the number of discounted fee-years the attracted money stays. The exponential form $Q(e) = Q_0 e^{ke}$, which yields closed-form expectations but an unbounded variance appetite, is retained in Appendix C as a local approximation; every quantitative statement in the main text uses the logit form, with Λ matched to the exponential benchmark's marginal incentive at the center of the calibrated return distribution (Section 3.5).

Assumption A2. *(the mean–variance reduction).* The manager prices the risk in his payoff $W = f(e) + \Lambda Q(e)$ by a local linearization around $E[e]$: $\mathrm{Var}(W) \approx \omega^2\,\sigma^2(x)$ with

$\omega = dW/de$, and the coefficient ρ absorbs ω^2 together with career and contract costs.

The third term, $\frac{\rho}{2}\sigma^2(x)$, prices the risk the manager bears through his own payoff (Assumption A2). We are explicit about what ρ contains: the variance of fee and flow income, career and contract costs (including the reform's own clause cutting the pay of managers who trail the benchmark by ten points over three years) and whatever pure risk preference remains; it is a reduced-form price of bearing performance risk, not a psychological primitive, held constant across the performance state because the career-cost component plausibly *rises* in deep losses, offsetting any collapse of pure risk tolerance there. The exact CARA–normal problem, which prices the lottery in W without linearization, is solved numerically in Appendix B; the qualitative directions of our main results survive, and we flag the one comparative static that does not.

The fourth term is the deviation cost, $C(x) = c_1 x + \frac{c_2}{2}x^2$ with $c_1 = 5\,bp$ and $c_2 = 80\,bp$: the linear term captures the running marginal cost of an active research process, and the quadratic term embeds the scale-times-activity trading cost of Pástor et al. (2020). Writing the cost explicitly, rather than folding it into the alpha technology, is a deliberate identification choice (Section 3.4). We maintain one technical assumption.

Assumption A3. *(regularity).* U is strictly concave in x over the calibrated region, so first-order conditions characterize the global optimum, and the corner solution $x^* = 0$ is characterized by $U_x(0;\mu) \leq 0$.

Concavity is not free: the expected fee is a Gaussian-smoothed step and the flow term is convex in parts of the state space, so we treat Assumption A3 as a regional statement, verified numerically over the calibrated region, with an explicit sufficient lower bound on c_2 reported in the appendix. Where concavity fails, the model predicts jumps to the corner $\bar{x}$; this failure is itself economic content, not a mathematical nuisance.

3.4 Modeling choices

Four choices in this section shape everything downstream; we state each with its reason and its cost.

The novelty budget of this paper is spent entirely on the contract's shape; none of it is spent on preferences. Fee revenue and the flow incentive enter one objective together (Sheng et al., 2014; Basak et al., 2007; Dangl et al., 2008, at the company level), and the flow term reads as the discounted value of future fee income (Goetzmann et al., 2003; Dangl et al., 2008). These precedents price risk with explicit constant-relative-risk-aversion (CRRA) or CARA utility; we use the mean–variance penalty instead (the quadratic approximation of the CARA–normal problem) to keep every force in the objective separately interpretable: fee, flow, risk, cost. Its cost is real: the form absorbs the fee schedule's own lottery variance into ρ, treating all contracts symmetrically; Appendix B solves the exact problem, and we flag the one affected result at the point of statement.

The flow side of the model is where the contract can lose, so we deliberately give the market its strongest case. The exponential form is the exact counterpart of the empirical literature's log-linear flow regressions (a flow rate log-linear in performance is an AUM multiplier exponential in it), but it lets a star fund's inflows, and with them the manager's appetite for variance, grow without bound, while estimated flow relations are convex in *ranks* and ranks are bounded. The logit saturation form respects that bound: convex where chasing is fierce, concave where the market is full. The functional family has precedent by analogy: in the demand system of Koijen and Yogo (2019), portfolio weights take exactly this logistic form; their object is asset demand, not fund flows, and we borrow the shape as an analogy, not as a flow estimate. The choice is not neutral in level: saturation shrinks the flow incentive's slope at the calibration center by a factor of about four, and Λ is rematched to restore the benchmark incentive (Section 3.5), but it is conservative in direction: anything the fee schedule survives against a bounded flow incentive, it survives against an unbounded one.

One could fold the cost of activity into the alpha technology and write a net-of-cost production function; we do not. The empirical question this paper feeds is whether the contract changes behavior *beyond* the mechanical cost channel of Pástor et al. (2020), and the two channels must be separable in the model to be separable in the data. The quadratic term c_2 carries the Pástor et al. (2020) cost explicitly; every effect we attribute to the contract is net of the mechanical channel by construction, not by assertion.

Finally, vintage-contingency is deliberately shelved after Section 3.2: this section studies a single lot whose holding period clears the gate, because the manager does not choose x vintage by vintage. The company's aggregation problem, where the vintage structure earns its keep, is Section 4.

3.5 Calibration

The propositions below are proved in symbols; every number attached to them is computed under one baseline calibration, collected in Table 1. Contract parameters come from the prospectuses of the first batch of twenty-six products; the rest are chosen to span the empirically relevant range of Chinese active equity funds and are scanned in Appendix D.

Because the model mixes objects measured in different units, Table 2 collects each

Table 1: Baseline calibration.

Parameter	Value	Meaning	Source / basis
$f_L,\ f_M,\ f_H$	0.6%, 1.2%, 1.5%	fee tiers	prospectuses of the first twenty-six products (Section 3.2)
$\ell,\ h$	3%, 6%	downgrade / upgrade thresholds	prospectuses
$u,\ d$	30 *bp*, 60 *bp*	reward jump, penalty jump	implied by the tiers
σ_0	1% (scanned 0.5%–2%)	residual tracking error of a closet indexer (Assumption A1)	the empirical range is roughly 1%–3% per year
$s,\ \bar{x}$	5%, 2	slope of the tracking-error technology and the deviation bound; one unit of deviation adds 5% of tracking error, the cap corresponds to roughly 11% (Section 3.1)	a units convention, not an estimate; the cap marks the far tail of sustainable deviation
k	8	flow semi-elasticity	platform-era Chinese evidence, evidence band $[4.5, 10.5]$ (below)
c	1	saturation constant	normalization plus symmetry: at $c = 1$ the flow term's variance appetite vanishes exactly at the calibration center; scans over $c \in \{0.5, 1, 2\}$ are robust (Appendix D)
Λ	776.8 *bp*	flow weight	incentive-equivalence matching to the 180 *bp* anchor (below)
ρ	5	risk penalty	reduced-form (Assumption A2), validated against the exact CARA–normal problem of Appendix B
$c_1,\ c_2$	5 *bp*, 80 *bp*	deviation cost, linear research cost plus quadratic term	the quadratic term embeds the scale-times-activity trading cost of Pástor et al. (2020)
μ	0–10%	perceived annual alpha per unit of deviation	the U.S. industry-average gross alpha is estimated near zero (Berk and van Binsbergen, 2015; Carhart, 1997); Chinese outperformance lacks persistence (Lin et al., 2009); the grid's center of mass sits at the low end

Table 2: Units map.

Object	Units	Interpretation
x	normalized, benchmark deviation	units of departure from the benchmark; one unit adds 5% of tracking error
μ	annual return per unit x	perceived alpha productivity: expected annual alpha the manager believes each unit of deviation delivers
σ	annual return	tracking volatility, residual plus deviation-induced
$u,\ d$	fee rate	reward and penalty jumps of the fee schedule
$\ell,\ h$	excess return	downgrade and upgrade thresholds
Λ	fee-equivalent AUM multiple	discounted value of the flows a vintage attracts, in units of base-fee income

object's unit and economic reading before the propositions attach numbers to it.

Readings quoted in the text, such as the activity threshold's "2.15% of annual alpha per 5% of tracking error," are economic translations of the model's thresholds into these units (the reading of μ^* through the units of x and μ in Table 2), not the model's raw units.

Two calibrations need more than a table row. The flow semi-elasticity is calibrated to Chinese evidence. In the platform era (2013–2017), the flow–performance sensitivity of Chinese active equity funds was roughly 3–3.5 times its pre-platform level, with top-decile funds attracting quarterly net inflows of about 20% of total net assets (TNA) (Hong et al., 2025). Applying this multiplier to the top-tail flow–performance slopes estimated for the U.S. (Huang et al., 2007; Sialm et al., 2015) implies an annual log-slope of the AUM multiplier with respect to excess returns of approximately 4.5–10.5, and the bull-market coefficient in Xiao (2013) brackets the same range once ranks are converted into return units. We set $k = 8$, at the aggressive end of the implied band; sensitivity at $k = 4$ is reported in Appendix D.

The flow weight $\Lambda = 776.8\,bp$ is matched, not picked. The anchor is the exponential benchmark $\Lambda = 1.5\,f_M \approx 180\,bp$, corresponding to incremental money staying about one and a half discounted fee-years; the matching principle is incentive equivalence at the center of the calibrated return distribution, $e \sim N(0, \sigma)$ with $\sigma = 4\%$: $\Lambda \cdot E[kQ(1-Q)] = \Lambda_{\text{exp}} \cdot k\, e^{k^2\sigma^2/2}$ (Lemma A.1). At $k = 8$ saturation shrinks the slope by a factor of 4.315, and the rematched $\Lambda = 180 \times 4.315 = 776.8\,bp$ restores the benchmark incentive exactly. The anchor survives a sanity check: platform-era turnover implies incremental money stays about 1.25 years, worth $\approx 178\,bp$ of fee income at the pre-reform rate with mild discounting; pre-platform turnover implies an anchor five to eight times larger, so $180\,bp$

is, for our purposes, conservative. When k changes, Λ is rematched under the same principle (halving k to 4 lowers Λ only to $733.9\,bp$, because the Jensen corrections partly offset), and the $c = 1$ symmetry on which the closed-form results rest holds for every $k > 0$, so the result set has no boundary in k.

3.6 The activity threshold and risk compression

A manager who deviates from the benchmark must cover, per unit of deviation, the research and trading cost $C(x)$, the risk penalty, and a third price the 2025 contract introduces. Because the penalty tier sits closer and heavier than the reward tier, a deviation that produces only noise is charged for the variance it creates and paid nothing for the mean it does not move: for a low-skill manager, activity is a negative-expectation gamble before costs. A sufficiently unskilled manager therefore does not deviate at all; he is a closet indexer by arithmetic, not by career concerns or herding. The logic runs through the expected fee, whose shape we record first.

Lemma 1. *(the contract prices the information ratio). Under the fee schedule of Section 3.2 and $e \sim N(\alpha, \sigma^2)$, writing φ and Φ for the standard normal pdf and cdf and $\bar{\Phi} \equiv 1 - \Phi$ throughout,*

$$E[f(e)] = f_M + u\,\bar{\Phi}\left(\frac{h-\alpha}{\sigma}\right) - d\,\Phi\left(\frac{-\ell-\alpha}{\sigma}\right),$$

and the floating part $V(\alpha, \sigma) \equiv E[f] - f_M$ satisfies

$$\text{(i)}\quad \frac{\partial E[f]}{\partial \alpha} = \frac{1}{\sigma}\left[u\,\varphi(t_U) + d\,\varphi(t_D)\right] > 0;$$
$$\text{(ii)}\quad \left.\frac{\partial E[f]}{\partial \sigma}\right|_{\alpha=0} = \frac{1}{\sigma^2}\left[uh\,\varphi\left(\tfrac{h}{\sigma}\right) - d\ell\,\varphi\left(\tfrac{\ell}{\sigma}\right)\right],$$

where $t_U = (h-\alpha)/\sigma$, $t_D = (-\ell-\alpha)/\sigma$. If the schedule is penalty-dominant ($d\ell \geq uh$ and $h > \ell$), then (ii) is strictly negative for every $\sigma > 0$: at the center of the distribution, variance is strictly bad.

Proof in Appendix A. The institutional calibration sits exactly on the boundary of the dominance condition ($d\ell = 60\text{bp} \times 3\% = uh = 30\text{bp} \times 6\%$): the weak inequality holds with equality, and part (ii) is strict nevertheless, because with $h > \ell$ the two Gaussian densities differ. Away from $\alpha = 0$ the sign of the variance derivative can flip, since it is a race between two exponentials: a high-alpha manager whose distribution has cleared the penalty zone finds that variance mostly buys him upgrade probability. A grid scan finds the derivative negative in 81.2% of the calibrated (α, σ) region, the positive cells concentrated in that corner (Appendix D), the same force that amplifies the high-skilled manager's activity in Proposition 1(b) below.

Lemma 1 says the contract prices the information ratio, not deviation: alpha is always rewarded, and under penalty dominance variance is always taxed at the center. This asymmetry has a level consequence:

Lemma 2. *(the break-even alpha). Suppose $h > \ell$ and $d \geq u$. For each $\sigma > 0$ there is a unique $\bar{\alpha}(\sigma) > 0$ at which $E[f] = f_M$, and $E[f] < f_M$ below it.*

Proof in Appendix A. Calibrated values: $\bar{\alpha} = 1.77\%, 2.35\%, 3.06\%, 3.83\%$ at $\sigma = 2\%, 4\%, 6\%, 8\%$. A fund running 4% tracking error must expect more than 2.35% of annual alpha just to expect the base tier, and the active industry's typical delivered alpha sits well below that level: the U.S. average is estimated near zero (Berk and van Binsbergen, 2015; Carhart, 1997), and in China individual outperformance shows no persistence (Lin et al., 2009; Zhuang and Tang, 2004). The median manager's expected fee under the new schedule therefore lies below the 1.2% he used to collect for certain. The reform's transfer is not symmetric belt-tightening: it ends the guaranteed fee specifically for managers whose deviation does not produce alpha, small per fund (between 0.08 and 1.22 basis points of expected fee, positive for all μ below about 6%) but concentrated, moving from low-skill managers to their investors.

Write the manager's first-order condition at the corner $x = 0$. Differentiating the objective of Section 3.3 and evaluating at $(\alpha, \sigma) = (0, \sigma_0)$ gives

$$U_x(0;\mu) = \underbrace{\mu\, a'(0)\, [V_\alpha + \Lambda\, E[Q'(\sigma_0 z)]]}_{\text{skill channel } (>0)} + \underbrace{\sigma'(0)\, [V_\sigma + \Lambda\, \sigma_0\, E[Q''(\sigma_0 z)] - \rho\sigma_0]}_{\text{variance channel (sign indeterminate)}} - c_1,$$

where $V_\alpha \equiv \frac{1}{\sigma_0}[u\,\varphi(h/\sigma_0) + d\,\varphi(\ell/\sigma_0)] > 0$ and $V_\sigma \equiv \frac{1}{\sigma_0^2}[uh\,\varphi(h/\sigma_0) - d\ell\,\varphi(\ell/\sigma_0)] < 0$ are the fee schedule's skill price and variance price at the origin (Lemma 1), and the flow moments reduce to Λk and $\Lambda k^2 \sigma_0$ under the exponential form of Appendix C. The three components of the variance channel are the three forces of the model in miniature: the contract's variance tax, the market's variance appetite ($\Lambda \sigma_0 E[Q''] > 0$ in the convex region), and the manager's own risk penalty.

Proposition 1. *(the activity threshold and risk compression). Maintain Assumptions A1 and A3.*

(a) (Extensive margin.) $x^*(\mu) = 0$ *if and only if* $\mu \le \mu^*$*, where*

$$\mu^* \equiv \frac{c_1 - B}{A}, \qquad A \equiv a'(0)\left[V_\alpha + \Lambda E[Q'(\sigma_0 z)]\right] > 0,$$
$$B \equiv \sigma'(0)\left[V_\sigma + \Lambda\sigma_0 E[Q''(\sigma_0 z)] - \rho\sigma_0\right],$$

provided $c_1 > B$ *(otherwise* $\mu^* < 0$ *and every skill level is active). The floating contract raises the threshold relative to a flat fee* ($\mu^*_{\mathit{float}} > \mu^*_{\mathit{flat}}$) *if and only if the contract's variance tax at the origin outweighs its skill reward evaluated at the flat-contract marginal manager:*

$$\sigma'(0)\,|V_\sigma| \;>\; \mu^*_{\mathit{flat}}\, a'(0)\, V_\alpha.$$

(b) (Intensive margin.) Suppose the flat- and floating-contract problems both have interior solutions. Then $x^*_{\mathit{float}}(\mu) \gtreqless x^*_{\mathit{flat}}(\mu)$ *according as* $\mu\, a'(x^*_{\mathit{flat}})\, V_\alpha \gtreqless \sigma'(x^*_{\mathit{flat}})\, |V_\sigma|$*, and there exists a cutoff* $\tilde{\mu}$ *such that the floating contract compresses the activity of managers with* $\mu < \tilde{\mu}$ *and amplifies the activity of managers with* $\mu > \tilde{\mu}$*.*

(c) (The two convex benchmarks.) Suppose the flat-contract problem has an interior solution. (i) If the flat optimum's alpha lies below the upgrade line, $\alpha(x^*_{\mathit{flat}}) < h$*, the bonus-only schedule strictly raises activity above the flat benchmark:* $x^*_{\mathit{bonus}}(\mu) > x^*_{\mathit{flat}}(\mu)$*; a one-sided upside schedule pays for variance as well as for skill. (ii) If moreover the upgrade line sits more than one residual standard deviation above the flat optimum's alpha,* $(h - \alpha(x^*_{\mathit{flat}}))/\sigma(x^*_{\mathit{flat}}) > 1$*, the symmetric fulcrum schedule does the same:* $x^*_{\mathit{sym}}(\mu) > x^*_{\mathit{flat}}(\mu)$*. The ordering between the two convex schedules is unsigned: symmetrizing the bonus schedule adds a skill reward and a variance tax of equal size.*

Proof in Appendix A. Part (a) solves the corner condition $U_x(0;\mu) \le 0$ of Assumption A3 for μ; strict concavity makes the first-order comparison global. For part (b), the floating objective adds the fee term $V(\alpha(x), \sigma(x))$ to the flat-fee problem, with marginal $dV/dx = \mu a'(x) V_\alpha + \sigma'(x) V_\sigma$. Evaluated at the flat-fee optimum x^*_{flat}, the sign of this term governs the direction of the adjustment: a manager expands activity if the marginal skill reward exceeds the variance tax, and contracts it otherwise. Appendix A establishes that this trade-off yields a unique crossing $\tilde{\mu}$ under a stated sufficient condition, verified numerically over the calibrated region. Part (c) applies the same sign argument to the two convex benchmarks: a one-sided upside schedule prices variance positively, and the symmetric schedule does so exactly when its upgrade line sits more than one standard deviation above the flat optimum's alpha.

The intuition for condition (a) is a trade-off between two implicit prices at the origin:

the first step away from the benchmark generates tracking variance at rate $\sigma'(0)$, taxed at unit price $|V_\sigma|$, and expected alpha at rate $a'(0)$, rewarded at unit price V_α and scaled by the marginal manager's skill μ^*_{flat}. The reform raises the barrier to active management whenever the tax on pure noise outweighs the reward for marginal skill: μ^* rises in the penalty jump d and as the reward line h recedes, and falls in the reward jump u and as the penalty line ℓ recedes. Under the baseline calibration the inequality holds by an order of magnitude: at $\sigma_0 = 1\%$ the downgrade line sits three residual standard deviations away, where the Gaussian tail is still fat, while the reward line sits six standard deviations out, so the variance tax is large and the skill reward near the origin negligible. The comparative statics of μ^* are therefore governed almost entirely by the penalty parameters: the reward tier shapes how active managers invest, but the penalty tier dictates who gets to be active at all.

Calibration. Figure 1 plots $x^*(\mu)$ under the four contracts; each curve lifts off zero at the contract's activity threshold, the extensive margin. One unit of deviation is 5% of tracking error, so the floating threshold $\mu^*_{\text{float}} = 2.15\%$ reads as follows: a manager takes the first step away from the benchmark only if he believes he delivers about 2.15% of annual alpha per 5% of tracking error, an information ratio near 0.4 before costs. The flat-fee threshold is 1.93%. Managers between the two thresholds are active under the flat fee and closet indexers under the new contract, and because both thresholds sit far above the industry's typical delivered alpha, the extensive margin binds the median manager, not the tail.

Three further readings; the full grid is in Appendix D. First, the contract's thresholds live in units of residual noise (Assumption A1): at $\sigma_0 = 0.5\%$ they sit too many standard deviations away and the schedule loses all marginal bite, while the reform bites hardest at σ_0 of 1%–2%, where the flat–float gap widens to 22–73bp. Second, the gap is not monotone in the flow strength: at $\sigma_0 = 2\%$ and half-strength flows it nearly closes (7.13% against 7.12%), and at $k = 4$ the ordering in that corner reverses, an economically sensible configuration since a penalty line already in the money leaves the floating contract no extra participation advantage (Appendix D.4). Third, doubling the flow incentive shrinks both thresholds and the gap between them, which is Proposition 2 in preview: the contract's bite is inversely related to the heat of the market.

The extensive-margin prediction is not the cost channel of Pástor et al. (2020) restated. That channel runs through scale and is silent about small funds; the contract's threshold exists net of costs by construction, so small funds launched under the new contract should closet at the same skill threshold as large ones. The cost channel cannot produce that

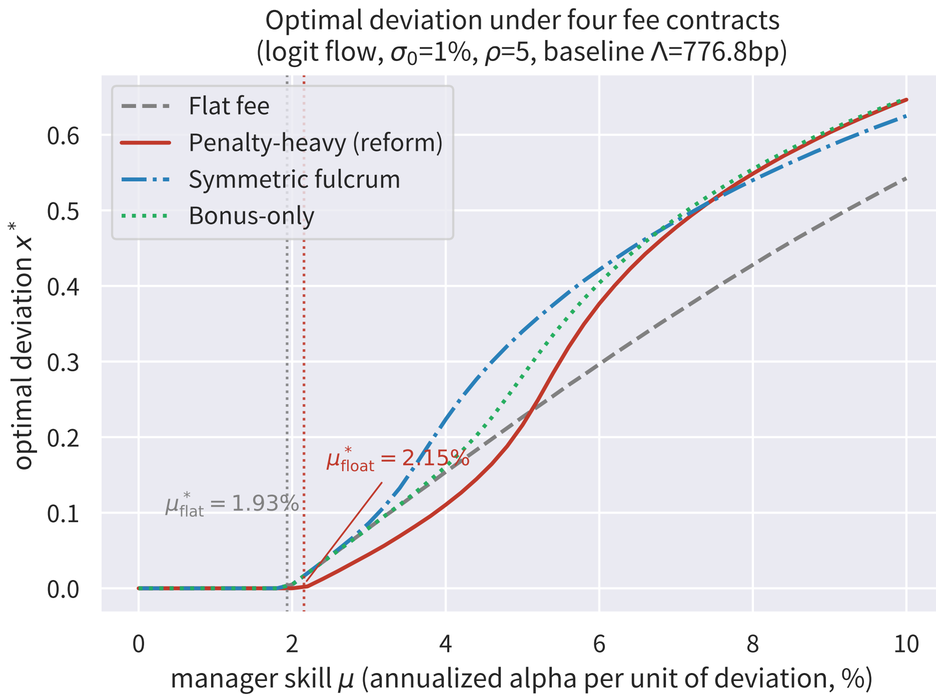

Figure 1: Optimal deviation $x^*(\mu)$ under the penalty-heavy contract, a flat fee, a symmetric fulcrum step, and a bonus-only schedule. Each curve lifts off zero at the contract's activity threshold of Proposition 1(a), marked for the flat fee and the reform contract; the vertical gaps between curves are the intensive margin of Proposition 1(b).

prediction, and it is the first handle the empirical section uses to separate the two channels.

Turn to managers who remain active. Under the floating contract, deviation moves the downgrade probability through two channels: the alpha channel shifts the return distribution right and escapes the zone, while the variance channel spreads mass into it. High-skill managers dominate the first channel, low-skill managers the second, so the penalty-heavy contract splits the active population, compressing the low-skilled and amplifying the high-skilled. We call the first half *risk compression.* Proposition 1(b) states the two halves together because they are one mechanism, the same two prices read at different skill levels.

The penalty comparative static takes the same form. A heavier penalty jump d lowers an interior manager's optimal deviation if and only if $\alpha'(x^*)\,\sigma(x^*) < \sigma'(x^*)\,(\ell + \alpha(x^*))$, which under the linear technologies collapses to the single skill comparison $\mu < \hat{\mu} \equiv s\ell/\sigma_0$. At the baseline calibration $\hat{\mu} = 15\%$, far above any plausible skill, so the comparative static is one-sided over the empirically relevant region: a heavier penalty compresses everyone. (The one-sidedness is a calibration statement: at $\sigma_0 = 2\%$ the cutoff falls to

$\hat{\mu} = 7.5\%$; the claim is stated for the baseline region $\sigma_0 \in [0.5\%, 1\%]$.)

Run across contracts rather than across parameters, the same logic delivers the ordering that organizes Figure 1:

$$x^*(\text{penalty-heavy}) \;<\; x^*(\text{flat}) \;\ll\; x^*(\text{symmetric fulcrum}) \;\approx\; x^*(\text{bonus-only}),$$

read over the bulk of the skill distribution below the cutoff $\tilde{\mu}$. The display mixes theorem and calibration. The first inequality is Proposition 1(b); the signs on the right are Proposition 1(c), under which the ordering between the two convex schedules is itself unsigned. What lies outside the theorem is magnitude: the wide gap marked $\ll$ and the near-tie marked $\approx$ are calibration readings. The first inequality is the result that matters: the penalty-heavy contract is the only schedule that drives low-skill activity below the flat-fee benchmark, and in the calibration it drives it to zero while the other three leave the same manager marginally active. Above the cutoff $\tilde{\mu} \approx 5.1\%$ (stable across $\sigma_0 \in \{0.5\%, 1\%, 2\%\}$; 8.1% at $k = 4$, the crossing structure unchanged; Appendix D.4) the first inequality reverses, the amplification half of Proposition 1(b), while the float contract remains below the bonus-only schedule throughout and overtakes the symmetric fulcrum near $\mu \approx 7.4\%$. The equilibrium activity distribution is reallocated toward skill: pseudo-active management at the bottom is eliminated into closet indexing, genuinely active management at the top is retained and, within the mean–variance benchmark, strengthened.

Remark. *(the amplification side is mean–variance-specific).* The two halves of Proposition 1(b) do not have the same epistemic status. The mean–variance formulation absorbs the fee schedule's own lottery variance (the variance of landing in one tier or another) into the reduced-form coefficient ρ, treating all contracts symmetrically. But that lottery variance is a contract-specific risk channel: a more active manager under the floating contract holds a fee lottery with real spread between tiers, and a manager who prices it exactly finds the floating contract less attractive at high activity than the mean–variance reduction suggests. The exact CARA–normal problem of Appendix B confirms this: for effective risk aversion above roughly 3, the crossing disappears within the plausible skill range and $x^*_{\text{float}} \leq x^*_{\text{flat}}$ holds everywhere: the amplification side is reversed once the fee lottery's own variance is priced, while the compression side survives in every specification examined. The compression half is therefore the robust prediction, and the crossing that separates the two halves is itself an economic observation: it locates the skill level at which the fee lottery's own variance, concealed by the mean–variance form, becomes large enough to reverse the comparison.

Remark. *(belief dispersion compresses through the intensive margin).* The skill-belief extension of Appendix A shows that when the manager holds beliefs $\hat{\mu}$ about his own skill with uncertainty τ, the model is isomorphic to the baseline with a larger effective variance, and the effect operates through the intensive margin: at $\mu = 3\%$, raising τ from 0 to 3% compresses x^* by about 17% while the activity threshold is exactly invariant. Young managers, whose skill is least known, should therefore compress fastest after the reform, a prediction testable in tenure groups, and one that again separates the contract channel from any story in which the response scales with fund size.

The ordering has a fee-revenue shadow, and it is where the reform's incidence lands. In the logit calibration, managers at $\mu = 2\%$–5% expect 118.8–119.9bp under the floating contract, below the 120bp they would collect for certain under the flat fee, and only around $\mu = 8\%$ does a premium show (128.5bp, earned as an interior optimum at $x^* = 0.548$). The expected transfer is concentrated on the low-skilled and skips the high-skilled: the incidence statement and the activity statement of Proposition 1(b) are two readings of one fact: the contract pays for information ratio, so it taxes exactly those who have less of it.

3.7 Concave contract, convex market

Both halves of Proposition 1 take the market's flow incentive as given; we now bring the market in. The two incentive systems the manager faces have opposite curvature. The contract is *concave*: the fee schedule is a bounded, saturating function of performance, with marginal bite concentrated near the thresholds and decaying to zero deep in the tails (Section 3.8 makes this precise). The market is *convex*: the flow–performance relation rewards variance wherever the ranking is still up for grabs, and its strength scales with Λ without bound. The question that disciplines the whole exercise is when the market's convex pricing drowns the regulator's concave contract.

Proposition 2. *(flow swamping). Fix all other parameters. As $\Lambda \to \infty$,*

$$|x^*_{\mathit{float}}(\mu) - x^*_{\mathit{flat}}(\mu)| \longrightarrow 0 \qquad \textit{and} \qquad \mu^*_{\mathit{float}} - \mu^*_{\mathit{flat}} \longrightarrow 0.$$

Proof in Appendix A. The contract term in the first-order condition is bounded (the Gaussian density is bounded and $\sigma \geq \sigma_0 > 0$) while the flow term grows without bound in Λ; dividing the first-order condition by Λ annihilates the contract asymptotically. The argument uses nothing about the flow function beyond growth in Λ, so the limit is form-free: it holds for the exponential and logit forms alike.

We call this flow swamping: when flow revenue is large enough relative to the fee at stake, the fee schedule's incentive content is diluted to nothing. The speed of the dilution is computable in closed form under the logit specification: at $c = 1$ the flow term's symmetry at the calibration center makes both thresholds rational functions of Λ, and the gap decays exactly like $1/\Lambda$. On the grid, the threshold gap is 22.0bp at baseline flow, half that at 2.2 times baseline, a quarter at 4.6, and a tenth at 12.0; both thresholds stay strictly positive at every finite Λ, so there is no finite death point: universal participation is the limit, not a crossing.[2]

What does not vanish is the last bit of bite. The expansion $g(\Lambda) = C_1'/\Lambda + C_2'/\Lambda^2 + O(\Lambda^{-3})$ (Appendix A) has leading constant $C_1' = -s\,V_\sigma/E[Q'] = 2.0 \times 10^{-4} > 0$, whose sign is exactly the penalty-dominance condition of Lemma 1: heat decides how much bite remains, never its sign. The intensive margin does not move monotonically as the market heats; flow amplification pushes both optima up the upgrade side, so the compression region first shrinks and then disappears.

Remark. *(the convergence is smooth).* Under the logit flow function the intensive-margin half of Proposition 2 is achieved exactly by the two optima gliding together. Both contracts' optima drift upward at an approximately logarithmic rate as flow strength rises, their behavioral difference disappears well before either reaches the deviation bound, and the corner itself is reached smoothly: at about 10 times baseline flow for a $\mu = 5\%$ manager, 15 times at $\mu = 3\%$, 22 times at $\mu = 2\%$. The economics of hot-money distortion does not disappear; it migrates to the within-year margin of Section 3.8, where the catch-up gambling zone and the reward peak survive under every flow form examined (Appendices A and C). Flow swamping does not restore the pre-reform equilibrium. It replaces one distorted margin with another.

The practical question is where real markets sit on this curve. The baseline $\lambda \approx 1.5$ discounted fee-years is conservative: incremental money that stays three to five fee-years corresponds to 1.7–2.7 times baseline flow, a range in which the threshold gap retains about half its bite. The dangerous case is joint inflation: in hot-money episodes the flow semi-elasticity k rises together with Λ, and both margins of the bite erode at once. The

[2]The local exponential approximation of the flow function (Appendix C) gives a larger baseline gap (24.1bp) and finite death points at 5.2 and 5.9 times baseline flow, and under it both optima jump to the deviation bound and bunch there at about 2.7 times baseline. Those coordinates are artifacts of that form's unbounded variance appetite, which lets return variance amplify the flow term without limit; the logit's saturation removes the amplifier, so dilution arrives without the jump and at essentially unchanged speed.

contract is a screening device for normal years: it is most needed exactly when it binds least, and it fails when hot money arrives.

That failure is not the end of the reform's design; it is the boundary at which the contract's other margin takes over. The vintage-contingent settlement of Section 3.2 does not depend on the fee tiers' marginal bite on the manager: it acts on the fund company's appetite for the hot money itself, lot by lot, at exactly the flow strength where the portfolio margin goes slack. Who governs the fund when the fee contract is swamped is the question Section 4 answers.

3.8 The two-peak risk profile

The analysis so far treats the evaluation year as a single draw. But the contract settles on the full-year excess return, and the manager learns his interim standing halfway through: given an interim result, how much risk should the second half carry? The tournament literature has asked this question of convex prize structures for thirty years, and its answer (trailing managers gamble to catch up, leading managers lock in) is organized entirely by the reward side (Brown et al., 1996; Chevalier and Ellison, 1997; Busse, 2001; Kempf and Ruenzi, 2008). The 2025 contract adds something that literature has never priced: a floor. The penalty tier is a step down to f_L and then nothing worse, and that floor changes the geometry of second-half risk in a way that is, to our knowledge, new.

The setup is the smallest one that carries the mechanism. Split the year at mid-year. The first-half excess return e_1 is realized and public: funds disclose performance semiannually, so the market and the manager update on the same number. The second-half excess return is $e_2 = \bar{\alpha}_2 + \sigma_2 z_2$, with expected alpha $\bar{\alpha}_2$ fixed (within a half-year the manager adjusts his risk exposure, not his stock-picking ability) and second-half risk $\sigma_2 \in [\underline{\sigma}, \bar{\sigma}]$ chosen at mid-year; the full-year result $e = e_1 + e_2$ settles the fee tier and attracts flows. The manager maximizes $E[f(e)] + \Lambda E[Q(e)] - \frac{\rho}{2}\sigma_2^2$, with $\bar{\alpha}_2 = 1.5\%$ (half of a 3% annual skill), the half-year flow weight $\Lambda = 90\text{bp}$, risk bounds $[0.5\%, 8\%]$, and $\rho = 5$ as in the static calibration. Two features of this objective deserve note. The research-cost function $C(x)$ of Section 3.3 does not appear: C priced the generation of alpha through deviation, and with $\bar{\alpha}_2$ fixed there is no such margin at mid-year; the only choice is risk, whose cost the variance penalty carries. The omission also sharpens the results: the two peaks and the deep-tail collapse below are produced by the contract's geometry and the exponential decay of its incentives, not by an exogenous convex cost holding risk down. And the variance penalty ρ now carries more than a preference parameter: over a tactical half-year it stands also for the fund's internal risk governance (tracking-error limits and turnover frictions that tighten after poor interim results) and for the career cost a manager bears if a speculative second half deepens into an unrecoverable drawdown. The two-period split is the coarsest grid on which an interim state exists; Appendix A.3.1 lets the manager move both margins, and nothing below flips when effort is allowed.

The whole answer is written in the sign of one object. Write $E_2 \equiv e_1 + \bar{\alpha}_2$ for the full-year performance center. The fee term's marginal value of risk is

$$\frac{\partial E[f]}{\partial \sigma_2} = \frac{G(E_2, \sigma_2)}{\sigma_2^2}, \qquad G(E_2, \sigma_2) \equiv u\,(h - E_2)\,\varphi\!\left(\tfrac{h-E_2}{\sigma_2}\right) \;-\; d\,(\ell + E_2)\,\varphi\!\left(\tfrac{\ell+E_2}{\sigma_2}\right).$$

The upgrade term is positive below the upgrade line (risk buys touch probability) and negative above it: a manager already above the line protects his tier by minimizing risk. The downgrade term is negative above the downgrade line, where risk deepens the left tail into the penalty zone, and positive below it. A manager who has already fallen through the downgrade line holds a fee that sits on the floor: falling further costs him nothing, and climbing back above $-\ell$ is worth the full 60bp step. He holds a free lottery (downside insured by the contract, upside a discrete jump) and the size of the incentive, $d\,|\ell + E_2|\,\varphi(|\ell + E_2|/\sigma_2)$, is maximized about one second-half standard deviation below the line.

Definition. *(catch-up gambling zone).* The catch-up gambling zone is the band of interim performance lying just below the downgrade threshold (about one second-half standard deviation) within which a manager optimally raises risk to gamble on climbing back above the line.

The floor that creates this zone also destroys it, further down. Both incentive sources decay exponentially in the deep tail: the climb-back value of the fee step dies with the Gaussian density of a distant threshold, and the flow term's convexity value dies with e^{kE_2}, while the risk cost does not decay. Deep enough in the left tail (calibrated at $e_1 \lesssim -6.75\%$), the marginal value of risk is negative everywhere on $[\underline{\sigma}, \bar{\sigma}]$, and the manager locks at the minimum. The naive option intuition (deep out of the money, raise volatility) fails because the option's strike recedes exponentially faster than volatility can chase it. A manager in the deep tail is not a gambler. He is, for incentive purposes, dead.

Definition. *(incentive dead zone).* The incentive dead zone is the region of sufficiently deep interim losses in which all performance-linked marginal incentives vanish and the manager's optimal risk equals its lower bound.

Assembling the signs of G over the state space gives the within-year risk profile its shape: two peaks, three locked regions, and an asymmetry between the peaks that is the fingerprint of the penalty-heavy shape.

Proposition 3. *(the two-peak risk profile). In the two-period problem with $d > 0$ and moderate risk aversion ($0 < \rho < \bar{\rho}$; calibrated $\bar{\rho} \approx 14$):*

1. *$\sigma_2^*(e_1)$ attains its* ***global*** *maximum in the catch-up gambling zone, near $e_1 \approx -\ell - \bar{\alpha}_2 - \sigma_2^*$;*

2. *$\sigma_2^*(e_1)$ attains a* ***local*** *maximum at the reward peak just below the upgrade line, near $e_1 \approx h - \bar{\alpha}_2 - \sigma_2^*$; this peak is floor-independent and exists under a bonus-only contract;*

3. *three regions are locked at $\sigma_2^* = \underline{\sigma}$: above the upgrade line (the strongest), the mid band, and the incentive dead zone ($e_1 \lesssim -6.75\%$ in the calibration);*

4. *under penalty-heavy parameters the catch-up peak exceeds the reward peak: the fee jump parameters governing the two peaks stand in ratio $d/u = 2$, and the flow term's concavity near the upgrade line depresses the reward peak further;*

5. *within each contract's catch-up gambling zone, the ordering is $\sigma_2^*[\text{penalty-heavy}] > \sigma_2^*[\text{symmetric step}] > \sigma_2^*[\text{bonus-only}] = \sigma_2^*[\text{fixed}] = \sigma_2^*[\text{linear fulcrum}] = \underline{\sigma}$;*

6. *both peaks survive up to $\bar{\rho} \approx 14$ and are flattened as $\rho \to \infty$.*

Proof in Appendix A. The regional sign analysis of G and the flow term (Lemmas A.8–A.11 in Appendix A) assembles parts 1–3; the peak comparison in part 4 reads off the fee jump parameters ($d/u = 2$) and the flow term's concavity near the upgrade line; part 5 notes that the catch-up incentive comes only from the penalty step, which the bonus-only, fixed, and linear contracts do not have.

Figure 2 shows the calibrated profile. At $e_1 = -6\%$ the penalty-heavy manager runs second-half risk of 1.70%, more than three times the 0.50% floor every other contract chooses at the same state; at $e_1 = +3\%$ the reward peak reaches 1.15%, shared exactly by the bonus-only contract. The peak tracks the threshold, one standard deviation below it, contract by contract: the symmetric-step contract, whose full-year downgrade line sits at -4.5%, is still gambling at 1.20% at $e_1 = -7\%$, where the penalty-heavy profile has already fallen back to the floor. A manager sitting exactly on his downgrade line does not gamble at all; he locks, because the free-lottery logic is sharpest one standard

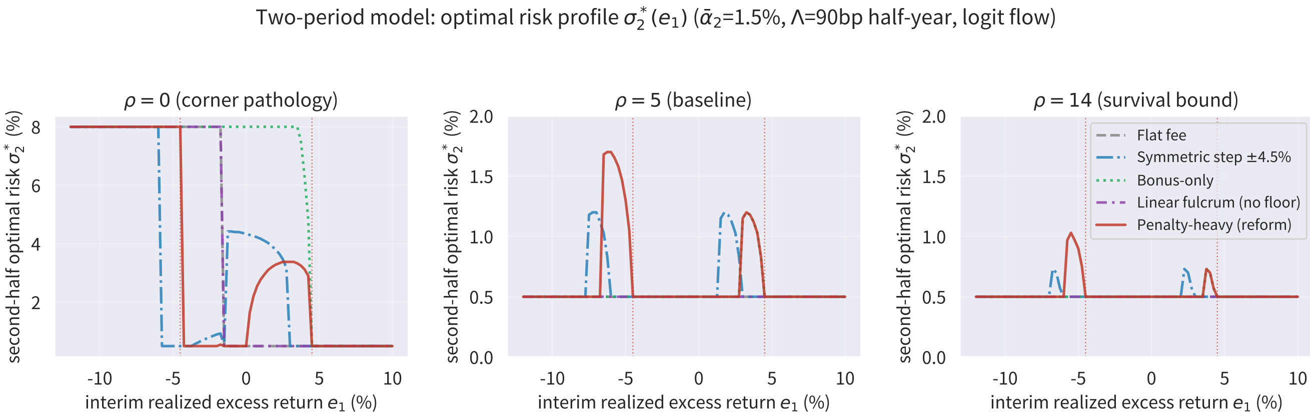

Figure 2: Second-half optimal risk as a function of interim performance, by contract. The penalty-heavy profile shows the catch-up gambling zone near $e_1 = -6\%$, the reward peak near $+3\%$, and the three locked regions; the linear fulcrum profile is flat at the risk floor throughout. Panels carry independent risk scales: the $\rho = 0$ corner pathology reaches the 8% bound, while the baseline and survival-bound panels are drawn on a 2% scale.

deviation below the line. (All quoted peak locations are threshold-aligned grid readings; unrestricted argmax readings differ only by the grid's alignment to the thresholds.)

Three implications give the profile its place in the paper. First, the asymmetric fingerprint: the dominant peak sits on the penalty side, and a within-year risk response stronger below the downgrade line than below the upgrade line is a pattern that reward-side accounts do not produce; the U.S. tournament evidence is organized by distance to the top (Brown et al., 1996; Chevalier and Ellison, 1997). Second, the fulcrum mirror: the linear fulcrum fee (the symmetric, floorless schedule that the 1970 U.S. amendments (United States Congress, 1970) effectively mandated) has expected fee $f_M + \beta E_2$, independent of σ_2, so its manager minimizes risk in every state: symmetric linearity with no floor removes both peaks together. The 2025 contract keeps the steps and tilts them, and in doing so contains a gambling zone that no contract in the U.S. experience has contained: the regulator wanted to punish laggards, and the floor of the punishment is where gambling breeds. This is, as far as we know, the first formal statement of what the fulcrum design was avoiding, and of what abandoning it costs (Elton et al., 2003). Third, the floor is the dividing line, and the dead zone is identification. A continuous performance-based schedule (the smooth fulcrum family of the Chinese PBF literature (Sheng et al., 2014)) keeps a nonzero marginal fee incentive at every performance level, so even a deeply underwater manager retains a reason to try; the step-plus-floor geometry kills this, and incentives do not merely weaken, they die. Every behavioral story of gambling in losses (loss aversion, the disposition effect, house-money logic) predicts a risk response that deepens monotonically as losses accumulate; the contractual mechanism predicts an inverted-U with a deep-tail collapse, and the deep tail is the placebo region that separates them: diff-in-diff on floating- versus fixed-fee funds, binned by distance to the downgrade line.

Three robustness checks bound the result; details are in Appendices A and D. Risk aversion: the baseline $\rho = 5$ sits deep inside the survival range, and the dead zone is not an artifact of constant risk aversion: ρ bundles the career and contract costs of losses, which *rise* in the deep tail (the reform's companion clause cuts the pay of managers trailing by ten points over three years), so the dead zone sits where the contractual component of ρ is largest. If the behavioral component nonetheless overwhelms it, the dead zone dies first, and observably: the behavioral stories predict rising risk exactly where we predict collapse (Kempf and Ruenzi, 2008; Busse, 2001). Effort: Appendix A.3.1 lets the manager buy expected alpha as well as risk, and the catch-up peak survives for every effort-cost curvature (Figure 12): calibrated height 1.70% at $\eta \geq 20$ and 1.55% at $\eta = 5$, its position shifting left as effort cheapens, while where effort can climb back over the line it substitutes for gambling, and the deep tail neither strives nor gambles. Flow form: the two peaks and three locked regions survive under the exponential benchmark of Appendix C as well; saturation changes their calibration, not their existence.

The static margins describe which managers deviate and by how much; Proposition 3 describes when a given manager's risk-taking ignites and when it shuts off, as a function of where he stands within the evaluation year. The first redistributes activity across skills; the second concentrates risk in states (one standard deviation below a threshold) that the contract's own geometry defines. Both are the shape of the incentives of the same three-tier schedule: its cross-sectional shape and its time-series shape.

4 The Vintage Structure

Section 3.6 closed with a boundary: in hot-money episodes the convex flow incentive swamps the fee schedule, and the contract's bite on the manager's portfolio choice decays to nothing (Proposition 2). The 2025 contract has a second margin that does not depend on the tiers' marginal bite on the manager at all. The fee is settled lot by lot, over each lot's own holding window (Section 3.2), and that settlement rule acts directly on the fund company's revenue accounting: on what the company collects from the money it has already gathered, and therefore on its appetite for gathering more. The reform also rewrote a market share, not only a contract: leading companies must issue at least sixty percent of new active equity products on floating fees, mandatory in the aggregate but discretionary at the margin, since the company chooses which of its managers get them. The vintage structure prices gathered money year by year; the quota allocates the contract across managers. Together they are the reform's company-side content.

Three results carry the single-fund argument. Window by window the realized fee is a lottery; averaged over a holding period it converges to a deterministic stair in true skill, at exponential speed (Proposition 4). The penalty on a bust fund attaches not to the hot money that came and left but to the sticky money that stayed, repriced toward the floor year after year (Proposition 5). And the marginal fee incentive of any lot dies at super-Gaussian speed as its window ages, unless the lot sits exactly on a threshold (Proposition 6). At the industry level the quota binds with room to spare (Lemma 3), the company floats its strongest managers so that a floating appointment is an endorsement (Proposition 7), and the aggregates settle with activity concentrated in the genuinely skilled segment.

Setup. Time is discrete years $t = 1, 2, \ldots$. The fund's excess return in year t is $e_t = \alpha + \sigma z_t$ with z_t iid $N(0, 1)$, and (α, σ) taken as given, the outcome of the manager's Section 3.6 activity choice, which this section returns to only where the vintage structure feeds back into it. Each year a cohort of money enters, with two components throughout: chased money $I_0(e_{\text{lag}})$, increasing in lagged performance, and the company's marketing spending m at strictly convex cost $\Psi(m)$ with $\Psi'(0) = 0$, so realized inflow is $I = I_0(e_{\text{lag}}) + m$. Both components are held at baseline until Section 4.2, where the company chooses m to maximize fee revenue net of marketing cost; Section 4.5 adds the company's second instrument, the contract assignment. A unit of money redeems after N complete years, with $q_n \equiv P(N = n)$ and a measure q_0 redeeming within the year; the baseline is geometric survival, $q_n = r(1 - r)^{n-1}$ with annual redemption hazard r. A lot held $n \geq 1$ years is charged, at exit, the rate $f(W_n)$ applied retroactively to all n years, where $W_n \equiv \frac{1}{n} \sum_{s=1}^{n} e_s$ is the annualized excess return over the lot's own window; money redeemed within the year pays f_M. Two institutional facts drive everything below: the window grows from entry, so a bad year never drops out of a lot's evaluation, it is only diluted at weight $1/n$; and the window mean is an iid average, $W_n \sim N(\alpha, \tau_n^2)$ with $\tau_n \equiv \sigma/\sqrt{n}$, so independent return noise enters the applicable fee rate with standard deviation $\sigma/\sqrt{n}$, which we call window averaging.

The results below are finite-n statements with explicit rates: the speed at which each lot's fee resolves is itself the economic object. The revenue side needs one accounting identity. A unit redeemed at age n has paid cumulative revenue $A_n(\beta) \cdot f(W_n)$ in present value, where $A_n(\beta) \equiv \sum_{a=0}^{n-1} \beta^a$ is the annuity factor at annual discount factor β, and within-year money is recorded as accruing one year at f_M. The expected lifetime value of a unit of new money is

$$LTV_\beta(\alpha, \sigma; q) \equiv q_0 f_M + \sum_{n \geq 1} q_n \, A_n(\beta) \, E[f(W_n)],$$

and the fee rate per yuan-year is its undiscounted version per billed year,

$$\bar{g}(\alpha, \sigma; q) \equiv \frac{LTV_1}{E[\bar{N}]} = \frac{q_0 f_M + \sum_{n \geq 1} q_n \, n \, E[f(W_n)]}{q_0 + \sum_{n \geq 1} q_n n},$$

where $\bar{N} \equiv \max(N, 1)$ is the lot's billed life, within-year money being billed a single year. Setting $\beta = 1$ in $\bar{g}$ is not an approximation: an average rate per yuan-year has numerator and denominator both measured in yuan-years, and discounting enters only where present values are compared, in the bust fine of Section 4.2 ($\beta = 0.97$). Under constant annual inflow I the stock is $I \cdot E[\bar{N}]$ and annual revenue is $I \cdot LTV_1$, so $\bar{g}$ is simultaneously one yuan's lifetime average rate and a steady-state fund's average rate. Scale throughout this section is measured in units of the steady-state inflow rate: a fund with inflow A carries the stock $E[\bar{N}]\, A$, the capacity dilution of Section 4.2 loads on scale in these units, and steady-state revenue $E[\bar{N}]\, \bar{g}\, A$ reads equally as the stock times the rate and as the lifetime value of one year's cohort. By Lemma 1, each term is $E[f(W_n)] = f_M + V(\alpha, \tau_n)$.

Remark. *(the filter: the contract bites only through seasoned money).* Money redeemed within the year contributes $q_0 f_M$ irrespective of (α, σ), since the one-year gate exempts it from the schedule entirely. Substituting $E[f(W_n)] = f_M + V(\alpha, \tau_n)$ gives $\bar{g} = f_M + \sum_{n \geq 1} w_n \, V(\alpha, \tau_n)$ with yuan-year weights $w_n \equiv q_n n / (q_0 + E[N])$ summing to at most one; the sensitivity of revenue to skill is strictly positive if and only if seasoned money has positive measure, and a flat-fee contract is exactly this contract at $q_0 = 1$. Two consequences organize the empirical content. The clientele's holding-period structure is the first moderating variable of the reform: products dominated by short-term retail churn face a weaker contract, whatever the prospectus says. And because $w_n \propto q_n n$, money that stays longer weighs more in the revenue function, the dual of the load-fee screening logic of Chordia (1996) and Nanda et al. (2009): they used fee structures to sort investor types; here the stock distribution of investor types shapes what the contract is to the manager.

4.1 Window averaging

A single year's performance reveals skill barely: a young lot's fee tier is a lottery, mostly noise, penalty-tilted in expectation. A ten-year window answers the same question in kind, not just in degree: the noise has averaged out at the rate $1/\sqrt{n}$, the tier is nearly decided, and what remains is the fund's true skill mapped into its tier. How fast the lottery resolves, and from which direction each type of fund's fee converges, is the content of Proposition 4.

Proposition 4. *(window averaging: the fee is an asymptotically deterministic stair in true skill). Let $\alpha \notin \{-\ell, h\}$ and $\delta(\alpha) \equiv \min(|\alpha - h|, |\alpha + \ell|)$ be the distance from skill to the nearest threshold.*

(a) (Limit and rate.) The applicable fee rate converges to the stair in true skill,

$$\lim_{n\to\infty} E[f(W_n)] = f(\alpha),$$

with a uniform exponential bound,

$$\big|E[f(W_n)] - f(\alpha)\big| \;\le\; \frac{u+d}{2}\,\exp\Big(-\frac{n\,\delta(\alpha)^2}{2\sigma^2}\Big).$$

(b) (Luck contamination.) The probability that a lot of age n is priced into a wrong tier satisfies the same exponential bound; the set of skill levels whose mispricing probability exceeds ε (the confusion band around each threshold) has total width at most $2\sigma\sqrt{2\ln(1/\varepsilon)}/\sqrt{n} = O(1/\sqrt{n})$.

(c) (The age profile.) As a function of n, $E[f(W_n)]$ is: strictly increasing toward f_H from below if $\alpha > h$; strictly decreasing toward f_L from above if $\alpha < -\ell$; and converging to f_M if $\alpha \in (-\ell, h)$, ultimately from below if $\alpha < \alpha_{be}$ and from above if $\alpha > \alpha_{be}$, where $\alpha_{be} \equiv (h-\ell)/2$ is the asymptotic break-even skill.

Proof in Appendix A. The window mean is Gaussian with scale τ_n, so each threshold sits at standardized distance at least $\delta(\alpha)\sqrt{n}/\sigma$; the Gaussian tail bound delivers (a) and (b) case by case, and for (c) the derivative's sign reads off Lemma 1's variance term, with the downgrade side dominating exactly when $\alpha < (h-\ell)/2$.

Two features organize the numbers. The probability of a mispriced tier dies exponentially in n, but the confusion band's width improves only at $1/\sqrt{n}$. And in the mid band a fund with $\alpha = 0$ pays strictly less than the base tier on every finite window: under penalty dominance $V(0,\tau) < 0$ for all $\tau > 0$, with the gap -11.6bp at $n = 1$ and -0.081bp at $n = 16$ at $\sigma = 4\%$. Young money of a mid-band fund pays a noise tax (the

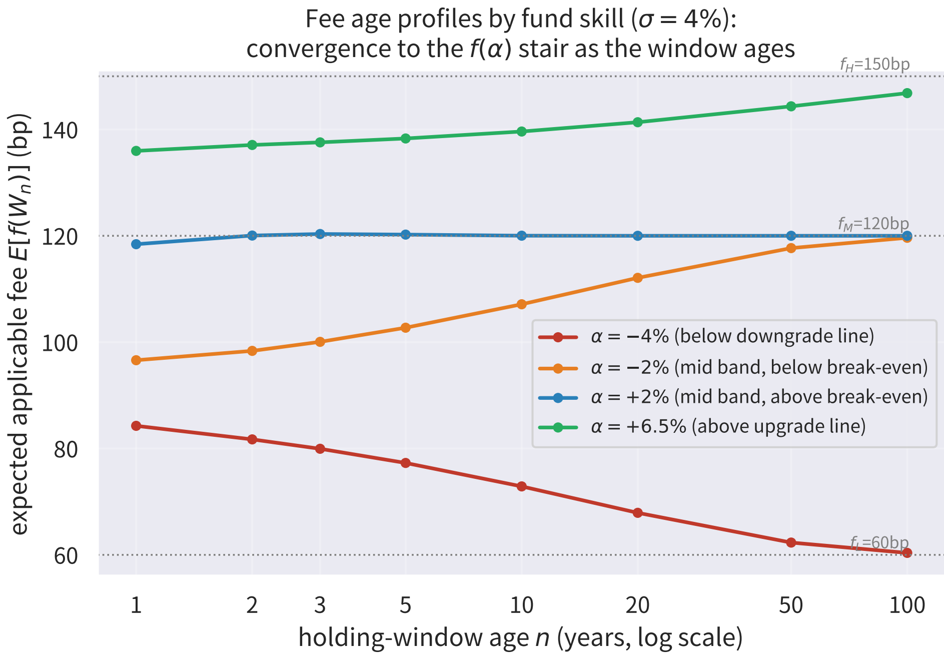

Figure 3: The applicable fee rate as a function of holding-window age, by fund skill ($\sigma = 4\%$): convergence down to f_L for $\alpha < -\ell$, up to f_H for $\alpha > h$, and to f_M from the side determined by the break-even skill $\alpha_{be} = 1.5\%$ for the mid band.

expected fee shortfall created by the penalty-heavy tilt on a window too noisy to carry signal), refunded as the window ages; a fund with $\alpha \in (\alpha_{be}, h)$ enjoys the mirror-image noise subsidy, which evaporates with age. In the limit of a very noisy window the lottery's expectation is $(u - d)/2 = -15$bp before any alpha is added.

Calibration. Figure 3 plots the fee age profile at $\sigma = 4\%$: a genuinely bad fund $(\alpha = -4\%)$ converges from 84bp at age one down toward the 60bp floor; a mid-band fund below break-even $(\alpha = -2\%)$ is refunded the noise tax as the window ages, climbing back to 120bp; a fund above the upgrade line $(\alpha = 6.5\%)$ climbs monotonically toward 150bp. The longer a unit of money stays, the closer its applicable rate sits to the rate the fund's true skill earns, and the profile is observable, since the gap between young-money and seasoned-money realized fee rates signs and sizes the noise tax. The long-run fee depends on no one's beliefs: a seasoned lot's rate converges to $f(\alpha_{\text{true}})$ however the manager or the investor attributes the luck that filled the window.

Corollary. *(asymptotic fixed-fee-ization of the mid band). Fix (α, σ) and let the clientele age in the sense of first-order stochastic dominance: under geometric survival, $r \to 0$ and $q_0 \to 0$, so the yuan-year weight $w_n = r^2 n(1-r)^{n-1}$ concentrates at $n \approx 2/r \to \infty$. Then $\bar{g}(\alpha; r) \to f(\alpha)$ for every $\alpha \notin \{-\ell, h\}$.*

In the long run the contract is a stair pricer of realized skill: 1.5% for $\alpha > h$, 0.6% for $\alpha < -\ell$, 1.2% between. A sustained annualized excess return above $h = 6\%$ is near the long-run ceiling of even the best active managers, so in the actual skill distribution the upgrade tier is realized almost only by young money's noise: the contract's long-run shape is, to a first approximation, the mid band fixed at 1.2% and only persistently significant losers paying 0.6%. Penalty-heavy degenerates asymptotically into punish-only with a passive middle. What the design durably buys is the punishment of persistent failure and the laundering of luck, not a sustained marginal incentive in the mid band, and the prediction is directly testable: as products mature, realized management fee rates should concentrate on the $f(\alpha)$ stair, thick in the middle, thin at both ends, with upgrades realized far less often than downgrades.

Remark. *(the closet safe harbor).* Closet indexing is the unique strategy whose long-run fee is deterministically the base tier: at $\alpha = 0$ the fee lottery dies with window age, while every $\alpha \neq 0$ strategy exposes its young money to it. This is the dynamic reinforcement of Proposition 1's activity threshold: the vintage model adds a long-run certainty premium to the static safe harbor, with one asymmetry: part (c) prices the closet indexer's young money below f_M, so only a company that intends to run a long-lived seasoned base cashes the harbor in full.

Everything in Section 3.6 ran on the sign properties of V; since $\bar{g}$ is a positive weighted average of V across window scales, those sign properties (penalty dominance at $\alpha = 0$, strict monotonicity in skill, the direction of the age profiles) survive aggregation (Appendix A). What aggregation does not preserve is the manager's marginal incentive, the subject of Section 4.3.

4.2 The bust fine and capacity internalization

In the old world, a fund that rode a bubble up and burst kept collecting 1.2% on the money that stayed; fee revenue had zero exposure to the fund's own bust. The new contract changes the accounting of the money that stays: a genuinely bad fund's sticky base is repriced year after year toward the 0.6% floor as each lot's window ages, and the bad years that caused the bust never leave the window. Part (a) of Proposition 5 makes this a present value; part (b) adds that once scale dilutes alpha, the contract converts the dilution into an immediate accounting loss on the fee line. The comparison isolates the revenue accounting: two worlds differing only in the contract, with returns, flows, and the redemption distribution identical.

Definition. *(bust fine).* For a fund with $\alpha < -\ell$, the bust fine is the present value, per unit of sticky money, of the fee shortfall relative to the flat-fee world:

$$BF(\alpha) \;\equiv\; \sum_{n\geq 1} q_n\, A_n(\beta)\left[f_M - E[f(W_n)]\right] \;=\; -\sum_{n\geq 1} q_n\, A_n(\beta)\, V(\alpha,\tau_n).$$

Proposition 5. *(the bust fine and capacity internalization). (a) (The bust fine.) Let $\alpha < -\ell$. Then (i) every term of $BF(\alpha)$ is strictly positive (the fine exists from the first year and is not an asymptotic mirage); (ii) the annual fine $f_M - E[f(W_n)]$ is strictly increasing in window age and converges to d; and (iii) $BF(\alpha) \geq \left[d\,\Phi(t_D(1)) - u\right]\cdot E[A_N] > 0$.*

(b) (Capacity internalization.) Let per-unit alpha decline in scale, $\alpha(A) = \alpha_0 - \kappa A$ with $\kappa > 0$. Under the flat fee, the fee-revenue value of a marginal yuan of AUM is f_M at every skill and scale. Under the floating contract it is

$$\frac{d}{dA}\left[\bar{g}(\alpha(A))\cdot A\right] = \bar{g}(\alpha) - \kappa A\,\bar{g}'(\alpha), \qquad \bar{g}'(\alpha) = \sum_n w_n\,\frac{u\,\varphi(t_U(n)) + d\,\varphi(t_D(n))}{\tau_n} > 0,$$

so the marginal value is strictly below the flat-fee value for every fund with $\bar{g}(\alpha) \leq f_M$, all but the high-skill tail, and there exists a parameter region in which it is negative: gathering one more yuan costs more, through dilution of the existing base's fee rate, than the yuan's own fee brings in.

Proof in Appendix A. For (a), $\alpha < -\ell$ makes $t_D(n)$ positive and increasing, so the annual fine exceeds $d\,\Phi(t_D(1)) - u > d/2 - u = 0$; monotonicity and the limit are Proposition 4(c). For (b), the chain rule gives the display, and $\bar{g}' > 0$ follows from skill-price positivity under positive weights.

The stick is nailed to the sticky base: hot money pays f_M in both worlds, insulated under the new contract by the one-year gate, so the reform leaves the revenue from hot money untouched and concentrates the entire punishment on the money that stayed. The fine cannot be outrun: the departure of hot money does not terminate it, and its size is set by the size and duration of the sticky base, not by the hot money that made the fund famous. At $\alpha = -4\%$, $\sigma = 4\%$, $\beta = 0.97$, and $r = 0.15$, the fine is about 266bp in present value per yuan of sticky money, 39% of what the same money would have paid over its lifetime in the flat-fee world. The headline number is the share: three to four tenths of a bad fund's lifetime fee revenue on the money it gathered in the boom and kept. The share is robust (32%–34% under performance-sensitive redemption); the basis-point level is not, shrinking by up to two-thirds under the same sensitivity, because bad performance accelerates exit and shortens the repricing time (Appendix D).

Part (b) is the strong form of the anti-hot-money result. Per-unit alpha declining in scale (the equilibrium dilution of Berk and Green (2004) and the size diseconomies of Chen et al. (2004) and Pástor et al. (2015, 2020)) is externalized under the flat fee, hurting the company only through future flows; under the floating contract the same dilution is internalized as a current accounting loss, every yuan gathered lowering the rate collected on the whole base through $\bar{g}'(\alpha) > 0$. The regulator did not prohibit gathering; it mailed the bill to the fee line. The sensitivity $\bar{g}'$ peaks at the downgrade threshold, at about 1,980bp per unit of alpha. At normal scale the negative region does not appear: the marginal value of AUM falls from 120bp to 91–104bp at the low-skill end, a discount of 13%–25%, inflow revenue discounted, not destroyed. Negative marginal values require blockbuster scale: at $A = 10\times$ base, $\kappa = 0.5\%$, and $\alpha_0 \in [1.5\%, 2\%]$ (dilution landing alpha exactly on the downgrade line), the marginal yuan of inflow is worth -7 to -9bp, precisely where the 2020–21 blockbuster funds lived.

Two behavioral consequences follow, both testable without any investor sophistication. Comparing marketing first-order conditions, $\Psi'(m^*_{\text{new}}) = E[\bar{N}] \cdot (\bar{g} - \kappa A\, \bar{g}')$ against $\Psi'(m^*_{\text{flat}}) = E[\bar{N}] \cdot f_M$: every fund with $\bar{g} \leq f_M$ markets strictly less under the new contract, and when the marginal value turns negative the company has an active motive to refuse money: star-fund subscription caps, historically read as investor protection or capacity self-discipline, become revenue maximization. High-skill funds, with $\bar{g} > f_M$, market harder, so marketing resources reallocate from low-skill to high-skill products. Investors need not recognize skill; the company acts on its own α. The contract makes the fund company the regulator's sorting agent.

4.3 The intertemporal structure of incentives

Propositions 4 and 5 describe the pricing map, not the manager's marginal incentive: a lot's settled rate and a lot's responsiveness to this year's return are different objects, and the difference grows with window age. A cohort has held for $n-1$ years with accumulated window mean R_{n-1}; year-n performance is $e_n \sim N(\alpha, \sigma^2)$, so

$$W_n = \frac{(n-1)R_{n-1} + e_n}{n} \;\sim\; N\Big(\frac{(n-1)R_{n-1} + \alpha}{n},\ \frac{\sigma^2}{n^2}\Big).$$

Two forces kill an old lot's marginal incentive: this year's return moves the window mean by only $1/n$, and each threshold, measured in the window's own standard deviations, recedes linearly in n, so the Gaussian density at the threshold dies faster than any polynomial. The exception is exact and permanent: a lot sitting precisely on a threshold keeps an $O(1)$ stake forever.

Definition. *(threshold-riding money).* A lot of money is threshold-riding if its accumulated window mean sits exactly on a fee threshold: $R_{n-1} = h$ or $R_{n-1} = -\ell$.

Proposition 6. *(the intertemporal structure of incentives). The cohort's expected fee rate responds to the current year's expected performance with sensitivity*

$$\mathrm{Sens}(n; R_{n-1}) \;\equiv\; \frac{\partial E[f(W_n)]}{\partial \alpha} \;=\; \frac{1}{\sigma}\Big[u\,\varphi\big(\tilde{t}_U(n)\big) + d\,\varphi\big(\tilde{t}_D(n)\big)\Big],$$

where the standardized threshold distances

$$\tilde{t}_U(n) = \frac{nh - (n-1)R_{n-1} - \alpha}{\sigma}, \qquad \tilde{t}_D(n) = \frac{-n\ell - (n-1)R_{n-1} - \alpha}{\sigma}$$

are linear in n, with slopes $(h - R_{n-1})/\sigma$ and $(-\ell - R_{n-1})/\sigma$ respectively. Hence:

(a) (Non-threshold-riding money dies.) If $R_{n-1} \notin \{-\ell, h\}$, at least one $|\tilde{t}|$ diverges linearly while the nearest threshold remains at positive distance, so $\mathrm{Sens}(n; R_{n-1}) = O(e^{-cn^2})$ *for some $c > 0$: the marginal fee incentive of a locked-in lot decays at super-Gaussian speed, faster than any polynomial, not $1/n$ and not $1/\sqrt{n}$. Money already settled into the mid band cannot have its tier changed by any feasible effort; money deep in the penalty zone cannot climb back.*

(b) (Threshold-riding money never dies.) If $R_{n-1} = h$ or $R_{n-1} = -\ell$, the corresponding $\tilde{t}$ is constant in n, so that side's sensitivity is $O(1)$ forever. The sprint-or-defend incentive at a threshold is permanent.

(c) (The catch-up gambling zone is a young-window phenomenon.) Let $R_{n-1} = -\ell - \varepsilon$ with small $\varepsilon > 0$, Section 3.8's catch-up gambling zone read across window ages. Climbing back above the downgrade line requires $e_n \geq -\ell + (n-1)\varepsilon$, linearly harder each year. Equivalently, $\tilde{t}_D(n)$ rises from below zero at slope ε/σ and crosses zero at

$$n^* = \frac{R_{n-1} - \alpha}{R_{n-1} + \ell},$$

so the catch-up incentive first rises to its peak at the "last salvageable year" n^* and then dies super-Gaussianly. The risk-sensitivity inherits the same decay: the sign-flip structure that powers the catch-up zone lives only in young cohorts whose $\tilde{t}_D$ has not yet diverged.

Proof in Appendix A. Substituting the conditional window distribution into Lemma 1's skill-price formula and applying the chain rule gives the sensitivity display; every rate statement reads off the linearity of $\tilde{t}$ and the Gaussian tail of φ.

At $\alpha = 0$, $\sigma = 4\%$, a cohort sitting one point below the downgrade line ($R = -4\%$) sees its fee sensitivity peak at 598bp at $n^* = 4$ (the last salvageable year), then fall below its year-one level from age seven; climbing back demands a single-year excess return of +1% at $n = 5$ and +6% at $n = 10$. A cohort exactly on the line holds a constant 452bp forever, per part (b). Aggregated to the fund level, the fund's marginal fee incentive is dominated by young money and threshold-riding money, and this is an explicit qualification on the empirical tests of Section 3.8: catch-up-side risk shifting should concentrate where young money is (funds shortly after large subscriptions, newly launched products); old funds with locked-in bases should show no catch-up zone at all, and whatever gambling remains there should tilt toward the upgrade-and-flow side, since the flow term does not decay with window age. This cross-sectional cut on the location of the catch-up zone separates the mechanism from any generic tournament story, which has nothing to say about window age. The contract is simultaneously an incentive device at the entry margin and a classification device at the stock margin, the strength of each set by the velocity of the fund's own money.

4.4 Limits of the vintage structure

The vintage structure also touches the flow–performance relation itself, but in a bounded way, through the marketing margin of the section's setup: inflow is $I = I_0(e_{\text{lag}}) + m$, and scale responds to inflow, $A = A(I)$ with $A' > 0$. The marketing first-order condition $\Psi'(m^*) = E[\bar{N}]\, v(A)$, with $v(A) \equiv \bar{g}(\alpha(A)) - \kappa A\, \bar{g}'(\alpha(A))$ the marginal value of AUM from Proposition 5(b), delivers both contracts' flow–performance sensitivities in one line. Under the flat fee $v \equiv f_M$, so m^* is performance-inert and $dI/de_{\text{lag}} = I_0'$: the company's margin adds nothing to the investor-side chase. Under the floating contract, total differentiation of the first-order condition gives

$$\frac{dI}{de_{\text{lag}}} \;=\; I_0'\, \frac{\Psi''(m^*)}{\Psi''(m^*) - E[\bar{N}]\, v'(A)\, A'(I)}, \qquad v'(A) \;=\; \kappa^2 A\, \bar{g}''(\alpha) \;-\; 2\kappa\, \bar{g}'(\alpha)$$

(the denominator is positive by the second-order condition wherever $v'(A) \leq 0$). The sign of the wedge is the sign of $v'(A)$: its second term is strictly negative wherever the fee schedule has marginal bite, and it dominates the first exactly where $\bar{g}'$ is large, the neighborhood of the thresholds; only deep in the tails, where $\bar{g}'$ and $\bar{g}''$ both vanish and nothing moves, is the sign nominally ambiguous. Where the contract bites, $v'(A) < 0$, and company-side flow–performance sensitivity is strictly lower than under the flat fee: chased inflow dilutes the base, dilution lowers the marginal value of AUM, and the company optimally cuts its own marketing against the chase. The magnitudes demote the channel (Appendix D): swinging lagged performance by $\pm 10\%$ moves optimal marketing by only 1%–2%, while the level effect is the large one: low-skill companies cut marketing by 17%–30%, high-skill companies raise it by 1%–5%. And the whole channel is conditional on capacity: at $\kappa = 0$ the formula gives $v' = 0$, the vintage rate does not respond to lagged performance at all, and the two contracts' sensitivities coincide. The empirical lead is therefore the level reallocation and subscription-cap behavior of Section 4.2, not the slope difference, which should be detectable only in funds with both tight capacity and large flow swings.

4.5 The quota binds

The industry analysis keeps three frictions and discards everything else: companies know more about their managers than investors do; investors come in two types that respond to contracts differently; and active management dilutes with scale. Competition between companies, investors' full Bayesian inversion of the assignment rule, and the overlapping-generations dynamics of old and new contracts are deferred to Appendix F. There is a continuum of new funds $i \in [0,1]$ with manager skill $\mu_i \sim G$, continuous on $[0, \bar{\mu}]$, calibrated as lognormal with median 2.5% and 90th percentile 6%. The company observes μ_i; its managers run on its research platform for years; investors observe the contract assignment and the public track record. Each new fund carries $c \in \{\text{float}, \text{flat}\}$ subject to the quota $\int \mathbf{1}\{c_i = \text{float}\}\, di \geq q = 0.6$; a stock of grandfathered fixed-fee funds coexists with the new issuance, and legacy contracts cannot be re-created. The fund-level building blocks are imported: a manager of skill μ chooses $x^*(\mu, c)$ as in Section 3.6 (under the calibration, activity thresholds $\mu^*_{\text{flat}} = 1.93\%$ and $\mu^*_{\text{float}} = 2.15\%$; the contracts cross at $\tilde{\mu} \approx 5.1\%$), and the company's revenue from a fund is the vintage-aggregated rate $\bar{g}(\alpha, \sigma)$ at clientele parameters $r = 0.15$, $q_0 = 0.3$. Smart money is an informed-capital sector in the Berk–Green tradition, allocating against capacity dilution κ with the return impact of its own size internalized (Berk and Green, 2004; Chen et al., 2004; Shleifer and Vishny, 1997); trusting money is fee-insensitive money (Barber et al., 2005), flowing on reputation and marketing, $A^T = I_0 + m$ at convex cost $\Psi(m)$. The representative company chooses an assignment rule $c(\cdot)$ and marketing policy $m(\cdot)$ to maximize steady-state fee revenue net of marketing cost,

$$\max_{c(\cdot),\, m(\cdot)} \int \Big[E[\bar{N}]\, \bar{g}\big(\alpha(\mu, c_\mu)\big) \cdot \big(A^S(\mu, c_\mu) + A^T(m(\mu))\big) - \Psi\big(m(\mu)\big) \Big]\, dG(\mu)$$
$$\text{s.t.} \qquad \int \mathbf{1}\{c_\mu = \text{float}\}\, dG \geq q,$$

An equilibrium is a tuple $(c^*, m^*, A^S, A^T, x^*)$ satisfying each side's optimality conditions and the quota; smart money takes the assignment rule as given and does not invert it, so this is a partial equilibrium with sorting, and the omitted feedback is shown below to reinforce the result.

Before asking how the industry responds to the quota, ask whether the industry needed it. The answer is in the company's own revenue account, whose float-minus-flat difference decomposes into three channels readable segment by segment.

Lemma 3. *(the three-channel decomposition). Let $R(\mu, c)$ be the company's steady-state revenue from a manager of skill μ under contract c, net of marketing cost: $R(\mu, c) = E[\bar{N}]\, \bar{g}_c \left(A^S_c + A^T_c\right) - \Psi(m^*_c)$, equivalently the lifetime value of one year's cohort of money. Define $\Delta(\mu) \equiv R(\mu, \mathit{float}) - R(\mu, \mathit{flat})$. Then $\Delta(\mu)$ decomposes into three additive channels:*

$$\Delta(\mu) = \underbrace{E[\bar{N}]\, \big[\bar{g}_L - f_M\big]\, A_L}_{\textit{(i) fee level}} + \underbrace{E[\bar{N}]\, f_M \left[A^S_L - A^S_F\right]}_{\textit{(ii) smart money}} + \underbrace{E[\bar{N}]\, f_M \left[A^T_L - A^T_F\right] - \left[\Psi_L - \Psi_F\right]}_{\textit{(iii) trusting money, net of marketing}},$$

where subscripts L and F denote the floating and flat contracts and $\Psi_c \equiv \Psi(m^*_c)$. The decomposition is an identity. In the low-skill segment ($\mu \leq \mu^*_{\mathit{flat}}$) channels (i) and (iii) are both strictly negative and $A^S_L = A^S_F = 0$: $\Delta(\mu) < 0$. Above the contracts' activity crossing $\tilde{\mu}$, channel (ii) is strictly positive and channel (i) is near-neutral through the mid band: $\Delta(\mu) > 0$ provided the quantity gain dominates.

The signs are not assumptions. In the low-skill segment both contracts induce closet indexing, so what remains is the noise tax, $\bar{g}_L(0, \sigma_0) < f_M$, in the fee-level channel and the marketing level effect of Section 4.2 in the third: the flat contract's marketing choice maximizes the flat-rate trusting-money program, so repricing the floating world's smaller marketing choice at the same flat rate and netting out the cost saving can only lose (the envelope argument of Appendix A, Section A.5): two negatives, same sign. Above $\tilde{\mu}$, Proposition 1(b) gives $x^*_{\text{float}} > x^*_{\text{flat}}$, and mid-band fixed-fee-ization keeps $\bar{g}_L \approx f_M$, so more alpha at almost the same fee rate draws strictly more smart money. The middle segment is where the algebra runs out and the calibration takes over: scanned over the full skill support at $\kappa \in \{0.1\%, 0.2\%, 0.5\%\}$ with the dilution feedback solved to a fixed

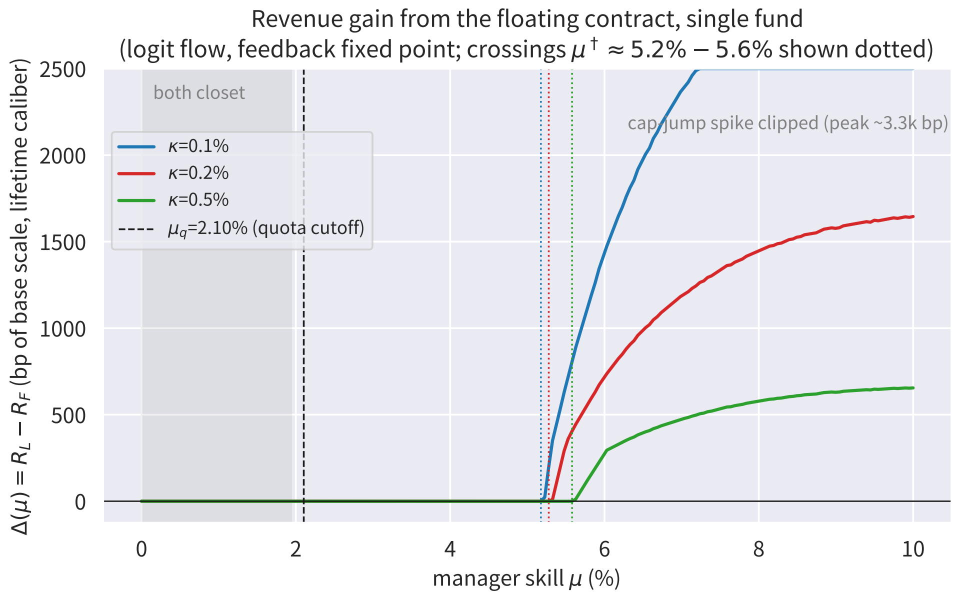

Figure 4: The company's steady-state revenue difference between floating and flat contracts as a function of manager skill, for three capacity parameters $\kappa \in \{0.1\%, 0.2\%, 0.5\%\}$. The crossing $\mu^\dagger$ and the floating share above it are marked.

point, $\Delta(\mu)$ crosses once, from below, at $\mu^\dagger \approx 5.2\%$–$5.6\%$, and at $\mu = 6\%$ the revenue difference runs from +260bp to +1,408bp of steady-state revenue, larger the smaller is κ (Figure 4). The company's unconstrained choice is to float the managers above $\mu^\dagger$ and keep the rest flat, a voluntary floating share of

$$1 - G(\mu^\dagger) \;\approx\; 12\%\text{–}14\% \;\ll\; q = 60\%.$$

The industry would float about one product in seven; the mandate requires three in five. The quota binds, with a wedge of more than forty-five points between what companies would do and what they must do, and the company's choice of *which* managers to sacrifice to that wedge is the content of the next subsection.

4.6 Sorting and the endorsement reversal

A binding quota turns the company's problem into a triage: forty percent of new products may keep the flat contract, and the question is who gets them. The natural guess, natural because the floating contract carries a penalty side, is that the scarce fixed-fee slots are the prize reserved for stars, the managers the company can least afford to expose. The guess is wrong, and the revenue account of Lemma 3 is why. For a high-skill manager the floating contract is not a punishment but an asset: it makes him more active, the activity draws smart money, and the fee level concedes almost nothing. For a low-skill manager it is a double loss. Facing a quota, the company floats its strongest managers and shelters its middle. The fixed-fee forty percent is not the prize shelf; it is the protection shelf, and the managers on it are, by the company's own revealed preference, the ones not good enough to float.

Proposition 7. *(sorting under the quota: floating appointments are endorsements). Suppose $\Delta(\mu)$ is single-crossing at $\mu^\dagger$ with $1 - G(\mu^\dagger) < q$ (the quota binds). Then the company's optimal assignment is a top-down cutoff rule:*

$$c^*(\mu) = \textit{float} \iff \mu \geq \mu_q, \qquad \mu_q = G^{-1}(1-q).$$

Proof in Appendix A. The company's problem separates manager by manager up to the quota constraint. With multiplier $\nu > 0$ on the binding quota, manager μ is floated exactly when $\Delta(\mu) + \nu \geq 0$; that is, the floating contracts go to the managers with the largest Δ, the greatest gains or the smallest losses. Single crossing makes that set the top quantile interval $[\mu_q, \bar{\mu}]$, and the binding quota fixes its boundary at $G^{-1}(1-q)$.

The signaling content survives the mandate in a specific, weakened form. Das and Sundaram (2002) showed that when the choice of fee structure is free, incentive fees can signal skill. Here the choice is not free (sixty percent must float), so the level of the contract carries no signal, but the allocation within the quota is free and observable, and it carries the entire signal: appointment to a floating-fee product certifies $\mu \geq \mu_q$, the floating segment is the company's top sixty percent by its own assessment, and the reserved fixed-fee segment is its middle and lower-middle. A mandate does not abolish signaling; it relocates it from the contract's form to the assignment's position.

The three-layer industry. Post-reform new issuance stratifies into three endogenous layers. The top layer, $\mu \geq \mu^\dagger$: genuinely active, floating, and voluntary: the company would have floated these managers anyway. The middle layer, $\mu_q \leq \mu < \mu^\dagger$: managers pushed into floating contracts by the quota, whose optimal response is to closet or to compress activity (Proposition 1), a quasi-indexing floating layer, products that carry the floating label and charge like indexers. The bottom layer: the reserved fixed-fee forty percent.

Under the calibration, the cutoff at $q = 0.6$ is $\mu_q = 2.10\%$ per year per unit of deviation, which sits between the flat and floating activity thresholds, $\mu^*_{\text{flat}} = 1.93\% < \mu_q < \mu^*_{\text{float}} = 2.15\%$. Between the two thresholds, the quota-marginal manager is active under the flat contract and a closet indexer under the floating one. The quota does not merely relabel the marginal product; it converts a manager who would have been modestly active into one who optimally does nothing. The mandate's cost is concentrated exactly in the middle band, and Section 5 will show that the active management the reform eliminates is precisely this band's.

Corollary. *Managers appointed to floating-fee products should have systematically stronger pre-appointment track records (prior performance, tenure, and platform standing) than managers appointed to fixed-fee products by the same company in the same vintage.*

The prediction needs no waiting for the new contracts to mature; Section 6 reports the test, and the appointment-date cross-section supports the sorting direction. One equilibrium consideration sits outside the proposition. If investors learn to read the rule (a floating appointment reveals $\mu \geq \mu_q$), floating products attract additional reputation-driven inflow, which raises the company's return to assigning its best managers to float: the feedback reinforces the sorting direction rather than unraveling it. The reverse manipulation, stuffing low-skill managers into floating slots to harvest the endorsement premium, is limited from two sides: the quota is a floor, not a ceiling, so misassignment merely swaps one floating body for a better one the company already had to supply; and window averaging washes out luck at an exponential rate, so a falsely endorsed manager's realized record falsifies the signal within a few lot-years.

Two further margins have clean directions but not the quantitative standing of Proposition 7. The first is the scarcity of legacy star funds. The reform constrains new issuance and leaves the stock untouched, and that asymmetry manufactures an asset: the grandfathered fixed-fee star fund. The price channel is real but small: a star with sustained alpha above break-even would, if issued today on the floating contract, earn a vintage rate above the base tier (at $\alpha = 8\%$, $\sigma = 4\%$, $\bar{g} \approx 146.5$bp against $f_M = 120$bp), so the legacy fixed rate gives smart money a net-alpha advantage measured in basis points. The reputation channel is the heavy one: a long realized track record is an asset no new product can replicate, and Section 4.2's marketing reallocation raises the legacy star's relative inflow further, with subscription caps as the equilibrium companion, the mirror image of Proposition 5(b), protecting the net alpha that is smart money's reason to come. The scarcity is small in basis points, real in trust; legacy star fixed-fee funds should gain share against matched same-performance controls after May 2025, and cap more often. The second margin is go-private. The manager's utility difference between contracts, $\Delta U(\mu) \equiv U(\mu, \text{float}) - U(\mu, \text{flat})$, is a V over the skill distribution: negative but small at the bottom (-0.08bp at $\mu = 0$), floored at -1.26bp near $\mu \approx 4.8\%$ where activity compression and the below-break-even fee stack, and positive above roughly $\mu \approx 5.9\%$. Combined with the convexity of the private-fund outside option (carry shares realized alpha with no upside cap), the exit pressure lands in the mid-high band: managers with enough skill to carry a private book whose public-fund activity and pay are compressed twice. The V's shape is structural, built from Proposition 1 and the break-even fee; its coordinates are left to the empirical calibration (Appendix F) and are not delivered here: under the illustrative W_{priv} the exit band is empty, public-fund utility carrying a flow-revenue floor of roughly 500bp.

4.7 Demand closure

Sorting determines who holds which contract, not how large the industry is; the demand side must be made explicit. Smart money is an informed-capital sector in the limits-of-arbitrage tradition (Berk and Green, 2004; Shleifer and Vishny, 1997): allocating against capacity dilution, the sector internalizes the return impact of its own size, so its position in fund i is

$$A_i^S = \max\left(0,\ \frac{\alpha_i(A_i) - \bar{g}_i}{\kappa}\right), \qquad \alpha_i(A_i) = \alpha_i^0 - \kappa A_i,$$

with the vintage fee rate evaluated at the diluted alpha. Trusting money $A_i^T = I_0 + m_i^*$ does not optimize at the margin, and total size is $A_i = A_i^S + A_i^T$. The sector's marginal unit retains a rent $\kappa A_i^S > 0$ in equilibrium: informed capital is not perfectly elastic, and the rent is what keeps the company's marketing margin alive, since trusting inflow is crowded out only partially rather than one for one. The vintage fee rate and the fund's size are jointly determined at this fixed point: smart-money inflow dilutes alpha, the diluted alpha reprices the vintage fee, and the repriced fee feeds back into the sector's position. Existence is proved and uniqueness verified numerically in Appendix A.6.

The per-unit net return of an interior fund is κA_i^S, earned by smart and trusting yuan alike; trusting money is the infra-marginal bearer of that return and of its reform-induced change, the account in which Section 5's welfare question lives. The aggregates below are computed at the fixed point.

Seen through the sorting rule, the reform's industry-level footprint is not dominated by the pushed layer, because size is concentrated where smart money concentrates, and smart money concentrates at the top. Three aggregates carry the industry-level argument; of the three, the activity decline reads off the manager block alone and is closure-free, while the fee transfer and the concentration reading are calibrated at the fixed point.

The fee transfer is positive, but small. Aggregate fee revenue relative to the flat-fee counterfactual, $\int [\bar{g}_i - f_M] \, A_i \, di$, is positive in the calibrated industry, between +0.06bp and +0.53bp per unit of normalized base scale across the capacity grid. The source is the top layer: genuinely skilled floating funds approach the upgrade tier in their vintage-aggregated rates, and they are large; their fee premium outweighs the noise-tax savings investors collect from the closeted bottom. The level is caliber-sensitive: it shrinks when the dilution feedback is switched on, and further under the formal logit flow function; the sign, not the level, is the robust reading.

Aggregate active management falls slightly. Total industry activity, $\int x^*(\mu, c^*(\mu)) \, dG$, moves from 0.089 to 0.085 across the reform, at all three capacity parameters. The quota pushes the wide middle band (μ between 2.1% and $\tilde{\mu} \approx 5.1\%$, where most of the skill distribution's mass sits) from flat-fee activity into floating-fee compression, while the voluntary top layer's amplification is bounded under the logit flow function. The decline is driven by the compression half of Proposition 1, which survives exact risk pricing (Appendix B), so the sign reported here is the robust one.

The industry's center of gravity moves to the top. Because size follows smart money and smart money follows gross alpha, the post-reform industry is more concentrated in its genuinely skilled segment: the top layer holds more assets, runs more activity, and earns a fee premium, while the pushed middle layer shrinks toward the benchmark in everything but its label. The reform's stated intent (concentrate the industry on real skill) is delivered by the company's sorting logic, not by the quota's arithmetic. The quota supplies the coverage; the endorsement reversal supplies the aim.

The equilibrium hands Section 5 its welfare inputs: the three aggregates above, together with investor net return separated by money type and manager surplus with its exit margin, both tabulated at the fixed point in Appendix D. On one side stand the gains already established: the suppression of tournament gambling, the end of the rain-or-shine fee, the laundering of luck; on the other, a quasi-indexing floating layer where the pushed middle band used to be active, and an exit pressure aimed at the mid-high skill the industry can least afford to lose. Whether the sixty percent quota's gains justify its costs is a question the first-round welfare account alone cannot settle; Section 5 takes it up.

5 Incidence and Institutional Interpretation

Section 4.7 closed by handing this section its welfare inputs: the three industry aggregates, investor net return by money type, and manager surplus with its exit margin. This section reads their incidence across the skill distribution and interprets the contract's geometry against the institutional record of Section 2: who gains and who loses when the flat fee becomes penalty-heavy (Proposition 8), and which institutional concerns point in the direction of the observed shape. The quota q (the minimum share of new active equity issuance on floating contracts, set at 60%) and the vintage-contingent settlement enter as institutional facts, the former binding the company's allocation of contracts across managers, the latter entering welfare through the vintage-aggregated rate $\bar{g}$ of Section 4.

5.1 The welfare basis

We measure **first-round incidence at fixed scale**: the reform's effect on investors of the existing stock of funds, at their existing sizes, before money reallocates. Under Section 4.7's closure, an interior fund's per-unit net return is κA^S, shared by smart and trusting yuan alike; but that return itself moves once scale adjusts, so the reform's direct incidence must be read one step earlier, on money that has not yet reoptimized. The money that has not reoptimized is trusting money, fee-insensitive, slow-moving, and infra-marginal by construction, which is why the static account and the equilibrium account answer different questions: the first-round measure isolates the contract's direct impact on the investors the reform was written for, before the market's own reallocation dilutes or offsets it. The equilibrium-scale counterpart, computed at Section 4.7's fixed point, is reported in Appendix D.

The object of measurement is the change in a unit-scale investor's first-round welfare, relative to the flat-fee world, from a fund of skill μ under contract c:

$$\Delta W_I(\mu, c) = \underbrace{\mu\,[x^*(\mu, c) - x^*(\mu, F)]}_{\text{(i) gross alpha delivered}} - \underbrace{[\bar{g}_c - f_M]}_{\text{(ii) expected fee rate}} - \underbrace{\frac{\gamma_I}{2}\,[\sigma^2(\mu, c) - \sigma^2(\mu, F)]}_{\text{(iii) activity variance drag}} - \underbrace{\frac{\gamma_I}{2}\,\Delta\sigma^2_{\text{gamble}}(c)}_{\text{(iv) gambling cost}},$$

where x^* and $\sigma^2(\mu, c)$ are the manager's Section 3.6 choices under contract c, $\bar{g}_c$ is the vintage-aggregated fee rate of Section 4, $\Delta\sigma^2_{\text{gamble}}(c)$ is the expected within-window excess variance of Section 3.8's risk profile (zero under the flat contract) and γ_I is the risk weight of the investor the regulator speaks for, which we scan rather than estimate. Industry welfare aggregates over the skill distribution, $\Delta W_I = \int \Delta W_I(\mu, c(\mu))\, A(\mu)\, dG(\mu)$, with $A(\mu) = 1$ on the first-round basis.

5.2 Incidence across skill

Ask who gains and who loses when the flat fee becomes penalty-heavy, and the four terms of the decomposition answer separately, by skill band. At the bottom, both contracts induce closet indexing, so only the fee changes: the closet indexer's floating fee sits below the base tier on every finite window, and the difference is the noise tax, refunded to his investors. In the lower-middle band, the floating contract converts marginal managers into closet indexers: the lost alpha is small, the fee falls, and variance contracts. In the middle band the alpha loss is no longer small and dominates the fee and variance savings. At the top, everything reverses sign: the fee rises above the base tier, variance is amplified, the gambling exposure is largest, and the alpha term flips with the preference specification.

Proposition 8. *(incidence across skill). In the first-round welfare decomposition of the penalty-heavy contract relative to the flat fee, the four terms' signs are distributed across skill bands as follows:*

Skill band	*(i) alpha*	*(ii) fee*	*(iii) variance*	*(iv) gambling*
low ($\mu \leq \mu^*_{flat}$)	0	+	0	0
lower-middle ($\mu^*_{flat} < \mu \leq \mu^*_{float}$)	−	+	+	≈ 0
middle ($\mu^*_{float} < \mu < \tilde{\mu}$)	−	+	+	+ *small*
high ($\mu \geq \tilde{\mu}$)	±	−	−	+

Signs are proved term by term from earlier results (Appendix A), with two calibrated exceptions: the fee term in the lower-middle and middle bands, whose sign rests on compressed activity keeping delivered alpha below the break-even of Lemma 2, verified on the calibration grid (Appendix D); and the magnitude qualifiers ≈ 0 and "small" in column (iv), which are calibrated readings, not signs. Column (iv) records the presence of the gambling exposure; it enters the decomposition as a cost wherever it is nonzero.

Proof in Appendix A. The signs are inherited from earlier results: (i) and (iii) from the extensive and intensive margins of Proposition 1: both contracts closet the low band, the floating contract alone closets the lower-middle band, compresses the middle, and amplifies the top within the mean–variance benchmark; (ii) from the break-even fee of Section 3.6 and the window averaging of Proposition 4 (a fund below the break-even alpha pays an expected vintage rate below f_M), and (iv) from Proposition 3, whose gambling zones exist only in contracts with steps and floors, discounted by Proposition 6's young-window weighting.

Two properties of the decomposition carry its economic weight. The gain side is preference-robust; the top band is not. The compression direction of the table survives exact CARA pricing (Appendix B); the amplification direction need not. On a skill grid against four contracts under both preference specifications, the sign table is matched cell-by-cell in 93.5% of cells under mean–variance and 82.0% under exact CARA pricing; these match rates are numerical verification (Appendix D), not part of the proof, and every exception cell is interpretable. The divergence is concentrated exactly where Section 3.6 said it would be: the high band's alpha term, +78.0bp in band mean under mean–variance, reverses to -142.8bp under CARA, where the amplification side of Proposition 1(b) does not survive; the lower bands (the protection side) are untouched by the preference form. The headline welfare statement therefore stands on the suppression side alone: the contract transfers from the managers of noise to the investors of noise. What it does for or to the investors of genuinely skilled funds is preference-dependent, and we rest no claim on it.

The band means make the concentration quantitative. At the calibration, with risk weight $\gamma_I = 2$, the four terms (alpha, fee, variance, gambling) read, band by band, in basis points: low band (0, +0.002, 0, ≈ 0), the pure noise-tax refund, small per fund but certain; lower-middle band (-1.97, +0.002, +0.10, ≈ 0); middle band (-12.0, +0.02, +0.51, small), the alpha loss dominates and the net is negative; high band (+78.0, -4.4, -3.3, largest gambling exposure) under mean–variance, with the first term reversing under CARA. Assembled, the contract's gains concentrate in the lower-middle of the skill distribution and its costs at the top: a protection device aimed at the investors of mediocre funds, not an incentive device aimed at stars. What the reform durably buys is protection, concentrated where the investors least equipped to protect themselves sit; its cost is the middle band's alpha. This is a trade, and the next subsection reads the institutional concerns that make the trade understandable.

5.3 Design trade-offs

Why $d > u$ and $\ell < h$? Three institutional concerns, each documented in Section 2, point in the same broad direction: persistent failure should carry a visible cost; reward-induced gambling should be limited, the lesson the United States learned in 1970; and the active industry should remain viable, since the same action plan calls for raising the share of equity investment. Read against the earlier sections' results, each concern attaches to a different parameter of the step family. The political and investor-protection concern attaches to the penalty jump: the bust fine of Section 4.2 is built from d. The proved fund-level statement is Proposition 5(a)'s: for a fund with persistent negative alpha, the bust fine is strictly increasing in d, and to exponential order the reward side does not enter it (Appendix A). The risk concern attaches to the reward jump: the reward peak of the within-year risk profile scales with u (Proposition 3), and the reward side's only compensating function, the high band's activity amplification, is the preference-specific half of Proposition 1(b). The viability concern bounds the penalty side: a heavier penalty or a closer downgrade line raises the activity threshold μ^*_{float} and thins the active population the regulator is mandated to grow (Proposition 1(a)).

Aggregated over the industry, these attachments become quantitative statements with a two-part evidentiary status: the per-fund slopes are proved (Appendix A), in each case with the decisive sign verified at the baseline calibration, while the industry-level magnitudes are calibrated comparative statics, read off the numerical scans of Appendix D. The political surface, computed on a representative bust fund as the bust fine's share of lifetime fee revenue, runs from 0% at $d = 0$ to 38.8% at the institutional $d = 60$bp to 77.7% at $d = 120$bp. The risk surface is monotone in the reward jump and strictly positive even at $u = 0$: the penalty floor by itself breeds the catch-up gambling zone, so a punish-only contract is not riskless. The political constraint's aggregate content is carried differently from the other two. Proposition 5's bust fine is a pricing map: for any product that persistently delivers $\alpha < -\ell$ it is defined, strictly positive, and strictly increasing in d, whatever mechanism produced that product. The industry model's population is a different object: G is the distribution of optimally behaving managers in post-reform new issuance, and on it the constraint's integration domain is empty by construction, since a manager who would persistently lose closets instead (Proposition 1(a)), leaving $\alpha(\mu) = \mu\, x^* \geq 0$ on the support of G. The products the political concern actually targets (grandfathered legacy funds, style-drifted products whose realized record sits persistently below the downgrade line) lie outside the new-issuance slate the model sorts. The model therefore prices the per-fund penalty exactly, and the aggregate weight of the targeted population is a cross-vintage distributional question the new-issuance model does not answer (Appendix F).

The comparators of Section 2.4 illustrate the same trade-offs. The symmetric step halves the political force at $d = 30$bp. Bonus-only removes the penalty side and keeps the reward peak's gambling undiminished. The linear fulcrum manufactures no bust fine at all, since a floorless linear schedule cannot punish a persistent loser beyond what he loses symmetrically. The two historical shapes are not rival answers to one question. The 1970 fulcrum answered a regulator whose live fear was gambling; the 2025 schedule answered one whose fear was the guaranteed fee, and the observed penalty-heavy geometry can be read as a compromise among the three concerns. The reading is directional; it does not identify optimal parameters.

5.4 The quota and the overconfidence tax

The quota margin. What the binding wedge buys investors on the first-round account is read off Proposition 8: the marginal entrant at quota q is the manager at the cutoff $\mu_q = G^{-1}(1-q)$, and the quota's marginal first-round welfare equals his own welfare change (Lemma A.16). At the calibration ($\mu_q(60\%) = 2.10\%$, between the two activity thresholds, $\gamma_I = 2$), the marginal welfare change rises monotonically from -16.0bp at $q = 0.2$ toward 0 as $q \to 1$, and is -2.6bp at the institutional $q = 0.6$: the lower-middle band's decomposition in miniature, the alpha loss from closetizing the marginal manager dominating the fee refund and variance contraction. The totals are no kinder: the aggregate first-round gain at $q = 0.6$ is $+2.3$bp, positive but thin and decreasing in q, and the positive total is contributed by the high band's mean–variance amplification side; under exact CARA pricing the total turns negative throughout, -29.5bp at $q = 0.6$ (Appendix B). The case for any particular coverage is therefore not made by the first-round welfare account; coverage is what the political concern prices, and the model's welfare account does not.

The overconfidence tax. Attribution bias (a manager who believes his skill is $\hat{\mu} = \mu+\delta$) interacts with the contract in a way that should, in principle, tax overconfidence: the biased manager over-activates, lands in the penalty zone more often than he expects, and pays for the difference. But the tax's sign depends on the welfare basis. On the single-window basis the tax is collected: at $\delta = 2\%$ the incremental expected fee transfer is -0.39bp to the biased manager. On the vintage-aggregated basis the sign flips, $+0.24$bp: multi-year windows wash out the penalty exposure at the speed of Proposition 4, while the biased manager's higher activity delivers real alpha he would not otherwise have produced. The contract taxes overconfidence in the short run and refunds it in the long run; a complete treatment remains open (Appendix F).

5.5 What the model does not establish

The interpretation of this section is bounded in six specific ways. First, the threshold levels are not derived: $d = 60$bp, $u = 30$bp, $\ell = 3\%$, $h = 6\%$ are taken as the institutional fact they are, and no result in this paper optimizes them. Second, the quota $q = 60\%$ is an institutional input; the first-round arithmetic of Section 5.4 does not rationalize its level. Third, the political constraint's aggregate weight lies outside the model's population: the industry model's slate is optimally behaving new issuance, on which a persistent loser closets rather than goes bust, so the bust fine is priced per fund (Proposition 5) while the size of the population it targets, legacy and style-drifted products, is not modeled. Fourth, the model does not identify a complete regulatory welfare function: the three concerns are named and their directions computed, but the weights a real regulator places on them are outside the model. Fifth, first-round incidence is not long-run equilibrium welfare: under the demand closure of Section 4.7, second-round reallocation changes the account, and we do not sum the two. Sixth, the comparator ordering is partly preference-dependent: the compression direction survives exact CARA pricing and the amplification direction does not, so any comparison that leans on the reward side's compensating function inherits that dependence.

Whether the contract's predicted footprints appear in the data the reform is already generating is the empirical question of Section 6.

6 Suggestive Evidence from the First Cohort

Sections 3 through 5 built the theory one margin at a time; this section is not its proof, and it cannot be. The first cohort of floating-fee funds was launched in June 2025, and the data available as of September 2026 cover roughly one evaluation year for twenty-six products, in a strong bull market that is itself a conditioning event the theory speaks to (Proposition 2). No design run on twenty-six funds over one such year can identify the threshold mechanism of Section 3.6 or the vintage dynamics of Section 4; what it can check is whether the data point in their direction.

What the cohort can do is two things. First, it checks the theory's shape: where realized fee tiers pile up, who gets appointed to the new products, how unstable a young window's tier is; these are facts about the schedule's outputs, and they either match the contract's geometry or they do not. Second, it sets the targets: every test below is specified tightly enough that its failure mode is informative, several fail in directions the theory itself predicts, and the section's last part converts the failures and the data gaps into an explicit test plan for 2027–28, when the tests of Propositions 3 and 6 acquire the power they do not have today. Everything below is suggestive evidence; the section reports every test conducted, including null results, unidentifiable contrasts, and one finding the theory does not generate.

The subsections follow the contract's chain of action: who receives the contract (Section 6.2), how appointed managers invest (Section 6.3), what the first settlement delivers

(Section 6.5), and how investors' money responds (Section 6.6).

6.1 Data and design

The treatment group is the first cohort of floating-fee active equity funds: 26 products, all launched in June 2025 under the May 2025 action plan (China Securities Regulatory Commission (中国证券监督管理委员会), 2025). The control group is drawn from actively managed equity funds launched on the fixed-fee contract between June 2025 and May 2026, spanning the twelve months following the treatment launches. We exclude index, exchange-traded fund (ETF), feeder, enhanced-index, and C-class shares, require at least 200 trading days of history, and remove later-batch floating products, leaving 144 controls. The same-vintage design differences out the launch window, the market regime, and the new-product channel environment; its cost is that the two groups' managers differ by assignment, not by chance, a difference Section 6.2 measures directly rather than assumes away.

The core panel is daily fund net asset values merged with daily returns of each fund's leg-by-leg composite prospectus benchmark; excess returns are computed against each fund's own composite benchmark for the floating group, exactly as the contract computes them, and against the CSI 300 for controls. A fund's tracking error (TE) is the annualized standard deviation of its daily excess return over the stated benchmark, computed over the window named in each test. Auxiliary sources: daily NAVs of 664 pre-existing funds run by the same managers; quarterly share disclosures from F10; the Gildata and Eastmoney F10 announcement archives (9,142 announcements across 184 funds, aligning both groups' purchase-limit event histories); the 2026H1 full holdings and holder-structure disclosures; and the F10 fee-field screen that identifies later-batch floating products in the control list. Coverage extends beyond the 170-fund analysis sample where the source requires it: the announcement archive covers the treatment group and the pre-decontamination control list (184 funds), and the holdings archive additionally covers the later-batch floating products and four benchmark ETFs used to proxy index constituents (188 series). The main settlement window ends 2026-05-27; rolling 60-trading-day windows extend to 2026-09-07. Full data construction is in Appendix E.

The verdict table. Table 3 is the section in one view: each prediction, the test, the data, and the verdict, with paper numbering on the left and the subsection on the right.

Preliminaries. Four caveats apply to every number in this section. First, $n = 26$ funds and roughly one year. Second, the sample year was a strong bull market; tests whose identification needs losing funds are underpowered by construction, and Proposition 2

Table 3: Predictions, tests, and verdicts.

Prediction (paper numbering)	Test	Data	Verdict
Proposition 7, Corollary (Section 4.6): floating appointments are endorsements	Pre-appointment 2023–24 excess return and tenure, floating vs. control managers	664 legacy funds' daily NAVs; 25 vs. 122 managers	**Supported** (Section 6.2)
Proposition 1(b) risk compression	ΔTE $\sim$ pre-appointment skill	Manager legacy TE + new-fund TE	**Marginal (direction consistent)**: floating slope +0.19 ($t \approx 2.4$) vs. control +0.02; slope difference classical $t = 1.54$, HC2 $t = 1.64$, fund-level CR1 $t = 1.59$, label-permutation $p = 0.266$ (Section 6.3)
Proposition 1 closet margin, holdings side	Top-10 concentration and benchmark-top-50 overlap, 2026H1 full holdings	184 funds + 4 benchmark ETFs, single cross-section	**Direction consistent**: concentration 54.0% vs. 44.9% ($t = 2.42$); overlap weight 20.2% vs. 9.9% NAV ($t = 3.66$); truncated active-share proxy identical ($t = 0.37$) (Section 6.3)
Proposition 3 catch-up gambling zone	Interim state at window midpoint $\rightarrow$ second-half change in holdings distance, beta, excess volatility	26 floating + 86 control first-window events	**Untestable at current power**: separating bins carry 8 floating events (catch-up $n = 1$); no contrast significant on any measure (permutation p 0.29–0.38) (Section 6.4)
Proposition 4 young-window lottery + Proposition 2 dilution	First-year realized tier frequencies vs. zero-skill lottery ($\sigma = 4\%$)	26 funds, inception $\rightarrow$ 2026-05-27 annualized excess	**Supported**: upgrade 46.2% / downgrade 38.5% vs. predicted 6.7% / 22.7%; binomial $p = 3.1\text{e}{-8}$ (Section 6.5)
Beta decomposition	Cross-sectional regression, fund return $\sim$ benchmark return	Same	**Partial**: slope 1.84 but $R^2 = 0.13$; beta explains levels, not dispersion (Section 6.5)
Proposition 4 window averaging: young tiers are unstable	Rolling 60-day implied tier switches and dwell shares	Daily panel	**Supported as a two-group regime fact**: 8.7–9.4 switches per 100 trading days in both groups; base-tier dwell 8–10% vs. lottery model's 70% (Section 6.5)
Proposition 5: near-threshold purchase limits	Limit announcements $\times$ distance to threshold; current status and full event histories	Gildata + Eastmoney announcement archives (9,142 announcements, 184 funds)	**Near-threshold margin silent**: 0 of 21 limit events near any threshold in either group; caps cluster on extreme winners in both groups, the scale-driven capacity logic of Proposition 5(b) (Section 6.6)
Section 4.4: company-side flow–performance-sensitivity (FPS) attenuation	Quarterly net flow rate $\sim$ lagged return, by group	F10 quarterly shares, 170 funds $\times$ 3 quarters	**Direction consistent, method-sensitive**: slope difference clustered $t = 2.90$, fund-label permutation $p = 0.107$; large level effect (Section 6.6)
Section 4.4: subscription-side decomposition	Subscription-rate and redemption-rate FPS separately, clustered SE	Same	**Supported**: subscription slope +1.69 ($t = 3.22$) control vs. +0.42 ($t = 0.82$) floating; redemption slopes close (+0.96 vs. +0.70) (Section 6.6)

tells us the contract's marginal bite is weakest exactly in such years. Third, during the initial closed (subscription) period of about three months, funds disclose NAV weekly, so the early panel is weekly-sampled; event studies mitigate this with minimum-observation thresholds, but early states are measured with error. Fourth, the repeated-observation designs (the quarterly flow panel and the tier-switching panel) exhibit within-fund serial correlation, so every group contrast below is reported under fund-level clustered and label-permutation inference, and several verdicts are downgraded on exactly this margin. None of these caveats is disqualifying for the section's stated purpose (shape consistency and target-setting); all of them rule out any stronger claim.

6.2 Sorting

The Corollary of Section 4.6 makes the one prediction that needs no waiting: if companies allocate floating contracts top-down under the quota, managers appointed to floating products should have systematically stronger pre-appointment records than same-vintage fixed-fee appointees, a prediction available at appointment, before any contract incentive has operated. We reconstruct each manager's record from the legacy funds he ran in 2023–24 (664 existing funds across the two appointment groups) and compute his annualized average daily excess return over the CSI 300 and his tenure. The pre-appointment window is deliberately hostile to the prediction: the CSI 300 returned only +1.0% over 2023–24, a flat two years in which sustained excess return is hard to manufacture with beta.

The sorting is large and significant. Floating-contract managers (25 with reconstructable records) averaged +1.20% annualized excess in 2023–24; control-group managers (122) averaged −6.36%. The gap of 7.6 percentage points has a Welch p-value of 0.006 (Mann–Whitney $p = 0.009$). Tenure points the same way: 5.14 years against 3.79 ($p = 0.027$). Figure 5 shows the two distributions.

Three readings. First, the direction is the endorsement reversal of Proposition 7, observed on the appointment margin itself: the floating slot went to the stronger record. Second, the magnitude colors every subsequent treatment–control comparison in this section: better floating-group outcomes cannot be read as contract effects without netting out the sorting, and we do not so read them. Third, Das and Sundaram (2002) predict that incentive fees signal skill when the contract choice is free; the mandate removes that freedom, and the signal migrates into the assignment within the quota (Section 4.6), of which this cross-section is the first direct observation. The records are public, so the same cross-section is also what a pure reputation rule would produce: the direction is established, but sorting on the company's private assessment, the proposition's mechanism, is

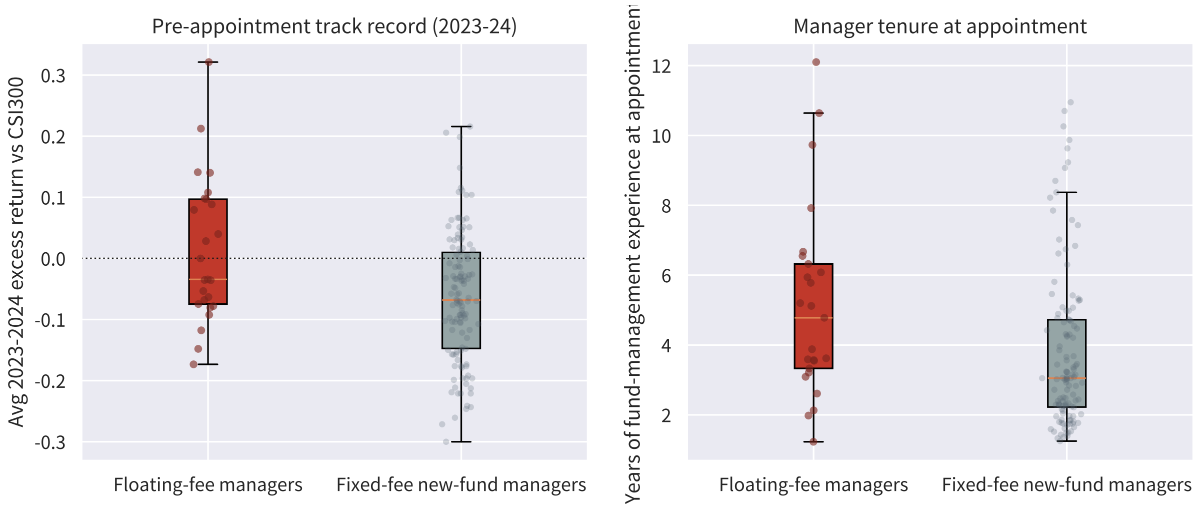

Figure 5: Pre-appointment (2023–24) annualized excess returns over the CSI 300, and tenure, for managers appointed to the first floating-fee cohort versus same-vintage fixed-fee appointments.

not separated from sorting on the record itself, and the designs that can separate them need more appointments than twenty-six (Section 6.7).

6.3 Risk response

Proposition 1(b) predicts the contract redistributes activity toward skill: activity moves with pre-existing skill more strongly under the floating contract than under the flat one. The test regresses the change in realized tracking error around the appointment (new-fund TE, computed inception to 2026-05-27, minus the manager's pre-appointment TE on his 2023–24 legacy funds) on pre-appointment skill, separately by group (Table 5, Panel A). The cross-section slopes apart: in the floating group, a manager with ten points more pre-appointment alpha expands his tracking error by about 1.9 points more than his lower-alpha colleague; among control appointments there is no such gradient (Figure 6). This is risk compression's intensive-margin signature in the only direction the first cohort can show it: not a level reduction (the bull market lifts everyone's TE) but a *re-grading* of activity toward skill, measured from each manager's own pre-reform baseline. The group contrast is marginal once inference is done properly: the slope difference is insignificant under fund-level clustered and label-permutation inference (the full grid is in Appendix E). Two qualifications keep the claim at "suggestive": the floating slope's statistical significance rests on individual observations (dropping the highest-$|\alpha|$ fund leaves the slope nearly unchanged at 0.18 but halves its precision, $t = 1.64$), and

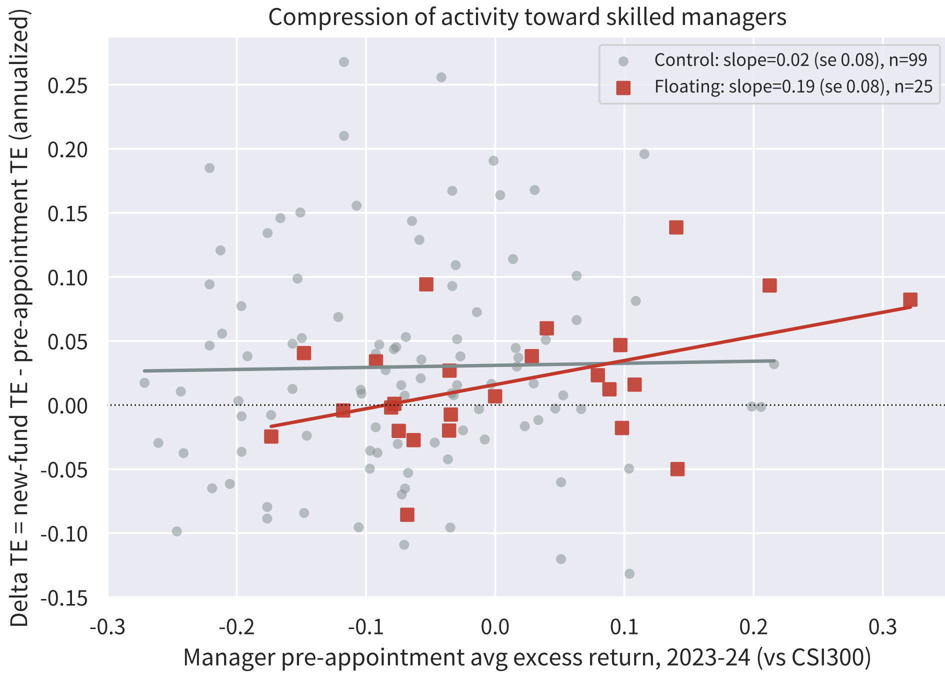

Figure 6: ΔTE (new fund minus manager's legacy-fund tracking error) against pre-appointment 2023–24 excess return, by contract group. The floating group's slope is +0.19; the control group's is +0.02, statistically indistinguishable from zero.

pre-appointment skill is Section 6.2's variable, so the regression partly re-reads the sorting result through a second margin. The compression is established in direction, marginal in the group contrast, in a sample where the contract's bite should be near its minimum.

The closet margin has a holdings-side signature that NAV data cannot show, and the 2026 interim reports supply a first cross-section: full holdings as of 2026-06-30 for the 184 funds and four benchmark ETFs, with the ETFs proxying index weights. The floating group's top-10 concentration averages 54.0% of NAV against 44.9% for controls (Welch $t = 2.42$); its top-10 holdings overlap the benchmark's top-50 constituents by 20.2% of NAV against 9.9% ($t = 3.66$). Yet the truncated active-share proxy is identical (82.3% against 81.6%, $t = 0.37$), and median new-fund tracking error is lower in the floating group (12.7% against 17.1%). The shape is coherent: core positions converge toward benchmark heavyweights while satellites stay highly active, the direction Proposition 1's compression logic predicts for the low-skill end. It is directional support, not confirmation: single-period, $n = 26$, ETF-proxied index weights (details in Appendix E).

6.4 The catch-up zone

Within the year, the risk response is state-dependent by construction: Proposition 3's catch-up zone raises risk and the dead zone compresses it, so a pooled behind-schedule bin mixes managers the theory sends in opposite directions. The within-year test therefore bins by the interim state, on the contract's own clock. Each fund's first evaluation window runs from inception to its first anniversary T; the interim state is the since-inception annualized excess return measured at the midpoint $M = T - 6$ months, the semiannual-report date of the model; the outcome window is the second half, $M \to T$. The outcome is measured three ways. The primary measure is the change in the holdings distance $D = \frac{1}{2}\sum_i |w_i^{\text{fund}} - w_i^{\text{bench}}|$ over A-share equities, computed from full holdings at the 2025 annual report (within three trading weeks of every midpoint) and at the 2026 interim report (at or just after every anniversary), with benchmark constituent weights proxied by the full holdings of the corresponding index ETFs (Appendix E). This is the manager's choice variable itself: raising risk means holding weights farther from the benchmark. The second measure is the change in market beta, estimated on daily returns over each half of the window; raising beta is the cheapest way to add risk without changing the portfolio's composition. The third is the change in annualized excess-return volatility, the realized tracking-error measure, reported with two caveats developed below. Controls receive the same own-anniversary clock. The sample is 26 floating events and 86 control events whose first anniversary falls inside the NAV panel; controls launched after early September 2025 have no observable anniversary and drop out. The prediction: floating funds just below the downgrade line (the bin at -4.5% to -1.5%) raise deviation most, funds deeper than about -8% sit in the dead zone and do not, controls show a flat profile.

The result is a null, and the null's geometry is precise (Table 4, Figure 7). The three states in which the contract theory and the behavioral alternative disagree carry eight floating events in total: one in the catch-up bin, two in the near-dead bin, five in the deep bin. No contrast between the groups is detectable at this mass: label-permutation p-values for the key separating gap (catch-up minus deep, floating versus control) are 0.38 for holdings distance, 0.33 for beta, and 0.29 for excess volatility. The directional residue is mixed rather than perverse: deep-bin floating funds raised beta more than deep-bin controls ($+0.42$ against $+0.07$) while trimming holdings distance (-1.1pp against -1.9pp), and no floating bin shows the predicted catch-up peak. Splitting the second half at $T-3$ months adds the timing dimension: only three floating funds stood behind the downgrade line at that date, and the sprint-segment response shows no intensification pattern.

Two measurement caveats qualify the volatility columns, and both argue for reading

Table 4: Second-half response by interim state at the window midpoint, floating versus control. ΔD: change in holdings distance (pp), 2025 annual-report to 2026 interim-report full holdings. $\Delta\beta$: change in market beta, second half minus first half of the evaluation window. ΔTE: change in annualized excess-return volatility (pp), same windows. n: funds. The 144 controls are truncated here to the 86 whose first anniversary falls inside the NAV panel.

	Floating				Control			
Interim bin	ΔD	$\Delta\beta$	ΔTE	n	ΔD	$\Delta\beta$	ΔTE	n
Deep, below −8%	−1.1	+0.42	+4.2	5	−1.9	+0.07	+5.8	32
−8% to −4.5%	+5.9	+0.26	−0.1	2	−5.6	+0.26	+0.4	1
Catch-up zone, −4.5% to −1.5%	+0.1	+0.55	+3.2	1	+0.6	+0.14	+3.3	7
−1.5% to +3%	−2.0	+0.38	+1.3	4	−0.9	+0.28	+4.4	5
Ahead, above +3%	−3.3	+0.06	−0.4	14	−1.4	+0.04	+4.3	41

only cross-group, cross-bin contrasts. First, realized tracking error mixes the manager's ex-ante position choice with the ex-post volatility of the exposure he holds: a fund that reached the deep bin did so by holding volatile exposures, and volatility clustering carries that level forward even if the manager changes nothing, so selection into the loss bins inflates forward realized tracking error under any behavior including complete passivity. The holdings distance and beta columns are free of this confound. Second, every control ΔTE cell is positive (+0.4 to +5.8pp): the first half of each window includes the closed subscription period, during which NAV is disclosed weekly and measured volatility is understated, shifting all ΔTE levels up. The same attenuation touches the first-half beta estimates.

The power statement is the substantive one. At the window midpoint only eight of twenty-six floating funds stood below −1.5% of annualized excess; the catch-up bin contains one fund. The benchmark strips out the market's level but not the cross-section of active tilts: the cohort shared an overweight of the year's winning sectors (12.1% of NAV in ChiNext constituents against 6.4% for controls at the 2025 annual report; Section 6.5), so excess returns rose together and the left tail of the excess distribution thinned. One catch-up event cannot distinguish any shape from any other.

The null matters because the two candidate mechanisms separate in exactly the states this sample does not contain. The contract theory of Proposition 3 and the behavioral-gambling alternative agree near both thresholds (trailing managers raise risk under either story) and disagree in between: in the contract's dead zone, where the marginal fee incentive decays to zero, the behavioral story still predicts rising risk as losses accumulate. The intermediate states are the placebo region in which the mechanisms can be told apart.

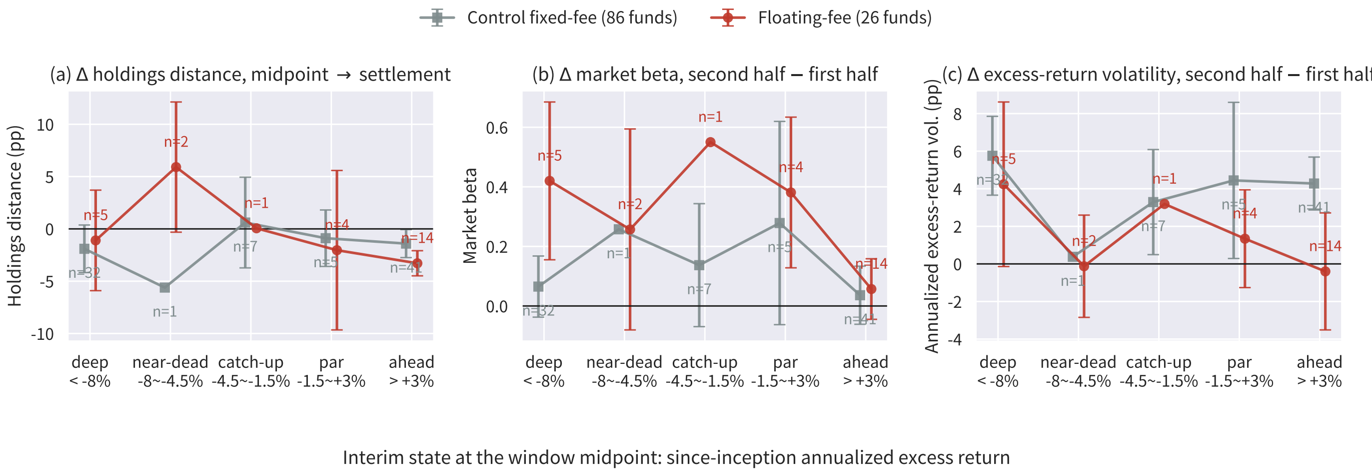

Figure 7: Second-half response by interim state at the window midpoint ($T - 6$ months), floating versus control: change in holdings distance from the 2025 annual report to the 2026 interim report (panel a), change in market beta between the two halves of the evaluation window (panel b), and change in annualized excess-return volatility over the same windows (panel c). Error bars are 95% intervals; counts are funds per bin.

The 2025–26 sample populates the separating rows only thinly. At the window midpoint, the floating cohort contributes five deep events, two in the −8% to −4.5% bin, and one in the catch-up bin (Table 4): a bull market lifts every cohort's window mean away from the downgrade line, so the states in which the two profiles differ barely occur. The test's current power is correspondingly low. Both attenuating forces are the model's own predictions: hot markets dilute the fee schedule's marginal bite (Proposition 2; the flow-swamping logic of Section 3.6), and the catch-up incentive lives only in windows that are young *and* near the threshold, decaying super-Gaussianly in window age (Proposition 6 and Section 4.3). The rerun on the separating rows, conditional on a market state containing losing funds, is on the falsification agenda of Section 6.7.

The null carries information, because the theory predicts its own failure here, twice. The first mechanism is Proposition 2: the catch-up zone is a marginal fee incentive, and in a hot-money episode the convex flow term swamps the concave fee term, flattening the within-year risk profile; the 2025–26 sample year was exactly such an episode, so the test was run in the state where the fee-side profile should be weakest. The second is Proposition 6(c): the catch-up incentive lives in young windows *near the threshold*, and a bull market lifts every cohort's window mean away from the downgrade line, so the cohort's windows, though young, were threshold-remote, outside the zone by distance, not by age. The correct test requires funds near the downgrade line with young windows (in practice a flat-to-down market) with the young-money share as the cross-sectional

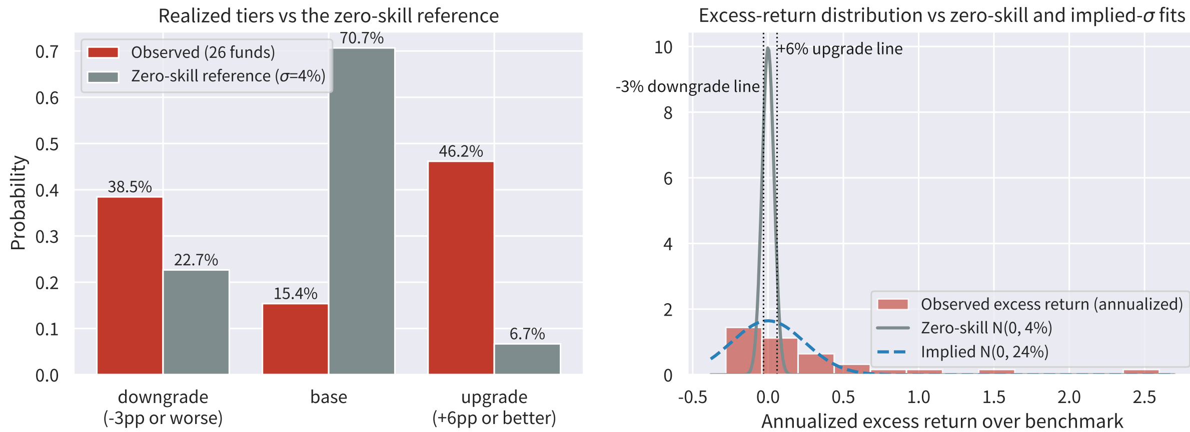

Figure 8: Realized first-year fee tiers of the 26-fund cohort against the zero-skill reference ($\sigma = 4\%$).

cut (Section 6.7). The epistemic position is exact: the outcome is not evidence against Proposition 3, because the test had no power in the direction of the prediction; it is not evidence for the behavioral alternative either, because the separating bins carry eight events in total, and the deep-bin volatility levels are inflated by the selection and disclosure confounds above. The null is consistent with the theory's own account of where this test should have no power.

6.5 Tier outcomes and window averaging

Proposition 4 says a young lot's fee tier is a lottery (mostly noise, penalty-tilted in expectation) and Proposition 2 says a hot market dilutes the schedule's marginal bite, widening activity dispersion. Read together, they predict the cohort's first-year tier distribution precisely: realized tiers should pile up at both corners, with the base tier thin. The contract's upgrade line is $+6\%$ and the downgrade line -3% of annualized excess over the fund's own benchmark. The reference point is the zero-skill lottery at $\sigma = 4\%$, under which 6.7% of funds would land above the upgrade line, 22.7% below it, and 70.6% in the base tier. The realized distribution, measured at the main settlement window ending 2026-05-27, sits far off this reference, in the predicted direction: upgrade 12 of 26 (46.2%), base tier 4 of 26 (15.4%), downgrade 10 of 26 (38.5%); a binomial test against the lottery's upgrade probability gives $p = 3.1\text{e}{-8}$. Figure 8 shows the realized histogram against the lottery reference.

The implied-volatility decomposition separates the two corners. The tracking error rationalizing the upgrade frequency under zero skill is 62.1%, the downgrade frequency

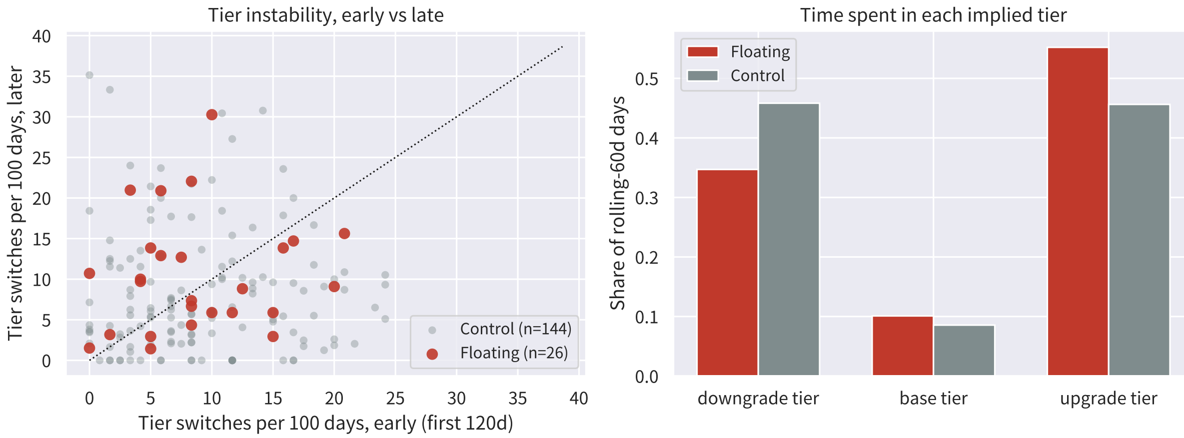

Figure 9: Implied tier dynamics on rolling 60-trading-day windows: switches per 100 trading days (overall and early vs. late) and dwell shares by tier, floating versus control.

10.2%; no single zero-skill distribution produces both. What reconciles them is exactly the composition the theory describes, a rightward mean shift plus fat tails: a shifted-normal fit gives $\mu = +3.8\%$ with $\sigma = 23.1\%$. The mean shift is the bull market and Section 6.2's sorting; the dispersion is Proposition 2's activity dispersion. The base tier's 15.4% dwell, against the zero-skill reference's 70.6%, is the summary statistic: the schedule's middle step, where the fee equals the old flat rate, is where almost no young fund actually lives.

The rolling version of the same test has the full daily panel: for each fund and day we compute the trailing 60-trading-day annualized excess, classify it into the contract's tiers, and count switches and dwell shares. The tiers flip about every eleven trading days: the floating group switches 9.4 times per 100 trading days, the control group 8.7; splitting each fund's history at day 120, the floating group's switch rate rises from 8.8 to 10.5 as the first-anniversary settlement approaches, while the control group's falls slightly, 9.2 to 8.2. Dwell shares confirm the snapshot: floating funds sit in the upgrade zone 55.2% of days, the base tier 10.1%, the downgrade zone 34.7%; controls split 45.6 / 8.6 / 45.8 (Figure 9). The group differences are statistically indistinguishable (switch rates, cross-fund Welch $p = 0.48$; base-tier dwell $p = 0.36$; upgrade dwell $p = 0.07$): the cohort's tier outcomes are those of the lottery regime (young windows, cornerized by a hot market) in both groups, not those of the deterministic-stair regime that Proposition 4 says emerges as windows age. The stair is the 2027–28 object; the lottery is the 2025–26 object, and the data look like it.

Two boundaries matter. The first is the control group: the two groups differ in contract but shared the bull market, the channel, and negatively selected managers on the control side, so the cornerization establishes the *regime* (young windows in a hot market produce lottery tiers) but cannot separate the contract's increment from the market's. Proposition 2 predicts the contract's bite is smallest exactly here, so the theory expects non-separation in 2025–26, and the non-separation is on record as such. The second concerns the mean shift. A cross-sectional regression of fund on benchmark cumulative returns gives slope 1.84 (se 0.97) with $R^2 = 0.13$: beta explains the cohort's level (mean fund +43.6% against +23.1%) but not its dispersion (residual SD 52pp against a benchmark cross-sectional SD of 10.8pp). And the obvious confound (measured alpha as style mismatch) is resolvable in the holdings. At the 2025 annual report the floating cohort held 12.1% of NAV in ChiNext-index constituents against 6.4% for controls (13.2% against 9.1% at the 2026 interim report). Across the 164 analysis-sample funds with a 2025 annual report on file, the ChiNext share alone explains 36% of the cross-sectional variance of first-window excess returns (+3.7pp of annualized excess per point of ChiNext weight, se 0.4), and controlling for it, the floating group's excess advantage disappears (floating dummy −6.1pp, se 9.8). The cohort's measured alpha is, to first order, the year's dominant style tilt, and this tilt is the mean shift behind the tier lottery's right corner. Section 6.2's sorting result is unaffected: it reads pre-appointment records from 2023–24, before this style year.

6.6 Flows

Section 4.4 is explicit about the expected sizes of the vintage structure's company-side leads: the slope effect (attenuated flow-performance sensitivity under the floating contract) should be small; the level and reallocation effects should be large. The first cohort's flow data read exactly that way, and add a level effect the theory does not generate.

From quarterly share disclosures we compute net flow rates, (subscriptions − redemptions) / lagged shares, drop the initial public offering (IPO) quarter, trim beyond ±200%, and regress the quarterly net flow rate on the lagged quarterly return, pooling 2025Q4–2026Q2 (Table 5, Panel B). The control group shows the standard chasing relation; the floating group's slope is insignificant and negative in point estimate. The slope difference is in the direction Section 4.4 predicts, but method-sensitive: significant under fund-level clustering, not under label permutation. The floating point estimate is also unstable across quarters, consistent with Section 4.4's calibration that the slope effect is small and visible only where capacity binds tightly, and its negative sign is better explained by the level effect below (winners being redeemed) than by incentive structure. The slope attenuation is direction-consistent with the predicted small effect and otherwise uninformative at this sample size; the mechanism is located by the gross flows below.

Table 5: Activity and flow regressions. Panel A: change in tracking error around appointment (new-fund TE, inception to 2026-05-27, minus the manager's 2023–24 legacy-fund TE) on pre-appointment annualized excess return, one observation per fund. Panel B: quarterly flow rates on the lagged quarterly return, pooled 2025Q4–2026Q2 fund-quarters. t-statistics in parentheses are fund-level CR1 cluster-robust; the difference column reports the interaction slope with its clustered t, and the 5,000-draw fund-label permutation p where computed.

	Floating	Control	Difference	Perm. p
Panel A. ΔTE on pre-appointment skill ($n = 25$ floating, 99 control funds)				
Slope	+0.19 (2.36)	+0.02 (0.22)	0.17 (1.59)	0.266
Panel B. Quarterly flow rate on lagged quarterly return ($n = 72$ floating, 304 control fund-quarters)				
Net flow rate	−0.28 (−1.14)	+0.72 (2.91)	1.00 (2.90)	0.107
Subscription rate	+0.42 (0.82)	+1.69 (3.22)	1.26 (1.73)	
Redemption rate	+0.70 (2.03)	+0.96 (2.76)	0.26 (0.53)	

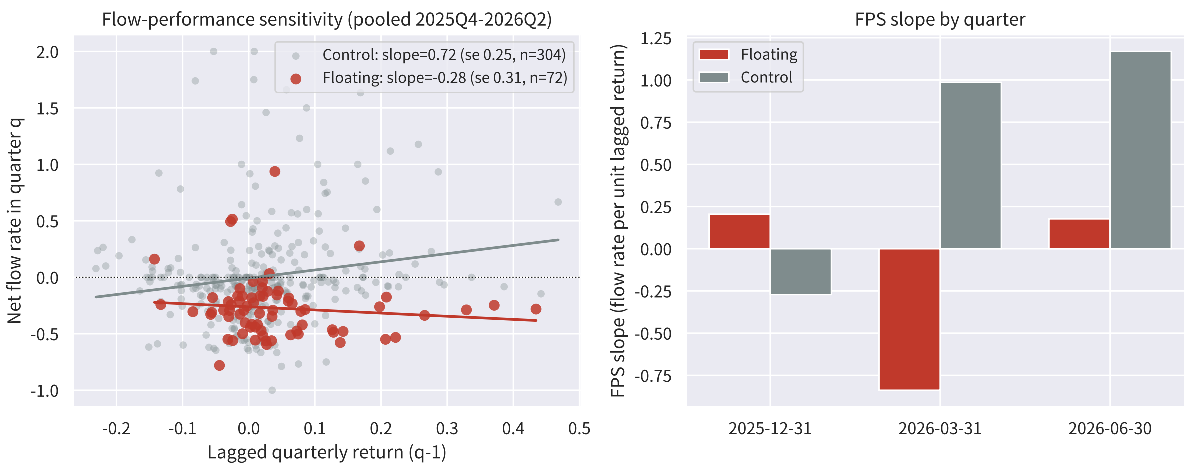

Figure 10: Pooled quarterly net flow rate against lagged quarterly return, floating versus control, 2025Q4–2026Q2.

Splitting the net flow into its gross components locates the mechanism at its working margin: the attenuation lives on the subscription side. The control group's subscription rate chases lagged returns strongly; the floating group's subscription slope is a quarter of the control group's and statistically indistinguishable from zero, while the redemption-rate slopes are close across groups (Table 5, Panel B). Investors redeem winners at essentially the same rate under both contracts; what weakens under the floating contract is the inflow that winning used to buy. This is the company-side flow-dilution mechanism read at the margin where it operates: the channel stops feeding new money to floating-contract winners, and it supports the vintage-attenuation channel more directly than the net-flow FPS, which dilutes the same fact with a redemption margin the contract does not touch.

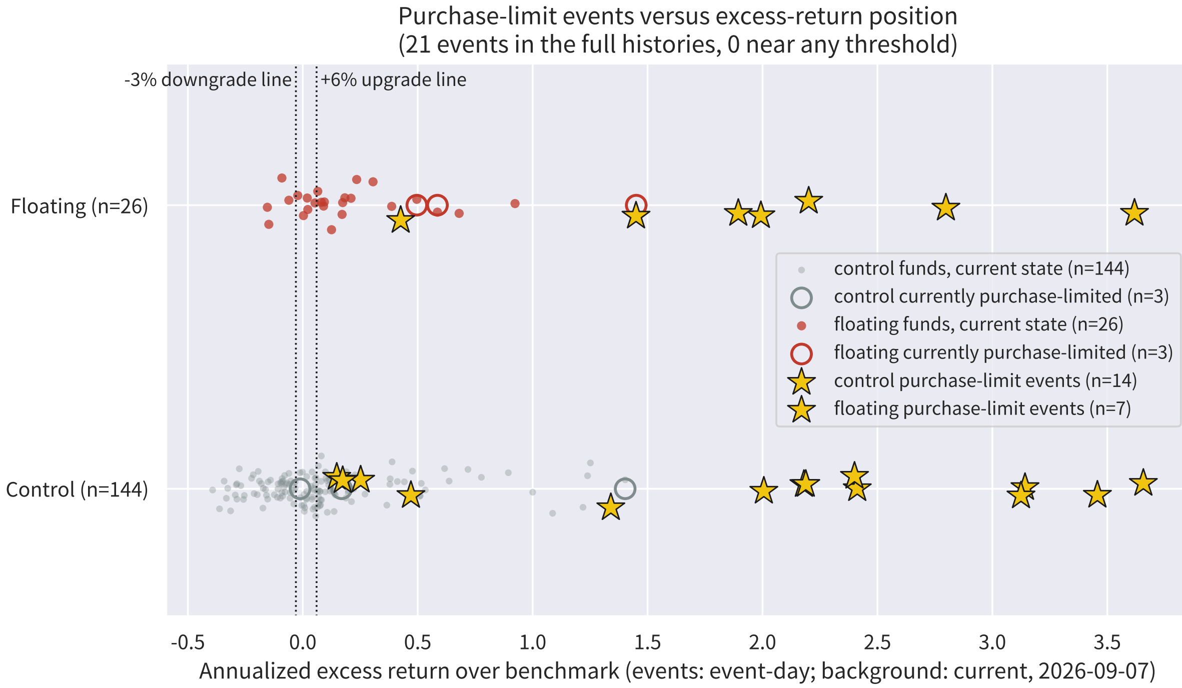

Figure 11: Purchase-limit events and current limit status, plotted against each fund's annualized excess return relative to the fee thresholds.

The marketing margin of Section 4.2 reads the subscription cap as revenue maximization: near-threshold funds should limit large subscriptions to defend the fee rate on the existing base. The completed announcement histories let the test run at full scale: 21 economically meaningful limit events, 7 in the floating group (3 of 26 funds, event-day annualized excess of +145% to +362%) and 14 in the control group (7 of 144 funds, +14.8% to +366%). No event in either group occurred anywhere near a fee threshold: within ±2 points of the upgrade or downgrade line the count is 0 of 21, and the lowest event-day excess in the archive is +14.8%, 8.8 points above the upgrade line. The group comparison comes in two calibers that agree in direction: the current-status cross-section (2026-09-07) gives 3 of 26 floating funds (11.5%) currently limiting large subscriptions (at excess returns of +59%, +50%, and +145%, all far from any threshold), against 3 of 144 controls (2.1%), Fisher $p = 0.047$; the event-history caliber gives 3 of 26 against 7 of 144 (4.9%), Fisher $p = 0.182$, weaker because the history includes limits since lifted (Figure 11).

The near-threshold mechanism is not supported, and the rejection is symmetric: zero observations in both contract groups. What the pattern fits is the flow-dilution narrative read from the company's side, identically in both groups: the caps arrived where hot money was arriving, on extreme winners at the height of a chasing episode (one control fund capped subscriptions five times between +201% and +346% annualized excess), as generic channel flow control indifferent to contract type. The floating group's higher limit rate is the mirror image of its having more extreme winners, not a contract effect. The scale dominance is rational given the contract's geometry: the fee state is a moving target (a third of all funds cross tier bins within a quarter; Section 6.5), while a purchase limit is a durable instrument, so threshold-targeted caps make sense only within weeks of settlement—and by then the 2025–26 cohort's winners sat far above the upgrade line, with nothing left to defend. We record the symmetric extreme-winner clustering as the capacity-internalization logic of Proposition 5(b) operating at blockbuster scale, exactly where Section 4.2 said revenue-destroying inflow appears; the near-threshold margin is silent at first-cohort scale.

The level effect in the flow data is the section's most surprising number. The floating cohort's mean quarterly net flow rate is −27.4% (median −29.3%); the control group's is +0.8%. The floating funds outperformed and were redeemed at a rate that would halve a fund in under three quarters; the cohort's flagship, at one point showing a triple-digit annualized excess, saw its shares fall from 665 million to 359 million. Investors in the best-performing new products of the year did not chase. They left.

The quarterly decomposition locates a tier signature: splitting the floating group's 2026Q2 flows by the tier settled on 2026-05-27, upgrade-tier funds ran roughly balanced net flows (−1.9%, hot money moving in both directions), downgrade-tier funds bled −28.1% with subscriptions at +0.9% (the channel abandoning losers entirely) and base-tier funds ran −21.5%. The surviving interpretation is the heat-driven version of Section 4's flow dilution, operating on investors rather than on the manager, coherent with the subscription-side FPS above and with the cohort's holder base: the floating group's median institutional share is 8.4% against 32.0% for controls, the young, retail, channel-driven money that Proposition 6(c)'s test design conditions on. The competitor reading (generic profit-taking on hot new products, a documented Chinese retail pattern) remains standing, and the two are separable only with daily or monthly flow data, which the public record does not supply.

6.7 The 2027–28 agenda

Everything above is one year of one cohort. The complete tests require the data that only time supplies: two to three settled windows per lot, a market state containing losing funds, and the second and third floating batches. We close with the test list in order of priority, both as a target for future work and as a placeholder claim on the designs.

1. **The anniversary-redemption test.** At quarterly resolution the first cohort rejects window concentration: redemption rates decline monotonically from the first full quarter (49.3% $\to$ 45.8% $\to$ 38.9%), with no anniversary spike (Section 6.6). The daily/monthly-resolution version requires data the public record does not supply; the feasible proxy is holder counts and structure from interim-report PDFs.

2. **The catch-up-zone rerun.** Section 6.4's test follows the contract's own clock and measures the choice variable in holdings; what is missing is the market state. The prediction survives or dies on a rerun conditional on funds near the downgrade line with young windows, binned jointly by distance-to-threshold, dwell time in the zone (in the first cohort, 35% of events spent less than half of the window to date in the zone their midpoint state assigns them to), and young-money share (the design of Proposition 6(c)), as later cohorts and a flat-to-down market supply the states.

3. **The stair test.** As second and third windows settle, realized management-fee rates should concentrate on the $f(\alpha)$ stair, thick in the middle, thin at the ends, upgrades rarer than downgrades (Section 4.3). The age-profile gap between young-money and seasoned-money realized rates signs and sizes the noise tax.

4. **Closet indexing in holdings.** A first single-period cross-section is in Section 6.3, and the holdings-distance panel has two full-holdings points (2025 annual report, 2026 interim report, same subsection); remaining are the extension to the 2026 annual report and the separation of the contract channel from the scale-cost channel, which holdings data make possible at the fund level.

5. **Style exposures beyond one axis.** The holdings-based check of Section 6.5 covers the year's dominant style axis; multi-axis exposures become computable as more holdings vintages accumulate.

6. **Sorting on private information.** Two designs, both needing more cohorts than the first: within-company appointment variation (companies appointing both floating and fixed managers in the same quarter), which controls for company-level appointment rules; and a record-conditional test, whether floating appointees subsequently outperform what their public records predict, which is the design that speaks to private assessment directly.

7. **Inference upgrades.** The within-year and flow tests carry fund-level clustered and label-permutation inference on the 144-fund control group. Remaining: weekly-

disclosure bias corrections for closed-period windows, and block-bootstrap designs for the event studies.

Each item on the list exists because a specific test in this section hit a specific wall (power, data, or identification), and each wall is dated: the cohort's second settlement window closes in mid-2027. The first cohort has matched the theory's geometry at the level of shapes (sorting, cornerized tiers, lottery windows), located the flow attenuation at the subscription margin where the company-side mechanism works, and shown, in the catch-up-zone null, that the theory's self-limiting predictions bite exactly where the propositions say they should. What it has not done, and could not have done, is measure the mechanisms at full power. That is 2027–28's work, and this section is the claim on it.

7 Conclusion

This paper explains a contract. In May 2025, China's fund regulator required newly issued active equity funds to charge fees that float against a benchmark, with a penalty twice the reward at half the distance, settled separately for every lot of shares over its own holding period (China Securities Regulatory Commission (中国证券监督管理委员会), 2025). We read the contract as what its design says it is: a concave, lot-by-lot repricing of the manager's revenue, built to stand against the convex flow incentive with which the market pays him. The theory then delivers the contract's logic one margin at a time. In normal years the fee schedule is a screening device: deviation becomes a negative-expectation gamble below a skill threshold, closet indexing emerges as an equilibrium consequence of the contract's geometry rather than of career concern or herding, and activity is redistributed toward skill rather than uniformly suppressed. Time itself works for the screen: window averaging launders luck out of a maturing lot's fee, so the realized charge converges to a stair in the fund's true skill, and only money sitting exactly on a threshold keeps its incentive forever. In hot-money episodes the screen fails: the bounded fee cannot race the unbounded flow, and the anti-hot-money function passes to the contract's second margin: the vintage structure reprices, year after year, the sticky money a company gathered at the peak, turning a bust from a marketing event into a bill mailed to the fee line. At the industry level the mandate binds, and the company's own revenue logic assigns the floating contracts top-down, so that appointment is endorsement and the quota's coverage is aimed by the party who knows the managers best, while the quota's cost lands where the contract's pressure does, in the middle skill band pushed toward the benchmark. The welfare account is correspondingly asymmetric: the robust effect is protection, not stimulation—investors of mediocre funds are refunded the noise tax as their managers closet, while the contract's incentive effect on genuinely skilled funds is limited and preference-dependent. And the shape itself (penalty-heavy, threshold-asymmetric, lot by lot) is not arbitrary: it is directionally consistent with the three concerns a regulator must balance when it must punish perceptibly, must not manufacture gambling, and must keep the industry alive at the same time.

The comparison with the earlier American mandate puts the design in relief. In 1970 the United States legislated the same contract in the opposite direction, forcing performance fees to symmetry (United States Congress, 1970); in 2025 China forced them past symmetry to the penalty side. The two schedules answer to different environments, and the analysis makes the difference legible. The 1970 fulcrum addressed a single concern, incentive-induced gambling among advisor-mediated investors, and it kept the incentive linear by construction. The 2025 schedule faces trusting retail investors no fee instrument can reach, platform-amplified flows, a trust rupture to repair, and an industry the regulator means to keep alive. Read through the model, the Chinese schedule's otherwise puzzling features (the penalty side, the floor, the lot-by-lot settlement) line up margin by margin with the concerns such an environment poses.

The mechanisms are not specific to one market. Any performance-fee schedule chooses how much of the manager's revenue to put at risk, where the thresholds sit, and whether the punishment binds, and the model traces each feature to the margin it controls; the dynamic analysis adds that a bounded penalty floor is where gambling incentives survive. The first evidence from the first large cohort in the Chinese public-fund industry provides several descriptive patterns consistent with the model, while the central behavioral mechanisms await higher-powered tests. The sorting has happened: floating appointments went to managers with stronger pre-appointment track records, the endorsement reversal visible in a single cohort. The young windows behave like the lotteries the theory says they should be: realized tiers pile up at the corners (46 percent of the cohort at the upgrade tier, 38 percent at the downgrade, against a zero-skill lottery's 7 and 23) and flip about every eleven trading days where a mature window's stair should be flat. And the flows read the way the vintage structure says they should: the flow-performance attenuation lives on the subscription margin, where the channel stops feeding new money to floating-contract winners, and the best-performing new products of a bull year were redeemed at a rate that halves a fund in under three quarters, the outflow bearing the settled tier's signature. The contract's remaining claims await the market states that will test them.

References

Admati, A. R. and Pfleiderer, P. (1997). Does it all add up? Benchmarks and the compensation of active portfolio managers. *Journal of Business*, 70(3):323–350.

Asset Management Association of China (中国证券投资基金业协会) (2026). Guidelines on performance appraisal management of fund management companies (in Chinese). Issued April 17, 2026. Original title: 《基金管理公司绩效考核管理指引》.

Barber, B. M., Odean, T., and Zheng, L. (2005). Out of sight, out of mind: The effects of expenses on mutual fund flows. *Journal of Business*, 78(6):2095–2120.

Basak, S., Pavlova, A., and Shapiro, A. (2007). Optimal asset allocation and risk shifting in money management. *Review of Financial Studies*, 20(5):1583–1621.

Berk, J. B. and Green, R. C. (2004). Mutual fund flows and performance in rational markets. *Journal of Political Economy*, 112(6):1269–1295.

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

Brown, K. C., Harlow, W. V., and Starks, L. T. (1996). Of tournaments and temptations: An analysis of managerial incentives in the mutual fund industry. *Journal of Finance*, 51(1):85–110.

Buffa, A. M., Vayanos, D., and Woolley, P. (2022). Asset management contracts and equilibrium prices. *Journal of Political Economy*, 130(12):3146–3201.

Busse, J. A. (2001). Another look at mutual fund tournaments. *Journal of Financial and Quantitative Analysis*, 36(1):53–73.

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

Carpenter, J. N. (2000). Does option compensation increase managerial risk appetite? *Journal of Finance*, 55(5):2311–2331.

Chen, J., Hong, H., Huang, M., and Kubik, J. D. (2004). Does fund size erode mutual fund performance? The role of liquidity and organization. *American Economic Review*, 94(5):1276–1302.

Chevalier, J. and Ellison, G. (1997). Risk taking by mutual funds as a response to incentives. *Journal of Political Economy*, 105(6):1167–1200.

Chevalier, J. and Ellison, G. (1999). Career concerns of mutual fund managers. *Quarterly Journal of Economics*, 114(2):389–432.

China Securities Regulatory Commission (中国证券监督管理委员会) (2025). Action plan for promoting high-quality development of publicly offered funds (in Chinese). CSRC Document [2025] No. 21, issued May 7, 2025. Original title: 《推动公募基金高质量发展行动方案》（证监发〔2025〕21 号）.

China Securities Regulatory Commission (中国证券监督管理委员会) (2026). Guidelines on performance benchmarks of publicly offered securities investment funds (in Chinese). Issued January 23, 2026, effective March 1, 2026. Original title: 《公开募集证券投资基金业绩比较基准指引》.

Chordia, T. (1996). The structure of mutual fund charges. *Journal of Financial Economics*, 41(1):3–39.

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

Cuoco, D. and Kaniel, R. (2011). Equilibrium prices in the presence of delegated portfolio management. *Journal of Financial Economics*, 101(2):264–296.

Dangl, T., Wu, Y., and Zechner, J. (2008). Market discipline and internal governance in the mutual fund industry. *Review of Financial Studies*, 21(5):2307–2343.

Das, S. R. and Sundaram, R. K. (2002). Fee speech: Signaling, risk-sharing, and the impact of fee structures on investor welfare. *Review of Financial Studies*, 15(5):1465–1497.

Dybvig, P. H., Farnsworth, H. K., and Carpenter, J. N. (2010). Portfolio performance and agency. *Review of Financial Studies*, 23(1):1–23.

Elton, E. J., Gruber, M. J., and Blake, C. R. (2003). Incentive fees and mutual funds. *Journal of Finance*, 58(2):779–804.

Goetzmann, W. N., Ingersoll, J. E., and Ross, S. A. (2003). High-water marks and hedge fund management contracts. *Journal of Finance*, 58(4):1685–1718.

Golec, J. and Starks, L. (2004). Performance fee contract change and mutual fund risk. *Journal of Financial Economics*, 73(1):93–118.

Grinblatt, M. and Titman, S. (1989). Adverse risk incentives and the design of performance-based contracts. *Management Science*, 35(7):807–822.

Hong, C. Y., Lu, X., and Pan, J. (2025). FinTech platforms and mutual fund distribution. *Management Science*, 71(1):488–517.

Huang, J., Sialm, C., and Zhang, H. (2011). Risk shifting and mutual fund performance. *Review of Financial Studies*, 24(8):2575–2616.

Huang, J., Wei, K. D., and Yan, H. (2007). Participation costs and the sensitivity of fund flows to past performance. *Journal of Finance*, 62(3):1273–1311.

Kempf, A. and Ruenzi, S. (2008). Tournaments in mutual-fund families. *Review of Financial Studies*, 21(2):1013–1036.

Koijen, R. S. J. and Yogo, M. (2019). A demand system approach to asset pricing. *Journal of Political Economy*, 127(4):1475–1515.

Lin, S., Li, X., Yang, X., and Tam, O. (2009). Are they really stars? Evidence from China's securities investment fund market (in Chinese). *Journal of Financial Research (金融研究)*, (5):107–120. Original title: 《他们真的是明星吗? ——来自中国证券基金市场的经验证据》.

Massa, M. and Patgiri, R. (2009). Incentives and mutual fund performance: Higher performance or just higher risk taking? *Review of Financial Studies*, 22(5):1777–1815.

Nanda, V., Narayanan, M. P., and Warther, V. A. (2000). Liquidity, investment ability, and mutual fund structure. *Journal of Financial Economics*, 57(3):417–443.

Nanda, V. K., Wang, Z. J., and Zheng, L. (2009). The ABCs of mutual funds: On the introduction of multiple share classes. *Journal of Financial Intermediation*, 18(3):329–361.

Ou-Yang, H. (2003). Optimal contracts in a continuous-time delegated portfolio management problem. *Review of Financial Studies*, 16(1):173–208.

Panageas, S. and Westerfield, M. M. (2009). High-water marks: High risk appetites? Convex compensation, long horizons, and portfolio choice. *Journal of Finance*, 64(1):1–36.

Pástor, L., Stambaugh, R. F., and Taylor, L. A. (2015). Scale and skill in active management. *Journal of Financial Economics*, 116(1):23–45.

Pástor, L., Stambaugh, R. F., and Taylor, L. A. (2020). Fund tradeoffs. *Journal of Financial Economics*, 138(3):614–634.

Scharfstein, D. S. and Stein, J. C. (1990). Herd behavior and investment. *American Economic Review*, 80(3):465–479.

Sheng, J., Wang, J., Wang, X., and Yang, J. (2014). Asymmetric contracts, cash flows and risk taking of mutual funds. *Economic Modelling*, 38:435–442.

Shleifer, A. and Vishny, R. W. (1997). The limits of arbitrage. *Journal of Finance*, 52(1):35–55.

Sialm, C., Starks, L. T., and Zhang, H. (2015). Defined contribution pension plans: Sticky or Discerning Money? *Journal of Finance*, 70(2):805–838.

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

Starks, L. T. (1987). Performance incentive fees: An agency theoretic approach. *Journal of Financial and Quantitative Analysis*, 22(1):17–32.

United States Congress (1970). Investment Company Amendments Act of 1970. Pub. L. 91–547, 84 Stat. 1413.

Xiao, J. (2013). Stock-market cycles and fund investors' choices (in Chinese). *China Economic Quarterly (经济学（季刊）)*, 12(4):1299–1320. Original title: 《股市周期与基金投资者的选择》.

Zhongzheng Jinniu (中证金牛座) (2025). The first batch of new floating-fee funds: The full list (in Chinese). 中证金牛座（中国证券报旗下）, May 23, 2025. Original title: 《首批新型浮动费率基金，全名单来了!》.

Zhuang, Y. and Tang, X. (2004). An empirical study of the persistence of fund performance (in Chinese). *Journal of Financial Research (金融研究)*, (5):20–27. Original title: 《基金业绩持续性的实证研究》.

Appendix A Proofs

Conventions. Notation is as in Section 3. Throughout, φ and Φ are the standard normal pdf and cdf, $\bar{\Phi} \equiv 1 - \Phi$, and $V(\alpha, \sigma) \equiv u\,\bar{\Phi}\big(\frac{h-\alpha}{\sigma}\big) - d\,\Phi\big(\frac{-\ell-\alpha}{\sigma}\big)$ is the floating part of the expected fee, with $t_U \equiv (h-\alpha)/\sigma$ and $t_D \equiv (-\ell-\alpha)/\sigma$. The fee slopes at the origin are $V_\alpha \equiv \partial E[f]/\partial\alpha$ and $V_\sigma \equiv \partial E[f]/\partial\sigma$ evaluated at $(\alpha, \sigma) = (0, \sigma_0)$. Quantitative statements use the formal logit flow function ($k = 8$, $c = 1$, $\Lambda = 776.8$bp); the exponential form $Q(e) = Q_0 e^{ke}$ appears only where its closed forms carry the argument, and is labeled each time.

Numerical verification. Steps that rest on computation rather than analysis are flagged in the text as *verified numerically over the calibrated region (Appendix D)*. Gaussian expectations under the logit form are computed by Gauss–Hermite quadrature (96 nodes).

A.1 Lemma 1, Proposition 1, and the shape of the expected fee

The exponential flow form admits a closed-form expectation, which the calibration matching of Section 3.5 and the robustness analysis of Appendix C both use; we record it first, then prove the main-text results in the order in which they appear.

Lemma A.1. *(the exponential flow expectation). Under $Q(e) = Q_0 e^{ke}$ and $e \sim N(\alpha, \sigma^2)$,*

$$E[Q(e)] = Q_0 \, \exp\!\Big(k\alpha + \tfrac{k^2}{2}\sigma^2\Big).$$

Proof. $E[e^{ke}]$ is the moment-generating function of $N(\alpha, \sigma^2)$ evaluated at k. □

Note that $E[Q]$ is increasing in *both* α and σ: the convex flow relation makes the manager love the mean and love the variance. This is one side of every tension in the model; the fee schedule of Lemma 1 is the other.

Lemma 1. *(the contract prices the information ratio). Under the fee schedule of Section 3.2 and $e \sim N(\alpha, \sigma^2)$,*

$$E[f(e)] = f_M + u\,\bar{\Phi}\left(\frac{h-\alpha}{\sigma}\right) - d\,\Phi\left(\frac{-\ell-\alpha}{\sigma}\right),$$

and the floating part $V(\alpha,\sigma) \equiv E[f] - f_M$ satisfies

$$\begin{aligned} &\textit{(i)} \quad \frac{\partial E[f]}{\partial \alpha} = \frac{1}{\sigma}\left[u\,\varphi(t_U) + d\,\varphi(t_D)\right] > 0; \\ &\textit{(ii)} \quad \left.\frac{\partial E[f]}{\partial \sigma}\right|_{\alpha=0} = \frac{1}{\sigma^2}\left[uh\,\varphi\big(\tfrac{h}{\sigma}\big) - d\ell\,\varphi\big(\tfrac{\ell}{\sigma}\big)\right], \end{aligned}$$

where $t_U = (h-\alpha)/\sigma$, $t_D = (-\ell-\alpha)/\sigma$. If the schedule is penalty-dominant ($d\ell \geq uh$ and $h > \ell$), then (ii) is strictly negative for every $\sigma > 0$: at the center of the distribution, variance is strictly bad.

Proof. Since f is piecewise constant on three intervals,

$$E[f] = f_L\,P(e \leq -\ell) + f_M\,P(-\ell < e < h) + f_H\,P(e \geq h).$$

With $e \leq -\ell \iff z \leq t_D$ and $e \geq h \iff z \geq t_U$, substitute $P(e \leq -\ell) = \Phi(t_D)$, $P(e \geq h) = \bar{\Phi}(t_U)$ and collect terms:

$$E[f] = f_M + (f_H - f_M)\,\bar{\Phi}(t_U) - (f_M - f_L)\,\Phi(t_D) = f_M + u\,\bar{\Phi}(t_U) - d\,\Phi(t_D).$$

(i) Both t_U and t_D have derivative $-1/\sigma$ in α. Hence

$$\begin{aligned} \frac{\partial}{\partial \alpha}\big[u\,\bar{\Phi}(t_U)\big] &= -u\,\varphi(t_U)\cdot\Big(-\frac{1}{\sigma}\Big) = \frac{u\,\varphi(t_U)}{\sigma}, \\ \frac{\partial}{\partial \alpha}\big[-d\,\Phi(t_D)\big] &= -d\,\varphi(t_D)\cdot\Big(-\frac{1}{\sigma}\Big) = \frac{d\,\varphi(t_D)}{\sigma}, \end{aligned}$$

and both terms are strictly positive.

(ii) For the variance derivative at a general point, $\partial t_U/\partial\sigma = -(h-\alpha)/\sigma^2$ and $\partial t_D/\partial\sigma = (\ell+\alpha)/\sigma^2$, so

$$\frac{\partial E[f]}{\partial \sigma} = \frac{1}{\sigma^2}\Big[u\,(h-\alpha)\,\varphi(t_U) - d\,(\ell+\alpha)\,\varphi(t_D)\Big],$$

which at $\alpha = 0$ reduces to the display, using that φ is even. For the dominance claim, the sign at $\alpha = 0$ is the comparison of $uh\,e^{-h^2/2\sigma^2}$ with $d\ell\,e^{-\ell^2/2\sigma^2}$. From $d\ell \geq uh$,

$$uh\,e^{-h^2/2\sigma^2} \;\leq\; d\ell\,e^{-h^2/2\sigma^2},$$

and from $h > \ell > 0$, $e^{-h^2/2\sigma^2} < e^{-\ell^2/2\sigma^2}$, which chained with the first inequality gives the strict sign. $\square$

The institutional calibration sits on the boundary of the dominance condition ($d\ell = 60\text{bp} \times 3\% = uh = 30\text{bp} \times 6\%$), and the Gaussian tail factor supplies the strict inequality. Away from $\alpha = 0$ the sign is a race between two exponentials and can flip: the positive cells are concentrated in the high-skill, low-noise corner. This step is verified numerically over the calibrated region (Appendix D): on a fine grid over $\alpha \in [-1\%, 5\%]$ and $\sigma \in [0.5\%, 10\%]$ the derivative is negative in 81.2% of cells, with the positive cells concentrated exactly in that corner. The flip boundary is approximately $\alpha \approx 1.5\%$–1.9% at $\sigma \in \{0.5\%, 1\%, 2\%\}$, and the derivative is negative everywhere on the grid for $\sigma \geq 4\%$.

Lemma 2. *(the break-even alpha). Let $h > \ell$ and $d \geq u$. For each $\sigma > 0$ there is a unique $\bar{\alpha}(\sigma) > 0$ with $E[f] = f_M$, and $E[f] < f_M$ if and only if $\alpha < \bar{\alpha}(\sigma)$.*

Proof. Write $E[f] - f_M = u\,\bar{\Phi}(t_U) - d\,\Phi(t_D)$. At $\alpha = 0$: since $h > \ell$, $\bar{\Phi}(h/\sigma) < \bar{\Phi}(\ell/\sigma) = \Phi(-\ell/\sigma)$, and since $u \leq d$, $u\,\bar{\Phi}(h/\sigma) < d\,\Phi(-\ell/\sigma)$, so $E[f] - f_M < 0$. As $\alpha \to \infty$, $\bar{\Phi}(t_U) \to 1$ and $\Phi(t_D) \to 0$, so $E[f] - f_M \to u > 0$. Continuity of Φ and the strict monotonicity of Lemma 1(i) give a unique root $\bar{\alpha}(\sigma) > 0$. $\square$

Calibrated values (Appendix D): $\bar{\alpha} = 1.77\%, 2.35\%, 3.06\%, 3.83\%$ at $\sigma = 2\%, 4\%, 6\%, 8\%$.

Proposition 1. *(the activity threshold and risk compression). Maintain Assumptions A1 and A3.*

(a) (Extensive margin.) $x^*(\mu) = 0$ if and only if $\mu \leq \mu^*$, where

$$\mu^* \equiv \frac{c_1 - B}{A}, \qquad A \equiv a'(0)\left[V_\alpha + \Lambda E[Q'(\sigma_0 z)]\right] > 0,$$
$$B \equiv \sigma'(0)\left[V_\sigma + \Lambda \sigma_0 E[Q''(\sigma_0 z)] - \rho\sigma_0\right],$$

provided $c_1 > B$ (otherwise $\mu^* < 0$ and every skill level is active). The floating contract raises the threshold relative to a flat fee ($\mu^*_{\mathit{float}} > \mu^*_{\mathit{flat}}$) if and only if the contract's variance tax at the origin outweighs its skill reward evaluated at the flat-contract marginal manager:

$$\sigma'(0)\,|V_\sigma| \;>\; \mu^*_{\mathit{flat}}\, a'(0)\, V_\alpha.$$

(b) (Intensive margin.) Suppose the flat- and floating-contract problems both have interior solutions. Then $x^*_{\mathit{float}}(\mu) \gtreqless x^*_{\mathit{flat}}(\mu)$ according as $\mu\, a'(x^*_{\mathit{flat}})\, V_\alpha \gtreqless \sigma'(x^*_{\mathit{flat}})\,|V_\sigma|$, and there exists a cutoff $\tilde{\mu}$ such that the floating contract compresses the activity of managers with $\mu < \tilde{\mu}$ and amplifies the activity of managers with $\mu > \tilde{\mu}$.

(c) (The two convex benchmarks.) Suppose the flat-contract problem has an interior solution. (i) If the flat optimum's alpha lies below the upgrade line, $\alpha(x^*_{\mathit{flat}}) < h$, the

bonus-only schedule strictly raises activity above the flat benchmark: $x^*_{bonus}(\mu) > x^*_{flat}(\mu)$; a one-sided upside schedule pays for variance as well as for skill. (ii) If moreover the upgrade line sits more than one residual standard deviation above the flat optimum's alpha, $(h-\alpha(x^*_{flat}))/\sigma(x^*_{flat}) > 1$, the symmetric fulcrum schedule does the same: $x^*_{sym}(\mu) > x^*_{flat}(\mu)$. The ordering between the two convex schedules is unsigned: symmetrizing the bonus schedule adds a skill reward and a variance tax of equal size.

Proof of (a). Differentiate the objective of Section 3.3 and evaluate at $x = 0$, where $(\alpha, \sigma) = (0, \sigma_0)$. The fee term contributes

$$\frac{dE[f]}{dx}\Big|_{x=0} = V_\alpha\, \alpha'(0) + V_\sigma\, \sigma'(0) = V_\alpha\, a'(0)\, \mu + V_\sigma\, \sigma'(0),$$

with $V_\alpha = \frac{1}{\sigma_0}[u\,\varphi(h/\sigma_0) + d\,\varphi(\ell/\sigma_0)] > 0$ by Lemma 1(i) and $V_\sigma = \frac{1}{\sigma_0^2}[uh\,\varphi(h/\sigma_0) - d\ell\,\varphi(\ell/\sigma_0)] < 0$ by the penalty-dominance half of Lemma 1(ii). The flow term contributes

$$\Lambda\, \frac{dE[Q]}{dx}\Big|_{x=0} = \Lambda\big[E[Q'(\sigma_0 z)]\, a'(0)\, \mu + \sigma_0\, E[Q''(\sigma_0 z)]\, \sigma'(0)\big],$$

since $\partial E[Q]/\partial\alpha = E[Q']$ and $\partial E[Q]/\partial\sigma = E[Q'\, z] = \sigma_0\, E[Q'']$ by Stein's lemma. (Under the exponential form of Lemma A.1 these moments reduce to Λk and $\Lambda k^2 \sigma_0$; under the logit form at $c = 1$ the second moment vanishes exactly, Lemma A.4 below.) The risk and cost terms contribute $-\rho\sigma_0\sigma'(0)$ and $-c_1$. Collecting,

$$\begin{aligned} U_x(0;\mu) = {} & \underbrace{\mu\, a'(0)\, [V_\alpha + \Lambda\, E[Q'(\sigma_0 z)]]}_{\text{skill channel } (>0)} + \underbrace{\sigma'(0)\, [V_\sigma + \Lambda\, \sigma_0\, E[Q''(\sigma_0 z)] - \rho\sigma_0]}_{\text{variance channel (sign indeterminate)}} \\ & - \; c_1 \;=\; A\mu + B - c_1. \end{aligned}$$

By Assumption A3, U is strictly concave, so $x^* = 0 \iff U_x(0;\mu) \le 0 \iff \mu \le (c_1 - B)/A$, with $A > 0$ because $V_\alpha > 0$ and Q is strictly increasing.

For the comparison, write the flat-contract constants $A_{\text{flat}} = a'(0)\, \Lambda E[Q']$ and $B_{\text{flat}} = \sigma'(0)[\Lambda\sigma_0 E[Q''] - \rho\sigma_0]$, so that $\mu^*_{\text{flat}} = (c_1 - B_{\text{flat}})/A_{\text{flat}}$ and

$$\mu^*_{\text{float}} = \frac{c_1 - B_{\text{flat}} - \sigma'(0)V_\sigma}{A_{\text{flat}} + a'(0)V_\alpha}.$$

Set $N \equiv c_1 - B_{\text{flat}} = \mu^*_{\text{flat}} A_{\text{flat}}$. The inequality $\frac{N - \sigma'(0)V_\sigma}{A_{\text{flat}} + a'(0)V_\alpha} > \frac{N}{A_{\text{flat}}}$ cross-multiplies (both denominators positive) to $-\sigma'(0)V_\sigma\, A_{\text{flat}} > N\, a'(0)V_\alpha$; substituting $N = \mu^*_{\text{flat}} A_{\text{flat}}$ and cancelling $A_{\text{flat}} > 0$ gives the stated condition, since $V_\sigma = -|V_\sigma|$ implies $-\sigma'(0)V_\sigma =$

$\sigma'(0)|V_\sigma|$. $\square$

The closed-form thresholds on the nine-cell ($\sigma_0 \times \Lambda$) grid are reported in Section 3.6; the closed form is the exact solution of $U_x(0;\mu) = 0$, and a fine-grid global argmax check (step 0.05%) coincides with it up to grid rounding (Appendix D). Two boundary facts are part of the same computation: at $\sigma_0 = 0.5\%$ the thresholds sit too many standard deviations away and the flat and floating thresholds coincide to three digits; and as $\sigma_0 \to 0$ both $V_\alpha, V_\sigma \to 0$ and the contract loses all marginal bite near $x = 0$, the boundary of Assumption A1.

The next lemma is the penalty comparative static used in Section 3.6's one-sidedness discussion.

Lemma A.2. *(the penalty comparative static). Let $x^*(\mu) \in (0, \bar{x})$ be interior. Then*

$$\operatorname{sign}\left(\frac{\partial x^*}{\partial d}\right) = \operatorname{sign}\Big(\alpha'(x^*)\,\sigma(x^*) - \sigma'(x^*)\,\big(\ell + \alpha(x^*)\big)\Big).$$

Under the linear technologies $a(x) = x$, $\sigma(x) = \sigma_0 + sx$, the condition collapses to a single skill comparison: $\partial x^*/\partial d > 0 \iff \mu > \hat{\mu} \equiv s\ell/\sigma_0$.

Proof. An interior optimum satisfies $U_x = 0$; by the implicit function theorem $\partial x^*/\partial d = -U_{xd}/U_{xx}$, and $U_{xx} < 0$ by Assumption A3, so the sign is that of U_{xd}. The parameter d appears only in the fee term $-d\,\Phi(t_D)$, hence

$$U_{xd} = \frac{\partial}{\partial d}\frac{dE[f]}{dx} = -\frac{d}{dx}\Phi(t_D) = -\varphi(t_D)\,t'_D(x).$$

With $t_D(x) = \frac{-\ell - \alpha(x)}{\sigma(x)}$,

$$t'_D(x) = \frac{-\alpha'(x)\,\sigma(x) + \sigma'(x)\,(\ell + \alpha(x))}{\sigma^2(x)},$$

so $U_{xd} = \dfrac{\varphi(t_D)}{\sigma^2}\big[\alpha'\sigma - \sigma'(\ell + \alpha)\big]$, and $\varphi > 0$, $\sigma^2 > 0$. Under the linear technologies the bracket is $\mu(\sigma_0 + sx) - s(\ell + \mu x) = \mu\sigma_0 - s\ell$, whose sign is that of $\mu - \hat{\mu}$. $\square$

The one-sidedness of the static is a calibration statement, not a theorem: $\hat{\mu} = 15\%$ at $\sigma_0 = 1\%$ (above any plausible skill) but $\hat{\mu} = 7.5\%$ at $\sigma_0 = 2\%$, near the top of the plausible range; the main text restricts the claim to $\sigma_0 \in [0.5\%, 1\%]$. This step is verified numerically over the calibrated region (Appendix D): halving d from 60bp to 30bp raises x^* at every interior grid point, as the lemma predicts.

Proof of (b). The flat-contract optimum satisfies $U_x^{\text{flat}}(x^*_{\text{flat}}) = 0$. The floating objective differs by the floating part of the fee, so $U_x^{\text{float}} = U_x^{\text{flat}} + dV/dx$, where

$$\frac{dV}{dx} = \mu\, a'(x)\, V_\alpha\big(\alpha(x), \sigma(x)\big) + \sigma'(x)\, V_\sigma\big(\alpha(x), \sigma(x)\big),$$

with the slopes evaluated at the induced return distribution. Evaluating at $x = x^*_{\text{flat}}$ gives $U_x^{\text{float}}(x^*_{\text{flat}}) = dV/dx$; strict concavity makes the comparison monotone, so the optimum moves in the sign of this derivative, and with $V_\sigma < 0$,

$$x^*_{\text{float}} \gtreqless x^*_{\text{flat}} \iff \mu\, a'(x^*_{\text{flat}})\, V_\alpha \;\gtreqless\; \sigma'(x^*_{\text{flat}})\, |V_\sigma|,$$

which is the stated sign condition.

For the existence and uniqueness of the cutoff $\tilde{\mu}$ we give a sufficient condition and a numerical bound. Freeze the fee slopes (V_α, V_σ) at their calibration values (their values at the calibration center) so that they do not move with μ. Under the linear technologies the comparison then reads

$$\operatorname{sign}\Big(\frac{dV}{dx}\Big) = \operatorname{sign}\big(\mu\, V_\alpha + s\, V_\sigma\big),$$

which is affine and strictly increasing in μ (since $V_\alpha > 0$), negative at $\mu = 0$ (since $V_\sigma < 0$), and positive for μ large. Hence, under the frozen-slope condition, there is exactly one crossing,

$$\tilde{\mu} = \frac{s\, |V_\sigma|}{V_\alpha},$$

with compression below and amplification above. Unfreezing the slopes perturbs this exactly: $V_\alpha(\alpha, \sigma)$ and $V_\sigma(\alpha, \sigma)$ are smooth functions of the induced distribution, which moves with μ through $x^*_{\text{flat}}(\mu)$, and their movement enters the comparison only at second order: the slopes multiply first-order terms and their own drift is $O\big((\alpha, \sigma - \sigma_0)\big)$ along the comparison path. This step is verified numerically over the calibrated region (Appendix D): on a skill grid with step 0.05%, the sign of dV/dx evaluated at x^*_{flat} crosses zero exactly once, at $\tilde{\mu} \approx 5.1\%$ under the logit flow function ($\approx 4.2\%$ under the exponential approximation), stable across $\sigma_0 \in \{0.5\%, 1\%, 2\%\}$, and the second-order movement of (V_α, V_σ) over the grid does not change the sign of the crossing. $\square$

Proof of (c). Both comparisons reuse the sign argument of part (b): the flat optimum satisfies $U_x^{\text{flat}} = 0$, the alternative contract's objective adds its fee schedule net of f_M, and strict concavity moves the optimum in the sign of $\mu\, a'(x^*_{\text{flat}})\, \Delta V_\alpha + \sigma'(x^*_{\text{flat}})\, \Delta V_\sigma$, where

ΔV is the alternative contract's expected fee net of the base tier.

For the bonus-only schedule, $\Delta V = u\,\bar\Phi\big((h-\alpha)/\sigma\big)$, so

$$\Delta V_\alpha = \frac{u}{\sigma}\,\varphi(t_U) > 0, \qquad \Delta V_\sigma = \frac{u\,(h-\alpha)}{\sigma^2}\,\varphi(t_U),$$

with $t_U = (h-\alpha)/\sigma$. Under the stated condition $\alpha(x^*_{\text{flat}}) < h$ both terms of the comparison are nonnegative, and the first is strictly positive for $\mu > 0$; hence $x^*_{\text{bonus}} > x^*_{\text{flat}}$. The content is that a one-sided upside schedule pays for variance: deviation buys upgrade probability even when it buys no skill.

For the symmetric fulcrum, $\Delta V = u\,\bar\Phi(t_U) - u\,\Phi(t_D)$ with $t_D = (-h-\alpha)/\sigma$, so

$$\Delta V_\alpha = \frac{u}{\sigma}\big[\varphi(t_U) + \varphi(t_D)\big] > 0, \qquad \Delta V_\sigma = \frac{u}{\sigma^2}\big[(h-\alpha)\,\varphi(t_U) - (h+\alpha)\,\varphi(t_D)\big].$$

The function $t \mapsto t\,\varphi(t)$ is strictly decreasing on $(1,\infty)$. Since φ is even and an interior flat optimum at $\mu > 0$ has $\alpha(x^*_{\text{flat}}) > 0$, the bracket compares $t\,\varphi(t)$ at $(h-\alpha)/\sigma$ against its value at the larger $(h+\alpha)/\sigma$, and it is positive whenever $t_U > 1$, which is the stated condition $(h-\alpha(x^*_{\text{flat}}))/\sigma(x^*_{\text{flat}}) > 1$.

For the last claim, symmetrizing the bonus schedule changes the slopes by $\Delta V_\alpha = u\,\varphi(t_D)/\sigma > 0$ and $\Delta V_\sigma = -u\,(h+\alpha)\,\varphi(t_D)/\sigma^2 < 0$: the comparison between the two convex schedules is a skill reward against a variance tax of the same size u and has no uniform sign. $\square$

We close the section with the standing concavity statement behind Assumption A3.

Lemma A.3. *(concavity: a regional statement with an explicit cost bound). A sufficient condition for $U_{xx} < 0$ at a point x is*

$$c_2 \;>\; \frac{d^2}{dx^2}\Big[E[f] + \Lambda E[Q]\Big](x) \;-\; \rho\big[\sigma'(x)^2 + \sigma(x)\,\sigma''(x)\big],$$

where the fee term's second derivative is finite and bounded uniformly in x: each second partial of $E[f]$ is a sum of terms of the form (jump size) $\times$ (a Gaussian density or one of its first two derivatives, bounded by $\varphi(0)$ or $\varphi(1)$) $\times$ (a polynomial in the thresholds) / (a power of σ), and $\sigma \geq \sigma_0 > 0$; explicitly, $|\partial^2 E[f]/\partial\alpha^2| \leq (u+d)\,\varphi(1)/\sigma_0^2$, with analogous bounds for the mixed and variance partials; and the flow term's curvature is bounded on any compact parameter region (under the logit form, globally, since $|Q''| \leq k^2/4$). Hence, on any region where the right-hand side is bounded, a large enough c_2 makes U strictly concave throughout the region; the supremum over a given region is estimable.

Proof. Direct computation: $U_{xx} = \frac{d^2}{dx^2}[E[f] + \Lambda E[Q]] - \rho[\sigma'^2 + \sigma\sigma''] - C''(x)$ with $C'' = c_2$; the displayed condition rearranges $U_{xx} < 0$. For the fee bound, $\partial^2 E[f]/\partial\alpha^2 = -\frac{1}{\sigma^2}[u\,\varphi'(t_U) + d\,\varphi'(t_D)]$ with $|\varphi'| \leq \varphi(1)$; the mixed and variance second partials expand into terms of the stated form using $|\varphi| \leq \varphi(0)$, $|\varphi''| \leq \varphi(0)$ and $\sigma \geq \sigma_0$, and the chain rule with $a'' \leq 0$, $\sigma'' \geq 0$ collects them. Under the logit form $Q'' = k^2 Q(1-Q)(1-2Q)$, so $|Q''| \leq k^2/4$ and $E[Q'']$ is bounded independently of (α, σ). $\square$

The bound is a sufficient condition, not a description; this step is verified numerically over the calibrated region (Appendix D). At baseline flow strength the objective is strictly concave over the whole choice set for $\mu \lesssim 5\%$; convex regions appear at high deviation for larger μ or larger Λ, and where they appear the model predicts a jump to the corner $\bar{x}$—economic content (hot-money gambling), not a mathematical nuisance. Every parameter point at which Propositions 1–3 are evaluated lies inside the verified concave region.

The last result of this section is not a proof of a text proposition but the formal content behind two remarks in Section 3: skill beliefs enter the model only as an isomorphic variance.

Extension. *(skill beliefs are an isomorphic variance). Suppose the manager's belief about his per-unit skill is Gaussian, $\hat{\mu} \sim N(\mu, \tau^2)$, independent of the return noise, and he maximizes his expected objective under the belief. Then his problem is the baseline problem with the residual standard deviation map replaced by*

$$\tilde{\sigma}(x) \;=\; \sqrt{\sigma(x)^2 + \tau^2 x^2}$$

inside the fee and flow terms. Two implications. (i) The activity threshold is exactly invariant: $\tilde{\sigma}(0) = \sigma_0$ and $d\tilde{\sigma}^2/dx = d\sigma^2/dx$ at $x = 0$, so $U_x(0;\mu)$, and with it Proposition 1(a)'s μ^*, does not depend on τ. (ii) Belief dispersion acts only through the intensive margin: it raises the effective marginal variance of deviation to $d\tilde{\sigma}^2/dx = d\sigma^2/dx + 2\tau^2 x$, strictly so for $x > 0$, and where the fee's variance price is negative (the penalty-dominance sign of Lemma 1) interior optima fall.

Proof. Conditional on the belief, the upgrade probability is $P(\hat{\mu}x + \sigma z > h) = P\big(\sqrt{\sigma^2 + \tau^2 x^2}\, z' > h - \mu x\big)$ by the convolution of independent Gaussians, so $E_{\text{belief}}[\bar{\Phi}((h - \hat{\mu}x)/\sigma)] = \bar{\Phi}((h - \mu x)/\tilde{\sigma})$; the downgrade probability and $E[Q]$ transform identically, and the risk and cost terms are unaffected. Part (i) evaluates the map at $x = 0$; part (ii) differentiates $\tilde{\sigma}^2$ and applies the sign argument of Proposition 1(b). $\square$

A.2 Proposition 2: flow swamping

This section contains the appendix's main proofs: under the formal logit flow function the swamping statement is delivered in closed form, because the flow term's variance appetite vanishes exactly at the calibration center. We prove that symmetry fact first, then the form-free limit, then the closed forms and the asymptotic expansion, and finally collect the exponential approximation's sharper coordinates, which are artifacts of that form's unbounded variance appetite, and are labeled as such.

Lemma A.4. *(the logit's variance appetite vanishes at the center). Let $Q(e) = e^{ke}/(e^{ke}+c)$. At $c = 1$ and $\alpha = 0$,*

$$\left.\frac{\partial E[Q]}{\partial \sigma}\right|_{\alpha=0} \;=\; E\big[Q'(\sigma_0 z)\, z\big] \;=\; \sigma_0\, E\big[Q''(\sigma_0 z)\big] \;\equiv\; 0,$$

while $E\big[Q'(\sigma_0 z)\big] > 0$. For $c \neq 1$ the identity fails.

Proof. Differentiate $E[Q] = \int Q(\alpha+\sigma z)\,\varphi(z)\,dz$ with respect to σ: $\partial E[Q]/\partial\sigma = \int z\,Q'(\alpha+\sigma z)\,\varphi(z)\,dz$, which at $(0, \sigma_0)$ is $E[Q'(\sigma_0 z)\,z]$. At $c = 1$,

$$Q'(e) = \frac{k\,e^{ke}}{(e^{ke}+1)^2} = \frac{k}{4\,\cosh^2(ke/2)},$$

an even function of e. Hence $Q'(\sigma_0 z)$ is even in z, the integrand $z\,Q'(\sigma_0 z)\,\varphi(z)$ is odd, and the expectation vanishes exactly. Stein's lemma, $E[z\,g(z)] = E[g'(z)]$ with $g(z) = Q'(\sigma_0 z)$, gives the second equality $E[Q'(\sigma_0 z)\,z] = \sigma_0\,E[Q''(\sigma_0 z)]$, so the curvature moment vanishes with it. $E[Q'] > 0$ since $Q' > 0$ everywhere. For $c \neq 1$, $Q'(e) = kc\,e^{ke}/(e^{ke}+c)^2$ is symmetric about $ke = \ln c$, not about zero, and the oddness argument fails. $\square$

The identity is exact, not approximate: numerically the moment evaluates to 4.4×10^{-17}, machine precision (Appendix D). The economic content is that the flow term's variance appetite (the term $\Lambda k^2 \sigma_0$ that powers the exponential form's death points below) is switched off at the calibration center by the logit's symmetry. The appetite is bounded in any case: $kQ(1-Q) \leq k/4$, so $|E[Q']| \leq k/4$ and $|E[Q'']| \leq k^2/4$ whatever the parameters. For $c \neq 1$ the identity fails but the boundedness remains: a scan of the saturation constant (Appendix D) finds $E[Q'z] = -4.7 \times 10^{-2}$ at $c = 0.5$ and $+4.7 \times 10^{-2}$ at $c = 2$, and even at $c = 2$ (where the sign matches the exponential form's appetite) the finite death points are postponed to 16.4 and 18.6 times baseline flow, more than three times the exponential coordinates.

Proposition 2. *(flow swamping). Fix all other parameters. As* $\Lambda \to \infty$*,*

$$|x^*_{\mathit{float}}(\mu) - x^*_{\mathit{flat}}(\mu)| \longrightarrow 0 \qquad \textit{and} \qquad \mu^*_{\mathit{float}} - \mu^*_{\mathit{flat}} \longrightarrow 0.$$

Proof. The contract term in the first-order condition is bounded independently of Λ: $|dV/dx| \le a'(x)\,|V_\alpha| + \sigma'(x)\,|V_\sigma|$ with $|V_\alpha| \le (u+d)\,\varphi(0)/\sigma_0$ and $|V_\sigma| \le (u+d)\,\varphi(1) \cdot C(h,\ell,\bar{\alpha})/\sigma_0^2$ (Lemma A.3's bounds), since φ is bounded and $\sigma \ge \sigma_0 > 0$. The risk and cost terms do not contain Λ at all. The flow term scales linearly in Λ. Divide the first-order condition $U_x(x;\mu,\Lambda) = 0$ by Λ:

$$\frac{1}{\Lambda}\Big[\frac{dV}{dx} - \rho\,\sigma\sigma' - C'(x)\Big] + \frac{dE[Q]}{dx} = 0,$$

and the bracket vanishes as $\Lambda \to \infty$, uniformly in x. Both contracts' first-order conditions therefore converge to the same limit equation (the flow-only problem), whose solution is independent of the contract. With the limit problem's optimum unique on the calibrated region (Assumption A3, Lemma A.3), continuity of the argmax gives $|x^*_{\text{float}} - x^*_{\text{flat}}| \to 0$. For the thresholds, Proposition 1(a) writes each as $\mu^*_c = (c_1 - B_c)/A_c$ with $A_c = a'(0)[V^c_\alpha + \Lambda E[Q']]$ growing linearly in Λ and B_c growing at most linearly (under the logit, B_c is Λ-free by Lemma A.4), so both thresholds tend to their common limit and $\mu^*_{\text{float}} - \mu^*_{\text{flat}} \to 0$. The argument uses nothing about the flow function beyond the linear growth of the flow term in Λ and the boundedness of the contract term: the limit is form-free, and holds for the exponential and logit forms alike. □

A.2.1 The logit closed forms: rational thresholds and the asymptotic expansion

Under the logit form at $c = 1$, Lemma A.4 collapses both thresholds of Proposition 1(a) into rational functions of Λ. Write $E[Q'] \equiv E[Q'(\sigma_0 z)]$ and specialize to the linear technologies ($a'(0) = 1$, $\sigma'(0) = s$). Since the variance channel's flow moment vanishes,

$$\mu^*_{\text{flat}}(\Lambda) = \frac{M_f}{\Lambda\,E[Q']}, \qquad \mu^*_{\text{float}}(\Lambda) = \frac{M_p}{V_\alpha + \Lambda\,E[Q']},$$
$$M_f \equiv c_1 + s\rho\sigma_0, \qquad M_p \equiv M_f - s\,V_\sigma \;>\; M_f \;>\; 0,$$

both **exact**: no approximation is involved, because at $x = 0$ the first-order condition evaluates the flow moments at the calibration center, where Lemma A.4 applies. Three consequences follow immediately.

No death point. Both thresholds are strictly positive at every finite Λ and tend to zero at the rate $1/\Lambda$ without crossing: universal participation is the limit, not a crossing. The exponential form's finite sign-change points (below) have no logit counterpart, because the variance-appetite term that manufactures them is identically zero here.

The gap. The threshold gap is itself a rational function,

$$g(\Lambda) \equiv \mu^*_{\text{float}}(\Lambda) - \mu^*_{\text{flat}}(\Lambda) = \frac{M_p}{V_\alpha + \Lambda E[Q']} - \frac{M_f}{\Lambda E[Q']}.$$

The asymptotic expansion. Let $L \equiv \Lambda E[Q']$ and expand $M_p/(V_\alpha + L) = \frac{M_p}{L}\big(1 - \frac{V_\alpha}{L} + O(L^{-2})\big)$. Collecting terms,

$$g(\Lambda) = \frac{C_1'}{\Lambda} + \frac{C_2'}{\Lambda^2} + O(\Lambda^{-3}), \qquad C_1' = \frac{-s\,V_\sigma}{E[Q']} \;>\; 0, \qquad C_2' = -\frac{M_p\,V_\alpha}{(E[Q'])^2} \;<\; 0.$$

The sign of the leading constant is exactly the penalty-dominance condition of Lemma 1: $C_1' > 0$ because the contract taxes variance at the origin ($V_\sigma < 0$), and $C_2' < 0$ because it pays for skill ($V_\alpha > 0$). Penalty dominance is therefore what makes the contract bind *in the same direction* however hot the market runs; heat decides how much bite remains, never its sign. The negative second-order constant means the simple $1/\Lambda$ formula systematically overstates the residual bite.

Calibration (Appendix D; Gauss–Hermite, 96 nodes). $E[Q'] = 1.9968$, $V_\alpha = 2.73 \times 10^{-3}$, $V_\sigma = -8.01 \times 10^{-3}$, $M_f = 3.000 \times 10^{-3}$, $M_p = 3.399 \times 10^{-3}$; hence

$$C_1' = 2.00 \times 10^{-4}, \qquad C_2' = -2.27 \times 10^{-6}.$$

At the baseline the exact gap is 22.02bp against the first-order value 25.72bp; the $1/\Lambda$ formula overstates the residual bite by 16.8% (7.8% at 2Λ, 3.0% at 5Λ). Measured on the grid, the gap falls to half its baseline value at 2.21 times baseline flow, to a quarter at 4.58 times, and to a tenth at 12.0 times. A fine-grid global argmax check (skill step 0.05%) coincides with the closed forms up to grid rounding: baseline thresholds 1.95%/2.20% on the grid against 1.934%/2.154% in closed form.

The intensive margin. The second half of Proposition 2's content (the convergence of the two contracts' optima) is smooth under the logit form. This step is verified numerically over the calibrated region (Appendix D): scanning Λ in steps of one baseline multiple, x^*_{float} and x^*_{flat} drift upward at an approximately logarithmic rate with no jump anywhere on $\mu \in \{2\%, 3\%, 5\%\}$, their behavioral difference vanishes well before either reaches the deviation bound, and the corner is reached smoothly: at about 10 times baseline flow for $\mu = 5\%$, 15 times for $\mu = 3\%$, and 22 times for $\mu = 2\%$. The intensive-margin gap is non-monotone along the way: at $\mu = 3\%$ it moves from -0.035 at baseline flow through zero near 2.5 times baseline to $+0.025$ at 4 times before converging, because flow amplification pushes both optima up the upgrade side of the schedule, so the compression region first shrinks and then disappears as the market heats.

A.2.2 The exponential approximation: closed forms, and why its sharper coordinates are artifacts

Under the exponential form of Lemma A.1 the same objects have closed forms with more dramatic coordinates. We record them for comparison and identify precisely which feature of that form manufactures them.

Proposition A.5. *(exponential closed forms). Under $Q(e) = Q_0 e^{ke}$, write $a_0' \equiv a'(0)$, $s_0 \equiv \sigma'(0)$, and M_f, M_p as above. Then:*

(i) (Flat threshold, exact.)

$$\mu^*_{\mathit{flat}}(\Lambda) = \bar{\mu} + \frac{M_f}{a_0'\, k\, \Lambda}, \qquad \bar{\mu} \equiv -\frac{s_0\, \sigma_0\, k}{a_0'} < 0.$$

(ii) (Gap expansion.)

$$g(\Lambda) = \frac{C_1}{\Lambda} + \frac{C_2}{\Lambda^2} + O(\Lambda^{-3}),$$
$$C_1 = \frac{s_0\big(\sigma_0 k\, V_\alpha - V_\sigma\big)}{a_0'\, k} > 0, \qquad C_2 = -\frac{V_\alpha\big(M_p + s_0 \sigma_0 k V_\alpha\big)}{a_0'\, k^2} < 0.$$

(iii) (Death points.) The thresholds change sign at explicit finite multiples:

$$\mu^*_{\mathit{flat}} < 0 \iff \Lambda > \Lambda^0_f \equiv \frac{M_f}{s_0\, \sigma_0\, k^2}, \qquad \mu^*_{\mathit{float}} < 0 \iff \Lambda > \Lambda^0_p \equiv \frac{M_p}{s_0\, \sigma_0\, k^2},$$
$$\Lambda^0_p - \Lambda^0_f = \frac{-V_\sigma}{\sigma_0\, k^2} > 0.$$

Proof. (i) For the flat contract $A = a_0' \Lambda k$ and $B = s_0(\Lambda k^2 \sigma_0 - \rho \sigma_0)$; substituting into $\mu^* = (c_1 - B)/A$ and splitting the numerator gives the display with no approximation. (iii) Each threshold's numerator is affine in Λ, so its sign change is explicitly solvable; the gap between the two death points is $(M_p - M_f)/(s_0 \sigma_0 k^2) = -V_\sigma/(\sigma_0 k^2)$. (ii) The floating threshold is $\mu^*_{\text{float}} = \big(M_p - s_0\sigma_0 k^2\Lambda\big)/\big(a'_0(V_\alpha + k\Lambda)\big)$. Expand $1/(V_\alpha + k\Lambda) = \frac{1}{k\Lambda}\big(1 - \frac{V_\alpha}{k\Lambda} + O(\Lambda^{-2})\big)$, multiply by the numerator, and collect the $1/\Lambda$ and $1/\Lambda^2$ terms; the $1/\Lambda$ coefficient is $\frac{M_p + s_0\sigma_0 k V_\alpha}{a'_0 k} - \frac{M_f}{a'_0 k} = \frac{s_0(\sigma_0 k V_\alpha - V_\sigma)}{a'_0 k}$ (the derivation does not use $a'_0 = 1$). $C_1 > 0$ is guaranteed jointly by $V_\alpha > 0$ (Lemma 1(i)) and $V_\sigma < 0$ (Lemma 1(ii)): penalty dominance is exactly the condition that the threshold gap is asymptotically sign-definite. □

Calibration and the artifact diagnosis (Appendix D). $C_1 = 5.12 \times 10^{-5}$, $C_2 = -1.42 \times 10^{-7}$; the two-term expansion is accurate to 0.33% already at baseline flow, and the one-term formula overstates the gap by 17.8% at baseline. The baseline gap is 24.1bp, with half-life 2.22 times baseline, a quarter at 4.61 times, and death points at 5.2 and 5.9 times baseline. On the intensive margin, both contracts' optima jump from interior values (≈ 0.67) to the corner $\bar{x} = 2$ and bunch there at about 2.6–2.7 times baseline flow: contract irrelevance arrives as universal corner gambling, not as gentle convergence.

These sharper coordinates are artifacts of the exponential form's *unbounded variance appetite*, and we label them as such wherever they appear. The mechanism is visible in the closed forms: the flow term's curvature, $\Lambda k^2\sigma\sigma'\, e^{k\alpha + k^2\sigma^2/2}$, grows without bound in the variance, so the objective develops a convex region at high deviation and the interior optimum is destroyed in a jump. The death points of part (iii) are the same term read on the extensive margin: the appetite $\Lambda k^2\sigma_0$ eventually covers the cost and risk terms by itself, and activity requires no skill at all. The logit's saturation removes the amplifier: $kQ(1-Q) \leq k/4$, and at the calibration center the appetite vanishes exactly (Lemma A.4), and with it the jump, the corner bunching, and the death points; the limit itself is form-free, as Proposition 2 states. The exponential coordinates are retained here, and only here, as the local-approximation benchmark against which the formal specification is the conservative reading.

A.3 Proposition 3: the two-peak risk profile

The setting is Section 3.8's: the year splits at mid-year, e_1 is realized and public, second-half performance is $e_2 = \bar{\alpha}_2 + \sigma_2 z_2$ with $\bar{\alpha}_2$ fixed and $\sigma_2 \in [\underline{\sigma}, \bar{\sigma}]$ chosen at mid-year, and the manager maximizes $U_2(\sigma_2; e_1) = E[f(e)] + \Lambda E[Q(e)] - \frac{\rho}{2}\sigma_2^2$ over the full-year result $e = e_1 + e_2 \sim N(E_2, \sigma_2^2)$, where $E_2 \equiv e_1 + \bar{\alpha}_2$. We prove the four regional lemmas and then assemble the proposition. The flow function in this section is the formal logit form.

Lemma A.6. *(the fee's marginal value of risk). For the step schedule* $f(e) = f_M + u\,\mathbf{1}\{e \geq$

$h\} - d\,\mathbf{1}\{e < -\ell\}$,

$$\frac{\partial E[f]}{\partial \sigma_2} = \frac{G(E_2, \sigma_2)}{\sigma_2^2}, \qquad G(E_2, \sigma_2) \equiv u\,(h - E_2)\,\varphi\!\left(\frac{h - E_2}{\sigma_2}\right) - d\,(\ell + E_2)\,\varphi\!\left(\frac{\ell + E_2}{\sigma_2}\right).$$

Proof. $E[f] = f_M + u\,[1 - \Phi((h - E_2)/\sigma_2)] - d\,\Phi((-\ell - E_2)/\sigma_2)$. With $a = (h - E_2)/\sigma_2$, $\partial a/\partial \sigma_2 = -(h - E_2)/\sigma_2^2$, so $\frac{\partial}{\partial \sigma_2}[1 - \Phi(a)] = (h - E_2)\varphi(a)/\sigma_2^2$. With $b = (-\ell - E_2)/\sigma_2$, $\partial b/\partial \sigma_2 = (\ell + E_2)/\sigma_2^2$, so $\frac{\partial}{\partial \sigma_2}\Phi(b) = \varphi(b)(\ell + E_2)/\sigma_2^2$. Collect, using that φ is even. □

The two terms of G are the two incentive sources. The upgrade term is positive below the upgrade line (risk buys touch probability) and negative above it (a manager already above the line protects his tier). The downgrade term is negative above the downgrade line (risk deepens the left tail) and **positive below it**: a manager who has already fallen through holds a fee on the floor: falling further costs nothing, climbing back is worth the full step.

Lemma A.7. *(the flow term's sign).* $\frac{\partial E[Q]}{\partial \sigma_2} = \sigma_2\,E[Q''(e)]$, *and* $Q''(e) \gtrless 0$ *as* $ke \lessgtr \ln c$*: the flow term encourages risk at low performance and discourages it near saturation. In the deep left tail,*

$$\frac{\partial E[Q]}{\partial \sigma_2} \approx \frac{k^2 \sigma_2}{c}\, e^{kE_2 + k^2\sigma_2^2/2} \longrightarrow 0^+ \qquad (E_2 \to -\infty).$$

Proof. Differentiate $E[Q] = \int Q(E_2 + \sigma_2 z)\varphi(z)\,dz$ in σ_2 and apply Stein's lemma with $g(z) = Q'(E_2 + \sigma_2 z)$: the derivative equals $\sigma_2 \int Q''(E_2 + \sigma_2 z)\varphi(z)\,dz = \sigma_2 E[Q''(e)]$. For the sign, $Q' = kQ(1-Q)$, so $Q'' = k^2 Q(1-Q)(1-2Q)$, positive iff $Q < 1/2$ iff $ke < \ln c$. In the deep tail $Q \approx e^{ke}/c$, so $E[Q] \approx \frac{1}{c}E[e^{ke}] = \frac{1}{c}e^{kE_2 + k^2\sigma_2^2/2}$ by Lemma A.1; differentiating gives the display, and the exponential factor dominates every polynomial factor as $E_2 \to -\infty$. □

The first-order condition is therefore

$$\frac{\partial U_2}{\partial \sigma_2} = \frac{G(E_2, \sigma_2)}{\sigma_2^2} + \Lambda\,\sigma_2\,E[Q''] - \rho\,\sigma_2 = 0,$$

with an interior solution where the first two terms balance the risk cost, and the floor $\underline{\sigma}$ (lock-in) or the ceiling $\bar{\sigma}$ (full-stakes gambling) determined by the sign. The regional analysis divides the state space by the position of E_2 relative to $-\ell$ and h.

Lemma A.8. *(the locked regions). (a) If $E_2 > h$: both terms of G are negative, and the flow term is negative as well (the distribution sits in the concave region), so $\partial U_2/\partial\sigma_2 < 0$ everywhere and $\sigma_2^* = \underline{\sigma}$, the strongest lock-in in the state space. (b) If $E_2 \in (-\ell,\, h-\delta)$, with δ of the order of σ_2: the upgrade term is exponentially negligible and the downgrade term is negative, so $G < 0$ and $\sigma_2^* = \underline{\sigma}$ or near it.*

Proof. (a) $E_2 > h$ gives $h - E_2 < 0$ (upgrade term negative) and $\ell + E_2 > 0$ (downgrade term negative); and $E_2 > h$ implies $kE_2 > kh > \ln c$; at the calibrated $c = 1$ this reduces to $kh > 0$, which holds automatically for every $k > 0$, so the distribution's mass sits in the concave region and $E[Q''] < 0$. All three terms of the first-order condition (FOC) are negative. (b) The downgrade term is $-d(\ell + E_2)\varphi((\ell + E_2)/\sigma_2) < 0$; the upgrade term's coefficient $u(h - E_2)$ is positive but, when E_2 is more than about $2\sigma_2$ below h, the Gaussian density makes the term arbitrarily small. □

Lemma A.9. *(the catch-up gambling zone). If the contract has a floor ($d > 0$), then for $E_2 \in (-\ell - 2\bar{\sigma},\, -\ell)$ the downgrade term of G is positive:*

$$-d\,(\ell + E_2)\,\varphi\!\left(\frac{\ell + E_2}{\sigma_2}\right) = d\,|\ell + E_2|\,\varphi\!\left(\frac{|\ell + E_2|}{\sigma_2}\right) > 0,$$

and as a function of $|\ell + E_2|$ it is maximized at $|\ell + E_2| = \sigma_2$, where it is proportional to $d\,\sigma_2\,\varphi(1)$. The gambling incentive is strongest about one second-half standard deviation below the downgrade line.

Proof. The sign is immediate from $\ell + E_2 < 0$. For the maximum, let $x = |\ell + E_2|$: $\frac{d}{dx}[x\,\varphi(x/\sigma_2)] = \varphi(x/\sigma_2)\big(1 - x^2/\sigma_2^2\big)$, which vanishes at $x = \sigma_2$, and the second-order condition is easily checked. The full G contains also the upgrade term, but at $E_2 \in (-\ell - 2\bar{\sigma},\, -\ell)$ it is bounded by $u(h + \ell + 2\bar{\sigma})\,\varphi\big((h+\ell)/\bar{\sigma}\big)$, of order $e^{-(h+\ell)^2/2\bar{\sigma}^2}$ (at the calibration the upgrade line sits about seven standard deviations away and the bound is of order 10^{-12}), so the maximizer located here is the maximizer of G up to that exponentially negligible tail term. □

Lemma A.10. *(the incentive dead zone). As $E_2 \to -\infty$, $G(E_2, \sigma_2)/\sigma_2^2 \to 0^+$ and $\Lambda\sigma_2 E[Q''] \to 0^+$, while the risk cost $-\rho\sigma_2 < 0$ does not decay. Hence for every $\rho > 0$ there is an $\underline{E}$ such that $E_2 < \underline{E}$ implies $\partial U_2/\partial\sigma_2 < 0$ on all of $[\underline{\sigma}, \bar{\sigma}]$: $\sigma_2^* = \underline{\sigma}$.*

Proof. $|G| \le u|h - E_2|\,\varphi((E_2 - h)/\sigma_2) + d|\ell + E_2|\,\varphi((\ell + E_2)/\sigma_2)$; as $E_2 \to -\infty$ both terms are (linear factor) $\times$ (exponentially decaying density) and tend to zero. The flow term tends to zero by Lemma A.7's deep-tail approximation. The risk cost is independent of E_2. □

The lemma holds for every fixed $\rho > 0$; the calibrated boundary is $e_1 \lesssim -6.75\%$ (Appendix D). The naive option intuition (deep out of the money, raise volatility) fails because both incentive sources decay exponentially while the cost does not: the option's strike recedes faster than volatility can chase it.

Lemma A.11. *(the reward peak). For $E_2 \in (h - 2\bar{\sigma},\, h)$ the upgrade term is positive and dominant: σ_2^* attains a local maximum in that region. The peak needs no floor and exists under a bonus-only contract.*

Proof. For $E_2 < h$ the upgrade term $u(h - E_2)\varphi((h - E_2)/\sigma_2) > 0$ and, by the computation of Lemma A.9, is maximized at $h - E_2 = \sigma_2$; the downgrade term is exponentially small because E_2 is far from $-\ell$. □

Proposition 3. *(the two-peak risk profile). In the two-period problem with $d > 0$ and moderate risk aversion ($0 < \rho < \bar{\rho}$; calibrated $\bar{\rho} \approx 14$–14.5):*

1. *$\sigma_2^*(e_1)$ attains its **global** maximum in the catch-up gambling zone, near $e_1 \approx -\ell - \bar{\alpha}_2 - \sigma_2^*$;*
2. *$\sigma_2^*(e_1)$ attains a **local** maximum at the reward peak just below the upgrade line, near $e_1 \approx h - \bar{\alpha}_2 - \sigma_2^*$; this peak is floor-independent and exists under a bonus-only contract;*
3. *three regions are locked at $\sigma_2^* = \underline{\sigma}$: above the upgrade line (the strongest), the mid band, and the incentive dead zone ($e_1 \lesssim -6.75\%$ in the calibration);*
4. *under penalty-heavy parameters the catch-up peak exceeds the reward peak: the fee jump parameters governing the two peaks stand in ratio $d/u = 2$, and the flow term's concavity near the upgrade line depresses the reward peak further;*
5. *within each contract's catch-up gambling zone, the ordering is $\sigma_2^*[\textit{penalty-heavy}] > \sigma_2^*[\textit{symmetric step}] > \sigma_2^*[\textit{bonus-only}] = \sigma_2^*[\textit{fixed}] = \sigma_2^*[\textit{linear fulcrum}] = \underline{\sigma}$;*
6. *both peaks survive up to $\bar{\rho} \approx 14$ and are flattened as $\rho \to \infty$.*

(The survival bound is computed on a refined risk-aversion grid extending to $\rho = 20$: the two-peak structure's upper bound is $\bar{\rho} \approx 14$–14.5 (the reward peak dies first, at $\rho = 14.5$) and the catch-up zone survives alone to at least $\rho = 20$, the grid's upper edge. Appendix D.)

Proof. Parts 1–3 assemble the regional lemmas: the global maximum sits in the catch-up gambling zone by Lemma A.9 (its location $E_2 \approx -\ell - \sigma_2^*$ reads, in the interim state, as $e_1 \approx -\ell - \bar{\alpha}_2 - \sigma_2^*$), the local maximum below the upgrade line is Lemma A.11, and the three locked regions are Lemma A.8(a), A.8(b), and A.10. Part 4: the marginal fee incentive driving each peak is the step size times the common factor $\sigma_2\varphi(1)/\sigma_2^2$ (Lemma A.9's maximum), so the two peaks are driven by fee jump parameters in ratio $d/u = 2$; evaluated at the respective optima ($\sigma_2^* = 1.70\%$ at the catch-up peak and 1.15% at the reward peak) the realized marginals stand in ratio $(d/u)\,(1.15/1.70) \approx 1.35$; on the upgrade side the flow term sits near its concave region ($ke > \ln c$ once the distribution approaches the line) and subtracts from the reward peak (Lemma A.7), widening the gap. Part 5: the catch-up incentive comes only from the penalty step (Lemma A.9); the bonus-only and fixed contracts have no step, and the linear fulcrum has expected fee $f_M + \beta E_2$, independent of σ_2 (since $E[e] = E_2$), so its manager minimizes risk in every state, $\sigma_2^* = \underline{\sigma}$; the symmetric step's penalty jump is half as large and its threshold further away. Part 6: a peak survives iff $\max_{\sigma_2} \left[G/\sigma_2^2 + \Lambda\sigma_2 E[Q'']\right] > \rho\sigma_2$ has a solution; the right-hand side grows linearly in ρ while the left does not depend on it, so all peaks are flattened as $\rho \to \infty$, and the survival bound $\bar{\rho}$ is a numerical question; this step is verified numerically over the calibrated region (Appendix D), with the coordinates of the note above. □

Calibrated profile readings (Appendix D; threshold-aligned grid cells, $\rho = 5$): at $e_1 = -6\%$ the penalty-heavy manager runs second-half risk of 1.70%, more than three times the 0.50% floor every other contract chooses at the same state; at $e_1 = +3\%$ the reward peak reaches 1.15%, shared exactly by the bonus-only contract; at $e_1 = -7\%$ the penalty-heavy profile has fallen back to the floor (the dead zone, whose calibrated boundary is $e_1 \lesssim -6.75\%$) while the symmetric-step contract, whose downgrade line sits at -4.5%, still gambles at 1.20%: the peak tracks the threshold, one standard deviation below it, contract by contract; and a manager sitting exactly on the downgrade line ($e_1 = -4.5\%$) locks, confirming that the peak lies one standard deviation *below* the line. Unrestricted argmax readings, -6.25% and $+3.25\%$ with heights 1.70% and 1.20%, differ from the threshold-aligned readings only by the grid's alignment to the thresholds; the text quotes the threshold-aligned convention throughout. At $\rho = 0$ the zone hits the risk bound, a corner pathology of vanishing risk cost, not a calibrated case; at the other end the reward peak dies first, at $\rho = 14.5$, and the catch-up zone survives alone past $\rho = 20$ (Appendix D).

A.3.1 The effort extension: "strive or gamble"

The baseline fixes $\bar{\alpha}_2$ and lets the manager move only risk. The extension lets him choose effort (moving the drift) and risk jointly: at mid-year he picks (a, σ_2) with $a \in [\underline{a}, \bar{a}]$, $E_2 = e_1 + a$, maximizing

$$U_2(a, \sigma_2;\, e_1) = W(E_2, \sigma_2) - \frac{\eta}{2}(a - a_0)^2, \qquad W(E_2, \sigma_2) \equiv E[f(e)] + \Lambda E[Q(e)] - \frac{\rho}{2}\sigma_2^2,$$

where a_0 is the natural drift (1.5% calibrated), η the effort-cost curvature, and $\bar{a}$ the alpha-supply bound. The baseline is $\eta \to \infty$. The extension is positioned as a **bound**, not a robustness result of the usual kind: realistic η is large (alpha supply within half a year is nearly inelastic), which is exactly the baseline's "pure gambling" case; small η moves the zone without removing it.

The outer problem collapses to one dimension, $\max_a\ V(e_1 + a) - \frac{\eta}{2}(a - a_0)^2$ with $V(E_2) \equiv \max_{\sigma_2} W(E_2, \sigma_2)$, and two facts about the reduction do the work. First, e_1 and a enter W only through the sum $E_2 = e_1 + a$, so for any given E_2 the conditional risk problem is pointwise the baseline's and the regional sign analysis of Lemmas A.8–A.11 holds verbatim in E_2-space: effort translates E_2 and does not deform the G function. Second, the envelope theorem gives

$$V'(E_2) \;=\; \frac{\partial E[f]}{\partial E_2} + \Lambda\,\frac{\partial E[Q]}{\partial E_2} \;=\; \frac{u}{\sigma_2}\,\varphi\Big(\frac{h - E_2}{\sigma_2}\Big) + \frac{d}{\sigma_2}\,\varphi\Big(\frac{\ell + E_2}{\sigma_2}\Big) + \Lambda\,\frac{\partial E[Q]}{\partial E_2} \;>\; 0:$$

the first term is the upgrade probability gained, the second the downgrade probability avoided, by a higher mean, and Q is strictly increasing; hence $a^* \geq a_0$ everywhere (the manager never shirks), and since the densities peak near the thresholds, the effort response is strongest in the catch-up zone.

Proposition A.12. *(strive first, then gamble: the catch-up zone survives in the translation sense). Let $\bar{a} < \infty$.*

(a) (Existence is η-free.) For every η, including the baseline $\eta \to \infty$, the set of states in which the catch-up peak survives is non-empty: every state with $e_1 + \bar{a} < -\ell$ ("cannot strive back") keeps $E_2^* < -\ell$, so the inner problem sits in the catch-up zone and $\sigma_2^* > \underline{\sigma}$.

(b) (The peak translates with η; its height does not.) Smaller η pushes E_2^* higher at given e_1, so the state at which the zone occurs, $e_1^*(\eta) \approx -\ell - \sigma_2^* - a^*(\eta)$, shifts left as effort

cheapens; the peak height is determined by the local structure of G and is invariant to η.

(c) (The strive-instead-of-gamble band.) If η is small and e_1 lies in the "can strive back" region ($e_1 + \bar{a} \geq -\ell$, with effort cheap enough), the manager pushes E_2^* across $-\ell$ into the locked region: $a^* > a_0$ and $\sigma_2^* = \underline{\sigma}$: effort substitutes for gambling.

(d) (The dead zone does not die.) As $e_1 \to -\infty$, $V'(E_2) \to 0$ (both densities and Q' decay exponentially), so $a^* \to a_0$ and $\sigma_2^* \to \underline{\sigma}$: the deep-tail manager neither strives nor gambles.

Proof. (a) is immediate from $\bar{a}$ finite and Lemma A.9. (b): the outer FOC is $V'(e_1+a^*) = \eta(a^*-a_0)$; smaller η supports a larger a^* at the same marginal value, raising E_2^*; the inner peak sits at $E_2 \approx -\ell - \sigma_2^*$ (Lemma A.9), so $e_1^* \approx -\ell - \sigma_2^* - a^*(\eta)$, decreasing in η. (c): when $E_2^* > -\ell$, Lemma A.8(b) applies. (d): the decay rates of Lemma A.10 applied to V''s display. $\square$

This step is verified numerically over the calibrated region (Appendix D): on the joint grid (a, σ_2) the catch-up peak survives at every $\eta \in \{5, 20, 100, \infty\}$, with height 1.70% at $\eta \geq 20$ and 1.55% at $\eta = 5$ (where the surviving gambling state sits at the alpha-supply bound); the peak position shifts from $e_1^* = -6.5\%$ toward -7.9% as η falls to 5, the strive-instead band appears exactly where predicted ($\eta = 5$, $e_1 = -6\%$: $a^* = 3.9\%$, $\sigma_2^* = \underline{\sigma}$, against $a^* = 1.9\%$ under the fixed contract), and the deep tail is flat at the floor at every η.

A.4 Propositions 4–6: the vintage structure

The setting is Section 4's: years are discrete, $e_t = \alpha + \sigma z_t$ iid, a lot held n years is charged $f(W_n)$ on its own window mean $W_n = \frac{1}{n}\sum_{s=1}^{n} e_s \sim N(\alpha, \tau_n^2)$ with $\tau_n = \sigma/\sqrt{n}$, and within-year money pays f_M. By Lemma 1, $E[f(W_n)] = f_M + V(\alpha, \tau_n)$ with standardized threshold distances $t_U(n) = (h-\alpha)/\tau_n$, $t_D(n) = (-\ell-\alpha)/\tau_n$. We first record the filter (the Remark of Section 4's setup), then prove Propositions 4, 5, 6 in order, with the aggregation lemma between.

Lemma A.13. *(the filter: the contract bites only through seasoned money). (i) Within-year money contributes $q_0 f_M$ irrespective of (α, σ). (ii) Substituting $E[f(W_n)] = f_M + V(\alpha, \tau_n)$ into the definitions,*

$$\bar{g} = f_M + \sum_{n \geq 1} w_n \, V(\alpha, \tau_n), \qquad w_n \equiv \frac{q_n \, n}{q_0 + E[N]},$$

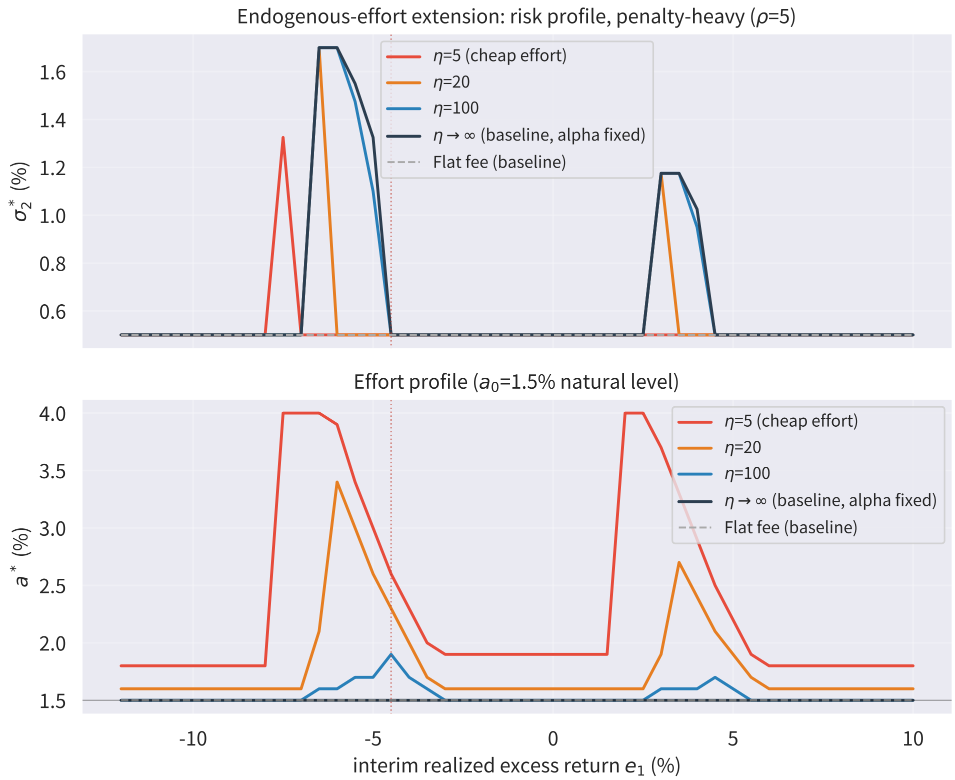

Figure 12: The catch-up gambling zone under endogenous effort: peak height essentially unchanged, position shifting left as effort cheapens; the deep tail flat throughout.

where the yuan-year weights sum to at most one and the remainder $w_0 = q_0/(q_0 + E[N])$ *is attached to* $V \equiv 0$*. (iii) The revenue sensitivity to skill,*

$$\frac{\partial LTV_1}{\partial \alpha} = \sum_{n \geq 1} q_n \, n \, \frac{u \, \varphi(t_U(n)) + d \, \varphi(t_D(n))}{\tau_n},$$

is strictly positive if and only if seasoned money has positive measure, and is strictly decreasing in q_0*. A flat-fee contract is exactly this contract at* $q_0 = 1$*.*

Proof. (i) is the institutional rule: money redeemed within the year pays f_M regardless of performance. (ii): the numerator of $\bar{g}$ is $q_0 f_M + \sum_{n\geq 1} q_n n[f_M + V(\alpha, \tau_n)] = (q_0 + E[N])f_M + \sum_{n\geq 1} q_n n \, V(\alpha, \tau_n)$; dividing by $q_0 + E[N]$ gives the display. (iii): differentiate $LTV_1 = q_0 f_M + \sum_{n\geq 1} q_n n \, E[f(W_n)]$ term by term and apply Lemma 1(i) at scale τ_n; every seasoned term is strictly positive, the q_0 term contributes zero, and raising q_0 (rescaling the redemption distribution toward within-year money) removes weight from the positive terms. □

Proposition 4. *(window averaging: the fee is an asymptotically deterministic stair in true skill). Let $\alpha \notin \{-\ell, h\}$ and $\delta(\alpha) \equiv \min(|\alpha - h|,\, |\alpha + \ell|)$ be the distance from skill to the nearest threshold.*

(a) (Limit and rate.) The applicable fee rate converges to the stair in true skill,

$$\lim_{n\to\infty} E[f(W_n)] = f(\alpha),$$

with a uniform exponential bound,

$$\bigl|E[f(W_n)] - f(\alpha)\bigr| \;\le\; \frac{u+d}{2}\,\exp\Bigl(-\frac{n\,\delta(\alpha)^2}{2\sigma^2}\Bigr).$$

(b) (Luck contamination.) The probability that a lot of age n is priced into a wrong tier satisfies the same exponential bound; the set of skill levels whose mispricing probability exceeds ε (the confusion band around each threshold) has total width at most $2\sigma\sqrt{2\ln(1/\varepsilon)}/\sqrt{n} = O(1/\sqrt{n})$.

(c) (The age profile.) As a function of n, $E[f(W_n)]$ is: strictly increasing toward f_H from below if $\alpha > h$; strictly decreasing toward f_L from above if $\alpha < -\ell$; and converging to f_M if $\alpha \in (-\ell, h)$, ultimately from below if $\alpha < \alpha_{be}$ and from above if $\alpha > \alpha_{be}$, where $\alpha_{be} \equiv (h-\ell)/2$ is the asymptotic break-even skill.

The proofs use the standard Gaussian tail bound $\bar\Phi(x) \le \frac{1}{2}e^{-x^2/2}$ for $x \ge 0$, which we record with its proof: define $H(x) = \frac{1}{2}e^{-x^2/2} - \bar\Phi(x)$; then $H(0) = 0$, $H(x) \to 0$ as $x \to \infty$, and $H'(x) = e^{-x^2/2}\bigl(\frac{1}{\sqrt{2\pi}} - \frac{x}{2}\bigr)$ changes sign once, from positive to negative, so $H \ge 0$ throughout.

Proof of (a). $E[f(W_n)] = f_M + V(\alpha, \tau_n)$, and both standardized distances satisfy $|t_U(n)|, |t_D(n)| \ge \delta(\alpha)\sqrt{n}/\sigma$, since $\tau_n = \sigma/\sqrt{n}$. Take the three cases in turn. If $\alpha > h$: $E[f] - f_H = -u\,\Phi((h-\alpha)/\tau_n) - d\,\Phi((-\ell-\alpha)/\tau_n)$, both arguments negative, so each Φ is a lower tail bounded by $\frac{1}{2}\exp(-n\,\delta^2/2\sigma^2)$; the limit is $f_M + u = f_H$. If $\alpha < -\ell$: $E[f] - f_L = u\,\bar\Phi(t_U(n)) + d\,\bar\Phi((\ell+\alpha)/\tau_n)$, both upper tails with the same bound; the limit is $f_M - d = f_L$. If $\alpha \in (-\ell, h)$: $E[f] - f_M = u\,\bar\Phi(t_U(n)) - d\,\Phi(t_D(n))$, an upper tail and a lower tail, each with the same bound; the limit is f_M. In each case $|E[f(W_n)] - f(\alpha)| \le \frac{u+d}{2}\exp(-n\delta(\alpha)^2/2\sigma^2)$. □

Proof of (b). For $\alpha \in (-\ell, h)$ the mispricing probability is $\bar\Phi(t_U(n)) + \Phi(t_D(n)) \le \exp(-n\delta^2/2\sigma^2)$ (the two half-bounds add); the other two cases are analogous with a single tail. Mispricing probability above ε requires $\delta(\alpha) < \sigma\sqrt{2\ln(1/\varepsilon)/n}$, i.e., α within that distance of a threshold: total width at most $2\sigma\sqrt{2\ln(1/\varepsilon)}/\sqrt{n}$. Note the division of labor between the two rates: the mispricing *probability* dies exponentially in n, while the *skill resolution* (the confusion band's width) improves only at $1/\sqrt{n}$. $\square$

Proof of (c). Write $\tau = \sigma n^{-1/2}$, so $d\tau/dn = -\tau/(2n) < 0$ and $\operatorname{sign}(\partial E[f]/\partial n) = -\operatorname{sign}(\partial E[f]/\partial \tau)$. By Lemma 1's variance derivative at general (α, τ), using that φ is even,

$$\frac{\partial E[f]}{\partial \tau} = \frac{1}{\tau^2}\Big[u\,(h-\alpha)\,\varphi\Big(\frac{h-\alpha}{\tau}\Big) - d\,(\ell+\alpha)\,\varphi\Big(\frac{\ell+\alpha}{\tau}\Big)\Big].$$

If $\alpha > h$: $h - \alpha < 0$ and $\ell + \alpha > 0$, so the bracket is negative for every τ, hence $\partial E[f]/\partial n > 0$; and $E[f] - f_H < 0$ from part (a)'s display (convergence from below). If $\alpha < -\ell$: $h - \alpha > 0$ and $\ell + \alpha < 0$, the bracket is positive, $\partial E[f]/\partial n < 0$, and $E[f] - f_L > 0$ (from above). If $\alpha \in (-\ell, h)$: as $\tau \to 0$ the sign is decided by the race of the two exponentials,

$$\frac{\varphi\big((\ell+\alpha)/\tau\big)}{\varphi\big((h-\alpha)/\tau\big)} = \exp\Big[\frac{(h-\alpha)^2 - (\ell+\alpha)^2}{2\tau^2}\Big] = \exp\Big[\frac{(h-\ell-2\alpha)(h+\ell)}{2\tau^2}\Big],$$

which tends to $+\infty$ when $\alpha < (h-\ell)/2$ (the downgrade side is nearer and dominates: the bracket is ultimately negative, $\partial E[f]/\partial n > 0$, convergence from below) and to 0 when $\alpha > (h-\ell)/2$ (convergence from above). $\square$

Calibrated profile readings (Appendix D; $\sigma = 4\%$): the asymptotic break-even skill is 1.5%; $\alpha = -4\%$: about 84bp at age one, 77bp at age five, converging down to the 60bp floor; $\alpha = -2\%$ (mid band, below break-even): 97bp at age one, rising back to 120bp as the noise tax is refunded; $\alpha = +2\%$ (above break-even): 118bp at age one (still below f_M, because young windows carry mass in both tails) crossing 120bp from age three and converging back from above; $\alpha = 6.5\%$: 136bp at age one, climbing monotonically toward 150bp. The mid-band noise tax at $\alpha = 0$ is -11.6bp at $n = 1$ and -0.081bp at $n = 16$; the noisy-window limit of the lottery's expectation is $(u-d)/2 = -15$bp.

Corollary. *(asymptotic fixed-fee-ization of the mid band). Fix (α, σ) and let the clientele age in the sense of first-order stochastic dominance: under geometric survival, $r \to 0$ and $q_0 \to 0$, so the yuan-year weight $w_n = r^2 n(1-r)^{n-1}$ concentrates at $n \approx 2/r \to \infty$. Then $\bar{g}(\alpha; r) \to f(\alpha)$ for every $\alpha \notin \{-\ell, h\}$.*

Proof. Under geometric survival the yuan-year weight is $w_n = r^2 n(1-r)^{n-1}$, a size-biased geometric distribution with mean of order $2/r$, which concentrates on large n as $r \to 0$. Proposition 4(a)'s bound is uniform on $\{\alpha : |\alpha - \text{threshold}| \geq \delta_0\}$ for any $\delta_0 > 0$: split the weighted sum $\bar{g} - f_M = \sum_n w_n V(\alpha, \tau_n)$ at n_0 large enough that the bound is below a given tolerance, note $|V| \leq (u+d)/2$ on the remaining initial segment, and let $r \to 0$ so that segment's weight vanishes. $\square$

Remark. *(the closet safe harbor; proof).* Closet indexing is the unique strategy whose long-run fee is deterministically the base tier. At $\alpha = 0$: (i) every finite window prices strictly below f_M ($V(0,\tau) < 0$ for all $\tau > 0$ by Lemma 1's penalty-dominance corollary) with the gap dying exponentially by Proposition 4(a) at $\delta(0) = \ell$; and (ii) $\bar{g} \to f_M$ with probability-one certainty, since $W_n \to 0 \in (-\ell, h)$ almost surely by the strong law. Every $\alpha \neq 0$ strategy exposes its young money to the fee lottery, since $V(\alpha, \tau_n) \neq 0$ at finite n. $\square$

Lemma A.14. *(aggregation preserves the sign facts). $\bar{g}(\alpha) = f_M + \sum_n w_n V(\alpha, \tau_n)$ is a positive weighted average of V across window scales. The sign properties used by Propositions 1–3 survive any positive weighting: (i) $V(0, \tau_n) < 0$ for all n (penalty dominance) implies $\bar{g}(0) < f_M$; (ii) $\partial V(\alpha, \tau_n)/\partial\alpha > 0$ for all n (Lemma 1(i)) implies $\bar{g}'(\alpha) > 0$; (iii) the strict age-profile monotonicities of Proposition 4(c) ($V(\alpha, \tau_n)$ strictly decreasing in n for $\alpha < -\ell$, strictly increasing for $\alpha > h$) are inherited by the weighted profile.*

Proof. A positive weighted average (weights summing to at most one, with at least one positive weight on a strict term) preserves strict inequalities and strict monotonicity. $\square$

What aggregation does *not* preserve is the manager's marginal incentive, which is a different object from the pricing map; that is Proposition 6.

Proposition 5. *(the bust fine and capacity internalization). (a) (The bust fine.) Let $\alpha < -\ell$. Then (i) every term of $BF(\alpha)$ is strictly positive, the fine exists from the first year and is not an asymptotic mirage; (ii) the annual fine $f_M - E[f(W_n)]$ is strictly increasing in window age and converges to d; and (iii) $BF(\alpha) \geq \left[d\,\Phi(t_D(1)) - u\right] \cdot E[A_N] > 0$.*

(b) (Capacity internalization.) Let per-unit alpha decline in scale, $\alpha(A) = \alpha_0 - \kappa A$ with $\kappa > 0$. Under the flat fee, the fee-revenue value of a marginal yuan of AUM is f_M at every skill and scale. Under the floating contract it is

$$\frac{d}{dA}\left[\bar{g}(\alpha(A)) \cdot A\right] = \bar{g}(\alpha) - \kappa A\, \bar{g}'(\alpha), \qquad \bar{g}'(\alpha) = \sum_n w_n \frac{u\,\varphi(t_U(n)) + d\,\varphi(t_D(n))}{\tau_n} > 0,$$

so the marginal value is strictly below the flat-fee value for every fund with $\bar{g}(\alpha) \leq f_M$, all but the high-skill tail, and there exists a parameter region in which it is negative: gathering *one more yuan costs more, through dilution of the existing base's fee rate, than the yuan's own fee brings in.*

Proof of (a). For $\alpha < -\ell$, $t_D(n) = (-\ell - \alpha)\sqrt{n}/\sigma$ is positive and increasing in n, so $\Phi(t_D(n)) \geq \Phi(t_D(1)) > 1/2$. The annual fine is

$$f_M - E[f(W_n)] = d\,\Phi(t_D(n)) - u\,\bar{\Phi}(t_U(n)) \;\geq\; d\,\Phi(t_D(1)) - u \;>\; \frac{d}{2} - u = 0,$$

using $\bar{\Phi} \leq 1$, $\Phi(t_D(1)) > 1/2$, and the institutional $d/2 = u$. Every term of $BF(\alpha) = \sum_{n\geq 1} q_n A_n(\beta)\,[f_M - E[f(W_n)]]$ is therefore strictly positive, which is (i), and summing the uniform lower bound against $\sum_n q_n A_n(\beta) = E[A_N]$ gives (iii). (ii) is Proposition 4(c) in the case $\alpha < -\ell$: $E[f(W_n)]$ is strictly decreasing toward f_L, so the fine $f_M - E[f(W_n)]$ is strictly increasing toward d. $\square$

Proof of (b). Under the flat fee, revenue is $f_M A$ and the marginal value is f_M. Under the floating contract, the chain rule with $\alpha'(A) = -\kappa$ gives

$$\frac{d}{dA}\big[\bar{g}(\alpha(A)) \cdot A\big] = \bar{g}(\alpha) + A\,\bar{g}'(\alpha)\,\alpha'(A) = \bar{g}(\alpha) - \kappa A\,\bar{g}'(\alpha),$$

and $\bar{g}' > 0$ is Lemma A.14(ii) under positive weights. For every fund with $\bar{g}(\alpha) \leq f_M$ the marginal value is then strictly below $\bar{g}(\alpha) \leq f_M$; and whenever $\kappa A\,\bar{g}'(\alpha) > \bar{g}(\alpha)$ (a non-empty parameter region, since $\bar{g}'$ is amplified near thresholds by the $1/\tau_n = \sqrt{n}/\sigma$ factor on seasoned money) the marginal value is negative. $\square$

Calibration and its boundary (Appendix D). At $\alpha = -4\%$, $\sigma = 4\%$, $\beta = 0.97$: the fine is about 266bp in present value per yuan of sticky money at annual redemption hazard $r = 0.15$ ($E[N] \approx 6.7$ years), 39% of the same money's lifetime fee revenue in the flat-fee world ($f_M \cdot E[A_N] \approx 684$bp), rising to about 523bp at $r = 0.07$ ($E[N] \approx 14$ years), near-linearly in clientele duration. The **share** is robust and is the headline; the **level** is not: under performance-sensitive redemption with sensitivity $\theta = 2$ (applied identically in both worlds), the level falls from 266bp to 89bp (-67%) while the share falls only to 32%–34% across hazards, because the denominator shrinks under the same redemption rule. For part (b): $\bar{g}'$ peaks at the downgrade threshold $\alpha = -\ell$ at about 1,980bp per unit of alpha ($r = 0.15$); at normal scale the marginal value of AUM falls from 120bp to 91–104bp at the low-skill end (a 13%–25% discount, inflow revenue discounted, not destroyed); negative values require $\kappa A \gtrsim 4\%$ (at $\kappa = 0.5\%$ per unit of base, scale of roughly 8.5 times base), and at $A = 10\times$, $\kappa = 0.5\%$, $\alpha_0 \in [1.5\%, 2\%]$ the marginal yuan is worth -7 to -9bp.

Proposition 6. *(the intertemporal structure of incentives). The cohort's expected fee rate responds to the current year's expected performance with sensitivity*

$$\mathrm{Sens}(n; R_{n-1}) \;\equiv\; \frac{\partial E[f(W_n)]}{\partial \alpha} \;=\; \frac{1}{\sigma}\Big[u\,\varphi\big(\tilde t_U(n)\big) + d\,\varphi\big(\tilde t_D(n)\big)\Big],$$

where the standardized threshold distances

$$\tilde t_U(n) = \frac{nh - (n-1)R_{n-1} - \alpha}{\sigma}, \qquad \tilde t_D(n) = \frac{-n\ell - (n-1)R_{n-1} - \alpha}{\sigma}$$

are linear in n, with slopes $(h - R_{n-1})/\sigma$ and $(-\ell - R_{n-1})/\sigma$ respectively. Hence:

(a) (Non-threshold-riding money dies.) If $R_{n-1} \notin \{-\ell, h\}$, at least one $|\tilde t|$ diverges linearly while the nearest threshold remains at positive distance, so $\mathrm{Sens}(n; R_{n-1}) = O(e^{-cn^2})$ for some $c > 0$: the marginal fee incentive of a locked-in lot decays at super-Gaussian speed, faster than any polynomial, not $1/n$ and not $1/\sqrt{n}$. Money already settled into the mid band cannot have its tier changed by any feasible effort; money deep in the penalty zone cannot climb back.

(b) (Threshold-riding money never dies.) If $R_{n-1} = h$ or $R_{n-1} = -\ell$, the corresponding $\tilde t$ is constant in n, so that side's sensitivity is $O(1)$ forever. The sprint-or-defend incentive at a threshold is permanent.

(c) (The catch-up gambling zone is a young-window phenomenon.) Let $R_{n-1} = -\ell - \varepsilon$ with small $\varepsilon > 0$, Section 3.8's catch-up gambling zone, read across window ages. Climbing back above the downgrade line requires $e_n \geq -\ell + (n-1)\varepsilon$, linearly harder each year. Equivalently, $\tilde t_D(n)$ rises from below zero at slope ε/σ and crosses zero at

$$n^* = \frac{R_{n-1} - \alpha}{R_{n-1} + \ell},$$

so the catch-up incentive first rises to its peak at the "last salvageable year" n^* and then dies super-Gaussianly. The risk-sensitivity inherits the same decay: the sign-flip structure that powers the catch-up zone lives only in young cohorts whose $\tilde t_D$ has not yet diverged.

Proof. The conditional window is $W_n = \frac{(n-1)R_{n-1} + e_n}{n} \sim N(m, s^2)$ with $m = \big((n-1)R_{n-1} + \alpha\big)/n$ and $s = \sigma/n$. Apply Lemma 1(i) at (m, s): $\partial E[f]/\partial m = \frac{1}{s}\big[u\,\varphi\big(\frac{h-m}{s}\big) + d\,\varphi\big(\frac{-\ell-m}{s}\big)\big]$, and the chain rule gives $\partial m/\partial\alpha = 1/n$, so

$$\text{Sens} = \frac{1}{ns}\Big[u\,\varphi\Big(\frac{h-m}{s}\Big) + d\,\varphi\Big(\frac{-\ell-m}{s}\Big)\Big] = \frac{1}{\sigma}\big[u\,\varphi(\tilde{t}_U) + d\,\varphi(\tilde{t}_D)\big],$$

since $\frac{h-m}{s} = \frac{nh-(n-1)R_{n-1}-\alpha}{\sigma} = \tilde{t}_U(n)$ and likewise for $\tilde{t}_D$. Both $\tilde{t}$ are affine in n with the stated slopes.

(a) If $R_{n-1} \in (-\ell, h)$: $\tilde{t}_U \to +\infty$ and $\tilde{t}_D \to -\infty$, both linearly. If $R_{n-1} > h$: $\tilde{t}_U \to -\infty$ linearly (slope $(h - R_{n-1})/\sigma < 0$) and $\tilde{t}_D \to -\infty$. If $R_{n-1} < -\ell$: both diverge to $+\infty$. In every case each $|\tilde{t}|$ diverges linearly, so each density term is $O(e^{-c_j n^2})$ with c_j half the squared slope, and the sum is $O(e^{-cn^2})$ for the smaller c_j.

(b) If $R_{n-1} = h$: $\tilde{t}_U(n) = (h-\alpha)/\sigma$, constant. If $R_{n-1} = -\ell$: $\tilde{t}_D(n) = (-\ell-\alpha)/\sigma$, constant. That side's density term never decays.

(c) With $R_{n-1} = -\ell-\varepsilon$: $W_n \ge -\ell \iff e_n \ge -n\ell - (n-1)R_{n-1} = -\ell + (n-1)\varepsilon$. And $\tilde{t}_D(n) = \big((n-1)\varepsilon - \ell - \alpha\big)/\sigma$, rising from below zero at slope ε/σ and crossing zero at $n^* = 1 + (\ell+\alpha)/\varepsilon = (R_{n-1}-\alpha)/(R_{n-1}+\ell)$. The downgrade-side density $\varphi(\tilde{t}_D(n))$ therefore rises to its maximum at n^* (the last year the window is exactly salvageable) and dies super-Gaussianly beyond it. The risk-sensitivity inherits the same decay: by Lemma 1's variance-derivative form evaluated at (m, s), the same $\tilde{t}$ enter, and the sign-flip structure that powers the catch-up zone (Lemma A.9's term changing sign as the distribution crosses the line) exists only while $\tilde{t}_D$ has not diverged. □

Calibrated death schedule (Appendix D; $\alpha = 0$, $\sigma = 4\%$): a cohort at $R = -4\%$ (one point below the downgrade line, squarely in a young window's catch-up zone) sees its sensitivity rise to a peak of 598bp at $n^* = 4$, about 1.1 times its year-one level, then fall below the year-one level from age seven and accelerate to zero; the required single-year catch-up return at $\varepsilon = 1\%$ is $e_n \ge -2\%$ at $n = 2$, $+1\%$ at $n = 5$, $+6\%$ at $n = 10$; a cohort hugging the line at $R = -2.9\%$ retains 327bp at age fifteen; a cohort exactly on the line holds a constant 452bp forever.

A.5 Proposition 7: sorting under the quota

We first prove the sign structure of Lemma 3's three-channel decomposition, on which the proposition's premise rests, then the proposition.

Lemma 3. *(the three-channel decomposition). Let $R(\mu, c) = E[\bar{N}]\,\bar{g}_c\left(A_c^S + A_c^T\right) - \Psi(m_c^*)$ be the company's steady-state revenue from a manager of skill μ under contract c, net of marketing cost, and define $\Delta(\mu) \equiv R(\mu, \text{float}) - R(\mu, \text{flat})$. Then $\Delta(\mu)$ decomposes into* *three additive channels:*

$$\Delta(\mu) = \underbrace{E[\bar{N}]\left[\bar{g}_L - f_M\right] A_L}_{\textit{(i) fee level}} + \underbrace{E[\bar{N}]\, f_M \left[A_L^S - A_F^S\right]}_{\textit{(ii) smart money}} + \underbrace{E[\bar{N}]\, f_M \left[A_L^T - A_F^T\right] - \left[\Psi_L - \Psi_F\right]}_{\textit{(iii) trusting money, net of marketing}},$$

where subscripts L and F denote the floating and flat contracts and $\Psi_c \equiv \Psi(m_c^*)$. In the low-skill segment ($\mu \le \mu^*_{\mathit{flat}}$) channels (i) and (iii) are both strictly negative and $A_L^S = A_F^S = 0$: $\Delta(\mu) < 0$. Above the contracts' activity crossing $\tilde{\mu}$, channel (ii) is strictly positive and channel (i) is near-neutral through the mid band: $\Delta(\mu) > 0$ provided the quantity gain dominates.

Proof. With $\bar{g}_F = f_M$, expanding $A_L = A_L^S + A_L^T$ and collecting terms shows the display is an identity. For the signs. In the low-skill segment both contracts induce closet indexing (Proposition 1(a)), so $\alpha = 0$ and $\sigma = \sigma_0$ under either contract; smart money, which requires non-negative marginal net alpha, stays away from both, $A_L^S = A_F^S = 0$. Channel (i) reduces to $E[\bar{N}]\,[\bar{g}_L(0, \sigma_0) - f_M]\, A_L$, strictly negative by Lemma A.14(i), the noise tax. Channel (iii): the marketing first-order conditions are $\Psi'(m_L^*) = E[\bar{N}]\,(\bar{g}_L - \kappa A_L \bar{g}'_L)$ against $\Psi'(m_F^*) = E[\bar{N}]\, f_M$ (Proposition 5(b)'s FOC, with the flat contract's marginal value exactly f_M); since $\bar{g}_L < f_M$ and Ψ' is increasing with $\Psi'(0) = 0$, $m_L^* < m_F^*$, hence $A_L^T = I_0 + m_L^* < A_F^T$. The cost saving does not overturn the sign: the flat contract's marketing choice maximizes the flat-rate trusting-money program $E[\bar{N}]\, f_M\,(I_0 + m) - \Psi(m)$, so $E[\bar{N}]\, f_M\, A_F^T - \Psi(m_F^*) > E[\bar{N}]\, f_M\, A_L^T - \Psi(m_L^*)$ (strictly, because Ψ is strictly convex and the two maximizers differ), which is exactly channel (iii) < 0. Two negatives, same sign, no ambiguity: $\Delta(\mu) < 0$. Above $\tilde{\mu}$: Proposition 1(b) gives $x^*_{\text{float}} > x^*_{\text{flat}}$ (within the mean–variance benchmark), so $\alpha_L > \alpha_F$; the asymptotic fixed-fee-ization of the mid band (Proposition 4's Corollary) keeps $\bar{g}_L \approx f_M$ for performance inside $(-\ell, h)$, so channel (i) is near-neutral while channel (ii) is strictly positive: $A_L^S = \max(0, (\alpha_L - \bar{g}_L)/\kappa) > \max(0, (\alpha_F - f_M)/\kappa) = A_F^S$. The middle segment ($\mu$ between μ^*_{float} and $\tilde{\mu}$, where the floating contract compresses activity and the fee-level sign is ambiguous) is where the algebra runs out and the calibration takes over. $\square$

Whether $\Delta(\mu)$ crosses zero exactly once is a numerical question, and the premise of Proposition 7. This step is verified numerically over the calibrated region (Appendix D): scanned over the full skill support at $\kappa \in \{0.1\%, 0.2\%, 0.5\%\}$ per unit of normalized scale, with the dilution feedback between smart-money demand and the vintage fee rate solved to a fixed point (1,200 fixed points, all cleanly convergent, no multiple roots), $\Delta(\mu)$ crosses once, from below, at $\mu^\dagger \in \{5.18\%, 5.28\%, 5.58\%\}$; the voluntary floating share above the crossing is $1 - G(\mu^\dagger) \in \{14.3\%, 13.7\%, 12.0\%\}$, far below $q = 0.6$.

Proposition 7. *(sorting under the quota: floating appointments are endorsements). Suppose $\Delta(\mu)$ is single-crossing at $\mu^\dagger$ with $1-G(\mu^\dagger) < q$ (the quota binds). Then the company's optimal assignment is a top-down cutoff rule:*

$$c^*(\mu) = \textit{float} \iff \mu \geq \mu_q, \qquad \mu_q = G^{-1}(1-q).$$

Proof. The company's problem,

$$\max_{c(\cdot),\, m(\cdot)} \int \Big[E[\bar{N}]\, \bar{g}\big(\alpha(\mu, c_\mu)\big) \cdot \big(A^S(\mu, c_\mu) + A^T(\mu)\big) - \Psi\big(m(\mu)\big) \Big]\, dG(\mu)$$
$$\text{s.t.} \qquad \int \mathbf{1}\{c_\mu = \text{float}\}\, dG \geq q,$$

separates manager by manager up to the quota constraint: the marketing policy optimizes pointwise given the assignment, and the assignment enters only through the integral constraint. Let $\nu \geq 0$ be the multiplier on the quota. The Lagrangian's assignment component is $\int \mathbf{1}\{c_\mu = \text{float}\}\, [\Delta(\mu) + \nu]\, dG(\mu)$ plus assignment-free terms, so manager μ is floated exactly when $\Delta(\mu) + \nu \geq 0$: the floating slots go to the managers with the largest Δ: the greatest gains or the smallest losses. Since the quota binds, $\nu > 0$ and the constraint holds with equality, so the floated set has measure exactly q. Single crossing of Δ from below at $\mu^\dagger$ makes the superlevel set $\{\mu : \Delta(\mu) \geq -\nu\}$ an upper interval $[\mu_q, \bar{\mu}]$: any candidate set of measure q that excludes some μ_1 with $\Delta(\mu_1) > \Delta(\mu_2)$ in favor of μ_2 is improved by swapping. The binding quota then fixes the boundary at $G(\mu_q) = 1 - q$, i.e., $\mu_q = G^{-1}(1 - q)$. □

Calibrated reading (Appendix D): the cutoff at $q = 0.6$ is $\mu_q = 2.10\%$ per year per unit of deviation, between the flat and floating activity thresholds, $\mu^*_{\text{flat}} = 1.93\% < \mu_q < \mu^*_{\text{float}} = 2.15\%$.

The Corollary of the text is the cross-sectional implication of the ordering for pre-appointment records; the endorsement certification and the three-layer partition are immediate readings, developed in the text of Section 4.6.

A.6 The industry fixed point of Section 4.7

The demand closure assigns to each fund, given its contract c and the manager's choice $(a^0, \sigma) = (\mu x^*, \sigma_0 + s x^*)$, a total size A split as $A = A^S + A^T$. Trusting money is $A^T = I_0 + m$ with the marketing policy of Proposition 5(b)'s first-order condition. Smart money is an informed-capital sector that internalizes its own dilution: allocating to maximize its total net surplus $A^S\big(\alpha(A) - \bar{g}\big)$ with $\alpha(A) = a^0 - \kappa A$ and the vintage fee $\bar{g}$ of Lemma A.14 evaluated at the diluted alpha, the sector's size is

$$A^S = \max\left(0,\ \frac{\alpha(A) - \bar{g}\big(\alpha(A), \sigma\big)}{\kappa}\right),$$

and the fund's size, fee rate, and marketing policy are jointly determined at the fixed point $A = A^S(A) + A^T(A)$, where

$$A^T(A) = I_0 + m(A), \qquad \Psi'\big(m(A)\big) = E[\bar{N}]\left[\bar{g}\big(\alpha(A)\big) - \kappa A\, \bar{g}'\big(\alpha(A)\big)\right]_+.$$

Two structural features of this closure do economic work in the text. First, the informed sector's marginal unit nets $\kappa A^S > 0$ in equilibrium: informed capital is not perfectly elastic, and the rent it retains is what keeps the company's marketing margin alive, since trusting inflow is then crowded out only partially ($dA/dA^T = \frac{1}{2}$ at an interior floating fund) rather than one for one. Second, the per-unit net return $\alpha(A) - \bar{g} = \kappa A^S$ is earned by every yuan in an interior fund, smart or trusting; trusting money, which does not optimize at the margin, is the infra-marginal bearer of that return and of its reform-induced change.

Lemma A.15. *(the industry fixed point exists). Let $F(A) \equiv A^S(A) + A^T(A) - A$ with A^S, A^T as above, and let $\bar{A} \equiv a^0/\kappa + I_0 + \bar{m}$, where $\bar{m}$ is any upper bound on the marketing policy (e.g., $\Psi'(\bar{m}) = E[\bar{N}] f_H$). Then F is continuous on $[0, \bar{A}]$, with $F(0) \geq I_0 > 0$ and $F(\bar{A}) < 0$, so a fixed point exists in $(0, \bar{A})$. Moreover $A^S(A)$ is nonincreasing: the map $A \mapsto \alpha(A) - \bar{g}\big(\alpha(A)\big)$ is strictly decreasing, since its derivative is $-\kappa\big(1 - \bar{g}'\big)$ and*

$$\bar{g}'(a, \sigma) = \sum_{n \geq 1} w_n \frac{u\, \varphi(t_U(n)) + d\, \varphi(t_D(n))}{\tau_n} \ \leq\ \frac{(u + d)\, \varphi(0)}{\sigma} \sum_{n \geq 1} w_n \sqrt{n},$$

which evaluates to 0.28 at the calibration ($\sigma = 4\%$; attained maximum 0.20, Appendix D), uniformly in a.

Proof. $\bar{g}$ is a positive weighted average of the continuous $V(\alpha, \tau_n)$ (Lemma A.14), hence continuous in A through $\alpha(A)$; the max and the marketing first-order condition preserve continuity, so F is continuous. $F(0) = A^S(0) + A^T(0) \geq I_0 > 0$. At $\bar{A}$: $\alpha(\bar{A}) = -\kappa\,(I_0 + \bar{m}) < 0 < f_L \leq \bar{g}$, so $A^S(\bar{A}) = 0$, and $A^T(\bar{A}) \leq I_0 + \bar{m} < \bar{A}$, giving $F(\bar{A}) < 0$; the intermediate value theorem delivers the root. For the monotonicity claim, $\bar{g}'$ is the weight-summed form of Lemma 1(i) (Lemma A.14(ii)); the displayed bound follows from $\varphi \leq \varphi(0)$ and $1/\tau_n = \sqrt{n}/\sigma$, with $\sum_n w_n\sqrt{n} = E[N^{3/2}]/\big(q_0 + E[N]\big)$ under geometric survival, and the calibrated evaluation is in Appendix D. □

Global uniqueness is not an analytic claim: the marketing policy's response to A need not be monotone in general, so F need not be. It is verified numerically over the calibrated region instead: each fund's fixed point is solved by damped iteration with a bracketing fallback, and a 400-point root scan per fund finds exactly one sign change in every cell of the grid (1,200 fixed points across contracts, skill levels, and the three capacity parameters, all cleanly convergent, no multiple roots; Appendix D). The atomistic-competition limit of the closure, in which perfectly elastic informed capital drives the marginal unit's net to zero, is discussed in Appendix F: crowding out is complete there, the marketing margin closes at interior funds, and the main text's informed-capital sector with positive rents is the specification that keeps both of the company's instruments live.

A.7 Proposition 8 and the calibrated comparative statics of Section 5.3

Proposition 8. *(incidence across skill). In the first-round welfare decomposition of the penalty-heavy contract relative to the flat fee, the four terms' signs are distributed across skill bands as follows:*

Skill band	*(i) alpha*	*(ii) fee*	*(iii) variance*	*(iv) gambling*
low ($\mu \leq \mu^*_{flat}$)	0	+	0	0
lower-middle ($\mu^*_{flat} < \mu \leq \mu^*_{float}$)	−	+	+	≈ 0
middle ($\mu^*_{float} < \mu < \tilde{\mu}$)	−	+	+	+ *small*
high ($\mu \geq \tilde{\mu}$)	±	−	−	+

Proof. The four terms of $\Delta W_I(\mu, c)$ are inherited from earlier results, term by term and band by band.

Terms (i) and (iii) (gross alpha delivered and activity variance) read off Proposition 1. In the low band both contracts closet ($x^* = 0$ under both, since $\mu \leq \mu^*_{\text{flat}} < \mu^*_{\text{float}}$), so both terms vanish. In the lower-middle band the floating contract alone closets (Proposition 1(a)): the manager's modest flat-fee activity disappears, so term (i) is negative and the variance contracts, term (iii) positive. In the middle band ($\mu^*_{\text{float}} < \mu < \tilde{\mu}$) both contracts are active and the floating contract compresses (Proposition 1(b)): term (i) negative, term (iii) positive. In the high band ($\mu \geq \tilde{\mu}$) the floating contract amplifies within the mean–variance benchmark: term (i) positive and term (iii) negative under mean–variance; the amplification half is the preference-specific half of Proposition 1(b) (Section 3.6's Remark), so the high band's (i) and (iii) are signed $\pm$.

Term (ii) (the expected fee rate) reads off Lemma 2 and the window averaging of Proposition 4. A fund whose delivered alpha sits below the break-even $\bar{\alpha}(\sigma)$ pays an expected vintage rate below f_M ($\bar{g} < f_M$), so the fee term, $-[\bar{g}_c - f_M]$, is positive: this covers the low band (closet indexers, whose $\alpha = 0 < \bar{\alpha}$; the refund is the noise tax of Proposition 4). In the lower-middle and middle bands the same sign holds because compressed activity keeps delivered alpha below break-even; this step is a calibrated fact, verified on the grid of Appendix D rather than proved, and is so labeled in the text's table. In the high band, delivered alpha clears $\bar{\alpha}(\sigma)$ once activity is amplified, so $\bar{g} > f_M$ and the fee term is negative, with the break-even boundary lying inside the band.

Term (iv) (the within-window gambling cost) reads off Propositions 3 and 6. The excess variance $[\sigma_2^*(e_1)^2 - \underline{\sigma}^2]_+$ is identically zero under the flat contract, whose profile sits on the floor in every state (Proposition 3, part 5); under the penalty-heavy contract it is positive exactly on the catch-up zone and the reward peak. In the low band both contracts closet, so the exposure is zero. The exposure grows with the fund's second-half risk capacity, hence with skill and activity, so it is approximately zero in the lower-middle band, small and positive in the middle band, and largest in the high band, whose tracking error reaches the gambling regions; and Proposition 6 discounts the whole term by the young-money share, since the catch-up incentive dies super-Gaussianly as windows age. □

This step is verified numerically over the calibrated region (Appendix D): on a skill grid against four contracts under both preference specifications, the sign table is matched cell-by-cell in 93.5% of cells under mean–variance and 82.0% under exact CARA pricing (Appendix B), and every exception cell is interpretable (the break-even boundary inside the high band's fee term, tolerance cells in the lower-middle variance term, and mid-band funds whose tracking error is too low to reach the catch-up zone). The high band's alpha term, +78.0bp in band mean under mean–variance, reverses to −142.8bp under CARA; the lower bands are untouched by the preference form. Two auxiliary checks bound the comparison across contracts: under the flat contract all four terms are identically zero by construction, and under the bonus-only contract the fee term is weakly negative everywhere (worst cell −18.5bp) while its gambling term is the undiminished reward peak.

Lemma A.16. *(coverage is priced at the quota-marginal manager). On the first-round basis of Section 5.1, aggregate investor welfare at coverage q is*

$$W(q) = \int_{\mu_q}^{\bar{\mu}} \Delta W_I(\mu)\, dG(\mu) \;+\; \textit{(terms independent of } q\textit{)}, \qquad \mu_q = G^{-1}(1-q),$$

since only floated managers' investors are affected and the floated set is $[\mu_q, \bar{\mu}]$ by Proposition 7. Then $W'(q) = \Delta W_I(\mu_q)$: the marginal first-round welfare of coverage is the quota-marginal manager's own welfare change.

Proof. Fixed-fee managers' first-round welfare change is identically zero, so only the floated interval enters. Differentiate under the integral sign: $W'(q) = -\Delta W_I(\mu_q)\, G'(\mu_q)\, \mu_q'(q)$, and differentiating $G(\mu_q) = 1 - q$ gives $\mu_q'(q) = -1/G'(\mu_q)$. □

Lemma A.17. *(signs of the constraint surfaces). Within the step family, the three constraint surfaces evaluated at the institutional point $(d, u, \ell, h) = (60\text{bp}, 30\text{bp}, 3\%, 6\%)$ have the following signs. Parts (a)–(c) are proved; the aggregation of part (a) over the industry is a calibrated comparative static for the reason recorded below.*

(a) The political constraint attaches to the penalty jump. $\mathcal{P}$ is strictly increasing in d, and, to exponential order, u does not enter $\mathcal{P}$.

(b) The risk constraint prices the reward side. $\mathcal{R}$ is increasing in u (the reward peak of Proposition 3 scales with the reward jump) while the reward side's only positive welfare function is the high band's activity amplification, the mean–variance-specific half of Proposition 1(b).

(c) The viability constraint bounds the penalty side. $\mathcal{S}$ is decreasing in d and increasing in ℓ; equivalently, decreasing in the proximity of the downgrade line: a heavier penalty or a closer downgrade line raises the activity threshold μ^*_{float} and shrinks the active population.

(d) Read together at the calibration, the three surfaces describe a trade-off directionally consistent with penalty-heavy asymmetry, $d > u$ with $\ell < h$: the political concern pulls the penalty jump up, the risk concern pulls the reward jump down, and the viability concern keeps the penalty side off the corner. This is a directional summary of (a)–(c), not an optimality or necessity claim. The comparators of Section 2.4 illustrate the same trade-offs: the symmetric step halves the political force at $d = 30bp$; bonus-only removes the

penalty side and worsens the risk side; the linear fulcrum's within-year profile is flat at the floor in every state (Proposition 3, part 5) and manufactures no bust fine.

Proof. (a) The bust fine of Proposition 5(a) is $BF(\alpha) = \sum_{n\geq 1} q_n A_n(\beta)\left[d\,\Phi(t_D(n)) - u\,\bar{\Phi}(t_U(n))\right]$, so

$$\frac{\partial BF}{\partial d} = \sum_{n\geq 1} q_n A_n(\beta)\,\Phi(t_D(n)) > 0$$

term by term, and $\mathcal{P} = \int_{\{\alpha(\mu)<-\ell\}} BF(\mu)\,A(\mu)\,dG(\mu)$ inherits the strict sign. The reward jump enters BF only through the upgrade-tail term $-u\,\bar{\Phi}(t_U(n))$; on the constraint's integration domain $\{\alpha < -\ell\}$ the upgrade threshold sits at standardized distance $t_U(n) = (h-\alpha)\sqrt{n}/\sigma \geq (h+\ell)\sqrt{n}/\sigma$, so $\bar{\Phi}(t_U(n)) \leq \frac{1}{2}\exp\bigl(-n(h+\ell)^2/2\sigma^2\bigr)$ vanishes exponentially: to exponential order, u does not enter $\mathcal{P}$.

(b) The aggregate excess variance $\mathcal{R}$ integrates $[\sigma_2^*(e_1)^2 - \underline{\sigma}^2]_+$, discounted by the young-money share (Proposition 6). The only u-dependent incentive in the within-year problem is the upgrade term $u(h-E_2)\varphi((h-E_2)/\sigma_2)$ of Lemma A.6, strictly increasing in u wherever it is positive; in the region where it binds, the implicit function theorem on the first-order condition (with the second-order condition strict) gives $\partial\sigma_2^*/\partial u > 0$, and elsewhere σ_2^* is u-free. The integrand is therefore weakly increasing in u state by state and strictly increasing on the reward-peak region, so $\mathcal{R}$ is increasing in u. Note the structural feature: $\mathcal{R}$ is strictly positive even at $u = 0$, because the penalty floor by itself breeds the catch-up gambling zone (Lemma A.9 needs only $d > 0$); a punish-only contract is not riskless. The reward side's compensating function is the high band's activity amplification, which is the mean–variance-specific half of Proposition 1(b) (Section 3.6's Remark): priced exactly, it does not survive, so the risk concern argues for the smallest u at which the upgrade tier exists.

(c) The two signs run through Proposition 1(a)'s threshold and Lemma A.2's intensive margin. For d: on the extensive margin, differentiating $\mu^*_{\text{float}} = (c_1 - B)/A$ (logit center, linear technologies) gives

$$\frac{\partial\mu^*_{\text{float}}}{\partial d} = \frac{\varphi(\ell/\sigma_0)}{\sigma_0^2\,(V_\alpha + \Lambda E[Q'])^2} \times \Bigl[s\ell\,(V_\alpha + \Lambda E[Q']) - \sigma_0\,(c_1 + s\rho\sigma_0 - sV_\sigma)\Bigr],$$

whose bracket is positive at the calibration (+0.0203; Appendix D): the variance-tax channel dominates the offsetting skill-reward channel (a heavier penalty also pays more for escaping the penalty zone, which works toward inclusion), so the activity threshold rises; on the intensive margin, Lemma A.2 gives $\partial x^*/\partial d < 0$ for all $\mu < \hat{\mu}$, one-sided over the calibrated skill range. Both margins give $\partial \mathcal{S}/\partial d < 0$. For ℓ: the variance tax weakens as the downgrade line recedes: $\partial V_\sigma/\partial \ell = \frac{d}{\sigma_0^2}\varphi(\ell/\sigma_0)\big(\frac{\ell^2}{\sigma_0^2}-1\big) > 0$ for $\ell > \sigma_0$, which lowers the threshold; the offsetting sub-channel runs through the skill reward, $\partial V_\alpha/\partial \ell = -\frac{d\ell}{\sigma_0^3}\varphi(\ell/\sigma_0) < 0$ (a more distant penalty line also pays less for escaping it), which raises the threshold. The variance channel dominates at the calibration, so μ^*_{float} falls as ℓ recedes and $\partial \mathcal{S}/\partial \ell > 0$: equivalently, $\mathcal{S}$ is decreasing in the *proximity* of the downgrade line. The calibrated elasticities are -0.028 per 100bp of penalty jump and $+0.016$ per percentage point of downgrade-line distance (Appendix D).

(d) A directional summary of (a)–(c): the political constraint pulls the penalty jump up and only the penalty jump (to exponential order, the reward side does not enter $\mathcal{P}$); the risk constraint pulls the reward jump down (its only compensating function is preference-fragile); the viability constraint stops the penalty logic from running to the corner (a penalty too heavy, or a downgrade line too close, thins the active industry). At the calibration the three directions are jointly consistent with heavy but bounded penalty, light but present reward, penalty line closer than the reward line: the $d > u$ with $\ell < h$ pattern of the 2025 contract. The comparators read as illustrations of the same trade-offs: the symmetric step ($d = u = 30\text{bp}$, $\ell = h$) halves the political force, since $\mathcal{P}$ is strictly increasing in d; bonus-only ($d = 0$) removes the penalty side ($\partial \mathcal{P}/\partial d > 0$ with $\mathcal{P} = 0$ at $d = 0$) and keeps the reward peak's gambling undiminished (Lemma A.11 needs no floor); the linear fulcrum's within-year profile is flat at the floor in every state (Proposition 3, part 5) and manufactures no bust fine, since a floorless linear schedule cannot punish a persistent loser. $\square$

The calibrated constraint surfaces (Appendix D): the political surface, computed on a representative bust fund as the bust fine's share of lifetime fee revenue, runs from 0% at $d = 0$ to 38.8% at the institutional $d = 60\text{bp}$ to 77.7% at $d = 120\text{bp}$ (33.8% under performance-sensitive redemption at $\theta = 1$); the risk surface runs from 2.3×10^{-5} at $u = 0$ to 1.4×10^{-4} at $u = 60\text{bp}$ on the raw measure, strictly positive at $u = 0$, as part (b) indicates. The scope statement of Section 5.3 applies here: the political surface is a per-fund pricing map, evaluated on a representative bust fund, and the population it aggregates over lies outside the industry model's new-issuance slate; the signs proved in (a)–(c) are statements about the constraint surfaces' slopes and are untouched by that omission.

Appendix B Exact Risk Pricing: the CARA–Normal Problem under Logit Flow

B.1 Setup: the exact problem, and what the linearization prices away

The manager's wealth is the sum of fee and flow income,

$$W(e) \;=\; f(e) + \Lambda\, Q(e), \qquad e = \mu x + (\sigma_0 + sx)\, z, \quad z \sim \mathcal{N}(0,1),$$

with f the three-tier step schedule of Section 3 and Q the flow function. The main text prices the lottery in W by a mean–variance reduction (Assumption A2): around $E[e]$, $\mathrm{Var}(W) \approx \omega^2\sigma^2(x)$ with $\omega = dW/de$, and the coefficient ρ absorbs ω^2 together with career and contract costs. This appendix solves the exact constant-absolute-risk-aversion problem instead:

$$\max_{x\in[0,\bar{x}]} \; \mathrm{CE}(x) \;=\; -\frac{1}{\gamma}\ln \mathbb{E}\big[e^{-\gamma W(e)}\big] - C(x), \qquad C(x) = c_1 x + \tfrac{c_2}{2}x^2,$$

with the **formal logit flow** $Q(e) = e^{ke}/(e^{ke}+1)$, $k = 8$, and $\Lambda = 776.8$bp (the matching constant of Section 3.3, recomputed here at $4.315\times$ the exponential benchmark). Wealth is measured as a fraction of AUM per year, so γ carries units of (unit wealth)$^{-1}$; we report each γ together with its *equivalent* ρ, defined by matching the second-order expansion $\mathrm{CE} \approx \mathbb{E}[W] - C - \frac{\gamma}{2}\mathrm{Var}(W)$ at a reference point, $\rho_{\text{equiv}} = 2(\mathbb{E}[W] - \mathrm{CE}_W)/\sigma^2(x_{\text{ref}})$.

What the linearization prices away is one specific channel. Under Assumption A2 the term $\frac{\rho}{2}\sigma^2(x)$ is a function of the return distribution alone; it treats the floating and the flat contract identically. But the step schedule makes W itself a lottery, landing in the penalty tier, the base tier, or the reward tier, and that **fee-lottery variance is contract-specific**: it is larger under the floating contract, it grows with activity, and it is exactly the risk a manager who prices his wealth lottery exactly charges to the floating contract. The mean–variance form absorbs this channel into the symmetric coefficient ρ; the CARA–normal problem restores it. A CARA problem with the exponential flow form embedded is the natural intermediate case; this appendix solves the joint problem (exact risk pricing *and* the formal bounded flow).

Calibration and computation. All parameters are the paper's calibration: $(f_L, f_M, f_H) = (60, 120, 150)$bp, $(\ell, h) = (3\%, 6\%)$, $\sigma_0 = 1\%$, $s = 0.05$, $(c_1, c_2) = (5, 80)$bp,

$\bar{x} = 2$; the mean–variance reference is $\rho = 5$. Expectations are taken by trapezoid quadrature over $z \in [-10, 10]$ with $dz = 0.002$ (10,001 nodes), the formal caliber of Appendix D; the optimizer is a two-stage grid search (coarse step 0.01, refined to 0.001) with an explicit comparison against $x = 0$. Two implementation disclosures. First, Gauss–Hermite quadrature (96 nodes) still misprices the certainty equivalent by 2.1–3.6bp on this integrand: the step discontinuities, not the unbounded flow, are what defeats Gauss–Hermite; boundedness of the logit flow does not cure it, so trapezoid quadrature remains the formal caliber and Gauss–Hermite is reported only as a cross-check. Second, matching γ to the mean–variance $\rho = 5$ requires choosing where to evaluate $\omega = dW/de$; we report both natural candidates, the origin slope $\omega = \Lambda Q'(0) = 2\Lambda \Rightarrow \gamma \approx 207.2$ and the calibration-center slope $\omega = \Lambda\,\mathbb{E}[Q'] \Rightarrow \gamma \approx 217.7$ (the latter is the exact analogue of how Λ itself was matched), plus a third candidate $\gamma \approx 241.1$. No conclusion below depends on the choice.

B.2 Results: compression is robust; amplification reverses; the reversal point brackets the text's "roughly 3"

The compression side holds at every risk aversion we compute. For all twelve values of γ on the grid (spanning equivalent risk aversion from 0.04 to 11.3), the penalty-heavy floating contract lies weakly below the flat contract everywhere below the crossing (and everywhere on the skill grid once the crossing disappears): $\max_\mu[x^*_{\text{float}}(\mu) - x^*_{\text{flat}}(\mu)] = 0$ on the pre-crossing region, achieved only where both managers closet. The low-skill suppression that carries the paper's headline does not depend on how risk is priced. The activity thresholds themselves rise with γ: exact pricing of the fee lottery widens the closet region for both contracts, and, as in the mean–variance benchmark, the floating contract's threshold exceeds the flat one's at every γ.

(At low γ the crossing sits below 5%, so the table's $x^*(5\%)$ columns show the float above the flat (the amplification side alive); at $\gamma = 60$ the crossing survives but migrates to $\tilde{\mu} = 6.2\%$, beyond the mean–variance $\tilde{\mu} \approx 5.1\%$ and at the edge of the calibrated skill distribution (median 2.5%, 90th percentile 6%), so the 5% columns already show suppression.)

The amplification side reverses, and the reversal point is where the text says it is. The crossing (the amplification half of Proposition 1(b)) survives exact pricing at low risk aversion and disappears at high risk aversion. On the grid, the last surviving crossing sits at equivalent risk aversion 2.40 ($\gamma = 60$, pushed out to $\tilde{\mu} = 6.2\%$) and the

Table B.1: CARA–normal exact solution under logit flow, by risk aversion. ρ_{equiv} is the second-order-equivalent mean–variance risk aversion at the reference point ($\mu = 5\%, x = 0.496$); μ^*_{flat} / μ^*_{float} are the activity thresholds (10bp grid); $\tilde{\mu}$ is the first skill at which $x^*_{\text{float}} > x^*_{\text{flat}}$; $x^*(5\%)$ gives the intensive margin at mid-grid skill.

γ	ρ_{equiv}	μ^*_{flat}	μ^*_{float}	$\tilde{\mu}$	$x^*_{\text{flat}}(5\%)$	$x^*_{\text{float}}(5\%)$	compression side
1	0.04	0.4%	0.6%	2.8%	0.794	0.898	holds
3	0.11	0.4%	0.7%	2.9%	0.785	0.881	holds
5	0.19	0.4%	0.7%	2.9%	0.775	0.869	holds
10	0.38	0.5%	0.7%	3.1%	0.751	0.836	holds
20	0.77	0.5%	0.8%	3.6%	0.706	0.775	holds
60	2.40	0.8%	1.2%	6.2%	0.547	0.504	holds
80	3.26	1.0%	1.4%	—	0.483	0.373	holds, reversed everywhere
100	4.17	1.1%	1.6%	—	0.428	0.264	holds, reversed everywhere
120	5.12	1.3%	1.9%	—	0.381	0.196	holds, reversed everywhere
207.2 (matched, origin)	9.56	2.0%	3.0%	—	0.235	0.059	holds, reversed everywhere
217.7 (matched, cal. center)	10.10	2.1%	3.1%	—	0.222	0.050	holds, reversed everywhere
241.1 (alternative matching)	11.29	2.2%	3.5%	—	0.195	0.035	holds, reversed everywhere

first full reversal at 3.26 ($\gamma = 80$): the reversal boundary in mean–variance-equivalent units lies in $(2.40, 3.26)$, bracketing the "effective risk aversion above roughly 3" of the Section 3.6 remark. At all three matched values of γ (the two logit matching candidates and 241.1), the crossing is gone and $x^*_{\text{float}} \leq x^*_{\text{flat}}$ holds over the entire skill grid, with the gap widening in skill: the float sits 0.17 of deviation below the flat at $\mu = 5\%$ and 0.20 below at $\mu = 8\%$ under the calibration-center matched γ (0.16 and 0.21 at $\gamma = 241.1$). The amplification half of Proposition 1(b) is mean–variance-specific: the mean–variance form prices away the fee schedule's own lottery variance; exact pricing restores that contract-specific channel and reverses the sign, while the compression half (the statement that the contract suppresses managers whose deviation buys noise) is invariant to how risk is priced.

Two contract-level observations. First, the penalty-heavy schedule is no longer the *unique* suppressor once risk is priced exactly: the symmetric-fulcrum schedule (itself a two-sided fee lottery) also falls below the flat contract at $\gamma \geq 20$ (by up to -0.059 of deviation at the matched γ), while the bonus-only schedule, which holds no downside lottery, remains weakly above flat throughout. The four-layer hierarchy of Section 3.6 is a mean–variance statement; under exact pricing, *every two-sided schedule suppresses*, and the penalty-heavy schedule suppresses most. This strengthens, rather than qualifies, the paper's reading of what the step-plus-tilt geometry does. Second, the corner solutions of the exponential-flow CARA version (x^* driven to the $\bar{x} = 2$ boundary from $\mu \approx 2\%$–4% upward at $\gamma \leq 10$) disappear entirely under logit flow: the cornering was the unbounded variance appetite of the exponential form, not a property of exact risk pricing.

B.3 Boundaries

What this appendix covers. (i) The static one-period problem of Sections 3–3.6, solved exactly for four contracts on a twelve-point risk-aversion grid, under the formal logit flow. (ii) The welfare decomposition of Section 5 and the quota margin on the CARA branch: every x^*-dependent term (the alpha term, the vintage-aggregated fee term, the variance term, and the gambling term through its dependence on $\bar{\alpha}_2 = \mu x^*/2$ and $\sigma(x^*)$) is recomputed with the joint solver. Headline readings at the calibration-center matched $\gamma = 217.7$: sign-table agreement 82.0% (656/800 cells, every exception of the same interpretable classes as before); band means (low/mid-low/mid/high) of the alpha term $0.0/-1.0/-42.1/-142.8$bp, so the high band's first term reverses as before; quota totals negative throughout, -29.5bp at $q = 0.6$, $\gamma_I = 2$ (marginal -1.5bp). Sensitivity across the three matched-γ candidates: agreement 78.4%–83.8%, high-band alpha term -138.4 to -150.1bp, quota total -27.9 to -30.0bp. The amplification side's welfare contribution is negative under exact pricing in every variant; the compression side's terms are untouched by the preference form, as Section 5 states.

What remains open. (i) **The two-period dynamic problem (Proposition 3) has no CARA version.** The gambling term (iv) enters this appendix only through x^*; the second-half variance choice $\sigma_2^*(e_1)$ is still optimized under the mean–variance form. A joint CARA version of the catch-up zone and the dead zone (where the fee lottery's variance is state-dependent and the concavity of the utility function interacts with the kink geometry directly, not through a linearized ρ) is the natural next step and is not approximated by anything computed here. (ii) **The industry equilibrium of Section 4.5 is mean–variance throughout**; the CARA branch reaches the welfare decomposition through x^* but was never propagated through the sorting and demand closure of the industry equilibrium. Given that the joint solver lowers activity levels at matched risk aversion, the industry aggregates would move in level; the compression-driven sign of the aggregate activity change (Section 4.5's revised reading) is carried by the side that this appendix shows to be preference-robust. (iii) **The reversal boundary**

is grid-bracketed, not characterized in closed form, and the equivalent-ρ mapping is a second-order diagnostic evaluated at one reference point; both are reported as intervals, not points. (iv) **Quadrature.** The formal caliber is trapezoid integration; Gauss–Hermite misprices the certainty equivalent by 2–3.6bp on this integrand even with bounded flow (the step discontinuities are the obstacle); we keep the schedule unsmoothed and report the mispricing. (v) The matching of γ to ρ is a second-order equivalence at a chosen evaluation point; we report both natural candidates and a third candidate, and no conclusion depends on the choice.

Appendix C The Exponential Benchmark Form

C.1 The exponential form as a robustness benchmark

The paper's formal flow function is the logit saturation form of Section 3.3, and every quantitative statement in the main text is computed under it. The exponential form $Q(e) = Q_0 e^{ke}$ is collected here as a robustness benchmark, for three reasons.

Closed-form intuition. The exponential is the only flow form under which the Gaussian expectations that drive the model are available in closed form: $E[e^{ke}]$ is the moment-generating function of $N(\alpha, \sigma^2)$ evaluated at k. One line of algebra then exposes the economic engine of the flow side: the expectation is increasing in *both* the mean and the variance of performance, so a convex flow relation makes the manager love variance, and makes the variance appetite, the term $\Lambda k^2 \sigma$ that fights the contract's variance tax, visible as an explicit object rather than a quadrature output. Appendix A uses this closed form at exactly one point (Lemma A.1) and reads the variance channel of Proposition 1(a) through it; the formal logit proofs then proceed by symmetry and boundedness arguments that the exponential intuition motivates.

Literature interface. The exponential is the reduced form most directly comparable to the empirical flow literature: the log-linear flow regressions of Chevalier and Ellison (1997) and Sirri and Tufano (1998) (percentage flow linear in performance or rank) correspond exactly to an AUM multiplier that is exponential in performance, and the reduced-form flow term $\Lambda E[Q]$ with $Q = Q_0 e^{ke}$ is that correspondence read into the manager's objective. Keeping the exponential benchmark's results collected here lets a reader translate every one of the paper's flow-side quantities (the activity thresholds, the dilution speed, the asymptotic constants) back into the units of those regressions, and shows that the paper's choice of a bounded form is the conservative direction: anything the fee schedule survives against a bounded flow incentive, it survives against an unbounded one (Section 3.4).

The local-approximation relation. The two forms are not rivals; the exponential is the logit's local approximation in the deep convex region. For $ke \ll \ln c$,

$$Q(e) = \frac{e^{ke}}{e^{ke} + c} \;\approx\; \frac{e^{ke}}{c},$$

and the curvature structure is identical: both are convex, both love the mean and the variance. The approximation fails exactly where the economics of saturation begins: at $ke \approx \ln c$ the logit turns concave while the exponential keeps accelerating. The calibrated region straddles that boundary: with $k = 8$ and $\alpha \le 5\%$, ke reaches 0.4 while $\ln c = 0$ at $c = 1$, so part of the calibrated state space sits on the logit's concave side, where the exponential form **overstates** the flow term's variance appetite. The calibration consequence is the matching constant of Section 3.3: saturation shrinks the flow incentive's slope at the calibration center by a factor of 4.315 ($E[Q'_{\text{exp}}] = 8.42$ against $E[Q'_{\text{logit}}] = 1.95$ at $e \sim N(0, 4\%)$), and the logit-form $\Lambda = 776.8$bp is chosen to restore incentive equivalence with the exponential benchmark's $\Lambda = 1.5\, f_M \approx 180$bp, incremental money staying about one and a half discounted fee-years, a deliberately conservative anchor.

C.2 The closed-form result set

The moment-generating-function moments. Under $Q(e) = Q_0 e^{ke}$ and $e \sim N(\alpha, \sigma^2)$, every moment the model uses is the same MGF evaluation:

$$E[Q] = Q_0 \, \exp\!\Big(k\alpha + \tfrac{k^2}{2}\sigma^2\Big), \qquad E[Q'] = k\, E[Q], \qquad E[Q''] = k^2\, E[Q],$$

since $Q' = kQ$ and $Q'' = k^2 Q$. The flow term's marginal value of variance is therefore

$$\frac{\partial E[Q]}{\partial \sigma} \;=\; k^2\, \sigma\, E[Q],$$

positive everywhere and growing without bound in σ, and self-amplifying, because $E[Q]$ itself grows exponentially in σ^2. This is the *unbounded variance appetite*: the object that powers every sharper coordinate below, and the precise feature that the logit's saturation removes ($kQ(1-Q) \le k/4$, with the appetite at the calibration center identically zero by the symmetry of Lemma A.4). At the calibration center the two flow moments that enter Proposition 1(a) reduce to $\Lambda E[Q'] = \Lambda k$ and $\Lambda \sigma_0 E[Q''] = \Lambda k^2 \sigma_0$ (at $Q_0 = 1$), which is how the exponential version reads the variance channel of the activity threshold.

Thresholds and dilution coordinates. Under the exponential form the flat-contract threshold is exact and affine in $1/\Lambda$,

$$\mu^*_{\text{flat}}(\Lambda) = \bar\mu + \frac{M_f}{a_0'\,k\,\Lambda}, \qquad \bar\mu \equiv -\frac{s_0\,\sigma_0\,k}{a_0'} < 0,$$

and both thresholds change sign at explicit finite multiples of baseline flow: the *death points*

$$\Lambda^0_f = \frac{M_f}{s_0\,\sigma_0\,k^2} = 5.2\times, \qquad \Lambda^0_p = \frac{M_p}{s_0\,\sigma_0\,k^2} = 5.9\times, \qquad \Lambda^0_p - \Lambda^0_f = \frac{-V_\sigma}{\sigma_0\,k^2} = 0.69\times,$$

with $M_f \equiv c_1 + s_0\rho\sigma_0$ and $M_p \equiv M_f - s_0 V_\sigma$ as in Appendix A. The threshold gap decays from a baseline of **24.1bp**, falling to half at 2.22 times baseline flow and to a quarter at 4.61 times:

Λ multiple	0.5	0.75	1.0	1.5	2.0	3.0	5.0	7.0
$\mu^*_{\text{float}} - \mu^*_{\text{flat}}$ (bp)	40.0	30.3	24.1	17.0	13.1	9.0	5.5	4.0

On the intensive margin, the two contracts' optima jump from interior values (≈ 0.67) to the corner $\bar x = 2$ and bunch there at about **2.6–2.7 times baseline flow**, contract irrelevance arriving as universal corner gambling rather than gentle convergence.

These coordinates are implications of the unbounded-variance-appetite form, and are labeled as such wherever they appear. The mechanism is visible in the displays above: the death points exist because the variance-appetite term $\Lambda k^2\sigma_0$ grows linearly in Λ until it covers the cost and risk terms by itself and activity requires no skill at all; the cliff exists because the flow term's curvature, $\Lambda k^2\sigma\sigma'\, e^{k\alpha + k^2\sigma^2/2}$, grows without bound in the variance and manufactures a convex region at high deviation, destroying the interior optimum in a jump. Appendix A.2.2 (Proposition A.5) proves the closed forms and carries the full artifact diagnosis, which we do not repeat here. Under the formal logit form none of these coordinates survives: both thresholds are rational functions of Λ, strictly positive at every finite multiple (no death point: universal participation is a limit, not a crossing), the baseline gap reads 22.0bp with milestones at 2.21/4.58/12.0 times baseline, and the intensive margin converges smoothly, the corner reached without a jump at 10–22 times baseline flow (Appendix A.2.1). The c-scan bounds the reading from the other side: even at $c = 2$, where the logit's symmetry fails in the appetite's direction, the bounded appetite postpones any death points to 16.4 and 18.6 times baseline, more than three times the exponential coordinates. The limit itself is form-free (Proposition 2); the role of the exponential coordinates in this paper is stated once, in Appendix A.2.2.

C.3 The asymptotic expansion (Proposition 4′)

The exponential form also delivers the threshold gap's two-term expansion in closed form (Proposition A.5(ii)):

Proposition 4′.

$$g(\Lambda) = \frac{C_1}{\Lambda} + \frac{C_2}{\Lambda^2} + O(\Lambda^{-3}),$$
$$C_1 = \frac{s_0\big(\sigma_0 k\, V_\alpha - V_\sigma\big)}{a_0'\, k} > 0, \qquad C_2 = -\frac{V_\alpha\big(M_p + s_0\sigma_0 k V_\alpha\big)}{a_0'\, k^2} < 0.$$

Calibrated values: $C_1 = 5.12\times10^{-5}$ and $C_2 = -1.42\times10^{-7}$. The two-term expansion is accurate to 0.33% already at baseline flow and to better than 0.1% beyond twice baseline; the one-term formula overstates the gap by **17.8%** at baseline (below 4% beyond five times baseline); and the convergence $g(\Lambda)\cdot\Lambda \to C_1$ is verified directly (5.103×10^{-5} against 5.119×10^{-5} at fifty times baseline). The signs are the content: $C_1 > 0$ is guaranteed jointly by the contract's skill price ($V_\alpha > 0$) and variance tax ($V_\sigma < 0$): penalty dominance is exactly the condition that the threshold gap is asymptotically sign-definite, and $C_2 < 0$ means the simple $1/\Lambda$ formula systematically overstates the residual bite.

Against the logit version of the same object (Appendix A.2.1):

object	exponential	logit ($c = 1$)
leading constant	$C_1 = 5.12\times10^{-5}$	$C_1' = -sV_\sigma/E[Q'] = 2.00\times10^{-4}$
second constant	$C_2 = -1.42\times10^{-7}$	$C_2' = -M_pV_\alpha/(E[Q'])^2 = -2.27\times10^{-6}$
one-term overstatement at baseline	17.8%	16.8% (7.8% at 2Λ, 3.0% at 5Λ)
sign structure	$C_1 > 0$, $C_2 < 0$	same, by the same penalty-dominance logic

The two expansions tell the same qualitative story: the last bit of bite survives with the penalty side's sign, and the naive formula overstates it. The constant levels differ, the logit's constants being larger because its matched Λ is.

The expansion's incremental content is limited. The leading constant's sign is already delivered by penalty dominance, a theorem-level fact that needs no expansion; and the intensive margin, the margin where an expansion would have been genuinely useful, provably admits none: the first-order implicit-function prediction achieves only order-of-magnitude accuracy in the deep interior (predicted-to-exact ratios of 1.4–1.7) and fails entirely in the corner-transition band, so the intensive margin carries only partition statements (Appendix A.2.2). Its role here is a compact summary of the exponential benchmark's dilution arithmetic.

Appendix D Calibration and Numerical Methods

D.1 The calibration, in one table

Table D.1 is the complete version of main-text Table 1: every parameter used anywhere in the paper or its appendices, its baseline value, its meaning, its source or basis, and the range over which it is scanned in the robustness exercises of Section D.3. Contract parameters come from the prospectuses of the first twenty-six products; all other parameters are chosen to span the empirically relevant range of Chinese active equity funds and are varied as shown.

Table D.1: Master calibration.

Parameter	Baseline	Scanned range	Meaning	Source / basis
f_L, f_M, f_H	0.6%, 1.2%, 1.5%	—	fee tiers	prospectuses
ℓ, h	3%, 6%	—	downgrade / upgrade thresholds	prospectuses
u, d	30bp, 60bp	d: 30bp–120bp; u: 0–60bp	reward / penalty jump	implied by the tiers; scanned along the constraint surfaces of Section 5.3
σ_0	1%	0.5%–2% (nine-cell grid; $\sigma_0 \to 0$ limit reported)	residual tracking error of a closet indexer (Assumption A1)	the empirical range is roughly 1%–3% per year
s	5%	—	slope of the tracking-error technology $\sigma(x) = \sigma_0 + s\,x$	a units convention, not an estimate; one unit of deviation is defined as +5pp of tracking error, anchoring the per-unit calibrations of μ, c_1, c_2 (Section 3.1)

continued on next page

Table D.1 (continued)

Parameter	Baseline	Scanned range	Meaning	Source / basis
$\bar{x}$	2	—	deviation bound	set so full deviation corresponds to $\approx 11\%$ tracking error, the far tail of sustainable deviation (Section 3.1)
$k,\ c$	8, 1	$c \in \{0.5, 1, 2\}$	flow semi-elasticity and saturation constant	k: platform-era Chinese evidence, band $[4.5, 10.5]$; c: normalization plus symmetry (Section 3.5)
Λ	776.8bp	0.5×–30× baseline	discounted flow incentive (logit calibration)	incentive-equivalence matched to the exponential benchmark's 180bp at ratio 4.315 (Section 3.5)
$c_1,\ c_2$	5bp, 80bp	—	deviation cost (linear research cost + quadratic term)	the quadratic term embeds the scale-times-activity trading cost of Pástor et al. (2020)
ρ	5	0–20 (refined grid)	risk penalty (scaled; Assumption A2)	reduced-form; validated against the exact CARA–normal problem of Appendix B
μ	grid 0–10% (step 0.05%)	—	annual alpha per unit of deviation (perceived skill)	grid spans the plausible range; the U.S. industry-average gross alpha is estimated near zero (Berk and van Binsbergen, 2015; Carhart, 1997); Chinese outperformance lacks persistence (Lin et al., 2009)
$\bar{\alpha}_2$	1.5%	effort extension: $a \in [\underline{a}, \bar{a}]$	fixed second-half expected alpha (Section 3.8)	effort-extension caliber

continued on next page

Table D.1 (continued)

Parameter	Baseline	Scanned range	Meaning	Source / basis
Λ (half-year, Section 3.8)	90bp	exponential-benchmark half-year caliber; logit-native caliber verified	half-year flow weight	—
ρ (Section 3.8)	5	0–20 (two-peak survival grid)	within-year risk penalty	the Section 3.5 calibration, shared by the within-year reduction
$[\underline{\sigma}, \bar{\sigma}]$	[0.5%, 8%]	—	second-half risk bounds	span the operative range of the within-year problem
η (effort cost)	∞ (baseline)	$\{5, 20, 100, \infty\}$	effort-cost curvature (Appendix A.3.1)	baseline ∞ removes the effort margin
r (redemption hazard)	0.15	0.07	annual redemption hazard of sticky money (Section 4)	—
β (discount factor)	0.97	—	annual discount factor for vintage fee revenue	—
θ (redemption sensitivity)	0	$\{0, 0.5, 1, 2\}$	performance-sensitive redemption	robustness control (Section 4)
q (quota)	0.6	margin traced over q	floating-share mandate (Section 4.5)	the 2025 action plan (China Securities Regulatory Commission (中国证券监督管理委员会), 2025)
κ (capacity dilution)	0.2%	$\{0.1\%, 0.2\%, 0.5\%\}$ per unit of normalized scale	smart-money dilution (Sections 4–4.5)	illustrative
G (skill distribution)	lognormal, median 2.5%, p90 6%	—	industry skill support $\mu \geq 0$ (Section 4.5)	illustrative lognormal fit
γ (CARA coefficient)	—	12-point grid {1, 3, 5, 10, 20, 60, 80, 100, 120, 207.2, 217.7, 241.1}	exact risk pricing (Appendix B); equivalent ρ spans 0.04–11.3	—
τ (skill-belief dispersion)	0 (baseline)	$\{0, 1\%, 2\%, 3\%\}$	the skill-belief extension of Appendix A (isomorphism check)	—

One caliber deserves an explicit statement, because it is the one place where the paper carries two numerical conventions at once. The half-year flow weight of 90bp (Section 3.8) is the $\Lambda \approx 180$bp annual benchmark read on a half-year window; the two-peak structure of Proposition 3 is a product of the fee schedule's steps and is robust to both the logit-native caliber and the c scan, so that section is computed on the 90bp grid. The exponential benchmark, separately, is retained as a local approximation, with its own coordinates collected in Appendix C and its artifact diagnosis in Appendix A.2.2; nothing in the main text is computed under it.

D.2 Numerical methods

Gaussian quadrature. Expectations of smooth functions against the Gaussian density (every flow moment under the logit form, including the rational closed forms of Appendix A.2.1) are computed by Gauss–Hermite quadrature with 96 nodes (GH-96). On smooth integrands this is machine-precision accurate: the vanishing variance appetite of Lemma A.4 evaluates to 4.4×10^{-17}.

Certainty equivalents: trapezoid is the formal caliber. The certainty-equivalent integrals of Appendix B, whose integrand contains the step schedule's discontinuities, are computed by trapezoid quadrature over $z \in [-10, 10]$ with $dz = 0.002$ (10,001 nodes). We do not smooth the schedule to accommodate quadrature; the alternative's error is on record: GH-96 misprices the certainty equivalent by 2.1–3.6bp on this integrand. The step discontinuities, not the flow form, are what defeats Gauss–Hermite: the logit's boundedness does not cure it, so trapezoid remains the formal caliber and Gauss–Hermite is reported only as a cross-check. This is the disclosure of Sections B.1 and B.3, recorded here as the paper's standing quadrature policy.

Global optimization. Static optima are found by grid argmax, never by local first-order solving alone: the manager's problem is known to develop convex regions at high deviation (Lemma A.3), and a local solver would mistake the manufactured convexity for a non-binding constraint. Where a closed form exists (the activity thresholds of Proposition 1(a), exact solutions of $U_x(0; \mu) = 0$), a fine-grid global argmax check at skill step 0.05% coincides with it up to grid rounding: logit baseline thresholds 1.95%/2.20% on the grid against 1.934%/2.154% in closed form. The CARA solver of Appendix B uses a two-stage grid (coarse step 0.01, refined to 0.001) with an explicit comparison against the corner $x = 0$.

Fixed points. The industry equilibrium's dilution feedback (smart-money inflow dilutes alpha, the diluted alpha reprices the vintage fee rate, the repriced fee feeds back into the informed sector's position) is solved by damped iteration, with a 400-point root scan per fund as the uniqueness check. Across the full verification suite, 1,200 fixed points were computed, all convergent under damping (maximum 34 iterations, maximum absolute residual 10^{-10}, exactly one sign change in every root scan, the bracketing fallback never invoked). The informed-capital rent identity $\alpha(A) - \bar{g} = \kappa A^S$ holds to 3.5×10^{-18} at every interior fixed point.

Equilibrium-scale incidence. Section 5.1's first-round decomposition is measured at fixed scale; its counterpart at Section 4.7's fixed point is as follows, in basis points of aggregate net return across the skill distribution. The informed sector's aggregate net rises across the reform, by +75.5, +35.6, +11.6bp at $\kappa = 0.1\%, 0.2\%, 0.5\%$: the rent κA^S grows where the floating contract amplifies skilled funds' size, which is the mean–variance amplification side of Proposition 1(b) operating at industry level and inherits that side's preference dependence. Trusting money's aggregate net moves by -1.2, -1.2, -0.8bp on the same grid: the pushed layer's activity compression lands on captive money, partly offset by the closet segment's fee refund. The two accounts answer different questions: the first-round decomposition prices the reform's direct incidence on money that has not reoptimized; the fixed-point account shows what scale adjustment then does with it.

Monte Carlo. The performance-sensitive-redemption robustness of the bust fine (Section 4) is computed by simulation: 500,000 redemption paths per parameter cell, horizon capped at 150 years, redemption rule identical across contracts so that only the contract differs; standard errors are below 1bp on every reported level.

D.3 Robustness scans, in one table

Each robustness statement in the paper is one row here.

D.4 Sensitivity to the flow semi-elasticity ($k = 4$)

Section 3.5 calibrates $k = 8$, at the aggressive end of the evidence band $[4.5, 10.5]$. This section reruns every affected calibration coordinate at $k = 4$, the conservative end. The method is one line: Λ is rematched under the incentive-equivalence rule of Section 3.5: $\Lambda(4) = 733.9$bp against $\Lambda(8) = 776.8$bp, a decline of only 5.5%, because the Jensen corrections on the two sides of the matching partly offset, and every other parameter is unchanged. Multiples of Λ below are multiples of each calibration's own $\Lambda(k)$.

Every qualitative statement survives. The activity threshold exists and is finite; the gap decays as C_1'/Λ with no finite death point; the crossing structure of Proposition 1(b)

Table D.2: Robustness scans, in one table.

Scan	Grid	Headline
Risk aversion, two-peak profile	$\rho \in \{5, 6, 7, 8, 9, 10, 12, 15, 20\}$	both peaks survive to $\rho = 14.0$; the reward peak dies at 14.5; the catch-up zone survives alone to $\rho \geq 20$
Saturation constant	$c \in \{0.5, 1, 2\}$	$E[Q'z] = -4.7{\times}10^{-2}/0/{+}4.7{\times}10^{-2}$; death points absent at $c \leq 1$, postponed to $16.4/18.6\Lambda$ at $c = 2$; two-peak profile and three locked regions robust throughout
Residual noise	$\sigma_0 \in \{0.5\%, 1\%, 2\%\}$	nine-cell threshold table; the $\sigma_0 = 0.5\%$ row degenerates (thresholds coincide to three digits, the contract has no jurisdiction over near-pure indexers); as $\sigma_0 \to 0$, $V_\alpha, V_\sigma \to 0$ and the contract loses all marginal bite near $x = 0$, the boundary of Assumption A1
Flow strength	$0.5\times$–$30\times$ baseline	gap decays 22.0bp $\to$ half at 2.21Λ $\to$ quarter at 4.58Λ $\to$ tenth at 12.0Λ; no finite death point; smooth corner arrival at $10/15/22\Lambda$ ($\mu = 5\%/3\%/2\%$)
Exact risk pricing	γ: 12-point grid, equivalent ρ 0.04–11.3	compression side holds at every γ; amplification side reverses, boundary bracketed in equivalent $\rho \in (2.40, 3.26)$
Redemption sensitivity	$\theta \in \{0, 0.5, 1, 2\}$	bust-fine level falls 266$\to$89bp (-67%); the share falls only to 32–34%; the headline is the share
Capacity dilution	$\kappa \in \{0.1\%, 0.2\%, 0.5\%\}$	$\Delta(\mu)$ single-crossing at $\mu^\dagger = 5.18/5.28/5.58\%$; voluntary floating share 12.0–14.3% $\ll q = 0.6$
Skill-belief dispersion	$\tau \in \{0, 1\%, 2\%, 3\%\}$	activity threshold exactly invariant in τ (first-order term vanishes at $x = 0$); the effect runs through the intensive margin: $x^*(\mu = 3\%)$ falls 17% at $\tau = 3\%$
Penalty jump	d: 60bp $\to$ 30bp	x^* rises at every interior grid point; Lemma A.2's one-sided static over $\sigma_0 \in [0.5\%, 1\%]$

Table D.3: Sensitivity to the flow semi-elasticity ($k = 4$).

Coordinate	$k = 8$ ($\Lambda = 776.8$bp)	$k = 4$ ($\Lambda = 733.9$bp)	Reading
Activity thresholds μ^*_{flat} / μ^*_{float}	1.934% / 2.154%	4.089% / 4.471%	both roughly double
Threshold gap at baseline	22.0bp	38.2bp	wider by 73%
Gap decay milestones (half / quarter / tenth)	2.21 / 4.58 / 12.0 Λ	2.50 / 5.37 / 13.93 Λ	decay slightly slower
Asymptotic constant $C'_1 = -sV_\sigma/E[Q']$	2.0×10^{-4}	4.0×10^{-4}	exactly doubles
Finite death point	none	none	unchanged ($c = 1$ symmetry)
Crossing skill $\tilde{\mu}$	5.1%	8.1%	shifts right by 3pp
Two-peak profile (Prop. 3)	catch-up peak 1.70% at $e_1 \approx -6.25\%$; reward peak 1.20% at +3.25%	identical heights and positions at $k = 4$	unchanged
Two-peak survival bound $\bar{\rho}$	14–14.5	14–14.5	unchanged
Dead-zone boundary	$e_1 \lesssim -6.75\%$	$e_1 \lesssim -6.75\%$	unchanged

is intact; the two-peak profile, the three locked regions, the catch-up peak's dominance over the reward peak, and the dead zone are all driven by the fee schedule's kink geometry and move by less than the grid resolution. The quantitative pattern is exactly the scaling the consistency rules of Section 3.5 anticipate: at $\sigma_0 = 1\%$ the $c = 1$ symmetry gives $E[Q'] \approx k/4$, and since Λ barely moves, the product $\Lambda E[Q']$ halves when k halves, so every threshold-type coordinate (μ^*_{flat}, μ^*_{float}, and the asymptotic constant C'_1) approximately doubles, scaling as $1/k$. The closed-form result set has no boundary in k; what moves is the calibration's level.

Two qualifications. First, the reversal zone in the $\sigma_0 = 2\%$ column of the nine-cell threshold table widens at $k = 4$. At $\sigma_0 = 2\%$ the downgrade line sits nearly in the money ($1.5\sigma_0$), the variance price V_σ is strongly negative, and the threshold ordering rests on $\Lambda E[Q']$ against the small skill price V_α; halving k halves $\Lambda E[Q']$, so the near-tie of the 0.5Λ cell at $k = 8$ (7.126% against 7.124%) widens into a clear reversal (15.01% against 10.92%) and spreads to the 1Λ cell (7.506% against 7.371%). The baseline cell and the entire $\sigma_0 =$ 1% column are robust in both calibrations (the ordering's zero-crossing sits at 0.13Λ and 0.27Λ respectively, far below the 0.5Λ column), and the nine-cell table of Section 3.6 carries a footnote marking this corner. Proposition 1(a)'s qualitative statement is made at the baseline calibration and is unaffected; the reversal itself is economically sensible, since a penalty already in the money leaves the floating contract no extra participation advantage. Second, the calibrated reading of $\tilde{\mu}$ is k-sensitive: existence and the compression-then-amplification structure are unchanged, but the amplification region begins at $\mu > 8.1\%$ rather than $\mu > 5.1\%$: over most of the plausible skill distribution, the floating contract sits in the compression region. The main text flags this at the $\tilde{\mu}$ statement (Section 3.6).

Appendix E Data Appendix

E.1 Sample construction

The treatment group: the first cohort of twenty-six. The sample frame is the first batch of floating-fee active equity products approved under the May 2025 action plan (China Securities Regulatory Commission (中国证券监督管理委员会), 2025). The list was obtained from the Gildata news archive, which returned the full twenty-six-name roster as published by Zhongzheng Jinniu (中证金牛座) (2025); fund codes were then matched name by name through the Eastmoney fund search interface (A-class shares; the codes concentrate in the new 024429–024474 range). The match is complete: 26 of 26 products are identified with code, inception date, management company, manager, and the full text of the prospectus benchmark. The products were launched between 2025-06-06 and 2025-07-23: twenty-four in June 2025, two in July (华商致远回报 07-15, 信澳 07-23).

The control group: same-vintage fixed-fee launches. The control group was drawn as ordinary equity and partial-equity hybrid funds launched on the fixed-fee contract between 2025-06-01 and 2026-05-31, excluding index, ETF, feeder, enhanced-index, and C-class shares, and requiring at least 200 trading days of history before 2026-05-27. The selection funnel: 922 candidates → 428 after type filters → 182 qualifying → **158** with complete data (failures are short windows or missing series). This is a same-vintage design: it differences out the launch window, the market regime, and the new-product channel environment; its cost is that the two groups' managers differ by assignment, not by chance, a difference Section 6.2 measures directly.

Decontamination: 158 → 144. The original control list was contaminated by later-batch floating-fee products, which public fund-classification fields do not distinguish from fixed-fee products. The screen that identifies them is zero-cost and decisive: on the

Eastmoney F10 basic-profile page, the "管理费率" (management fee rate) field displays a concrete number for a fixed-fee fund (e.g., "1.20%（每年）") but cannot display a single number for a three-tier floating fund and shows "详情" (details) instead. The signature was validated on the twenty-six known floating funds (26 of 26 show "详情"). Fourteen of the 158 controls were flagged and then verified contract by contract on the F10 fee-terms pages: all fourteen contain the three-tier clause (holdings past one year; upgrade to 1.5% at excess return above +6%; downgrade to 0.6% at or below −3%, −2% for 华泰柏瑞). All fourteen are confirmed new-type floating-fee funds (14/158 = 8.9%):

Batch	Funds
Second batch (approved 2025-07-24)	025058 中欧核心智选；025068 建信医疗创新；025060 国泰优质核心；025075 汇添富成长优选；025073 南方瑞景；025069 景顺长城高端装备；025065 平安研究驱动；025077 华泰柏瑞制造业主题
Normalized-registration batch (from 2025Q4)	025686 国泰半导体制造精选；025340 鹏华制造升级；025830 嘉实成长共享（approved 2025-10-15）；025758 华安新兴动力；025511 广发信息产业；025824 易方达产业优选

Four further second-batch products (中银品质新兴、摩根慧启成长、东方红医药创新、易方达价值回报) never entered the 158, having failed the window or type filters. An independent corroboration of the screen: the fourteen contaminated funds' mean quarterly net flow rate is −26.7%, essentially the floating group's −27.4% and entirely unlike the controls' +0.8%: the contamination, left in, would have systematically depressed the control group's flow level and diluted every group contrast. All contrasts in Section 6 use the 144 controls.

E.2 Data sources and variables

Daily net asset values. Fund unit NAVs come from Eastmoney's fund disclosure pages (pingzhongdata daily series). No dividend or split event occurred in any of the twenty-six treatment funds over the window (verified on the unitMoney field), so unit NAV is the total-return caliber. Control-group daily NAVs were fetched fund by fund (158 series).

Composite benchmarks, leg by leg. Every treatment fund's prospectus benchmark is composite: a main equity index at 55–80% weight, a Hong Kong index at 5–20%, and a bond index at 15–25%, and each was parsed into its leg weights. Index series come from two public sources: Tencent quotes (CSI 300, CSI 500, CSI 800, CSI A500, Hang

Seng, Hang Seng China Enterprises Index (HSCEI), ChiNext, STAR 50) and Gildata daily index quotes (中证港股通综合 930930, 中证 800 成长 H30355, 恒生综合 HSCI, and 中证全债 H11001 as the bond-leg proxy). Excess returns for the floating group are computed against each fund's own composite benchmark, exactly as the contract computes them; excess returns for the control group are computed against the CSI 300.

Quarterly shares. Subscription, redemption, and end-of-quarter share figures come from the Eastmoney F10 quarterly scale-change pages, for 182 funds over up to four quarters; the IPO quarter (2025Q3) is excluded from flow regressions because initial issuance contaminates the share base, and flow rates beyond ±200% are trimmed. After the exclusion and trimming, 170 funds over three quarters enter the flow regressions of Section 6.6.

Announcements. The announcement archive is the Eastmoney F10 fund-announcement (JJGG) interface, fetched fund by fund for all 184 funds (158 controls plus the 26 treatment funds, the latter used for route cross-validation): 9,142 announcements, 184 of 184 successful. Purchase-limit events are classified by title rules: titles containing "大额申购/限购" and not "恢复", excluding Hong-Kong-connect holiday suspensions and channel notices; the rule was cross-validated against the Gildata announcement archive for the floating group, recovering all three events found there plus four more, and gives both groups aligned event histories (21 economically meaningful limit events).

Holdings and holder structure. Holdings come from the F10 holdings interface at three vintages: the 2025 annual reports (full holdings as of 2025-12-31; 168 of 174 successful, the six failures being late-2025 control launches with no 2025 annual report), the 2026 first-quarter reports (top-10 positions as of 2026-03-31; 174 of 174), and the 2026 interim reports (full holdings as of 2026-06-30; 188 of 188 successful): the 184 funds plus four benchmark ETFs (510300 for CSI 300, 510500 for CSI 500, 515800 for CSI 800, 563360 for CSI A500) whose full holdings proxy index constituent weights (17,890 holding rows in the 2026H1 vintage). 中证 800 成长 has no listed ETF, so the two funds benchmarked to it (024433, 024455) are approximated with the CSI 800 and flagged. Holder structure (institutional versus individual shares, F10 cyrjg) was retrieved for 184 of 184 funds as of 2026-06-30; three floating funds disclose the institutional share as "—" and are treated as missing.

Manager histories. Current managers' complete appointment histories were parsed from the Eastmoney manager pages (3,224 appointment records). Pre-appointment records use the 2023-01-01 to 2024-12-31 window on actively managed equity products the manager already ran (appointment overlap $\geq$ 240 days): 724 fund×manager segments covering 664 legacy products, with each segment's cumulative return measured against the CSI 300 as a uniform yardstick, then averaged by manager. Tenure is first-appointment date to the floating/fixed appointment date.

E.3 Known limitations

1. **Closed-period weekly disclosure.** During the initial subscription (closed) period of about three months, funds disclose NAV weekly. The fact is quantified: within the first 95 days from inception, all 26 treatment funds show a median interval between successive NAV observations of exactly 7 days, with intervals of five days or longer accounting for 55–80% of observations (median 71%); the series turns daily at about three months. Event studies mitigate this with minimum-observation thresholds, but early states are measured with error.

2. **Serial correlation in repeated observations.** The repeated-observation designs (the quarterly flow panel and the tier-switching panel) exhibit within-fund serial correlation. Fund-level clustered standard errors run 1.3–2 times the classical ones; every group contrast in Section 6 is reported under clustered and label-permutation inference.

3. **Benchmark approximations.** Two funds' benchmark legs are approximated by the CSI 800 (no ETF proxy for 中证 800 成长; affects the holdings-overlap metrics); the bond leg is proxied by 中证全债 H11001 throughout (estimated error below ±0.5pp annualized; one fund's 5% demand-deposit leg is treated as zero-return).

4. **The Hong Kong leg is not FX-adjusted.** Prospectus benchmark texts specify valuation-FX adjustment; the construction uses HKD-denominated index returns directly. HKD/CNY moves about ±2–3% a year, so the distortion on a 5–20% leg is below ±0.6pp, a perturbation that touches exactly one fund's tier boundary (024446 at −3.10% annualized excess, 0.1pp from the downgrade line) and no fund's upgrade boundary.

5. **Quarterly flow resolution.** Shares are disclosed quarterly; no daily or monthly subscription/redemption series exists in the public record, which is why the anniversary-redemption hypothesis is testable only at quarterly resolution and the within-quarter timing remains open.

6. **Holder counts not yet extracted.** The cyrjg interface supplies shares but not holder counts, which live in the interim-report PDFs (extraction is the next step); three floating funds' institutional share is missing.

7. **Legacy-fund yardstick.** Pre-appointment records use the CSI 300 as a uniform yardstick rather than each legacy fund's own benchmark (parsing 664 prospectus benchmarks was not cost-feasible); manager names are not disambiguated across companies.

Appendix F Discussion of Limitations

The theory of Sections 3 through 5 is narrower than its subject. We modeled one contract shape within the step family, one company's allocation problem, one regulator's three constraints, and a representative holding window wherever the argument allowed it; welfare was measured at first-round incidence, and the evidence of Section 6 covers one cohort in one market state. These were choices, and each bought something: a readable manager problem, an identifiable sorting logic, an objective the institutional record can name. This appendix states what the narrowing costs. Each limitation below is given in three parts: the boundary itself, what survives inside it, and the question the boundary opens. None of them, in our judgment, touches the paper's main line: that a concave, lot-by-lot repricing of the manager's revenue is a screening device in normal years, fails against hot money, and is succeeded by its own vintage structure. Several of them are where we would start next.

The skill distribution has no negative segment. Section 5.3 stated this limitation at the point where it binds; we expand on it here because it is the revision we would run first. The political constraint aggregates over the funds it targets: products whose delivered alpha sits persistently below the downgrade line. Inside the industry model's population there are none, and the reason is behavioral rather than distributional: a manager who would persistently lose closets instead (Proposition 1(a)), so delivered alpha is $\mu x^* \geq 0$ on the support of G even if the support were extended below zero. The model's slate is optimally behaving new issuance; the products the constraint targets in reality (grandfathered legacy funds, style-drifted products with persistently negative realized records) lie outside it, so the political elasticities of Section 5.3 (0% to 38.8% to 77.7% as d runs from 0 to 60 to 120bp) are priced per fund, on a representative bust fund's curve, with the targeted population's aggregate weight unmodeled. What the omission does not touch is the surfaces' direction: Lemma A.17 of Appendix A signs the constraint surfaces' slopes, and slopes do not depend on the missing mass. The repair is a model of how bust funds come to be, not a wider support: skill that decays ex post, so that a formerly skilled fund drifts into the penalty region, or a screening failure that lets a negative-skill manager stay active. Either repair activates the political constraint's aggregation and turns the representative-fund reading of Section 5.3 into a statement about the industry's shape. It also sharpens a question the current model can only pose: how much of the penalty side's political value is punishment of the persistently unskilled, and how much is insurance against skill that decays?

The informed-capital closure is one of two natural specifications. Section 4.7's smart money is an informed sector that internalizes its dilution and retains the rent κA^S; the alternative, perfectly elastic informed capital bidding the marginal unit's net to zero, is the atomistic limit of the same Berk–Green logic. The choice governs the fate of the company's second instrument. Under the elastic closure, trusting inflow crowds informed capital out one for one at interior funds, total size is pinned by the zero-net condition, and the marketing margin closes exactly where the model uses it; the marketing statements of Sections 4.2 and 4.4 do not survive there, and the fee transfer of Section 4.7 flips sign, the upgrade-tier premium competed away in scale, leaving the noise tax. Under the main text's closure the sorting results are untouched: recomputed under the elastic closure, $\Delta(\mu)$ crosses at the same points and the quota binds by the same margin. We work with the specification in which both of the company's instruments are live.

Three equilibrium margins are deferred, not solved. Section 4.5 names its simplifications at the outset and flags what each costs at the point where it matters; each also opens a line of work. First, the signal feedback. Smart money in Section 4.5 takes the assignment rule as given information and does not invert it; the equilibrium is a partial equilibrium with sorting, not a full signaling equilibrium. Section 4.5 argues the omitted inversion reinforces the endorsement reversal: a market that reads floating appointments as endorsements concentrates on exactly the funds the company sorted to the top, but the argument is one-directional, and a fully closed feedback, in which appointment, inference, and flow determine each other, remains to be solved. Second, competition between companies. Our representative company allocates contracts within its own slate; it does not compete for managers, for trusting money, or for the legacy stars' scarcity rents. Competition should discipline the endorsement margin: a star under-assigned by one company is another company's hire, and how much of Proposition 7's sorting survives it is open. Third, the coexistence dynamics. The reform binds only new issuance, so old and new contracts will share the industry for years, and Section 4.6 priced the legacy stock in steady state. The transition path (a rolling vintage of mandated cohorts entering beneath an aging stock of fixed-fee funds, each cohort's windows maturing on Section 4's schedule) is an overlapping-generations problem we have not written down, and it is the version of the model that speaks to the reform's second and third years rather than to its limit.

The thresholds are parameters, not choices. We take the schedule's geometry ($\ell = 3\%$, $h = 6\%$, $u = 30$bp, $d = 60$bp, $q = 60\%$) as the institutional fact it is. Section 5.3 reads the geometry directionally: the three concerns point toward $d > u$ with $\ell < h$. That is a statement about directions, not about levels. The design problem (choose $(\ell^*, h^*, u^*, d^*, q^*)$ to maximize Section 5.1's objective subject to the three constraints) is the natural next step, and the paper already holds its first-order pieces: the constraint-surface elasticities of Section 5.3 are precisely the marginal rates at which a freer design would trade punishment against gambling against viability. Section 5.4 shows why the problem is not a formality. The quota's first-round arithmetic is thin in total and negative at the margin, so the case for any particular coverage must be carried by values outside that account; a design version of the model would have to say whose values those are. It would also have to decide what the political constraint prices. Perceptible punishment is a constraint on the public's beliefs about the regulator, not on fee revenue, and formalizing it as one rather than the other changes what the optimal thresholds buy. We have not made that choice on the regulator's behalf.

The overconfidence dynamics are open. Section 5.4 reported this extension's result without resolving it: the contract taxes an overconfident manager on the single-window basis and refunds him on the vintage-aggregated basis: -0.39bp against $+0.24$bp at $\delta = 2\%$, with 29.1% and 42.8% of managers pushed past the activity threshold at $\delta = 1\%$ and 2% respectively. The divergence is not a defect to be patched; it is the fingerprint of a missing model. The complete version needs two ingredients we have not built. The first is an attribution-bias dynamic, in which a lucky run inflates the manager's skill belief and the belief decays only as evidence accumulates against it. The second is factor mean reversion, which decides how long the luck can run before the market itself supplies the contradictory evidence. Together they define what one might call the half-life of a fake star (the horizon over which an inflated belief survives its own track record), and the contract's vintage stair interacts with that half-life in both directions: window averaging washes out the penalty exposure that does the taxing, while the biased manager's over-activity delivers real alpha against which the tax must be netted. Which basis governs the contract's incidence is a horizon question of the same kind as Section 5.1's first-round incidence decision, and it deserves the same treatment rather than an arbitrary one. We report the bracket and leave the dynamics to future work.

The outside option is uncalibrated. Section 4.6's go-private analysis defends the V shape of the utility gap: the contract difference is negative but small at the bottom (-0.08bp at $\mu = 0$), reaches its floor of -1.26bp in the mid-high band near $\mu \approx 4.82\%$, and turns positive above roughly $\mu \approx 5.93\%$. That shape is structural: it is built from Proposition 1 and the break-even fee, and it does not depend on how the private-fund outside option is parameterized. What the calibration does not deliver is the exit band's coordinates. Under our illustrative W_{priv} the band is empty: public-fund utility carries a flow-revenue floor of roughly 500bp, and binding the private option requires carry parameters about 7.5–8 times the illustrative values. The empirical calibration of W_{priv} (from observed go-private transitions, private-fund compensation structures, and the skill composition of leavers) is a data exercise this paper does not attempt. It is also the exercise that converts Section 4.6's qualitative pressure into a quantitative prediction about who leaves, and at what skill the industry-viability constraint of Section 5 actually binds.

Two inherited calibrations, recalled in one sentence each. The amplification side of Proposition 1(b) is mean–variance-specific: the fee schedule's own lottery variance is a contract-specific risk channel that the mean–variance form absorbs into ρ, and the exact CARA problem of Appendix B, solved jointly with the logit flow form, reverses the sign; the compression side, which carries the paper's headlines, survives every specification examined. The bust fine of Proposition 5 is headlined in shares because the share is robust (three to four tenths of a bust fund's lifetime fee revenue, 32%–34% under performance-sensitive redemption) while the basis-point level is not, shrinking by up to two-thirds under the same sensitivity. Both statements stand as made in Sections 3.6 and 4; nothing here revises them.

The evidence window is one bull-market year. Section 6's sample is twenty-six funds observed through a strong bull market, and the section's preliminaries state what that costs: tests whose identification needs losing funds are underpowered by construction, the early panel is weekly-sampled through the closed period, and the null results (the catch-up zone above all, with one event in its critical bin) are power statements, not rejections. Proposition 2 makes the window a conditioning event in the theory's own terms: the contract's marginal bite on the manager is weakest exactly in hot markets, so a bull-year null on the portfolio margin is what the model predicts, while the bull-year results on the company margin (the persistent net redemption with its tier signature, the subscription-side flow attenuation, the purchase limits on extreme winners in both contract groups) are the flow side working where the portfolio margin cannot. Section 6.7's dated agenda is therefore the empirical continuation of this paper rather than a separate project: the same tests, run on the second and third settlement windows, in a market state that contains the losing funds the theory is about. The cohort's second settlement window closes in mid-2027. The theory has staked its claims in advance, and the contract itself will generate the data that settle them.

None of these limitations is incidental. Each is the price of keeping one mechanism readable at a time, and each reads naturally as a research program: a distribution with a negative segment, an equilibrium with feedback and competition, a design problem with free thresholds, a belief dynamic with a half-life, a calibrated outside option, and a second settlement window. We list them in the order we would pursue them.