

Crowding in an Artificial World: AI and the Future of Investing

Ludwig Chincarini Conor Falvey Fabio Moneta

Draft: August 31, 2026

ABSTRACT

We give four independently operating AI agents identical data, instructions, objectives, and binding portfolio constraints—including a library of 209 published return signals—and ask each to design long-short equity strategies. The agents converge: their six distinct strategies draw on five signals, signal overlap sits roughly seven standard deviations above a family-matched selection null, and holdings, desired trades, and returns are highly correlated across agents. In a twelve-year locked test, similarity is essentially unchanged and return correlations rise. Holdings similarity is elevated under both selection benchmarks we specify; trade synchronization is elevated only under the benchmark preserving each agent’s own construction, so implementation choices materially affect measured crowding. Gross average returns remain positive for all strategies; net six-factor alphas are significant at 5% for two of six specifications. A counterfactual model shows similarity alone is insufficient for fragility: amplification depends jointly on similarity, capital, leverage—especially leverage utilization—and market depth.

JEL classification: G11, G12, G14, G23, C45, C58.

Keywords: Artificial intelligence, large language models, crowded trades, anomalies, portfolio choice, market efficiency, financial fragility.

We would like to thank ChatGPT 5.6 Sol through codex-cli 0.148.0 (Agent 1), Claude Fable 5 through claude-code 2.1.241 (Agent 2), DeepSeek V4 Pro through reasonix v1.31.3 (Agent 3), and Gemini 3.7 Flash through antigraivty 1.1.19 (Agent 4) for executing our research ideas. We also would like to thank the American Finance Association for making a call of papers entirely directed by artificial intelligence. We are still in the process of checking some of the results from AI, but we created agents to do our empirical work, the theoretical model, and the writing of our paper. Contact: Ludwig Chincarini, CFA, Ph.D., is an Professor of Finance at the University of San Francisco School of Management, Advisor to Wedge Capital and Advisor to Rukfunds, Office MH 117, University of San Francisco, School of Management, 101 Howard Street, Suite 500, San Francisco, CA 94105. Email: chincarini1@hotmail.com or lbchincarini@usfca.edu Phone: 415-422-6992. Conor Falvey is a research analyst at RUK Funds, Inc.; email: conor.falvey22@gmail.com. Fabio Moneta is a Professor of Finance at University of Ottawa, Telfer School of Management, 55 Laurier E, Ottawa, ON K1N 6N5, Canada; email: fmoneta@uottawa.ca.

I. Introduction

Asset managers are beginning to delegate strategy construction to artificial intelligence systems. Those systems are pointed at the same published menu of return predictors, and the underlying models may share training material—a possibility their providers do not document and we do not observe. This raises a question prior to any assessment of performance. If independently operating AI systems, given the same research library, construct the same strategies, then adoption at scale would concentrate capital in common positions without any coordination among the adopters—the classic precondition for crowded trades (Stein (2009)). Observational data cannot cleanly answer whether independent AI systems converge: in field data, similar portfolios may reflect similar tools, common mandates, correlated flows, or direct imitation, and the researcher observes the outcome only after capital has moved and prices have responded.

This paper answers the question with a controlled experiment. We gave four independently operating AI agents—built on ChatGPT 5.6 Sol, Claude Fable 5, DeepSeek V4 Pro, and Gemini 3.7 Flash—the Open Source Asset Pricing library of 209 published cross-sectional signals (Chen and Zimmermann (2022)), CRSP monthly stock data, and a hash-verified protocol fixing objectives and binding portfolio constraints. Each agent, working in an isolated workspace and told nothing about the other agents or the experiment’s purpose, designed two dollar-neutral long-short strategies on data through 2002: one maximizing the gross Sharpe ratio (Strategy A) and one maximizing the Sharpe ratio net of estimated transaction costs (Strategy B). The agents received identical data, instructions, objectives, and binding portfolio constraints, but independently implemented signal eligibility and portfolio construction. Consequently, differences across agents represent the combined effect of the underlying AI model, its research choices, and its implementation. After a limited, logged validation step on 2003–2012 data, the strategies were frozen—cryptographically hashed, with all inputs date-audited—and evaluated over a locked test period, January 2013 through December 2024,

that no agent, no strategy choice, and no selection decision ever touched. The design yields a property that, to our knowledge, no existing study of crowding possesses: strategy similarity is measured before any capital moves, against explicit randomization-based selection nulls, with a genuine out-of-sample period in which persistence can be tested.

The agents converged, most sharply in signal selection. Of the 126 signals eligible for all four agents, the six distinct frozen specifications—two agents froze one specification for both objectives—draw on five, and two of the five appear in every specification. The average Jaccard overlap of the agents’ signal sets is roughly 0.6, about seven standard deviations above a selection null that matches signal-family composition and further above nulls that do not. These are conditional randomization statements about these four agents, this library, and the realized strategy sizes; they are not sampling estimates for the population of AI systems.

The convergence propagates downstream and persists. Signed holdings cosines average near 0.6, roughly 45% of desired monthly trading is common and same-direction across agents, and pairwise strategy return correlations are near 0.8. Between the validation and locked test periods, holdings and trade similarity move by less than 0.02, while return correlations rise. Return correlations are downstream of the same holdings and signals, so we read them as a compact summary of the same convergence rather than as independent confirmation. That the convergence survives out of sample is the expected consequence of durable selection on a fixed shared library; the experiment’s contribution is to measure that persistence cleanly, before any capital moves.

Whether portfolio-level convergence is abnormal depends on the counterfactual. Relative to a benchmark preserving each agent’s actual portfolio-construction architecture, holdings, trades, and liquidity-scaled common trading are unusually similar in both evaluation periods. A standardized-construction benchmark produces substantially greater mechanical similarity, however, leaving holdings overlap as the most robust portfolio-level result. This contrast shows that implementation heterogeneity materially moderates the translation of common signal

choices into common trades. Section IV develops both benchmarks, the different economic questions they answer, their pre-specification status, and the multiple-testing caveats that attach to comparisons across measures, objectives, periods, and benchmarks.

One economically scaled quantity behaves differently. Liquidity-scaled common trading—the agents’ common desired trading per dollar of the traded securities’ volume—falls by roughly two thirds in the locked period. The decline appears to be driven primarily by higher dollar volume rather than lower common desired trading: unscaled common trading is essentially unchanged, and benchmark portfolios decline comparably. The decline is a market-wide statement about scale relative to measured trading activity, not evidence that the agents diverged—and not a measure of price impact, a term we reserve for the theoretical model.

Strategy performance supplies supporting evidence rather than the paper’s principal contribution. Gross average returns remain positive for all eight records, although Sharpe ratios decline and gross factor-adjusted significance does not survive uniformly. Costs matter in substance: the strategies rebalance monthly in several hundred names per leg, and the two objectives produce visibly different portfolios—the net-Sharpe strategies trade less and hold longer. Net of estimated costs, two of the six distinct frozen specifications have locked-period six-factor alphas significant at the 5% level, and one more at the 10% level. The locked results evaluate the persistence of a strategy frozen using information through 2002, not the performance of a contemporary rolling AI research process that periodically retrains on recent data. That stringency makes the observed pattern—positive gross performance, sharp attenuation after costs and factor adjustment—the familiar out-of-sample signature of published-anomaly investing (McLean and Pontiff (2016); Jacobs and Müller (2020)), now reproduced by machine selectors whose approved information set is the published signal library itself.

The experiment documents convergence in this setting; it cannot establish its market consequences. Four agents managing hypothetical capital exert no pressure on prices, and no

realized price impact is measured or claimed anywhere in the paper. A deliberately spare model therefore asks the counterfactual question: under what conditions would persistent similarity of the measured magnitude become economically consequential? Its central lesson is that similarity alone is not sufficient for fragility; amplification depends jointly on similarity, capital scale, leverage—especially leverage utilization—and market depth. The experiment’s measured similarity enters the model as a scenario anchor, not an estimate, and none of the model’s numerical exercises is a prediction.

The paper contributes to three literatures. First, the measurement of crowded trades. Existing work infers crowding from observed positions, flows, and returns: return correlations among anomaly stocks (Lou and Polk (2022)), hedge fund holdings overlap (Brown, Howard, and Lundblad (2022)), institutional trading in anomaly space (Calluzzo, Moneta, and Topaloglu (2019); Chincarini, Lazo-Paz, and Moneta (2026)), construction-induced commonality (Bruno, Chincarini, and Ohara (2018)), futures positions (Kang, Rouwenhorst, and Tang (2021)), currency style betas (Pojarliev and Levich (2011)), and momentum crowding proxies (Barroso, Edelen, and Karehnke (2022)). All of these designs face the attribution problem that motivated our experiment: similarity in observed positions cannot be decomposed into common inputs, common methods, and mutual response. The controlled design removes mutual response by construction, holds the common inputs fixed, and makes signal choice and construction separately observable—while the agent itself enters as a bundled model-harness-session treatment that the design does not further decompose. The benchmark-dependence finding sharpens a theme this literature has already encountered (Bruno, Chincarini, and Ohara (2018)): what one measures as crowding depends on what one holds fixed. Related theory studies trading when arbitrageurs cannot observe how many peers hold the same position (Stein (2009); Kyle, Obizhaeva, and Wang (2025)) and competition among factor investors (DeMiguel, Martín-Utrera, and Uppal (2025)); the experiment supplies a measured value of an input those models parameterize—the similarity of independently constructed strategies.

Second, the paper contributes to the literature on the post-publication behavior of return predictors. Predictor returns decline after publication and published-predictor portfolios become more correlated with one another (McLean and Pontiff (2016)); the decay varies with news, geography, data availability, and timing (Engelberg, McLean, and Pontiff (2018); Jacobs and Müller (2020); Tobek and Hronec (2021); Bowles, Reed, Ringgenberg, and Thornock (2024)); and much of the published cross-section fails to survive higher hurdles and standardized replication (Harvey, Liu, and Zhu (2016); Hou, Xue, and Zhang (2020)). Our agents are, in effect, the next generation of the publication-informed arbitrageurs this literature studies. Given the published record, they concentrate on a thin, overlapping subset of robust signals, earn gross returns that survive twelve years out of sample at compressed magnitudes, and generate directly the correlated positioning that McLean and Pontiff (2016) infer from post-publication return comovement.

Third, the paper connects measured strategy similarity to the theory of amplification and fragility. The model’s stress block inherits its economics from the limits-of-arbitrage and fire-sale traditions (Shleifer and Vishny (1997); Greenwood and Thesmar (2011); Vayanos and Woolley (2013)) and from the observation that arbitrage capital turns alphas into betas (Cho (2020)); its price-impact scaling is disciplined by demand-system and inelastic-markets evidence that flows move prices far more than frictionless theory admits (Koijen and Yogo (2019); Gabaix and Koijen (2022)). What the model adds is specific to our setting: it takes as input a *measured* degree of strategy similarity among independent algorithmic managers—a quantity these theories parameterize but do not observe—and maps that similarity, jointly with adoption, leverage, and depth, to the boundary between shock absorption and shock amplification. The broader backdrop is the economics of data in financial markets (Farboodi and Veldkamp (2023)): when many managers buy intelligence from the same technology, the diversity of strategies that market stability implicitly relies on is no longer guaranteed by the diversity of managers.

The paper also speaks directly to the use of generative AI in financial research. Generative

AI is not merely used to assist with drafting or computation; independently operated AI agents are the experimental subjects. Each agent receives the same data, research objective, and binding constraints but independently selects signals, develops portfolio-construction rules, writes executable code, and evaluates candidate strategies within a controlled search budget. This design allows us to observe whether generative-AI research workflows diversify investment ideas or instead converge when they operate on a common empirical literature. The complete prompts, decision histories, trial logs, frozen specifications, and evaluation procedures provide an auditable record of that workflow.

The remainder of the paper proceeds as follows. Section II describes the experimental design, the data, and the integrity controls. Section III reports the strategies the agents built and their performance. Section IV measures convergence against explicit benchmarks. Section V develops the model, Section VI interprets the joint evidence, and Section VII concludes.

II. Experimental Design, Data, and Strategy Construction

The experiment places four independently operating AI agents in identical research environments and asks each to design two implementable long-short equity strategies from a common library of published anomaly signals. Because the agents received identical raw research inputs and identical eligibility rules, convergence cannot be attributed to differences in what they were given. Everything downstream of those inputs—model architecture, training, harness behavior, and session-level research choices, including each agent’s own implementation of the common eligibility rules—enters as a bundle that the design does not further decompose. And because each agent’s final strategies were frozen—cryptographically hashed, with all inputs date-audited—before any post-2012 data was accessible in any agent’s environment, the 2013–2024 evaluation is a genuine out-of-sample test of both performance and convergence. Sections II.A through II.E describe the data, the agents and their information environment, the objectives and constraints, the three-period design and its integrity controls, and the

randomized benchmarks.

A. Data and Investment Universe

The signal library is the Open Source Asset Pricing (OSAP) library of Chen and Zimmermann (2022), supplied as 209 stock-month signal files together with the library’s documentation of each signal’s authors, publication outlet, economic category, construction, and original-sample boundaries. The library is a natural laboratory for this experiment: it is the common, public menu of published cross-sectional predictors that any quantitatively oriented researcher—human or machine—can access today, so convergence on its elements is precisely the channel through which AI-assisted research could crowd.

Return data are a local export of CRSP monthly returns containing, for each PERMNO-month, the total return, price, shares outstanding, volume, market value, monthly dollar volume, and monthly turnover. Total returns incorporate reported and imputed delisting returns under the pre-specified CRSP cleaning convention, so terminal distributions and delisting outcomes are reflected in strategy returns. Transaction-cost inputs are annual Hasbrouck (2009) effective-cost estimates at the PERMNO-year level and monthly three-factor idiosyncratic volatility used in cost imputation for stocks without an effective-cost estimate (Section II.C describes the cost model). Factor data are the five Fama-French factors (Fama and French (2015)) plus the momentum factor, at the monthly frequency. All input files were extracted on August 18, 2026, and the files supplied to the agents end in December 2012; no observation dated after December 2012 was present in any agent’s environment.

The investable universe and signal-eligibility rules are fixed by the protocol and are not search variables. Stocks must have a formation-date price of at least \$5. A signal is eligible only if it has nonmissing, point-in-time-eligible values for at least 1,000 stocks passing the common investability screens in at least 80% of development-period formation months (a signal whose nonmissing values take only two distinct values is ineligible), and it may be used in a formation month only if at least 1,000 eligible stocks have nonmissing values. The rules

are common, but each agent implemented them in its own code, and the realized eligible sets differ: 126 signals for Agent 1 and 142 for each of the others. The frozen per-agent reports attribute the difference to documented implementation choices—treatment of nonfinite values, formation-month denominator conventions, and whether unresolved signal-timing documentation rendered a signal ineligible—without a signal-by-signal decomposition. The intersection of the four sets contains 126 signals and coincides with Agent 1’s set. Every signal ultimately selected lies in this intersection, within which the Section IV selection comparisons are evaluated. The signal counts reduce to a short flow:

Stage	Signals
Supplied OSAP signal files	209
Eligible for Agent 1	126
Eligible for each of Agents 2–4	142
Intersection eligible for all four agents	126
Distinct signals in the frozen strategies	5

All five selected signals lie in the 126-signal intersection.

Timing conventions eliminate look-ahead at the formation date. A signal value dated month t may be used at the end-of-month- t formation date only if its underlying inputs were publicly available by that date, and each agent was required to verify and document the timing convention of every signal it selected. Portfolios are formed after the month- t close and earn returns during month $t + 1$. The experiment represents a researcher operating today with access to the current published library: a signal may use its historical observations before its academic publication date, provided those observations are constructed from information publicly available at the historical formation date. Publication dates are retained and reported as diagnostics but do not restrict historical eligibility. This choice matches the economic question—how similar are the strategies AI agents would build now, from the library that exists now—rather than the distinct question of real-time discoverability.

B. AI Agents and the Information Environment

The four agents are commercial large language models operating in agentic coding harnesses: ChatGPT 5.6 Sol through codex-cli 0.148.0 (Agent 1), Claude Fable 5 through claude-code 2.1.241 (Agent 2), DeepSeek V4 Pro through reasonix v1.31.3 (Agent 3), and Gemini 3.7 Flash through antigravity 1.1.19 (Agent 4). The labels are the providers’ public model names; the sessions accessed the models through an API proxy under the provider model identifiers `openai/gpt-5.6-sol`, `anthropic/claude-fable-5`, `deepseek/deepseek-v4-pro`, and `gemini/gemini-3.7-flash`, in August 2026. Each agent is a single realized session of a stochastic system, not a deterministic function of the prompt, and the archived run records preserve each session’s complete prompts, decision histories, and trial logs. Reasoning-effort and sampling-temperature settings were not separately recorded; each session ran at its harness’s default configuration, and for frontier models of this generation the defaults generally correspond to the provider’s highest reasoning setting.

Each agent ran in an isolated workspace with no communication channel to any other agent, no shared state, and no access to data outside the approved files of Section II.A. Each received the same protocol document—the SHA-256 hash of the delivered protocol is identical across the four sessions—specifying the data, objectives, constraints, search budget, and validation procedure, and disclosing neither the other agents’ existence nor the experiment’s crowding-measurement purpose. Within its workspace, each agent wrote and executed its own research code in a common Python environment, and every development-period backtest was recorded in an append-only trial log.

The empirical inputs approved for strategy construction were exactly the OSAP, CRSP, cost, and factor files of Section II.A. A read-only corpus of published papers was also present on disk, but it was not an approved strategy input and the protocol did not direct the strategy agents to it; it served the separately prompted theory work described in Section V. Validation evaluations ran only through a supplied evaluator that enforces the round limits of Section II.D and returns a fixed summary package; agents could not access stock-level or monthly

validation data.

Throughout the paper, an “agent” is one model-harness session executing this protocol, and statements that an agent “chose” or “designed” describe the logged outputs of that session. The experiment holds the supplied inputs and rules fixed and varies only the artificial researcher—understood as the bundled model-harness-session treatment, not as a random sample from the population of AI systems.

C. Objectives and Constraints

Each agent designed two strategies: Strategy A maximizes the annualized gross Sharpe ratio over the development period, and Strategy B maximizes the annualized net Sharpe ratio under the prescribed transaction-cost model. Monthly Sharpe is the mean monthly long-short return divided by its standard deviation, annualized by $\sqrt{12}$; because the portfolios are dollar-neutral, no risk-free rate is subtracted. The two strategies face the same universe, screens, rebalancing rule, constraints, and search budget, so the objective—ignoring costs versus internalizing them—is the only intended difference between A and B. The protocol permits A and B to be identical and forbids forced differentiation.

The construction constraints are fixed and common to all agents. A strategy may use at most five signals. Portfolios are long-short and dollar-neutral—long-leg weights sum to +1 and short-leg weights to −1—with no additional leverage or volatility scaling. Positions are capped at 5% of either leg in absolute value, with cap enforcement preserving the ± 1 leg exposures. Each leg must hold at least 100 distinct securities at every formation date; a specification that cannot meet this requirement in a month is infeasible for that month. Rebalancing is monthly, after the close on the final trading day, and is not a search variable. Within these constraints, the agents were free to choose signal transformations, combination rules, missing-signal rules, within-leg weighting, and the cap-redistribution method. These free dimensions matter: they are the degrees of freedom along which independent designers could have diverged, and Section IV measures the extent to which they did not.

Turnover and transaction costs are fixed by the protocol and are not AI-selected methodology. Following footnote 13 of Calluzzo, Moneta, and Topaloglu (2026), the return-drifted weight of stock i immediately before the rebalance at $t + 1$ is

$$\tilde{w}_{i,t+1} = \frac{w_{i,t} (1 + r_{i,t+1})}{1 + \sum_j w_{j,t} r_{j,t+1}},$$

where $w_{i,t}$ is the signed target weight set at the prior rebalance and $r_{i,t+1}$ is the return between the two formation dates. Each stock’s contribution to turnover at the rebalance is $q_{i,t+1} = \frac{1}{2}|w_{i,t+1} - \tilde{w}_{i,t+1}|$, with entering and exiting securities assigned zero weight on the side of the rebalance on which they are absent; portfolio turnover is the sum of $q_{i,t+1}$ across positions, and annualized turnover is 12 times the monthly average. Transaction costs at each rebalance are $TC_{t+1} = \sum_i q_{i,t+1} c_{i,t+1}$, where $c_{i,t+1}$ is the stock’s Hasbrouck effective-cost estimate for the calendar year of the trade, averaging across estimation samples when a PERMNO-year has multiple estimates. When c is missing, it is imputed by the nearest-donor method of Novy-Marx and Velikov (2016): the donor is the stock with observed c in the same calendar year nearest in Euclidean distance on cross-sectionally standardized log market capitalization and idiosyncratic volatility, falling back to size alone when volatility is missing and to the year’s cross-sectional median c when no donor exists. Net return is gross return minus TC . The same cost model applies to both strategies; Strategy A is selected on gross Sharpe but reports net performance, and Strategy B is selected on Sharpe computed from net returns.

D. Development, Validation, and Locked Testing

The sample is split into three fixed periods. The development period, January 1980 through December 2002, is the search sample. The validation period, January 2003 through December 2012, supports a limited amount of model selection under the round limits described below. The locked test period, January 2013 through December 2024, was never accessible to any agent: the raw files in every agent’s environment end at December 2012, and the locked-

period evaluation was performed once, by the evaluator rather than the agents, after all strategies were frozen. Locked-test results are the paper’s principal evidence; validation results influenced strategy selection and are reported separately, never pooled with locked results except in explicitly labeled descriptive statistics.

Search intensity was bounded and fully logged. Every eligible signal was first evaluated once under a common pre-specified standalone methodology; these required screening evaluations do not count against the search budget. Beyond the screen, each agent could evaluate at most 100 additional candidate specifications across both strategies combined, where every distinct specification for which a return series is computed counts as one trial. The four agents logged 100, 59, 88, and 80 counted trials—one used the entire budget, none exceeded it—and no unlogged development-period return calculation may inform selection.

Validation access was rationed to at most two evaluation rounds: one provisional-pair evaluation, followed by at most one validation-informed revision and, if a strategy was revised, one final evaluation. Before observing any validation result, each agent submitted a provisional pair of fully specified strategies and predeclared both a revision rule and the permissible replacement specifications, each already logged in the development search or generated mechanically by the predeclared rule. Each round returned only a fixed summary package per strategy—Sharpe ratios, mean returns, volatility, turnover and costs, maximum drawdown, and six-factor alphas—never stock-level or monthly data. The selection rule was fixed in advance: validation gross Sharpe for Strategy A and net Sharpe for Strategy B, with lower turnover and then lower complexity as tie-breakers, and a revision that validated below its provisional version had to be discarded.

The realized histories show the design operating as intended. Agents 1 and 3 used a single round: their predeclared rules did not trigger. Agent 2’s trigger—revise if the validation objective falls below half its development value—fired for both strategies, and its pre-listed backups became final. Agent 4 evaluated revisions but retained both provisional versions because they validated better. The two-round cap and predeclared rules allowed a bounded

response to validation evidence while foreclosing iterative fitting of the validation sample.

Upon final selection, each agent wrote a freeze manifest recording the SHA-256 hash of every input file read, every specification detail, the timing convention of each signal, the cost specification, the trial identifiers, the validation scores, and the hash of every deliverable. After the freeze, nothing changed. Before the locked data were introduced, an independent pre-unlock verification hashed all 132 empirical data inputs, confirmed that no dated series extends past December 2012, and reproduced all eight frozen portfolios and all validation-period crowding outputs byte-for-byte from the frozen code and inputs. Locked-period access was authorized on August 31, 2026, only after these checks passed, and the locked evaluation applied the frozen specifications, cost model, and measurement code to the 2013–2024 data exactly once. Two data limitations carry through the results: Agent 1’s December 2024 formation date fails the frozen 1,000-stock coverage requirement, and momentum factor data end in October 2024, so locked-period six-factor regressions use 142 months.

E. Benchmark Construction

Observed similarity between two AI strategies is not, by itself, evidence of behavioral convergence: four researchers drawing from the same 126-signal universe, subject to the same screens, caps, and leg requirements, would overlap to some degree by construction. The question is how much. We therefore compare the observed similarity measures (defined in Section IV) with their distributions under two randomized benchmarks in which signal identities are drawn at random but everything else is held fixed. Both benchmarks were fully specified, simulated, and frozen before any locked-period data were accessed.

Both benchmarks share the same signal-randomization design. For each objective, a pseudo-experiment replaces each agent’s selected signals with random draws from the common eligible library, preserving that agent’s number of signals within each economic-family category, so a pseudo-agent has the same family composition as the agent it shadows. Draws are without replacement within a pseudo-agent and independent across pseudo-agents; 1,000

feasible four-agent pseudo-systems are retained per objective. The benchmarks differ in how drawn signals become portfolios.

The standardized benchmark—pre-specified in the experiment’s time-stamped measurement protocol—maps each pseudo-agent’s signals into portfolios by a single common rule: canonical standalone portfolios are retrieved from a centralized library, averaged with equal weights, and renormalized and capped under the experiment’s rules. The agent-matched benchmark instead builds each pseudo-agent’s portfolio with that agent’s actual frozen construction code, so that only signal identities are randomized and construction is held at what the agents actually did. It was added after the standardized results were computed, and we label it *post hoc* throughout; the label refers to its pre-specification status, since its draws, seed, and implementation were likewise frozen before any locked-period access. A reconstruction diagnostic validates the agent-matched engines: feeding each agent’s actual signal set through its engine reproduces the observed holdings to numerical precision.

The two benchmarks answer different economic questions. The standardized benchmark asks whether observed similarity exceeds what common signals *plus common engineering* would produce; the agent-matched benchmark asks the economically closer counterfactual for these four agents—whether it exceeds what the four realized engineering architectures would produce under random signal selection. Neither is universally correct, and the difference between them is itself evidence that implementation choices materially affect measured crowding (Section IV.D). Because construction moves the null distributions materially, we describe each result as *elevated under both benchmark constructions* or *elevated under only one*, always at nominal, unadjusted 5% upper-tail thresholds without multiple-testing correction across measures, objectives, periods, and benchmarks; the description is a protocol-defined label, not a claim of statistical robustness. A third, simpler family of benchmarks—randomizations of signal identities alone, used to assess raw signal-selection overlap—is described with the results in Section IV.B.

III. AI-Designed Portfolios

This section describes what the agents built and how the frozen strategies performed. Because Agents 1 and 3 each froze identical specifications for both objectives, the eight strategy records comprise six distinct strategies produced by four independent agent sessions; the convergence statistics in Section IV are computed across the four agents within each objective, never across the eight records. We regard the performance record as supporting evidence rather than the paper’s principal contribution.

A. Final Specifications and the Development Process

Table I reports the frozen strategy definitions: the signals each agent selected, their orientations, and the construction rules that map composite scores into portfolio weights.

[Insert Table I about here]

The signal choices cluster tightly. All four agents selected cash-based operating profitability (CBOperProf) and the earnings announcement return (AnnouncementReturn) in both strategies, and the remaining choices come from just three signals: the change in net operating assets (dNoa), the change in financial liabilities (DelFINL), and the earnings surprise (EarningsSurprise). Signal descriptions and orientations follow the OSAP documentation supplied to the agents (Chan, Jegadeesh, and Lakonishok (1996); Ball, Gerakos, Linnainmaa, and Nikolaev (2016); Hirshleifer, Hou, Teoh, and Zhang (2004); Richardson, Sloan, Soliman, and Tuna (2005); Foster, Olsen, and Shevlin (1984)). Every position runs in the direction of the originally documented predictability; no agent constructed a contrarian position in any signal.

Construction is where the agents differ. Agent 1 froze a single strategy for both objectives—an equal average of oriented percentile ranks over its three signals, with decile cutoffs and equal weighting within legs—because the same logged candidate had both the highest development gross Sharpe and the highest development net Sharpe under its predeclared mechanical

selection rules. Agent 3 likewise froze one four-signal rank-average specification for both objectives. Agent 2’s two strategies differ in one signal and in one construction element—Strategy B adds a no-trade band aimed at the net objective—and, uniquely, Agent 2 weights the component ranks by each signal’s standalone development-period Sharpe ratio, fixed at freeze and never re-fit. Agent 4 selected the same two signals for both strategies but built them differently, with Strategy B adding a hysteresis buffer. Where A and B specifications differ within an agent, the difference runs in the direction the objectives predict: the net-Sharpe strategies trade less.

The development histories varied in intensity but not in outcome. The agents used 100, 59, 88, and 80 of their 100 counted specification trials, on top of the exempt standalone screen of every eligible signal. Development-period objective values were high—gross Sharpe ratios between 3.4 and 4.0, net-objective values between 2.6 and 2.9—but these are in-sample values produced by directed search and carry that search’s full selection bias. The relevant comparison is what survived: validation-period gross Sharpe ratios for the frozen strategies range from 0.77 to 1.23, roughly a quarter to a third of the development values, and the locked test compressed them further. That ordering is exactly what one expects when strategies are fit hard to one sample and then carried, frozen, into new data.

B. Portfolio Characteristics

The eight strategies are broad, dollar-neutral portfolios with modest concentration. Over the locked test period the average number of names per leg ranges from 182 to 410, and no position ever exceeds 0.79% of a leg in absolute weight, far inside the 5% cap. Annualized one-way turnover ranges from 4.6 to 9.0 times, essentially unchanged from validation, and annualized transaction costs under the Hasbrouck cost model range from 1.7% to 3.4% (Appendix Table D.II reports the strategy-level rows). The net-objective construction devices work as designed: within Agents 2 and 4, the Strategy B variants trade materially less than their Strategy A counterparts.

Factor exposures, estimated from six-factor regressions (the Fama and French (2015) five factors plus momentum), are reported alongside the alphas in Table III, Panels C and D. Three features stand out. Every strategy carries a small negative market beta in every period. Profitability loadings rise from near zero in validation to between 0.63 and 0.91 in the locked test—an estimated-exposure fact; the frozen outputs do not include a decomposition of why the loadings moved, and we do not offer one. And adjusted R-squared values rise from 0.25–0.45 to 0.44–0.55: the strategies were more factor-spanned out of sample, which matters for the alpha results below and for the return-commonality evidence in Section IV.

C. Performance in the Validation and Locked Test Periods

Table II reports gross and net performance by period, and Figure 1 plots cumulative returns with the validation–locked boundary marked and wealth restarted at one at the start of the locked test.

[Insert Table II about here]

[Insert Figure 1 about here]

Every strategy remained profitable, gross and net of costs, in the locked test period—the twelve years of data no agent, no strategy, and no selection decision ever touched. But every strategy’s Sharpe ratio declined: gross Sharpe ratios fell from 0.77–1.23 in validation to 0.39–0.87 in the locked test, and net ratios to 0.11–0.67. The only uniform component of the Sharpe decline was higher volatility: annualized volatility rose for every strategy, while mean returns fell for six records but increased for Agent 4’s two strategies. Maximum drawdowns deepened as well. The ranking was stable: Agent 3, the weakest validation performer, remained the weakest, and Agent 4’s Strategy B, the strongest validation net performer, remained the strongest.

Factor adjustment sharpens the decline. Table III reports alphas from the Fama and French (1993) three-factor model (FF3) and from the six-factor model (FF6, the Fama and

French (2015) five factors plus momentum), with heteroskedasticity- and autocorrelation-robust t-statistics.

[Insert Table III about here]

In the validation period, all eight gross alphas are significant under both models, and all eight net alphas are significant under FF6 (under FF3, which leaves more residual variance, Agent 3’s net alpha is not). In the locked test, FF6 gross alphas fell for every strategy; six of the eight remain significant at the 5% level, with Agent 3’s the exception. Net of costs, the FF6 evidence is weak. Among the six distinct records, only two locked net FF6 alphas are significant at 5%—Agent 2’s Strategy A at 4.4% per year [$t = 2.35$] and Agent 4’s Strategy B at 4.6% [$t = 2.48$]—and Agent 4’s Strategy A, at 2.6% [$t = 1.79$], clears only the 10% threshold. The remaining net FF6 alphas are insignificant, and Agent 3’s is negative. The three-factor model, which leaves the profitability, investment, and momentum exposures that dominate the locked period (Panels C and D) unadjusted, attributes correspondingly more of the locked returns to alpha: locked net FF3 alphas range from 2.3% to 7.3% per year and are significant at 5% for four of the six distinct records. The gap between the FF3 and FF6 alphas shows the sensitivity of estimated alpha to additionally controlling for profitability, investment, and momentum exposures; it is not a claim about which adjustment is correct. (The locked FF6 regressions use 142 of the available months because the momentum factor data end in October 2024; FF3 requires no momentum factor and uses all 144 months, 143 for Agent 2.)

The arithmetic of the decline is descriptive: gross means stayed positive and most gross alphas survived, while higher volatility, heavier factor spanning, and trading costs set against thinner risk-adjusted gross premia did the damage to the net numbers. Whether the positive gross returns reflect continued capture of the targeted premia, factor exposure, or chance is not separately identified by these regressions. One engineering result stands out: the strategies designed for the net objective delivered on the design, with lower turnover, lower costs, and (for Agent 4) the best net outcome in the experiment. For completeness, Appendix

Table D.I reports the combined 2003–2024 sample, which mixes validation with locked data; we treat it as descriptive only.

D. Interpretation and Limitations

The locked test supplies the confirmatory evidence in this paper, but the validation record carries information of its own: the validation decade is itself out of sample relative to the development period, and it already displays the central performance pattern before any locked data enters. The test is also stringent in a specific sense—it evaluates the persistence of strategies frozen using information through 2002, not a contemporary rolling AI research process that retrains on recent data—and the outcome is neither the “AI rediscovers alpha” result nor the “anomalies are gone” result. It is the familiar out-of-sample compression that meta-studies of published anomalies document, reproduced here under controlled conditions in which look-ahead and selective reporting are excluded by construction rather than argued away.

Four limitations bound the reading. The locked test is a single twelve-year window, and the strategies’ profitability and momentum tilts make their performance partly a bet on how those premia fared over 2013–2024. Net results depend on the cost model: the Hasbrouck effective-cost estimates with Novy-Marx and Velikov (2016)-style imputation are a standard but not innocuous choice, and costs at institutional scale would also reflect price impact, which the model excludes. The frozen protocol precludes the sensitivity analyses a discretionary researcher might add; no results beyond the frozen set exist, and we report none. Finally, minor feasibility events truncate some series—Agent 1’s strategies skip one formation month that fails a frozen coverage cutoff, Agent 2’s locked series has 143 monthly observations, and the FF6 regressions use 142 as noted—each disclosed under the frozen attrition rule rather than backfilled, diagnosed after unlock without any change to the frozen specifications, and immaterial to any conclusion above.

The performance record, however, is not the experiment’s main yield. Four agents

working in isolation from a common supplied library converged on five signals across their six distinct strategies, and two of those signals appear in every specification. What that convergence looks like in holdings, in trades, and against benchmarks that hold the common eligible signal universe fixed is the subject of the next section.

IV. Crowding Across AI-Designed Strategies

This section presents the paper’s central evidence. The agents converged sharply on signals and rankings, and the convergence carried into holdings, trades, and returns. Whether the portfolio-level similarity is *unusually* high depends on the counterfactual, and the two benchmarks of Section II.E answer different economic questions—a dependence that is itself a finding about implementation. One economic quantity moves sharply: the common trading’s liquidity-scaled footprint declines by roughly two thirds in the locked period even as similarity persists.

A. Measuring Crowding Among Frozen Strategies

All measures in this section were specified, implemented, and frozen on validation-period data before the locked test data were unlocked; the locked-period values apply the identical frozen code to 2013–2024. The measures operate at four layers, from ideas to order flow.

Signal choice. For each pair of strategies under the same objective, signal overlap is the Jaccard index of their selected signal sets—the ratio of the intersection to the union—computed on signal identities and on OSAP economic-family classifications, within the common eligible universe U of 126 signals defined in Section II.

Security rankings. Because agents may apply different monotonic transformations, ranking similarity is the cross-sectional Spearman correlation, each formation month, between two strategies’ composite scores over the securities with valid scores for both.

Holdings. The principal holdings measure is the signed cosine similarity between two strategies’ full weight vectors,

$$H_{mn,t} = \frac{\sum_i w_{i,m,t} w_{i,n,t}}{\sqrt{\sum_i w_{i,m,t}^2} \sqrt{\sum_i w_{i,n,t}^2}},$$

which ranges from -1 (exact inverse portfolios) to 1 (identical directional portfolios). The signing is intentional: opposite-side positions in the same security are offsetting, not crowded, exposure. Weighted long-leg and short-leg overlaps are reported as supplements.

Desired trades and their liquidity footprint. Crowding relates most directly to simultaneous desired trades. Each strategy’s desired trade in a security at rebalance is the gap between its new target weight and its return-drifted prior weight (the turnover convention of Section II.C). The principal trade measure is same-direction overlap, SD : the common component of two strategies’ desired buys plus the common component of their desired sells, normalized by the smaller strategy’s total absolute desired trade, so that SD lies in $[0, 1]$. Opposite-direction overlap (OD) and the signed trade cosine are supplements.

To capture the aggregate liquidity footprint, the analysis allocates hypothetical capital equally across the four agents, computes for each security the desired buying (and selling) that comes from two or more strategies—the “common” component—divides by the security’s dollar volume in the formation month, and sums across securities. Expressed per dollar of total system capital, this is the liquidity-scaled common trading measure, P . The measure imposes no price-impact relation, so we do not describe it as price pressure; that term is reserved for the model of Section V. The three principal outcomes— H , SD , and P —were designated in advance; all other measures are supplementary.

Two properties of these definitions matter throughout. They measure similarity of positions and desired trades, not price impact: P is a sum of trading-to-dollar-volume ratios under an assumed capital allocation, not an estimate of transaction costs or realized price effects. And “high” has no meaning without a reference distribution, which the two benchmark constructions of Section II.E supply.

B. Convergence in Signals and Rankings

Table IV reports pairwise signal overlap against three signal-selection null distributions.

[Insert Table IV about here]

Signal overlap is unusually high. Across the six agent pairs, the mean Jaccard index is 0.581 for Strategy A and 0.606 for Strategy B; Agents 1 and 2 selected literally identical Strategy B signal sets, and because each selected signal comes from a distinct economic family, family-level overlap equals signal-level overlap for every pair. The principal comparison is the family-matched null, which preserves each strategy’s economic-family composition and asks only whether the agents picked the same signals *within* families: it expects a mean Jaccard of 0.250, and the observed values lie more than seven null standard deviations above it ($p \approx 0.0003$). The two bounding models are further away still. A uniform draw from the 126-signal universe expects 0.014, some 26 standard deviations below the observed values; restricting draws to the top development-period quartile of standalone gross Sharpe ratios—one important selection pressure, though not the agents’ full search—expects 0.056 and still leaves the observed values 12 to 13 standard deviations above. Under every specified randomization model, independent selection rarely generates this much agreement.

These tests establish that the selection is concentrated relative to the specified randomization models; they do not identify why. The agents shared the data, the constraints, the development period, the objectives, and the anomaly literature embedded in the OSAP documentation, and similar evaluation criteria can make a small subset of signals attractive to independent optimizers. We return to the mechanism question in Section V; here the point is the fact of convergence under isolation.

The convergence carries through to rankings. The average pairwise Spearman correlation of composite scores is 0.82 for both objectives, defined for all six pairs in every validation formation month and confined to a narrow band. Four separately designed scoring systems ordered several thousand stocks nearly identically.

C. Holdings and Trades: Shared Destinations, Distinct Paths

Table V reports the holdings- and trade-based measures for both periods.

[Insert Table V about here]

Ranking agreement translates into substantial but not extreme portfolio agreement. The signed holdings cosine averages 0.596 (Strategy A) and 0.571 (Strategy B) in validation, positive in every month, with roughly half of each normalized leg held in common. That holdings similarity sits below score similarity (0.60 versus 0.82) is expected: portfolio rules convert continuous rankings into legs through different cutoffs, weights, and buffers, and Section IV.D shows just how much those choices matter.

Trades tell a more differentiated story. Same-direction desired-trade overlap averages 0.439 (A) and 0.477 (B), and opposite-direction overlap is nearly zero: when the strategies trade the same security in the same month, they almost never trade against each other. The agents' order flow is positively aligned—but far less aligned than their holdings would suggest if they also rebalanced identically. Common holdings describe a shared destination; trade overlap describes coordination in the path used to reach and maintain it. Different breadth, thresholds, and buffer rules stagger entries and exits, so the agents hold many of the same securities while adjusting them at different times and in different amounts.

D. Benchmark-Relative Evidence

Are these levels unusually high? Table VI reports each principal measure against both frozen benchmark distributions in both periods, and Figures 4 and 5 plot the observed values against the simulated densities.

[Insert Table VI about here]

[Insert Figures 4 and 5 about here]

Against the standardized benchmark—1,000 family-matched pseudo-agent groups whose portfolios are built by one common standalone procedure—holdings similarity is elevated: the observed H sits at the 97.6th (A) and 96.4th (B) percentiles in validation and at the 95.6th (A, $p = 0.045$) and 92.8th (B, $p = 0.073$) percentiles in the locked test. Desired-trade overlap and liquidity-scaled common trading, by contrast, fall *below* the standardized null—in validation, below every one of the 1,000 draws for both objectives. Because the pre-specified tests are upper-tail tests for excess similarity, the recorded p-values for those outcomes are 1.000; the defensible statement is descriptive: under a common construction, the observed agents did not exhibit excess trade synchronization.

The agent-matched benchmark reverses that conclusion. When the same kind of randomized signal draws are instead run through each agent’s own frozen construction code—preserving the heterogeneity in cutoffs, weights, buffers, and bands that the standardized procedure suppresses—the null distributions shift sharply downward (null mean H falls from about 0.46 to 0.33, null mean SD from about 0.65 to 0.27). Against these distributions the observed system is unusually similar on *all three* principal measures in *both* periods. In the locked test, observed H and SD exceed essentially all 1,000 draws for both objectives ($p \leq 0.004$), and observed P sits at the 99.7th (A) and 98.4th (B) percentiles ($p = 0.005$ and 0.017).

The locked P p-values are conditional on the 869 of 1,000 benchmark systems for which the measure remains defined in the locked period under the predetermined feasibility rule, with no replacement draws. A post-unlock attrition diagnostic, disclosed as such, compares retained and excluded draws on validation characteristics that predate the failures: the excluded draws are narrow, low-common-trade systems (validation breadth of roughly 200 versus 650 commonly traded securities per month), their signal composition is statistically indistinguishable from the retained draws’, and the retained draws have *higher* validation-period liquidity-scaled common trading. Retention therefore selects toward higher- P draws, which makes the observed upper-tail placements conservative (Appendix C.4).

How much does construction matter? The two benchmark designs differ in one element—how drawn signals become portfolios—and a paired exercise evaluates both constructions on identical signal draws. Imposing the common standardized procedure raises mean H by about 0.13 and mean SD by about 0.38, and the effect dominates the observed values themselves: running the agents’ *actual* selected signals through the standardized construction pushes H from roughly 0.6 to 0.81 and SD from 0.44–0.48 to 0.80 for both objectives. Had the four agents converged on construction as completely as they converged on signals, measured crowding on every dimension would have been far higher.

Summarizing across constructions at nominal, unadjusted 5% upper-tail thresholds: in the locked test period, Strategy A holdings similarity is elevated under both benchmark constructions; every other measure-objective combination is elevated under only one—the agent-matched null, not the standardized null. In validation, holdings similarity is elevated under both constructions for both objectives. These descriptions carry exactly the weight of their definition and no more: the thresholds are unadjusted for the multiplicity of measures, objectives, periods, and benchmark constructions, and the locked-period Strategy A holdings p-value under the standardized benchmark, 0.045, sits close to the cutoff.

The two benchmarks answer different economic questions, and we do not average them. The standardized benchmark—the pre-specified design benchmark—asks whether the observed system is more similar than randomized signal selection *plus common engineering* would produce. The agent-matched benchmark—post hoc in pre-specification status, frozen before locked-sample access—asks whether the observed system is more similar than the four realized engineering architectures would have been under random family-matched signal selection. Because the agents in fact differ in construction, we regard the agent-matched null as the economically closer counterfactual for these four agents; its pre-specification status remains junior, however, so headline inference reports both constructions rather than electing the winner. The difference between them is not a nuisance; it is one of the paper’s findings. Measured crowding depends jointly on what strategies select and on how selections become

portfolios, and crowding measures that implicitly standardize construction will understate the dampening effect of implementation heterogeneity.

E. Liquidity-Scaled Common Trading

The liquidity-scaled common trading measure aggregates the system’s common order flow and scales it by the contemporaneous dollar volume of the securities being traded. Dollar volume is a liquidity proxy, not structural market depth: the scaling says nothing about the price movement a given order would cause, and the empirical denominator should not be read as an estimate of the depth parameters in Section V’s model. In the validation period, P averages 8.9×10^{-6} (A) and 7.1×10^{-6} (B) per dollar of system capital monthly. These figures are sums of common-trade-to-dollar-volume ratios under a hypothetical equal allocation—not transaction costs, expected price changes, or estimates of realized impact—and we use them only comparatively, across benchmarks, objectives, and periods. (Appendix Table D.II reports the liquidity context alongside the turnover and breadth figures of Section III.B.)

The comparative facts are two. First, as Section IV.D reported, the observed value is below the standardized null but far above the agent-matched null in both periods. Second, observed P per dollar *fell* by 62.9% (A) and 64.2% (B) from validation to locked test—from 8.9×10^{-6} to 3.3×10^{-6} and from 7.1×10^{-6} to 2.5×10^{-6} —while the unscaled common trade per dollar of capital was essentially flat. The period comparison means little on its own: the benchmark distributions declined comparably (Section IV.F), so the decline is a market-wide feature of the measure in this window, not an agent-specific one.

The decline is driven primarily by higher dollar volume rather than lower common desired trading. Because P is a sum of security-level ratios, composition effects could in principle matter, so we verify the reading with a post-unlock decomposition, disclosed as such; it re-describes frozen quantities and alters no frozen number. Holding the common desired trades at their validation realizations and repricing each stock-month’s dollar volume at the same quantile of the locked-period distribution reproduces most of the observed decline

(−54.9 of the observed −62.9 percentage points for A; −60.9 of −64.2 for B); the converse counterfactual—locked trades at validation liquidity—yields only −11.8 and −10.1. The growth is within-security, not compositional: among securities commonly traded in both periods, median dollar-volume growth is $1.8\times$ in both objectives.

A companion breadth diagnostic, likewise post-unlock, shows the decline is not a narrowing of the common-trade set: the number of commonly traded securities falls only about 14% (755 to 647 per month for A; 700 to 602 for B), while per-name intensity falls 56% to 58% at the mean, median, and 95th percentile alike. Persistent strategic convergence coexisted with a shrinking dollar-volume-scaled footprint because measured trading activity in the traded names grew faster than the strategies’ essentially unchanged common demands. Whether the market’s capacity to absorb those trades without price movement grew correspondingly is a statement about depth and price-impact elasticities that the experiment does not measure.

F. Persistence Across Periods

Did the crowding measured in validation persist into twelve years of untouched data? Table V, Table VII, and Figures 2, 3, and 6 compare the two periods. (Crowding measures exist only for the validation and locked test periods—the measurement system was defined after development ended.)

[Insert Table VII about here]

[Insert Figures 2, 3, and 6 about here]

The similarity structure barely moved. Holdings cosine went from 0.596 to 0.588 (A) and from 0.571 to 0.561 (B); same-direction trade overlap and every supplementary measure are within 0.02 of their validation values; and the monthly series show no trend or break at the boundary. Return commonality, if anything, strengthened: mean pairwise gross return correlation rose from about 0.80 to 0.87, and six-factor residual correlations rose above 0.81—comovement not explained by the FF6 specification, though we do not claim the

remainder is idiosyncratic or uniquely AI-related. Return correlations are downstream of the same holdings and signals, so they summarize the same convergence rather than confirm it independently.

The one large movement is the decline in liquidity-scaled common trading of Section IV.E, visible month by month in Figures 2 and 3. The benchmark distributions moved with it: under the standardized null the locked observed value for Strategy A sits at the 20.8th percentile—no longer below the entire distribution, as in validation—because dollar-volume growth compressed the null too, while under the agent-matched null locked P remains unusually high for both objectives. This is why every period comparison of P in the paper is paired with the corresponding benchmark movement: the level declined market-wide, and only the benchmark-relative position is informative about the agents.

In sum, what the agents *are*—their signal sets, rankings, holdings structure, and mutual return dependence—carried into the locked test essentially unchanged, while their common demands became smaller relative to measured trading activity—a fact about a liquidity proxy, not a demonstration that potential market stress fell. Because the validation period is itself out of sample relative to development, the similarity evidence rests on two sequential out-of-sample windows, the first influencing model selection and the second untouched.

G. What the Evidence Does and Does Not Establish

The results support a three-tier reading, in declining order of evidentiary strength. *Strategy similarity*—in signals, rankings, holdings, trades, and returns—is established directly and persists out of sample; this is the paper’s empirical contribution. *Liquidity-scaled common trading* is characterized only under explicit hypothetical allocations and only relative to a liquidity proxy; its decline says the hypothetical common demands fell relative to that proxy, not relative to structural market depth. *Realized price impact* is not established: nothing in this experiment measures whether the agents’ commonality moved prices, and no result here should be quoted as evidence that AI-induced crowding has affected market prices. Section

V provides the framework for when the first tier would generate the third; the experiment identifies the first tier only.

Several measurement limitations bound the conclusions. The randomization inference is conditional on these four agents, this signal library, the specified null-selection mechanisms, and the realized strategy sizes and family compositions—strong evidence that the observed selections are unusual under those nulls, not sampling-based inference to a population of AI agents. All agents drew on a single supplied library, so convergence is measured within, not across, research libraries—though that is also the design’s point, since a common library is precisely the environment practitioner adoption would create. The agents’ realized eligible sets are not identical (126 signals for Agent 1, 142 for the others; the universe U is their intersection), and the signal-selection nulls draw from the intersection rather than each agent’s own set—a choice whose smaller common universe makes accidental overlap more likely and therefore works against finding unusually high overlap. The locked-period P p-values are conditional on benchmark attrition, in the conservative direction documented above. Finally, two measurement-layer events are preserved from the audit trail: one boundary-month holdings file was mechanically reconstructed from manifest-hashed code, and the centralized standalone signal library reconciled exactly to Agent 1 but showed documented low-level implementation differences for Agents 2–4. The latter is itself a substantive lesson the benchmark analysis quantifies: apparently identical strategy descriptions become different portfolios through low-level conventions, and that heterogeneity is a material determinant of measured crowding.

V. A Model of AI Adoption, Strategy Similarity, and Market Fragility

This section develops a deliberately transparent three-date equilibrium model in which AI adoption (a capital share x) and strategy similarity (a portfolio-geometry primitive η) enter separately. The model’s central lesson, and its contribution, is Proposition 3: similarity alone is not sufficient for fragility; amplification depends jointly on similarity, capital scale,

leverage—especially leverage utilization—and market depth. Under stress, the same similar, levered capital absorbs or amplifies a liquidation depending on its proximity to leverage constraints, with severity governed by a single feedback-intensity statistic $g = \lambda(x) \eta L^2 x A$.

Two secondary results, derived under stronger maintained assumptions, complete the picture and are subordinate to the stress result. Adoption of similar strategies displaces prices and decays alpha, with a two-sided effect on pricing accuracy: moderate adoption of informative similar strategies improves it, and deterioration sets in only past an interior threshold (Proposition 1). And adoption inflows temporarily inflate measured strategy performance even as expected future returns fall (Proposition 2). The model is a counterfactual laboratory disciplined by—but not estimated from—the empirical similarity measures of Section IV; all numerical exercises below are transparent scenario analyses, not calibrated predictions of crash probabilities or market instability.

A. Empirical Motivation, Environment, and Timing

Three findings from Section IV motivate the model’s two blocks. Holdings similarity is roughly 0.6 and stable out of sample. Desired trades are synchronized in a benchmark-dependent sense—well above the agent-matched null, below every standardized-construction draw. And liquidity-scaled common trading fell by roughly two thirds in the locked period. Two qualifications are maintained throughout: trade synchronization and liquidity-scaled common trading are elevated under only one of the two benchmark constructions, so both benchmarks inform the model’s scenario grids rather than any single point estimate; and the experiment measures portfolio and trade similarity, not causal price impact, so no result below is a causal claim about the experimental returns.

These findings shape the model in a specific way. What persists out of sample is the *geometry* of the agents’ portfolios—where they hold and where they want to trade—while the scale of their common trading relative to the liquidity proxy declined. The model’s role is therefore counterfactual: it asks under what combinations of adoption, leverage, and depth a

measured level of portfolio similarity would become economically consequential, not whether the experimental system exerted pressure. It distinguishes two channels: **routine flow crowding**—common rebalancing flow in normal times—and **latent stress exposure**—the capacity for synchronized liquidation, governed by holdings similarity together with capital scale, leverage, liquidity, and proximity to leverage constraints. Lower trade similarity relative to holdings similarity does not imply an absence of routine crowding, and a common shock can synchronize liquidation beyond its routine level. Routine common flow scales with $\tau\eta$ at synchronization $\tau < 1$; the formal stress model fixes $\tau_s = 1$ —forced liquidation compresses sales into the impact horizon—and imperfect stress synchronization is discussed qualitatively below.

Two scope statements sharpen what the formal model does with the empirical measures. First, the model formalizes the *adoption* margin (date-0 entry demand) and the *stress* margin (date-1 forced liquidation); recurring rebalancing flow in normal times is not separately modeled, and the paper offers no formal bridge from measured trade overlap to routine price pressure—that would require a model of rebalancing frequency against horizon-specific depth, which we deliberately do not build. Second, throughout the section the holdings cosine H serves as a *scenario counterpart* for the model’s similarity parameter η —anchoring the scenario grid $\{0.33, 0.58, 0.81\}$, which brackets the agent-matched null, the observed value, and the standardized-construction value—not as an estimate of it.

Assets and dates. There are three dates. At date 0, capital is allocated and prices are set; at date 1, a liquidation catalyst may arrive; at date 2, payoffs realize. The traded object is a single crowded long–short *composite* s_c —the common component of AI portfolios—normalized so that one unit is one dollar of per-side gross notional at inception. The composite is dollar-neutral, so its price p is a relative-price premium per unit, a pure number. The composite pays $v = \mu + \epsilon$ at date 2, where μ is the true alpha and ϵ is payoff noise with variance σ^2 . The remainder of each manager’s portfolio—its “sleeve”—is the manager-specific part of the book: positions unique to that manager that the other AI systems do not hold.

For example, if two agents both hold a profitability-and-momentum core but one adds an idiosyncratic tilt toward a distinct set of value names, the shared core belongs to s_c and the idiosyncratic tilt is that agent’s sleeve. Sleeve positions sit in perfectly deep side assets and do not affect the composite price.

Managers and AI systems. A mass A of active-manager equity is divided between AI-adopting managers (share $x \in [0, 1]$) and non-AI active managers (share $1 - x$). AI manager j , with equity E_j ($\sum_j E_j = xA$), delegates portfolio construction to one of M AI systems. System m delivers the portfolio

$$s_m = \sqrt{\eta} s_c + \sqrt{1 - \eta} e_m, \quad \eta \in [0, 1], \quad (\text{V.1})$$

where e_m is a system-specific sleeve orthogonal to s_c . The parameter η is the common-direction share of each AI book’s risk budget; under equal risk budgets and orthogonal sleeves, the expected pairwise cosine of AI portfolio-weight vectors equals η exactly, making the holdings measure H its natural empirical counterpart. A simple numerical example illustrates the geometry: at $\eta = 0.64$, each AI book loads $\sqrt{0.64} = 0.8$ on the common composite and $\sqrt{0.36} = 0.6$ on its own sleeve, so any two such books—whose sleeves are orthogonal to each other—have cosine similarity $0.8 \times 0.8 = 0.64 = \eta$, close to the holdings cosines near 0.6 measured in Section IV.

Information. System m perceives composite alpha $\mu + z_c + \zeta_m$: a *common* error $z_c \sim N(0, \sigma_c^2)$ (shared data, objectives, and training) and a *system-specific* error $\zeta_m \sim N(0, \sigma_\zeta^2)$, defined as the composite system-specific error absorbing the perceived alphas of the fairly priced sleeve directions exactly, so that the manager’s perceived alpha on the delivered portfolio (V.1) at composite price p is exactly $\hat{\alpha}_m = \sqrt{\eta}(\mu + z_c + \zeta_m - p)$. The error correlation across systems is $\rho_e = \sigma_c^2/(\sigma_c^2 + \sigma_\zeta^2)$: η is *where* AI capital pushes; ρ_e is *whether* the push is information or correlated noise. The two are separate primitives. Managers know (x, η) .

Residual sector. Prices are set against a reduced-form residual demand schedule

with slope (depth) k_R : dispersed non-AI active managers and market makers absorb the AI sector’s net demand at price concession $1/k_R$ per dollar. Residual suppliers do not observe μ and do not learn from prices—a maintained friction that makes the price-informativeness results below conditional benchmarks rather than full rational-expectations outcomes. Passive investors do not respond to mispricing absent flows and enter only as a wedge already reflected in k_R .

Assumption 1 (Capital-reallocation counterfactual). $k_R(x) = k_0 + (1 - x) a$, with $a = A/\theta$ and $k_0 \geq 0$ the depth supplied by market makers and other non-active-manager liquidity.

Assumption 1 says AI adoption *reallocates* capital away from dispersed non-AI managers who previously supplied residual liquidity: raising x both increases common AI demand and removes depth. This is a maintained counterfactual about the form adoption takes, not an empirical fact—adoption funded by new capital, or by capital that supplied no liquidity, leaves k_R unchanged. We therefore report the *fixed-depth* case $k_R = k_0 + a$ alongside every reallocation-path result (dashed lines in Figure 7), so the direct correlated-demand effect can be distinguished from the additional liquidity-depletion effect.

For reference, the model’s market and adoption primitives in words: x is the share of active-manager capital delegated to AI systems; A is total active-manager equity in dollars; $\eta \in [0, 1]$ is the share of each AI portfolio’s risk budget in the common composite—the model’s similarity parameter; $L \geq 1$ is the leverage cap on gross notional relative to equity; γ is risk aversion and σ^2 payoff variance, entering jointly through $\theta = \gamma\sigma^2$; and k_R is residual depth—dollars of net demand the residual sector absorbs per unit of price concession—with $\lambda = 1/k_R$ its inverse, the price impact per dollar.

The stress and information primitives are fewer: ϕ is the fraction of equity redeemed in the stress scenario; u_j is manager j ’s leverage utilization, the ratio of its desired gross notional to the maximum the cap permits; ρ_e is the correlation of forecast errors across AI systems; and τ is the fraction of common desired trading that lands within the impact

horizon. Appendix A tabulates the full notation.

B. The Manager's Problem

Manager j chooses a composite book b (dollars) to solve

$$\max_b b \hat{\alpha}_m - \frac{\Gamma_j}{2} \sigma^2 b^2 \quad \text{s.t.} \quad |b| \leq LE_j, \quad (\text{V.2})$$

a mean–variance objective over dollar profit and loss with a state-dependent absolute risk-aversion coefficient and a regulatory or contractual leverage cap $L \geq 1$ on gross notional relative to equity.

Assumption 2 (A-RC, risk-capital scale). $\Gamma_j = \gamma/(LE_j)$: absolute risk aversion is inversely proportional to permitted risk capital LE_j , with γ a pure number.

Under A-RC the first-order condition gives the interior book $b_j^* = LE_j u_j$ with dimensionless *leverage utilization* $u_j = \hat{\alpha}_m/(\gamma\sigma^2) \equiv \hat{\alpha}_m/\theta$, and the cap binds if and only if $u_j \geq 1$. A-RC is the single place the leverage limit enters *unconstrained* demand: risk tolerance scales with permitted risk capital, so a higher cap raises desired positions even away from the cap. Section V.D reports the sensitivity of every result to the alternative $\Gamma_j = \gamma/E_j$, under which leverage affects only constrained outcomes.

Two lemmas, proved in Appendix A, carry the model's mechanics. The first aggregates demand.

Lemma 1 (Demand aggregation). *Under A-RC, aggregate unconstrained AI demand for the composite at price p is $E_{AI}(p) = \Phi(\mu + z_c + \bar{\zeta} - p)$ with $\Phi = \eta x LA/\theta$ and $\bar{\zeta} = \sum_m w_m \zeta_m$, exactly.*

The single factor η (not $\sqrt{\eta}$) reflects two projections: the composite's loading on s_c and the conversion of composite positions into s_c exposure. If managers could trade s_c directly, η would not appear: the “system outputs one portfolio, the manager scales it” restriction is the key institutional assumption.

The second lemma characterizes each manager's behavior under stress. At date 1, with

probability π a catalyst arrives: investors redeem the fraction ϕ of every AI manager's equity at pre-decline marks (proportional redemption), and redemption-driven supply displaces the price by $\delta = p_0 - p_1 \geq 0$.

The intuition is the tension at the heart of every fire-sale model. A price decline creates two opposing forces: it makes the crowded strategy look cheaper, encouraging unconstrained managers to buy, but it also erodes the equity supporting levered positions, so sufficiently constrained managers must instead sell. The lemma makes this tension precise at the manager level.

Lemma 2 (The manager's stress program). *Under the proportional redemption and a subsequent decline δ : (i) $E_j^+ = (1 - \phi)E_j$, $b_j^+ = (1 - \phi)b_j$, and $u_j^+ = u_{j0}$; (ii) on the unconstrained branch, $u_j(\delta) = u_{j0} + \sqrt{\eta}\delta/\theta$ and $E_j(\delta) = (1 - \phi)E_j \exp\{-\sqrt{\eta}L(u_{j0}\delta + \sqrt{\eta}\delta^2/(2\theta))\}$, so the book $b_j = LE_j(\delta)u_j(\delta)$ satisfies $db_j/d\delta = \sqrt{\eta}LE_j(\delta)[1/\theta - Lu_j(\delta)^2]$: the manager buys if and only if $u_j(\delta) < u_b = 1/\sqrt{\theta L}$; (iii) manager j 's cap binds if and only if $\delta \geq \bar{\delta}_j = (1 - u_{j0})\theta/\sqrt{\eta}$, with the book continuous at $\bar{\delta}_j$; under A-MOD, $\bar{\delta}_j$ is simultaneously j 's buy/sell boundary and constraint boundary; without A-MOD the two split and an unconstrained-selling region appears; (iv) once at the cap, $E_j(\delta) = E_j e^{-\sqrt{\eta}L\delta}$ and the composite-direction sales rate per unit decline is $\eta L^2 E_j(\delta)$.*

Which force wins depends on a single state variable, the manager's leverage utilization. Proportionality of the redemption eliminates selection: surviving book composition is identical to date 0, $\bar{\zeta}$ carries through exactly, and the stress results below hold realization by realization.

C. Decentralized Equilibrium: Displacement, Alpha Decay, and Price Informativeness

Market clearing $E_{AI}(p_0) = k_R(x)p_0$ yields the unique date-0 price

$$p_0 = \frac{\Phi(\mu + z_c + \bar{\zeta})}{k_R(x) + \Phi}, \quad \mathbb{E}[v - p_0 \mid \mu] = \frac{\mu k_R}{k_R + \Phi}, \quad (\text{V.3})$$

the second expression (at $z_c = \bar{\zeta} = 0$) being the expected excess return—alpha decays as

similar capital crowds in. A simple benchmark illustrates the magnitude: when AI demand capacity equals residual depth ($\Phi = k_R$), half of the true alpha is already impounded into the date-0 price, and the expected excess return to a new dollar of AI capital is half what the strategy's backtest would suggest. Unconditional mispricing of the composite is

$$\text{MSE} = \frac{\Phi^2 S + k_R^2 \sigma_\mu^2}{(k_R + \Phi)^2}, \quad S \equiv \sigma_c^2 + \sigma_\zeta^2 \text{HHI}, \quad (\text{V.4})$$

where $\text{HHI} = \sum_m w_m^2$ is the concentration of AI-system usage: with few dominant systems, system-specific errors do not diversify, so vendor concentration matters even at moderate η .

Proposition 1 (Entry displacement and price informativeness). *(a) p_0 is increasing in x , η , and L (given $\mu + z_c + \bar{\zeta} > 0$), holding the other primitives and k_0 fixed. Under Assumption 1 the x -effect combines rising common demand and falling residual depth and strictly exceeds the fixed-depth benchmark. $dp_0/dA > 0$ if and only if $k_0 > 0$; at $k_0 = 0$, p_0 is invariant to A (scaling all active capital raises AI demand and residual absorption proportionally). (b) Along the adoption path (x moving Φ and k_R together),*

$$\frac{d\text{MSE}}{dx} = a\left(\eta L + \frac{\Phi}{k_R}\right) \frac{\partial \text{MSE}}{\partial \Phi} \Big|_{k_R}, \quad \text{sign} \frac{d\text{MSE}}{dx} = \text{sign}(\Phi S - k_R \sigma_\mu^2).$$

Adoption is neutral at $\eta = 0$; it monotonically improves pricing accuracy when correlated error vanishes ($S \rightarrow 0$). Otherwise, because $\Phi(0) = 0$, MSE always falls at small x , and the marginal term $\Phi S - k_R \sigma_\mu^2$ is strictly increasing along the path, so the derivative changes sign at most once on $x \in [0, 1]$: if $\Phi(1)S > k_R(1)\sigma_\mu^2$, MSE is U-shaped in x with a unique interior optimum x^ solving $\Phi(x^*)S = k_R(x^*)\sigma_\mu^2$; otherwise pricing error decreases throughout the feasible adoption range. When interior optima exist under both depth paths, the reallocation-path optimum lies strictly below the fixed-depth optimum.*

Proposition 1(b) is the model's price-informativeness result, and it is deliberately two-sided: moderate adoption of informative similar strategies *improves* pricing accuracy (lowers

MSE); the deterioration sets in only past x^* when that optimum is interior, and sooner along the reallocation path than under fixed depth. The economics is a race between two components of the same common demand. AI demand loads on $\mu + z_c + \bar{\zeta}$: the informative part pulls the price toward fundamentals, while the correlated-error part pushes it away, and the error push is impounded rather than diversified because it is common to the adopting capital. At low adoption the informative pull dominates and mispricing falls; once correlated forecast error becomes large relative to residual risk-bearing capacity ($\Phi S > k_R \sigma_\mu^2$), each further unit of common demand adds more error than information and mispricing rises. Under the capital-reallocation counterfactual, adoption also removes the residual depth that damps both components, so the crossover arrives at a lower adoption share. Both claims are conditional on the reduced-form residual sector and the maintained absence of price-based learning; they are benchmarks, not welfare statements.

The second equilibrium result concerns what an observer of returns would see while adoption is under way.

Proposition 2 (Inflow appreciation). *Along an adoption path with x increasing between dates (no catalyst), conditional on positive perceived composite alpha ($\mu + z_c + \bar{\zeta} > 0$, so that $p_0 > 0$ and entering demand is buying pressure), incumbents' realized composite return contains the price-pressure term $\Delta p_0 = p_0(\Phi', k'_R) - p_0(\Phi, k_R) > 0$. Adoption inflows thus temporarily inflate measured strategy performance—they add an endogenous price-pressure component to realized returns that an observer may attribute to strategy skill—while expected future returns (V.3) simultaneously fall, and the date-2 reversion of the price premium to fundamentals grows one-for-one with the accumulated displacement p_0 . Continued appreciation requires continued adoption growth. (Larger liquidation losses in a date-1 stress episode are a separate conclusion: they require the leverage-constraint mechanism and assumptions of Proposition 3, not this result.)*

Proposition 2 depends materially on the maintained assumption that price pressure persists over the relevant entry and rebalancing horizon: residual capital is slow-moving and

does not condition on prices, so displacement is not arbitrated away within the window. In the opposite limiting case of rapidly reversing (temporary) impact, the appreciation component decays within the measurement horizon and Proposition 2 weakens or disappears at, say, monthly horizons; the liquidation-feedback results of Section V.D survive in that limit, because forced flows interact with *contemporaneous* depth rather than with accumulated displacement. This is a statement about the theoretical counterfactual; the locked backtest cannot exhibit it and is not claimed to.

D. Stress: Absorption, Amplification, and Comparative Statics

The common proportional redemption is an *assumed* AI-sector stress shock, not a derived object, and it plays a specific role in the division of labor between the model’s primitives: similarity η determines *where* affected managers sell—their liquidations concentrate on the same crowded composite—while the common catalyst determines *synchronization*—that they sell at the same time. The baseline is deliberately adverse on the synchronization margin, and the natural dilutions can be described without solving a new model: partial participation shrinks the equity subject to forced liquidation; heterogeneous redemption fractions do the same on average but break exact selection-neutrality; and imperfect stress synchronization $\tau_s < 1$ —sales spilling beyond the impact horizon—weakens both the direct liquidation pressure and the induced-sales feedback. Each dilution weakens amplification relative to the baseline without changing the mechanism; we record them as scope qualifications rather than as solved variants.

Assumption 3 (A-MOD, moderate risk bearing). $\gamma\sigma^2L \leq 1$, equivalently $u_b \geq 1$.

A-MOD is an assumed sufficient parameter restriction (what the program derives is the buy condition $u < u_b$); under it, the buy region covers the entire unconstrained region, so slack survivors buy dips throughout. When it fails, an intermediate region appears in which unconstrained survivors sell.

Managers enter date 1 with heterogeneous utilizations u_{j0} , and manager j ’s cap binds

if and only if $\delta \geq \bar{\delta}_j \equiv (1 - u_{j0})\theta/\sqrt{\eta}$: there is no single aggregate regime boundary, and constraint entry is *staggered*, beginning at $\min_j \bar{\delta}_j$ with managers hitting their caps in decreasing order of u_{j0} . The stress analysis therefore distinguishes (i) the *unconstrained region* ($\delta < \min_j \bar{\delta}_j$), in which surviving managers may absorb sales; (ii) the *all-at-cap benchmark* ($u_{j0} = 1$ for all j), an adverse but analytically transparent stress benchmark—not a description of AI managers in normal conditions—for which the amplification result is analytically characterized; and (iii) the *heterogeneous intermediate region*, in which constraints bind progressively and the model provides manager-level comparative statics rather than a completely solved aggregate equilibrium. Fragility accordingly depends on the *distribution* of leverage utilization—not only on similarity and average leverage.

How strong is the resulting feedback? In plain terms, it strengthens with each of four ingredients: when price impact is larger (markets are shallower), when strategies are more similar, when more capital follows them, and when that capital is more levered. The model collapses the four into one statistic. Once a manager is at the cap, its induced sales per unit of further decline are ηL^2 per dollar of equity (Lemma 2(iv)): one $\sqrt{\eta}$ projects the price decline into the manager’s loss, L is the cap slope, the second $\sqrt{\eta}$ projects book sales back onto the composite, and LE is the book scale. Aggregating over surviving equity $(1 - \phi)xA$ and multiplying by inverse depth $\lambda(x) = 1/k_R(x)$ defines

$$g = \lambda(x) \eta L^2 xA, \quad \kappa(\phi) = g(1 - \phi), \quad (\text{V.5})$$

the pre-catalyst *feedback-intensity statistic* governing the severity of liquidation amplification, and the surviving-sector gain after a catalyst ϕ . The two powers of L in g and the single power in Φ measure different objects (cap dynamics versus date-0 demand); there is no double count.

The primary economic benchmark is the *exact* nonlinear date-1 equilibrium. In the all-at-cap benchmark, writing $y = \delta/p_0$ and using the clearing identities $k_R p_0 = \sqrt{\eta} L x A u_0$ and $g = \sqrt{\eta} L p_0 / u_0$, exact constrained clearing is

$$1 - y = (1 - \phi) e^{-gy}. \quad (\text{V.6})$$

Proposition 3 (Liquidation: absorption and amplification). *Under the catalyst (any realization with $p_0 > 0$):*

(a) *Absorption (region (i), under A-MOD). If all survivors are below their caps throughout ($\delta < \min_j \bar{\delta}_j$), each buys the dip, and the first-order (linear-impact) decline, obtained by aggregating the exact manager-level demand paths and then linearizing, is*

$$\delta_u = \frac{\phi p_0 k_R}{k_R + (1 - \phi) \Phi (1 - \gamma \sigma^2 L \bar{u}_E^2)} < \phi p_0, \quad \bar{u}_E^2 \equiv \frac{\sum_j E_j u_{j0}^2}{\sum_j E_j}, \quad (\text{V.7})$$

where the wealth-drag factor carries the equity-weighted second moment of utilization: $\bar{u}_E^2 \geq \bar{u}_0^2$ (Jensen), so at a common average utilization, cross-sectional dispersion weakly deepens the decline. Absorption ($\delta_u < \phi p_0$) holds if and only if $\gamma \sigma^2 L \bar{u}_E^2 < 1$, which A-MOD implies; without A-MOD, unconstrained survivors with $u_j > u_b$ sell and $\delta_u > \phi p_0$ is possible with no constraint binding.

(b) *Amplification (all-at-cap stress benchmark (ii)). Suppose the relevant AI managers enter the stress episode at their caps ($u_{j0} = 1$). Then: Exact characterization (primary). Equation (V.6) has a unique interior solution $0 < y^* < 1$ for every finite g and every $\phi \in (0, 1)$; y^* is strictly increasing in g and in ϕ , and full liquidation $y = 1$ is approached only asymptotically as $g \rightarrow \infty$. There is no sharp crash boundary at finite g : the decline deepens continuously as feedback intensity rises. The round-by-round deleveraging recursion (mark losses at the current price, collect the induced sales, let the price absorb them) converges to y^* monotonically from no initial decline, for every finite g ; the recursion is a constructive device for computing the static clearing equilibrium, not a model of real-time trading dynamics. Linear-feedback approximation. Linearizing the induced-sales schedule at $\delta = 0$ and adding the position floor (survivors cannot sell more than they hold) yields the piecewise solution*

$$\delta = \begin{cases} \frac{\phi p_0}{1 - \kappa(\phi)}, & g \leq 1, \\ p_0, & g > 1, \end{cases} \quad (\text{V.8})$$

continuous at $g = 1$, with first-order amplification $\mathcal{A} = 1/(1 - \kappa(\phi))$ increasing and convex in η , L , and x on the interior region. Within this approximation, $g = 1$ is the position-floor (admissibility) criticality boundary: the point at which the linear model's interior fixed point reaches the floor. It is a property of the approximation, not of the exact system, and not a universal crash threshold; the exponential decay of surviving equity, dropped by the linearization, is what keeps the exact path (V.6) interior.

(c) *Staggered constraint entry (region (iii)).* For each manager, $\bar{\delta}_j = (1 - u_{j0})\gamma\sigma^2/\sqrt{\eta}$, and under A-MOD $\bar{\delta}_j$ simultaneously separates j 's buying from selling and unconstrained from constrained. Aggregate constraint entry begins at $\min_j \bar{\delta}_j$ and amplification phases in across the utilization cross-section; small catalysts (with $\delta_u < \min_j \bar{\delta}_j$) are absorbed even by levered similar survivors. This part is a manager-level comparative static, not a solved heterogeneous equilibrium path.

Proposition 3 contains the model's central economics. *The same similar, levered capital absorbs or amplifies depending on proximity to constraints:* under A-MOD, unconstrained AI survivors buy fire sales—declines below $\min_j \bar{\delta}_j$ are bought, not amplified—while capped survivors are forced sellers whose sales deepen the decline that forces further sales.

The model's sharpest empirically discriminating implication follows directly and deserves emphasis. In a generic correlated-holdings theory, overlapping portfolios mechanically imply correlated selling under stress; here, whether a similar-strategy fund sells is *conditional on its pre-shock leverage utilization and proximity to constraints*, not on its holdings alone. The cross-sectional prediction: in liquidation episodes, low-utilization funds holding the crowded composite with no binding redemptions should *add* exposure while constrained funds cut; systematic selling by slack similar-strategy funds would reject—not recalibrate—the

leverage channel. The same observation disciplines measurement: an empirical fragility assessment built on holdings overlap alone, without the utilization cross-section, misses the state variable on which the absorb/amplify switch turns. Larger catalysts always deepen the decline ($d\delta/d\phi > 0$); in the linearized system, the same local amplification factor $\mathcal{A}$ applies to an exogenous date-1 fundamental price shock. At date 2 the price reverts to fundamentals: forced sellers realize the non-fundamental component as a transfer to liquidity suppliers, plus real risk-bearing costs.

Three comparative-static observations complete the analysis. First, one caution on convexity: the convexity claims attach to the amplification factor $\mathcal{A}$, and monotonicity to the interior decline; the *total* decline is not globally convex in L , because entry displacement p_0 is concave in capacity and dominates at small L . Second, adoption and similarity enter separately. x is a capital share; η is a portfolio-geometry primitive. Price-level objects depend on them through $\Phi = \eta x L a$ and separately on x through $k_R(x)$, so iso-displacement curves in (x, η) are not exact hyperbolas: higher x also removes liquidity suppliers under Assumption 1. Under unequal AI-system market shares the correlated-error term in (V.4) carries HHI, so system-level concentration matters even at moderate η —measurable from vendor market shares, whereas x is the AI share of active assets under management. Leverage enters twice: once in capacity ($\Phi \propto L$, via A-RC) and twice in feedback (L^2 in g , via the binding cap)—the model’s formalization of leverage-adjusted saturation.

Third, the results are attributable across the risk-capital assumption. Assumption 2 (A-RC) makes risk tolerance scale with *permitted* risk capital, so the leverage limit affects unconstrained date-0 demand: $\Phi \propto L$, and p_0 , alpha decay, and the capacity term ηL in Proposition 1(b) all inherit the cap. Under the natural alternative $\Gamma_j = \gamma/E_j$ (risk tolerance scales with equity alone), unconstrained demand loses its factor of L , and date-0 displacement is correspondingly smaller and L -invariant while managers are slack. What survives unchanged is everything generated by *binding* constraints: once at the cap the book is LE_j regardless of Γ , so the induced-sales rate $\eta L^2 E$, the feedback statistic g in (V.5), and the

entire amplification analysis of Proposition 3(b) are unaffected. The attribution is therefore: L 's effect on date-0 displacement and capacity (Propositions 1 and 2) depends on A-RC; L 's effect on liquidation feedback (Proposition 3(b), the L^2 in g) comes from the binding constraint and does not. Under $\Gamma_j = \gamma/E_j$, utilization is $u_j = \hat{\alpha}_m/(L\gamma\sigma^2)$ and the constraint region shifts accordingly, but the absorb/amplify dichotomy and the staggered-entry logic of Proposition 3(a) and (c) are qualitatively unchanged.

E. Externalities and the Scope for Welfare Analysis

This subsection is deliberately informal; the model contains no welfare proposition, and we do not state one. In the manager program (V.2), each manager takes the residual-demand and price-impact environment as external when choosing positions: the first-order condition contains no term for the manager's marginal contribution to aggregate price pressure (V.3) or to the feedback gain $\kappa(\phi)$ through which their date-1 sales deepen others' losses. This is a model-level wedge identified in the program itself, not a proved welfare result, and it motivates the mitigation discussion that follows. A formal welfare analysis and optimal policy design are left for future work: they would require an objective-probability welfare criterion and a careful separation of pecuniary transfers (forced sellers to liquidity suppliers) from real costs—risk-bearing disutility, insolvency deadweight when cascades impair intermediaries, and allocative costs of degraded price informativeness.

F. Mitigation Mechanisms

Because the wedge in Section V.E motivates responses at several levels of formality, we label each candidate mechanism by its epistemic status: formally analyzed within the model, empirically suggested by the experiment's measures, or economically motivated but outside both.

Formally analyzed. No mitigation mechanism is formally analyzed within the baseline model. The one formal adjacent result is the negative-similarity extension (Appendix

B): offsetting AI strategies—negative average pairwise similarity—partially internalize one another’s order flow before it reaches the residual sector, acting as added effective depth, shrinking the entry-displacement analogue of (V.3) and the feedback-statistic analogue of (V.5). The qualitative conclusion is one-directional and unsurprising—negative similarity reduces common-flow pressure on both the routine and stress margins—and no baseline result relies on it.

Empirically suggested and incentive-based. One qualification is required by the model itself: baseline managers are price takers, so within the model none of them perceives any private benefit from mitigating aggregate pressure—that is precisely the wedge. The mitigation incentives described here therefore attach to managers large enough to internalize part of their own price impact, a case outside the baseline program. Such managers would have incentives to differentiate (lowering effective η), to hold utilization buffers (raising $\bar{\delta}_j$), and to negotiate redemption gates that slow ϕ . Each maps to a lever the experiment’s measures make visible: the construction-effect decomposition of Section IV.D shows that implementation heterogeneity alone moved trade-overlap similarity by roughly 0.36 to 0.38 at fixed signals, so differentiation in portfolio construction is a candidate mitigation margin whose trade-level effect the experiment has measured—though whether exercising it would reduce realized fragility is a model statement, not an experimental one.

Economically motivated policy candidates. On the policy side, the same objects suggest liquidity-sensitive leverage limits or margins (targeting $L^2\lambda$ in g), capacity or position limits (targeting ηxA relative to depth), disclosure of common exposures, and stress tests built on the feedback statistic g rather than on any single similarity measure. None of these is derived from an optimal-policy problem; they are readings of which primitives the fragility object g depends on.

G. Proofs, Extensions, and Illustrations

Appendix A collects the model’s notation and the proofs of Lemmas 1–2 and Propositions 1–3, together with a remark fixing the economic status of the deleveraging recursion—the equilibrium object throughout is the static date-1 clearing condition, and the recursion is a constructive fixed-point device, not a model of real-time trading dynamics—and distinguishing the linear approximation’s admissibility boundary $g = 1$ from its iteration-divergence threshold $\kappa(\phi) = 1$. Appendix B develops the negative-similarity extension referenced in Section V.F. Every derivation is additionally verified symbolically and numerically in a companion verification script, which is part of the archived workflow record.

Figures 7 and 8 illustrate the model’s two blocks on the scenario grid anchored by the empirical measures.

[Insert Figure 7 about here]

[Insert Figure 8 about here]

Figure 7 plots date-0 price displacement and the fraction of alpha surviving displacement against the adoption share x at similarity scenarios $\eta \in \{0.33, 0.58, 0.81\}$ —the agent-matched null, the observed holdings cosine, and the standardized-construction value—under both the reallocation and fixed-depth paths. Figure 8 plots the exact liquidation decline y^* against the linear-feedback solution and shows the scenario map in leverage and AI capital share: the $g = 1$ locus is drawn, and labeled, as the linearized model’s admissibility boundary, while the exact outcome, which has no finite crash threshold, varies continuously across it. Both figures are transparent scenario analyses under illustrated parameterizations, not calibrated predictions. One numerical illustration conveys the magnitudes and the discipline required in reading them: with $\lambda = 1/(\zeta_{\perp} W)$ (W the composite-basket capitalization, $\zeta_{\perp} \approx 1$ the micro-elasticity), $g = \eta L^2(xA/W)/\zeta_{\perp}$ is dimensionless, and $\eta = 0.6$, $L = 3$, $\zeta_{\perp} = 1$ puts $g = 1$ at an AI share of roughly 18.5% of basket capitalization. That figure is not a critical

AI market share: the exact liquidation outcome (V.6) varies continuously through it, $g = 1$ marks only where the linearized formula stops being admissible, and the number moves with every input of the illustration. No crash-frequency or crash-probability claim attaches to it.

Seven limitations bound the model’s claims. (1) The residual sector is reduced-form: suppliers do not learn from prices, so the price-informativeness results are conditional benchmarks; introducing price-based learning would tend to dampen displacement. (2) The reallocation assumption (Assumption 1) is a maintained counterfactual; the fixed-depth case is reported alongside, and adoption financed by new capital lies between the two. (3) Proposition 2 requires persistent price pressure; under rapidly reversing impact it weakens or vanishes while the stress results survive. (4) The exact liquidation equilibrium is analytically characterized only for the all-at-cap stress benchmark; the heterogeneous intermediate region is characterized by manager-level comparative statics. (5) The exact locations of boundaries (e.g., where $g = 1$ falls in parameter space) are specific to linear impact and proportional redemption; the comparative statics of g are the structural content. Under nonlinear impact or non-proportional liquidation the boundary moves, but amplification still requires the local feedback gain to be controlled. (6) The model is single-asset (one crowded composite); cross-impact across several crowded composites is left to future work, where the scalar condition generalizes to a spectral-radius condition on the product of the price-impact and common-exposure matrices. (7) None of the model’s quantitative inputs is estimated from the experiment: the similarity parameter η is anchored to the measured holdings cosine as a scenario, not estimated; the impact and depth objects λ , $\zeta_\perp$, and k_R are illustrated, not identified; routine rebalancing flow is not formally modeled (Section V.A); and the stress block assumes full synchronization of forced sales.

Neither holdings similarity nor trade similarity alone is a sufficient statistic for fragility. The model’s fragility object is the feedback statistic g —empirically, a liquidity-weighted common-exposure ratio built from holdings, leverage, and depth. Routine-crowding measures govern rebalancing impact in normal times; latent-stress measures—holdings similarity

together with leverage, depth, and the utilization distribution—govern amplification under forced liquidation. Routine trade synchronization below its holdings-overlap level is therefore not evidence of low fragility. Section VI develops what this division of labor implies for reading the experiment’s evidence.

VI. Economic Interpretation and Implications

This section draws together what the experiment and the model jointly say. The synthesis is deliberately conservative: nothing here adds a result; everything is an interpretation of results already presented, labeled as such.

A. What the Crowding Measures Jointly Reveal

The measures tell a coherent story once one distinguishes the level at which agents converge from the level at which markets would feel it. Convergence is strongest upstream, in signal choices and security rankings, and attenuates—but only modestly—as it propagates downstream through portfolio construction. Each transformation, scores into weights and weights into trades, injects implementation heterogeneity that absorbs some of the upstream agreement. What remains is nonetheless the paper’s central empirical fact: a substantial share of what these systems hold, and of what they want to trade in a given month, is common across independently operating agents (Section IV.A).

The locked test period sharpens rather than softens this fact. Holdings and trade similarity were essentially unchanged out of sample, while pairwise return correlations rose (Section IV.C): the strategies became more alike in outcome even though they became no more alike in structure. Rising factor spanning accounts for part of the increase—common factor exposure transmits common positioning into common returns—but not all of it, since the six-factor residual correlations rose alongside the raw ones. We report both facts without a mechanism; the frozen outputs contain no decomposition of why the loadings moved.

Against this stability, the one large movement—the out-of-sample decline in liquidity-

scaled common trading—is a statement about markets, not about the agents. Common desired trading was essentially flat; what changed is that the securities the agents commonly trade traded at far higher dollar volume, and benchmark portfolios declined comparably over the same window (Section IV.E). This is the paper’s clearest illustration of why similarity and market footprint must be kept conceptually separate: strategic convergence is a property of the agents and persisted; the liquidity-scaled measure divides by a market quantity—dollar volume, a liquidity proxy, not structural depth—and that denominator grew.

B. Benchmark Dependence as a Measurement Lesson

Whether the observed convergence is abnormal depends on what one holds fixed, and the paper’s two benchmarks hold different things fixed. Homogenized construction is itself a powerful synchronizer, so trade-level similarity that looks unremarkable against a standardized-construction null looks exceptional against a null that preserves each agent’s own architecture. The paired construction effects quantify the wedge: at fixed signals, standardizing construction alone accounts for more than three quarters of observed trade-level similarity (Section IV.D).

Two lessons follow, and we regard both as findings rather than nuisances. First, for the empirical literature that measures crowding from holdings or trade data, implementation heterogeneity is a first-order determinant of measured trade-level similarity: crowding statistics computed under standardized portfolio constructions will overstate the trade synchronization attributable to shared information and understate the dampening role of implementation diversity. Second, the two benchmarks answer different questions, and headline inference belongs to the classification they jointly produce rather than to either alone. The agent-matched benchmark speaks most directly to the question this paper asks—did independent optimization converge?—and by that benchmark the answer is yes at every level measured, in both periods; but its pre-specification status is junior, and under the pre-specified standardized benchmark only holdings similarity is elevated. Section IV.D records the resulting classification; signal overlap, upstream of construction, is exceptionally high under every

selection null considered.

C. Similarity Is Not Yet Fragility: Connecting the Measures to the Model

The model disciplines what the measures can and cannot be asked to say about fragility, and Table A.I in Appendix A summarizes the mapping. The model’s fragility object is not holdings or trade similarity but the feedback statistic $g = \lambda\eta L^2 xA$: a liquidity-weighted common-exposure ratio in which similarity is one input alongside adoption scale, leverage, and depth. Neither headline similarity measure is a sufficient statistic for it. Holdings similarity disciplines the model’s similarity parameter—as a scenario anchor, not an estimate—and speaks to latent stress exposure: where synchronized liquidation would concentrate if it occurred. Trade-overlap and common-trading measures discipline routine-flow scenarios. The experiment identifies neither adoption, nor capital, nor leverage, nor structural depth; those enter only as scenario grids from external sources.

Three implications of this division of labor deserve emphasis. First, routine trade synchronization sitting below holdings similarity is not evidence of low fragility: in the model, stress compresses liquidation into the impact horizon, so the relevant similarity under stress is the holdings overlap. Second, the declining liquidity-scaled common trading does not retire the fragility question, because that measure is a routine-flow object; amplification turns on leverage utilization and depth under forced liquidation, quantities the experiment does not measure. Third, and symmetrically, nothing in the experiment establishes that the measured convergence *is* economically consequential today: the common trading absorbs a small share of contemporaneous dollar volume, the experiment involves four agents, and no realized price impact is measured or claimed. The model’s contribution is precisely to convert this ignorance into an explicit counterfactual statement—which combinations of adoption, similarity, leverage, and depth would make convergence of the measured magnitude matter, and which would not.

D. How AI Crowding Could Scale: The Institutional Adoption Channel

How could AI-generated portfolios come to constitute a large common exposure in practice? This subsection discusses plausible deployment paths and is deliberately not added to the formal model. Scalable vehicles—ETFs, model portfolios, separately managed accounts, and index-like products built on AI recommendations—could translate similar AI outputs into substantial common holdings and synchronized flows across many end investors, without any coordination among them. The model organizes the channel: an *unlevered* AI product primarily raises deployed AI capital, moving the economy along the displacement and price-informativeness results of Proposition 1, while *leverage* raises exposure and potential feedback through g . Proposition 2 adds a reinforcing loop that operates through measurement itself: adoption inflows temporarily inflate measured performance, and strong measured performance during the inflow phase could attract further assets—an observer attributing the price-pressure component to skill would be misled at exactly the moment the strategy’s expected future return is falling. One caveat: leveraged ETFs are not literal implementations of the model’s balance-sheet mechanism—their derivative-based daily rebalancing differs from the redemption and leverage-constraint channel—although both generate correlated, procyclical order flow.

The experiment speaks to the front end of this channel and only the front end. It shows that the “similar AI outputs” premise is not hypothetical: four independently operating systems produced strategies with overlapping signals, holdings, and desired trades, and the overlap survived twelve years out of sample. Whether the back end materializes—adoption at scale, in levered form, in markets whose depth does not grow commensurately—is not something a controlled experiment can determine.

E. Responses by Managers and Policymakers

For managers, the margin the experiment can actually speak to is implementation: construction heterogeneity alone moved trade-level similarity by more than a third at fixed signals, so

differentiation in portfolio construction—not only in signal research—is a candidate margin for reducing measured crowding. Whether exercising it would reduce realized fragility is a model statement, not a measured one. The model adds the incentive qualification: atomistic managers capture no private benefit from reducing aggregate pressure, so differentiation, utilization buffers, and redemption gates are privately rational only for managers large enough to internalize part of their own price impact. For policymakers, the model implies that single-statistic monitoring can be incomplete. Holdings overlap alone misses the state variable—leverage utilization—on which the absorb-or-amplify switch turns; trade overlap alone misses latent stress exposure. Monitoring built on the composition of g —common exposures, leverage, and depth jointly—follows from the model’s structure. These are model-implied candidate monitoring margins, not recommendations grounded in measured harm; no formal optimal-policy analysis is offered.

F. Boundaries of the Evidence

We close by restating what the paper does not establish, in the order the claims might otherwise be overread. The experiment does not measure realized price impact; no return in the locked test period is attributed to the agents’ own trading, which never occurred in markets. The trade-synchronization and liquidity-scaled common-trading findings are elevated under only one of the two benchmark constructions, and we report both. The model’s fragility results are counterfactual scenario analyses under maintained assumptions, and none of its numerical illustrations is a calibrated prediction. The experiment involves four agents, six pairs, one signal library, one asset class, and one locked twelve-year window; each is a dimension along which replication could strengthen or overturn the findings. Within those boundaries, the evidence supports a specific and, we believe, durable conclusion: in this experiment, independent AI agent sessions pointed at the same published cross-section of expected returns produced substantially overlapping strategies rather than diversifying across the supplied opportunity set—overlapping on signals most of all, and on portfolios

and trades to a degree that survives their heterogeneous engineering—and the agreement did not decay out of sample. What that agreement costs, and whom, depends on scale, leverage, and liquidity that lie outside the laboratory.

VII. Conclusion

This paper asks what happens when independently operating AI agents are given the same published cross-section of expected returns and told to build the best portfolio they can. We ran the experiment under controls strong enough to make the answer evidence rather than anecdote: four frozen agents, isolated workspaces, an identical hash-verified protocol, a capped and logged search budget, and a twelve-year locked test period that no agent, no strategy, and no selection decision ever touched.

The agents converged, and the convergence persisted. Their six distinct frozen strategies drew on five signals, two of which appear in every specification—overlap far beyond what any of the specified selection nulls generates—and their holdings, desired trades, and returns remained aligned through the locked test. The strategies remained profitable gross of costs out of sample, with the familiar post-publication compression after costs and factor adjustment. The locked test evaluates the persistence of strategies frozen with information through 2002; it does not test whether an AI system given data through 2024 could construct a successful strategy today.

The most important qualification is benchmark dependence: trade-level synchronization is unusually high only relative to benchmarks that preserve each agent’s own construction architecture, holdings similarity is the most robust downstream result, and implementation heterogeneity materially moderates measured crowding. And convergence is not market impact: the strategies never traded in markets, and no realized price effect is measured or claimed.

What the experiment provides is direct evidence, in one controlled setting, for the premise that fragility arguments about AI in asset management have so far assumed: independent

optimization over shared inputs can produce durable strategic convergence. The model converts that premise into explicit counterfactuals, and its central lesson is that similarity alone is not sufficient for fragility; amplification depends jointly on similarity, capital scale, leverage—especially leverage utilization—and market depth, with severity summarized in a single feedback statistic $g = \lambda\eta L^2 x A$. None of the model’s numerical exercises is a prediction. All of them say which combinations of adoption, leverage, and depth would make convergence of the measured magnitude matter.

The limitations bear repeating in one place: four agents, one supplied signal library, one asset class, one locked window; cost estimates from a model rather than realized execution; crowding measured relative to explicit but contestable nulls; and a theory whose stress block rests on maintained assumptions.

What remains unknown is consequence. Replication—more agents, model families, libraries, asset classes, and vintages—would establish whether the convergence is a property of this generation of systems or of optimization over shared inputs generally. And the model’s fragility object is built from observable ingredients—common exposures, leverage, and depth—so constructing its empirical counterpart for actual fund populations would move the fragility question from scenario to estimate; the multi-asset generalization and a formal welfare analysis remain open.

Machines built to find the best strategy found the same strategy. Whether that matters depends on how much capital follows them, how levered it is, and how deep the markets are that it trades. The first fact is now documented—for four agents, one library, one controlled design. The consequences remain to be measured.

References

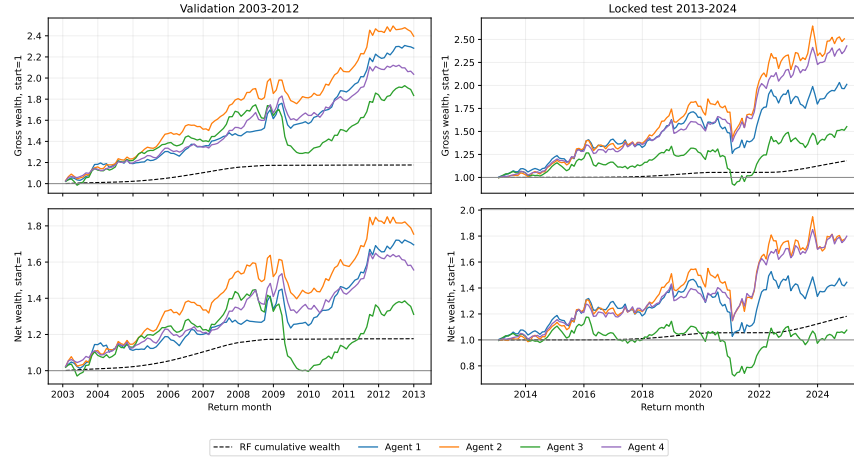
- Ball, Ray, Joseph Gerakos, Juhani T. Linnainmaa, and Valeri Nikolaev, 2016, Accruals, cash flows, and operating profitability in the cross section of stock returns, *Journal of Financial Economics* 121, 28–45.
- Barroso, Pedro, Roger M. Edelen, and Paul Karehnke, 2022, Crowding and tail risk in momentum returns, *Journal of Financial and Quantitative Analysis* 57, 1313–1342.
- Bowles, Boone, Adam V. Reed, Matthew C. Ringgenberg, and Jacob R. Thornock, 2024, Anomaly time, *Journal of Finance* 79, 3543–3580.
- Brown, Gregory W., Philip Howard, and Christian T. Lundblad, 2022, Crowded trades and tail risk, *Review of Financial Studies* 35, 3231–3271.
- Bruno, Salvatore, Ludwig Chincarini, and Frank Ohara, 2018, Portfolio construction and crowding, *Journal of Empirical Finance* 47, 190–206.
- Calluzzo, Paul, Fabio Moneta, and Selim Topaloglu, 2019, When anomalies are publicized broadly, do institutions trade accordingly?, *Management Science* 65, 4555–4574.
- Calluzzo, Paul, Fabio Moneta, and Selim Topaloglu, 2026, Momentum at long holding periods, Working paper, Queen’s University and University of Ottawa (March 2026).
- Chan, Louis K. C., Narasimhan Jegadeesh, and Josef Lakonishok, 1996, Momentum strategies, *Journal of Finance* 51, 1681–1713.
- Chen, Andrew Y., and Tom Zimmermann, 2022, Open source cross-sectional asset pricing, *Critical Finance Review* 11, 207–264.
- Chincarini, Ludwig, Renato Lazo-Paz, and Fabio Moneta, 2026, Crowded spaces and anomalies, *Journal of Banking and Finance* 182, 107579.
- Cho, Thummim, 2020, Turning alphas into betas: Arbitrage and endogenous risk, *Journal of Financial Economics* 137, 550–570.
- DeMiguel, Victor, Alberto Martín-Utrera, and Raman Uppal, 2025, Can competition increase profits in factor investing?, *Management Science* 71, 5552–5571.

- Engelberg, Joseph, R. David McLean, and Jeffrey Pontiff, 2018, Anomalies and news, *Journal of Finance* 73, 1971–2001.
- Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.
- Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.
- Farboodi, Maryam, and Laura Veldkamp, 2023, Data and markets, *Annual Review of Economics* 15, 23–40.
- Foster, George, Chris Olsen, and Terry Shevlin, 1984, Earnings releases, anomalies, and the behavior of security returns, *The Accounting Review* 59, 574–603.
- Gabaix, Xavier, and Ralph S. J. Koijen, 2022, In search of the origins of financial fluctuations: The inelastic markets hypothesis, Working paper (May 2022).
- Greenwood, Robin, and David Thesmar, 2011, Stock price fragility, *Journal of Financial Economics* 102, 471–490.
- Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, ... and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.
- Hasbrouck, Joel, 2009, Trading costs and returns for U.S. equities: Estimating effective costs from daily data, *Journal of Finance* 64, 1445–1477.
- Hirshleifer, David, Kewei Hou, Siew Hong Teoh, and Yinglei Zhang, 2004, Do investors overvalue firms with bloated balance sheets?, *Journal of Accounting and Economics* 38, 297–331.
- Hou, Kewei, Chen Xue, and Lu Zhang, 2020, Replicating anomalies, *Review of Financial Studies* 33, 2019–2133.
- Jacobs, Heiko, and Sebastian Müller, 2020, Anomalies across the globe: Once public, no longer existent?, *Journal of Financial Economics* 135, 213–230.
- Kang, Wenjin, K. Geert Rouwenhorst, and Ke Tang, 2021, Crowding and factor returns, Working paper, Yale School of Management (March 2021).

- Koijen, Ralph S. J., and Motohiro Yogo, 2019, A demand system approach to asset pricing, *Journal of Political Economy* 127, 1475–1515.
- Kyle, Albert S., Anna A. Obizhaeva, and Yajun Wang, 2025, Trading in crowded markets, *Journal of Financial and Quantitative Analysis*, forthcoming.
- Lou, Dong, and Christopher Polk, 2022, Comomentum: Inferring arbitrage activity from return correlations, *Review of Financial Studies* 35, 3272–3302.
- McLean, R. David, and Jeffrey Pontiff, 2016, Does academic research destroy stock return predictability?, *Journal of Finance* 71, 5–32.
- Novy-Marx, Robert, and Mihail Velikov, 2016, A taxonomy of anomalies and their trading costs, *Review of Financial Studies* 29, 104–147.
- Pojarliev, Momtchil, and Richard M. Levich, 2011, Detecting crowded trades in currency funds, *Financial Analysts Journal* 67, 26–39.
- Richardson, Scott A., Richard G. Sloan, Mark T. Soliman, and İrem Tuna, 2005, Accrual reliability, earnings persistence and stock prices, *Journal of Accounting and Economics* 39, 437–485.
- Shleifer, Andrei, and Robert W. Vishny, 1997, The limits of arbitrage, *Journal of Finance* 52, 35–55.
- Stein, Jeremy C., 2009, Presidential address: Sophisticated investors and market efficiency, *Journal of Finance* 64, 1517–1548.
- Tobek, Ondřej, and Martin Hronec, 2021, Does it pay to follow anomalies research? Machine learning approach with international evidence, *Journal of Financial Markets* 56, 100588.
- Vayanos, Dimitri, and Paul Woolley, 2013, An institutional theory of momentum and reversal, *Review of Financial Studies* 26, 1087–1145.

Panel A

Strategy A: validation and locked cumulative performance
Risk-free line is the cash context for these dollar-neutral strategies



Panel B

Strategy B: validation and locked cumulative performance
Risk-free line is the cash context for these dollar-neutral strategies

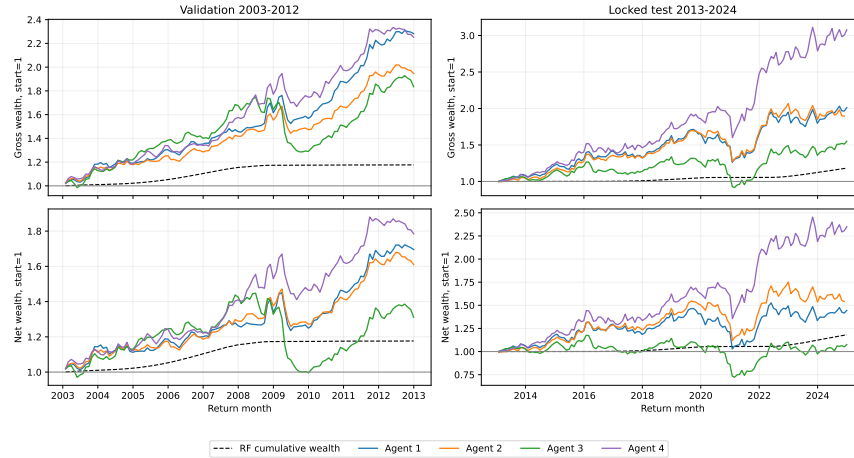


Figure 1. Cumulative performance of the frozen strategies. Cumulative wealth (initial value 1) of each agent’s frozen strategy in the validation period (January 2003–December 2012, left) and the locked test period (January 2013–December 2024, right), gross of transaction costs (top) and net of estimated transaction costs (bottom). Panel A: Strategy A (gross-Sharpe objective). Panel B: Strategy B (net-Sharpe objective). The dashed line plots cumulative total-return wealth in the risk-free asset (the Fama–French one-month Treasury bill rate, RF), compounded monthly. RF is shown as a contextual capital-growth benchmark for these dollar-neutral strategies; it is not a risk-matched benchmark, and it is not subtracted from the strategy returns, which are plotted as raw gross and net returns. The near-flat RF segments — 2009–2012 in the validation panel and 2013–2015 in the locked panel — reflect the near-zero policy-rate period (monthly T-bill returns of 0.00–0.02%), not missing data. Wealth is reset to 1 at the start of each evaluation period and separately compounded; the two periods are not a continuous investment path.

Strategy A: principal crowding outcomes by separate sample
Common vertical scale across periods for each measure

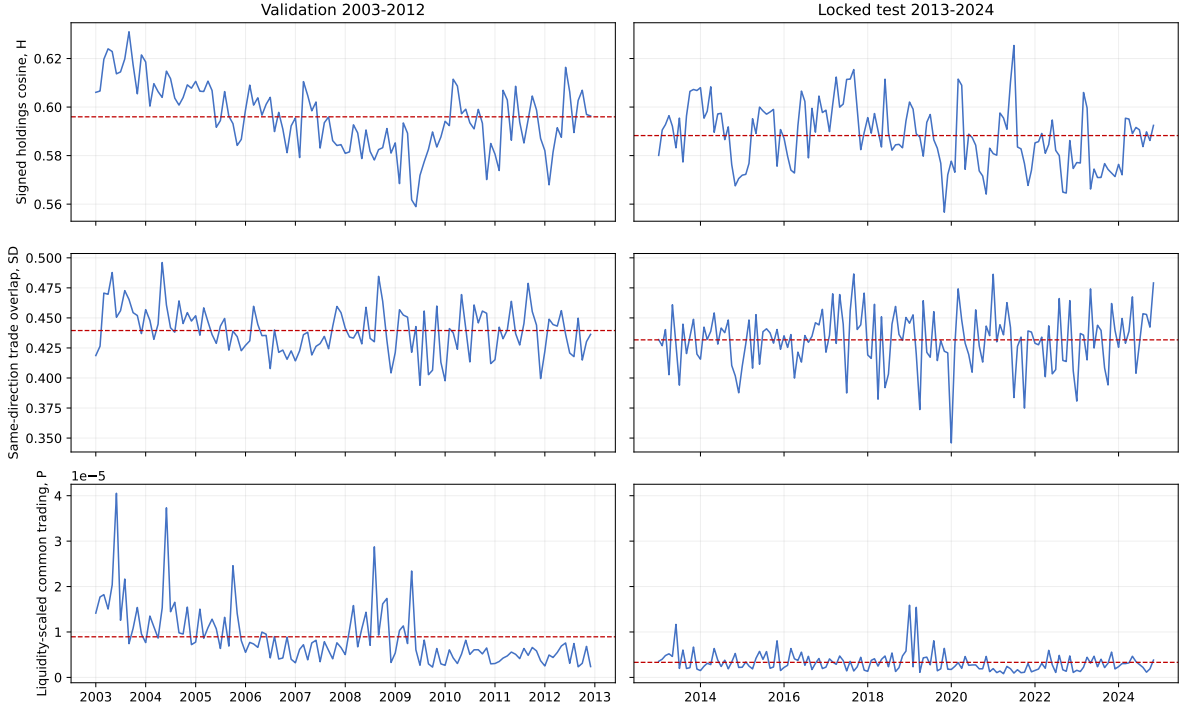


Figure 2. Monthly principal crowding measures, Strategy A. Monthly cross-agent means (over the six agent pairs) of the three principal crowding measures for Strategy A: the signed holdings cosine H (top), same-direction desired-trade overlap SD (middle), and liquidity-scaled common trading per dollar P (bottom, $\times 10^{-5}$), in the validation period (left) and the locked test period (right). Dashed horizontal lines mark the period means. Each measure uses a common vertical scale across the two periods: the validation-period mean of P (8.9×10^{-6}) is roughly three times the locked-test mean (3.3×10^{-6}), while H and SD are essentially unchanged.

Strategy B: principal crowding outcomes by separate sample
Common vertical scale across periods for each measure



Figure 3. Monthly principal crowding measures, Strategy B. As Figure 2, for Strategy B (net-Sharpe objective): monthly cross-agent means of the signed holdings cosine H , same-direction desired-trade overlap SD , and liquidity-scaled common trading per dollar P ($\times 10^{-5}$), in the validation period (left) and the locked test period (right), with period means as dashed lines and a common vertical scale across periods for each measure.

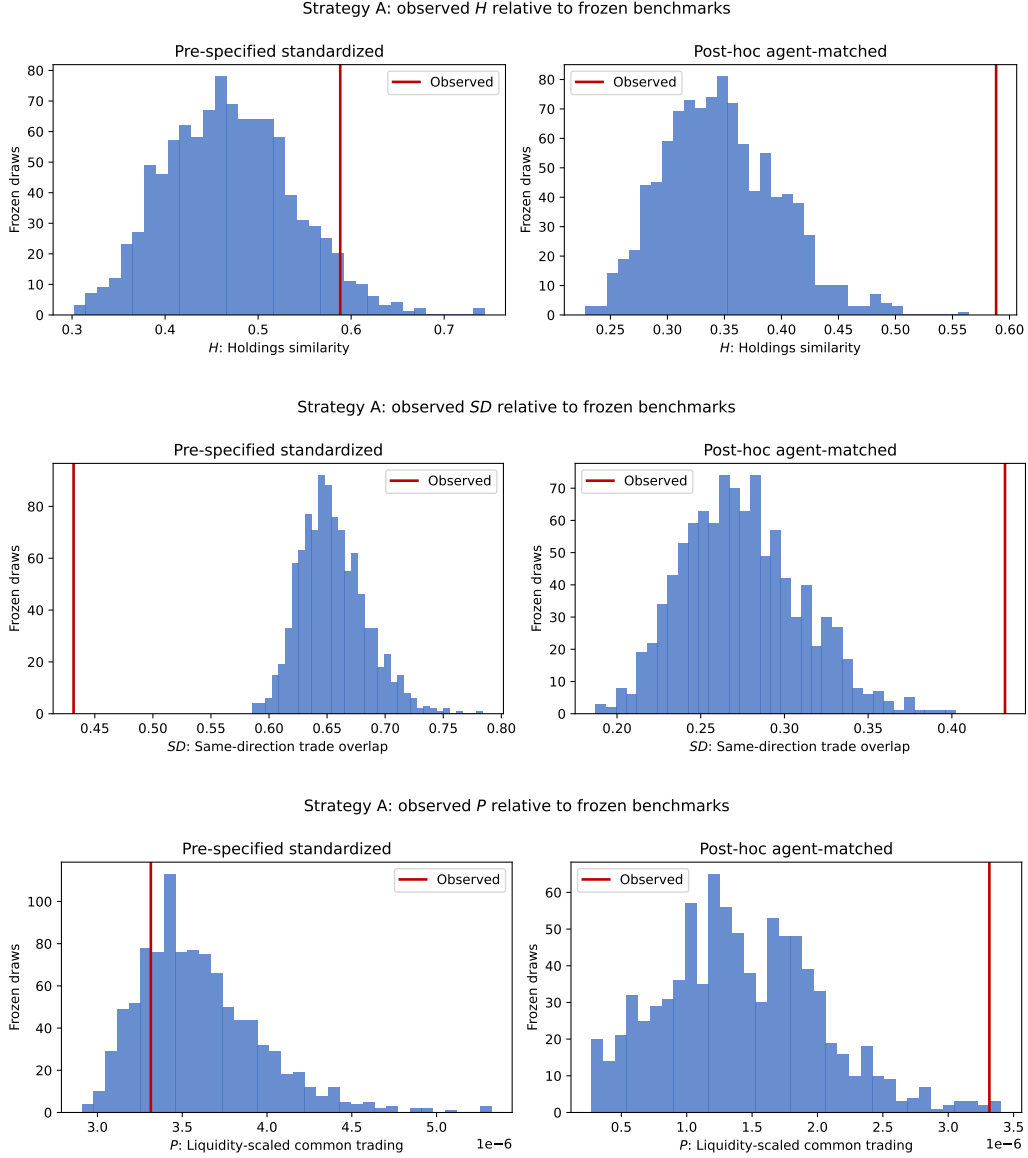


Figure 4. Observed crowding measures relative to frozen benchmark distributions, Strategy A, locked test period. Each row plots the locked-test-period observed value (vertical line) of one principal crowding measure for Strategy A against the distribution of the same measure across frozen benchmark draws: the signed holdings cosine H (top), same-direction trade overlap SD (middle), and liquidity-scaled common trading P (bottom). Left column: the pre-specified standardized-construction benchmark, which redraws signal sets and builds all benchmark portfolios under a single common construction (1,000 draws). Right column: the agent-matched benchmark, which redraws signal sets while holding each agent’s own frozen construction code fixed (1,000 draws; 869 feasible draws for P). Observed H lies in the upper tail of both distributions; observed SD and P lie in the upper tail of the agent-matched distribution but at or below the mass of the standardized distribution.

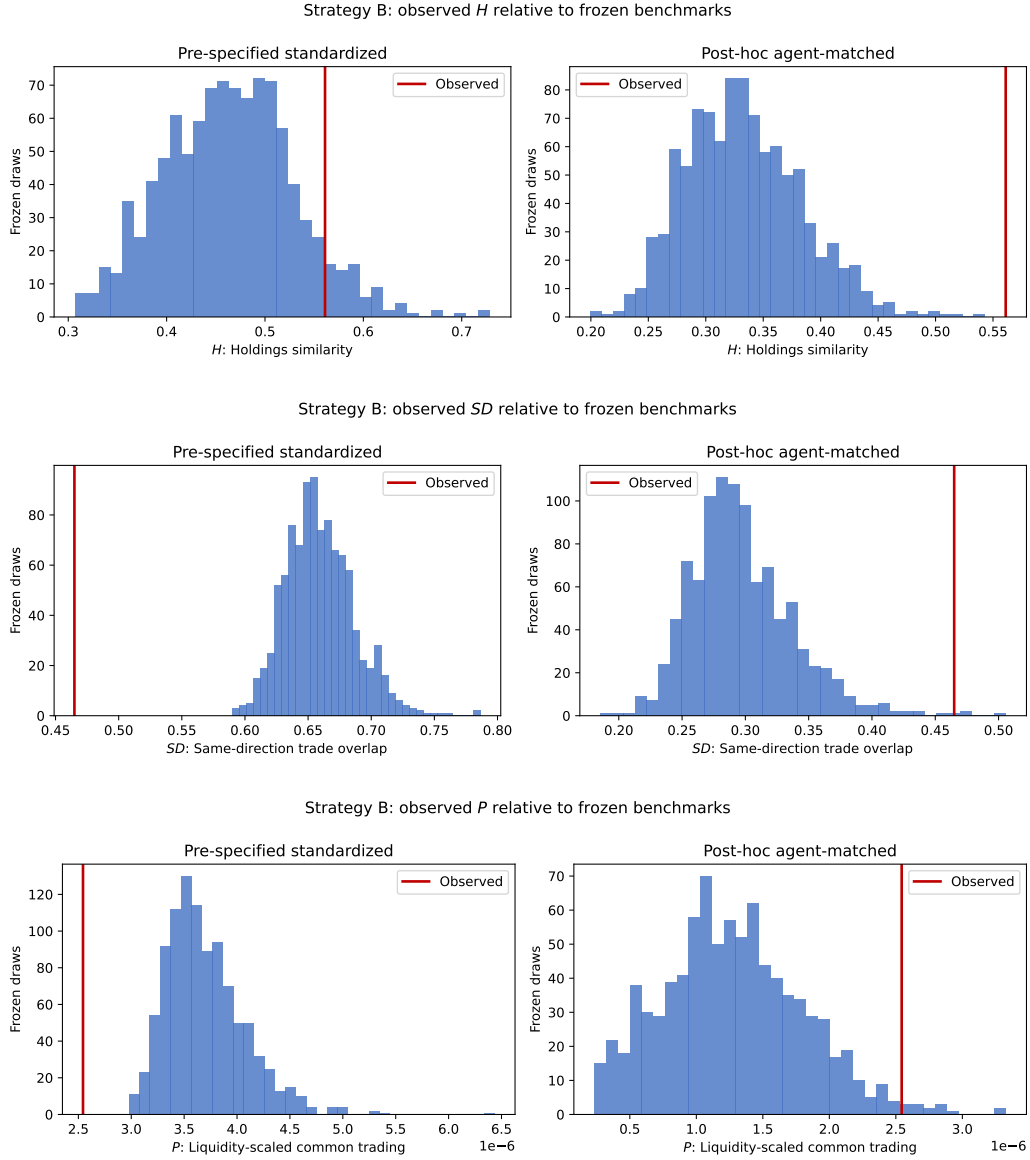


Figure 5. Observed crowding measures relative to frozen benchmark distributions, Strategy B, locked test period. As Figure 4, for Strategy B: locked-test-period observed values of H , SD , and P (vertical lines) against the frozen standardized-construction (left) and agent-matched (right) benchmark distributions.

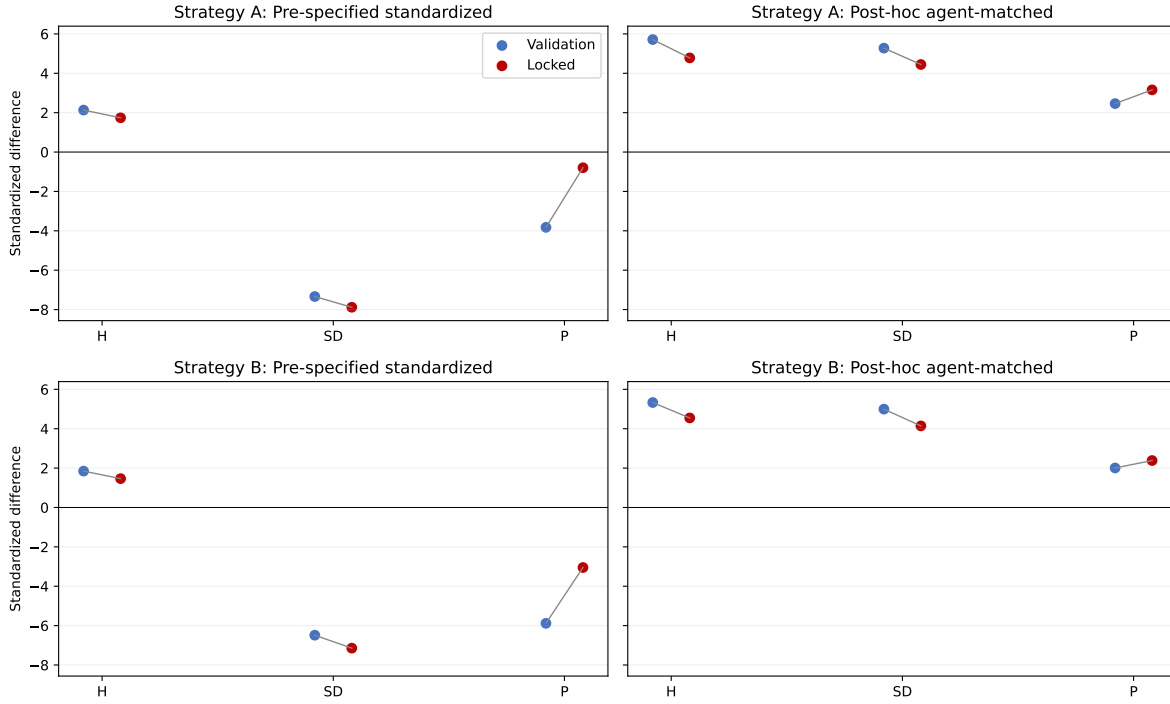


Figure 6. Benchmark-relative crowding in the validation and locked test periods. Standardized differences $((\text{observed} - \text{null mean})/\text{null SD})$ of the three principal crowding measures relative to each frozen benchmark, in the validation period (blue) and the locked test period (red). Top row: Strategy A; bottom row: Strategy B. Left column: pre-specified standardized-construction benchmark; right column: agent-matched benchmark. Connecting segments trace the movement of each standardized difference from the validation to the locked test period. Against the agent-matched benchmark all three measures remain well above zero in both periods; against the standardized benchmark only H is elevated, SD is far below every draw in both periods, and the standardized difference for P is negative in both periods.

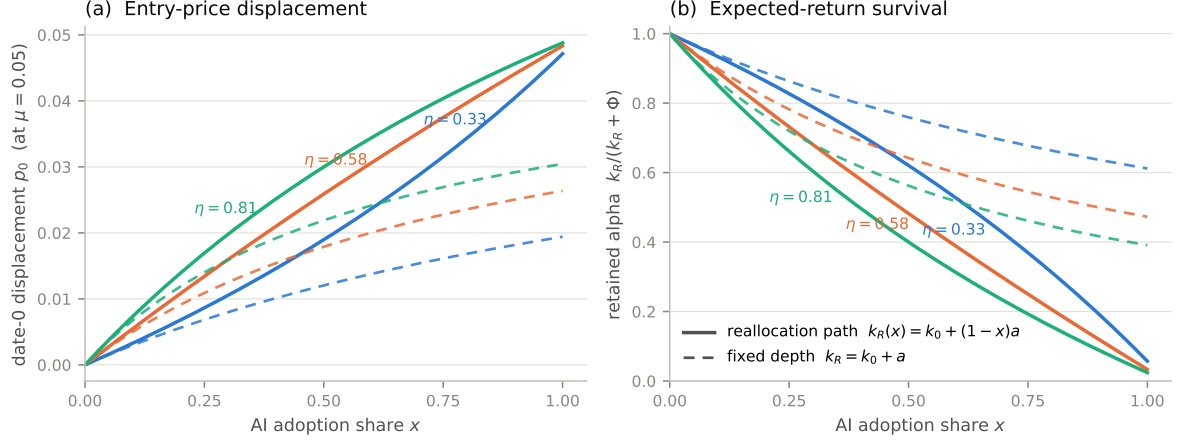


Figure 7. Entry displacement and expected-return decay across adoption and similarity scenarios. Left: date-0 price displacement p_0 as a function of the AI adoption share x , at similarity scenarios $\eta \in \{0.33, 0.58, 0.81\}$ (agent-matched null, observed holdings cosine, standardized construction). Right: the fraction of alpha surviving displacement, $\mathbb{E}[v - p_0 \mid \mu]/\mu = k_R/(k_R + \Phi)$. Solid lines: capital-reallocation path $k_R(x) = k_0 + (1-x)a$; dashed lines: fixed residual depth $k_R = k_0 + a$, isolating the direct correlated-demand effect from the additional liquidity-depletion effect. Illustrative scenario parameterization ($\gamma\sigma^2L = 0.16$; $k_0 = 0.5$, $a = 12.5$, so residual depth falls from $k_0 + a = 13$ at $x = 0$ to $k_0 = 0.5$ at $x = 1$); not a calibrated prediction. In particular, the near-total alpha decay as $x \rightarrow 1$ on the solid path is a property of the reallocation counterfactual with small retained market-maker depth k_0 —when almost all residual capital has been reallocated, almost nothing damps the common demand—not a generic prediction about full AI adoption.

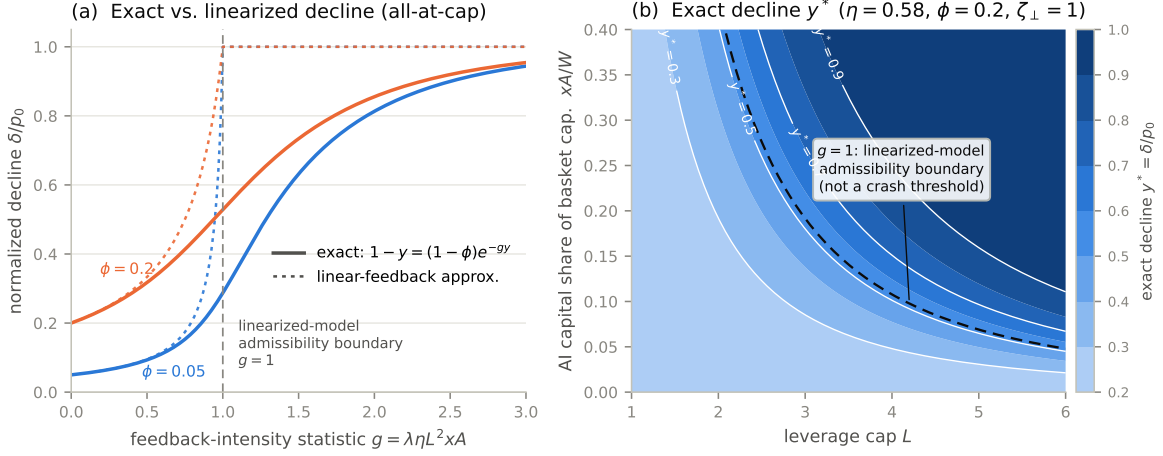


Figure 8. Liquidation amplification: exact nonlinear outcome versus the linearized-model criticality boundary. Left: the date-1 decline δ/p_0 in the all-at-cap stress benchmark as a function of the feedback-intensity statistic g , for catalyst sizes $\phi \in \{0.05, 0.20\}$. Solid: the exact nonlinear equilibrium, which has a unique interior solution for every finite g and deepens continuously. Dotted: the linear-feedback approximation, whose interior fixed point reaches the position floor at its admissibility boundary $g = 1$ (vertical line). Right: illustrative scenario map in leverage L and AI capital share of basket capitalization xA/W at $\eta = 0.58, \zeta_{\perp} = 1, \phi = 0.20$: filled contours show the exact decline y^* ; the dashed curve is the $g = 1$ locus, labeled as the *linearized model's admissibility boundary*, not a crash threshold—the exact outcome varies continuously across it. The elasticity mapping placing the axes in these units, $\lambda = 1/(\zeta_{\perp} W)$ with $\zeta_{\perp} = 1$, is itself illustrative ($\zeta_{\perp}$ plausibly 0.5–2), and the map rescales accordingly. Transparent scenario analysis; not an estimate of market instability or crash probability.

Table I

Frozen Strategy Definitions and Construction Rules

All values are taken directly from the frozen locked-run and measure tables in the archived workflow record; no value is recomputed.

This table reports the eight frozen strategy specifications, as locked before the opening of the January 2013–December 2024 locked test period. Signals are Open Source Asset Pricing (OSAP) predictors; orientation gives the sign applied to each signal in the composite. All strategies share the protocol-mandated constraints: dollar-neutral long and short legs scaled to +1 and −1, position weights capped at $|w| \leq 5\%$ with within-leg redistribution, at least 100 names per leg, a \$5 minimum price screen, and monthly rebalancing at month-end formation with returns earned in the following month. Agents 1 and 3 froze identical specifications for Strategies A and B; SHA-256 hashes of the frozen specification files are archived in the workflow record. The selected signals are: AnnouncementReturn, the cumulative return around earnings announcements; CBOperProf, cash-based operating profitability; dNoa, the change in net operating assets (selected with negative orientation); DelFINL, the change in net financial liabilities (negative orientation); and EarningsSurprise, standardized unexpected earnings. In the combination rules, “Development-Sharpe-weighted” means that component ranks are weighted by each signal’s standalone development-period Sharpe ratio, fixed at the freeze. The trade-suppression rules reduce turnover by retaining incumbent positions: Agent 2’s no-trade band (3.0) admits a new position only at the decile cutoff but holds an incumbent security while its composite rank remains within 3.0 times the entry decile (the top or bottom 30% of the ranking); Agent 4’s hysteresis buffer (15%) likewise admits at the decile cutoff and holds an incumbent while its composite score remains within 15 percentile points of the entry cutoff.

Agent	Strategy	Signals (orientation)	Combination rule	Leg cutoffs	Within-leg weight-ing	Trade-suppression rule
1	A = B	AnnouncementReturn (+), CBOperProf (+), dNoa (−)	Equal average of oriented cross-sectional percentile ranks	Top/bottom decile	Equal	None
2	A	CBOperProf (+), AnnouncementReturn (+), DelFINL (−)	Development-Sharpe-weighted average of oriented ranks	Top/bottom decile	Equal	None
2	B	CBOperProf (+), AnnouncementReturn (+), dNoa (−)	Development-Sharpe-weighted average of oriented ranks	Top/bottom decile	Equal	No-trade band (3.0)
3	A = B	AnnouncementReturn (+), CBOperProf (+), DelFINL (−), EarningsSurprise (+)	Equal average of oriented ranks	Top/bottom decile	Equal	None
4	A	CBOperProf (+), AnnouncementReturn (+)	Linear score weights on oriented percentile ranks	Top/bottom quintile	Score-proportional	None
4	B	CBOperProf (+), AnnouncementReturn (+)	Linear score weights on oriented percentile ranks	Top/bottom decile	Equal	Hysteresis buffer (15%)

Table II

Strategy Performance in the Validation and Locked Test Periods

This table reports annualized performance of the eight frozen strategies in the validation period (January 2003–December 2012, 120 months) and the locked test period (January 2013–December 2024, 144 months; 143 for Agent 2, for which one locked month is infeasible under its frozen construction and is disclosed under the frozen attrition rule). Gross returns are before transaction costs; net returns subtract effective-cost estimates from the Hasbrouck (2009) trading-cost model, applied to turnover measured against prior-month weights drifted by realized returns (Section II.C). Volatility (vol.) is the annualized standard deviation of monthly returns. Sharpe ratios are monthly mean over monthly standard deviation, annualized by $\sqrt{12}$, with no risk-free adjustment (dollar-neutral portfolios). Maximum drawdown (Max. DD) is the largest peak-to-trough decline of the cumulative arithmetic return path over the period. Turnover is annualized one-sided turnover; Cost is the annualized average transaction-cost drag. Agents 1 and 3 hold identical A and B strategies; their rows are repeated for completeness.

Panel A: Validation period (2003–2012)

Agent	Strategy	Gross mean	Net mean	Gross vol.	Gross Sharpe	Net Sharpe	Gross max. DD	Net max. DD	Turnover	Cost
1	A	8.52%	5.53%	6.93%	1.23	0.80	−13.50%	−14.83%	7.41	2.99%
1	B	8.52%	5.53%	6.93%	1.23	0.80	−13.50%	−14.83%	7.41	2.99%
2	A	9.04%	5.90%	7.39%	1.22	0.80	−11.89%	−14.70%	7.76	3.14%
2	B	6.87%	4.96%	6.24%	1.10	0.80	−13.73%	−14.57%	4.63	1.91%
3	A	6.44%	3.07%	8.40%	0.77	0.36	−26.38%	−31.22%	8.74	3.37%
3	B	6.44%	3.07%	8.40%	0.77	0.36	−26.38%	−31.22%	8.74	3.37%
4	A	7.34%	4.64%	6.52%	1.13	0.71	−12.67%	−14.17%	6.72	2.70%
4	B	8.41%	6.08%	7.38%	1.14	0.82	−14.27%	−15.52%	5.60	2.34%

Panel B: Locked test period (2013–2024)

Agent	Strategy	Gross mean	Net mean	Gross vol.	Gross Sharpe	Net Sharpe	Gross max. DD	Net max. DD	Turnover	Cost
1	A	6.41%	3.65%	10.69%	0.60	0.34	−26.37%	−29.95%	7.63	2.76%
1	B	6.41%	3.65%	10.69%	0.60	0.34	−26.37%	−29.95%	7.63	2.76%
2	A	8.46%	5.56%	11.98%	0.71	0.46	−23.18%	−26.06%	7.91	2.90%
2	B	5.91%	4.19%	10.42%	0.57	0.40	−25.08%	−27.55%	4.66	1.73%
3	A	4.28%	1.23%	11.02%	0.39	0.11	−31.69%	−38.63%	9.03	3.05%
3	B	4.28%	1.23%	11.02%	0.39	0.11	−31.69%	−38.63%	9.03	3.05%
4	A	7.88%	5.36%	9.56%	0.82	0.56	−16.46%	−19.60%	6.82	2.52%
4	B	10.08%	7.82%	11.61%	0.87	0.67	−21.05%	−22.34%	5.73	2.26%

Table III

Factor-Adjusted Performance

This table reports annualized alphas (Panels A and B) and factor loadings (Panels C and D) from regressions of monthly strategy returns on the Fama–French three-factor model (FF3; Fama and French, 1993) and on the Fama–French five factors plus momentum (FF6), in the validation and locked test periods. t-statistics, in brackets, use heteroskedasticity- and autocorrelation-consistent (HAC) standard errors with six lags. The factors are MKT (market excess return), SMB (small minus big, size), HML (high minus low, value), RMW (robust minus weak, profitability), CMA (conservative minus aggressive, investment), and MOM (momentum); FF3 uses MKT, SMB, and HML only. The validation sample is 120 months for both models. The locked-period FF6 sample is 142 months because the momentum factor is unavailable for November–December 2024; FF3 requires no momentum factor, so its locked sample is 144 months (143 for Agent 2, whose one infeasible locked month is disclosed under the frozen attrition rule). Panels C and D report loadings from the gross-return FF6 regressions; the corresponding net-return loadings differ from the gross loadings by at most 0.02 in every cell and are omitted. Agents 1 and 3 hold identical A and B strategies; their rows are repeated for completeness. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

Panel A: Gross returns

Agent	Strategy	Model	Validation α	Locked α	Validation adj. R^2	Locked adj. R^2
1	A	FF3	9.92%*** [5.63]	7.55%*** [3.08]	0.15	0.30
1	A	FF6	9.34%*** [5.56]	4.90%** [2.51]	0.30	0.44
1	B	FF3	9.92%*** [5.63]	7.55%*** [3.08]	0.15	0.30
1	B	FF6	9.34%*** [5.56]	4.90%** [2.51]	0.30	0.44
2	A	FF3	11.01%*** [5.87]	9.99%*** [4.36]	0.29	0.30
2	A	FF6	10.80%*** [5.85]	7.19%*** [3.89]	0.32	0.49
2	B	FF3	8.10%*** [5.41]	6.95%*** [3.13]	0.14	0.37
2	B	FF6	7.86%*** [4.67]	4.53%** [2.57]	0.25	0.51
3	A	FF3	8.04%*** [2.73]	5.33%** [2.02]	0.18	0.31
3	A	FF6	7.93%*** [3.77]	2.07% [1.01]	0.45	0.50
3	B	FF3	8.04%*** [2.73]	5.33%** [2.02]	0.18	0.31
3	B	FF6	7.93%*** [3.77]	2.07% [1.01]	0.45	0.50
4	A	FF3	9.17%*** [6.39]	7.64%*** [4.19]	0.38	0.28
4	A	FF6	9.13%*** [5.73]	5.13%*** [3.54]	0.40	0.52
4	B	FF3	10.39%*** [6.64]	9.57%*** [4.33]	0.34	0.32
4	B	FF6	10.19%*** [6.10]	6.86%*** [3.74]	0.36	0.55

Panel B: Net returns

Agent	Strategy	Model	Validation α	Locked α	Validation adj. R^2	Locked adj. R^2
1	A	FF3	6.93%*** [3.85]	4.81%* [1.95]	0.15	0.30
1	A	FF6	6.39%*** [3.80]	2.23% [1.12]	0.30	0.44
1	B	FF3	6.93%*** [3.85]	4.81%* [1.95]	0.15	0.30
1	B	FF6	6.39%*** [3.80]	2.23% [1.12]	0.30	0.44
2	A	FF3	7.87%*** [4.11]	7.12%*** [3.11]	0.28	0.31
2	A	FF6	7.67%*** [4.12]	4.39%** [2.35]	0.32	0.49
2	B	FF3	6.19%*** [4.12]	5.24%** [2.34]	0.13	0.38
2	B	FF6	5.95%*** [3.59]	2.84% [1.59]	0.25	0.51
3	A	FF3	4.69% [1.54]	2.26% [0.85]	0.17	0.32
3	A	FF6	4.62%** [2.13]	-0.96% [-0.46]	0.45	0.50
3	B	FF3	4.69% [1.54]	2.26% [0.85]	0.17	0.32
3	B	FF6	4.62%** [2.13]	-0.96% [-0.46]	0.45	0.50
4	A	FF3	6.48%*** [4.49]	5.08%*** [2.79]	0.38	0.28
4	A	FF6	6.48%*** [4.08]	2.61%* [1.79]	0.41	0.53
4	B	FF3	8.06%*** [5.14]	7.28%*** [3.28]	0.34	0.32
4	B	FF6	7.90%*** [4.75]	4.60%** [2.48]	0.37	0.55

Panel C: FF6 loadings, gross returns, validation period

Agent	Strategy	MKT	SMB	HML	RMW	CMA	MOM
1	A	-0.10	-0.12	-0.09	-0.05	0.44	0.12
1	B	-0.10	-0.12	-0.09	-0.05	0.44	0.12
2	A	-0.12	-0.17	-0.10	0.05	-0.05	0.10
2	B	-0.08	-0.14	-0.07	-0.06	0.28	0.11
3	A	-0.06	-0.12	-0.05	-0.02	0.05	0.29
3	B	-0.06	-0.12	-0.05	-0.02	0.05	0.29
4	A	-0.10	-0.18	-0.19	-0.01	0.02	0.08
4	B	-0.10	-0.19	-0.18	0.04	-0.02	0.10

Panel D: FF6 loadings, gross returns, locked test period

Agent	Strategy	MKT	SMB	HML	RMW	CMA	MOM
1	A	-0.13	-0.24	0.22	0.66	-0.06	0.19
1	B	-0.13	-0.24	0.22	0.66	-0.06	0.19
2	A	-0.20	-0.28	0.17	0.84	-0.23	0.15
2	B	-0.12	-0.33	0.20	0.63	-0.04	0.16
3	A	-0.10	-0.25	0.19	0.65	-0.03	0.31
3	B	-0.10	-0.25	0.19	0.65	-0.03	0.31
4	A	-0.08	-0.23	0.12	0.77	-0.23	0.14
4	B	-0.10	-0.35	0.12	0.91	-0.26	0.13

Table IV

Signal and Economic-Family Overlap

This table reports pairwise overlap of the agents' frozen signal sets and compares observed overlap with three frozen selection nulls. Panel A reports pairwise Jaccard similarity of signal sets by objective—the number of signals two sets share divided by the number of signals in their union; economic-family Jaccard (based on the signal-family classification of the Open Source Asset Pricing (OSAP) library) is identical to signal-level Jaccard for every pair. Panel B compares the observed mean pairwise Jaccard with 100,000-draw Monte Carlo selection nulls: uniform draws from the common eligible universe of 126 signals; draws matched to the observed economic-family composition; and draws restricted to signals passing a development-period performance screen. The standardized difference (Std. diff.) is $(\text{observed} - \text{null mean})/\text{null SD}$, where the null mean and null SD are the mean and standard deviation of the measure across the null draws; the upper-tail p-value is the fraction of draws at or above the observed value, computed with the add-one estimator (one plus the number of such draws, divided by one plus the number of draws).

Panel A: Pairwise Jaccard overlap of frozen signal sets

Agent pair	Strategy A	Strategy B
1–2	0.50	1.00
1–3	0.40	0.40
1–4	0.67	0.67
2–3	0.75	0.40
2–4	0.67	0.67
3–4	0.50	0.50
Mean	0.58	0.61

Panel B: Observed mean overlap relative to frozen selection nulls (100,000 draws each)

Objective	Null	Null mean	Null SD	Observed	Std. diff.	Upper-tail p
A	Uniform	0.014	0.022	0.581	26.27	<0.0001
A	Family-matched	0.250	0.046	0.581	7.18	0.0003
A	Performance-screened	0.056	0.042	0.581	12.50	<0.0001
B	Uniform	0.014	0.022	0.606	27.43	<0.0001
B	Family-matched	0.252	0.048	0.606	7.35	0.0003
B	Performance-screened	0.056	0.042	0.606	13.10	<0.0001

Table V

Holdings and Desired-Trade Similarity Across Periods

This table reports the time-series means of the cross-agent crowding measures in the validation period (2003–2012) and the locked test period (2013–2024), averaged over the six agent pairs. H is the signed holdings cosine, the cosine similarity between two strategies’ signed portfolio weight vectors. A strategy’s desired trade in a security at a rebalance is the gap between its new target weight and its return-drifted prior weight; SD (OD) is the same-direction (opposite-direction) desired-trade overlap, the common component of the pair’s desired buys plus the common component of their desired sells (buys of one against sells of the other, for OD), normalized by the smaller strategy’s total absolute desired trade; T is the cosine similarity of the signed desired-trade vectors; leg overlaps are minimum-overlap coefficients of long and short memberships (shared names divided by the smaller leg’s count). Common trade per dollar is the same-direction desired trade coming from two or more strategies, per dollar of capital; P (liquidity-scaled common trading) additionally scales each security’s common desired trade by its formation-month dollar volume, per dollar of capital, under the pre-specified hypothetical equal allocation of assets under management (AUM) across the four agents. Principal measures are those designated in the pre-specified measurement plan, which is frozen in the archived workflow record.

Panel A: Strategy A

Measure	Class	Validation	Locked test	Difference	Relative change
H (signed holdings cosine)	Principal	0.596	0.588	−0.008	−1.3%
SD (same-direction trade overlap)	Principal	0.439	0.432	−0.008	−1.8%
P (liquidity-scaled common trading, $\times 10^{-6}$)	Principal	8.94	3.31	−5.62	−62.9%
Common trade per dollar	Supplementary	0.518	0.524	+0.006	+1.1%
T (trade cosine)	Supplementary	0.463	0.455	−0.007	−1.6%
OD (opposite-direction overlap)	Supplementary	0.016	0.016	−0.000	−0.2%
Long-leg overlap	Supplementary	0.484	0.481	−0.004	−0.7%
Short-leg overlap	Supplementary	0.509	0.494	−0.015	−3.0%

Panel B: Strategy B

Measure	Class	Validation	Locked test	Difference	Relative change
H (signed holdings cosine)	Principal	0.571	0.561	−0.010	−1.8%
SD (same-direction trade overlap)	Principal	0.477	0.465	−0.012	−2.6%
P (liquidity-scaled common trading, $\times 10^{-6}$)	Principal	7.10	2.54	−4.56	−64.2%
Common trade per dollar	Supplementary	0.418	0.421	+0.002	+0.6%
T (trade cosine)	Supplementary	0.427	0.419	−0.008	−1.9%
OD (opposite-direction overlap)	Supplementary	0.020	0.020	+0.001	+3.2%
Long-leg overlap	Supplementary	0.489	0.482	−0.007	−1.3%
Short-leg overlap	Supplementary	0.509	0.492	−0.017	−3.4%

Table VI

Crowding Measures Relative to Frozen Benchmarks

This table compares the observed principal crowding measures with two frozen randomization benchmarks in both evaluation periods. The standardized benchmark (pre-specified in the measurement protocol) redraws signal sets and builds all benchmark portfolios under a single common construction; the agent-matched benchmark (specified post hoc but frozen before locked-sample access) redraws signal sets while holding each agent’s own frozen construction code fixed. The principal measures are defined in Table V: H is the signed holdings cosine, SD the same-direction desired-trade overlap, and P the liquidity-scaled common trading per dollar of capital (reported $\times 10^{-6}$); Obj. denotes the strategy objective (A, gross Sharpe; B, net Sharpe). Each benchmark uses 1,000 frozen draws per objective; for the agent-matched liquidity-scaled common trading measure P in the locked test period, 869 draws are feasible (draws with no common trading in a month are infeasible for the dollar-volume denominator). Std. diff. is the standardized difference (observed – null mean)/null SD; the percentile is the observed value’s location in the null distribution; the upper-tail p-value is computed with the add-one estimator (one plus the number of draws at or above the observed value, divided by one plus the number of draws). Panel C reports the protocol-defined classification (Appendix C.4): *elevated under both* if the observed value lies in the 5% upper tail of both null distributions, *elevated under one* if in the upper tail of exactly one; the 5% thresholds are nominal and unadjusted for the multiplicity of measures, objectives, periods, and benchmarks.

Panel A: Standardized-construction benchmark

Period	Obj.	Measure	Observed	Null mean	Null SD	Std. diff.	Percentile	Upper-tail p
Validation	A	H	0.596	0.456	0.066	2.13	0.976	0.025
Validation	A	SD	0.439	0.649	0.028	−7.34	0.000	1.000
Validation	A	P ($\times 10^{-6}$)	8.94	11.87	0.77	−3.83	0.000	1.000
Validation	B	H	0.571	0.452	0.065	1.85	0.964	0.037
Validation	B	SD	0.477	0.653	0.027	−6.49	0.000	1.000
Validation	B	P ($\times 10^{-6}$)	7.10	12.23	0.87	−5.88	0.000	1.000
Locked test	A	H	0.588	0.469	0.068	1.74	0.956	0.045
Locked test	A	SD	0.432	0.655	0.028	−7.88	0.000	1.000
Locked test	A	P ($\times 10^{-6}$)	3.31	3.60	0.36	−0.80	0.208	0.792
Locked test	B	H	0.561	0.464	0.066	1.47	0.928	0.073
Locked test	B	SD	0.465	0.659	0.027	−7.15	0.000	1.000
Locked test	B	P ($\times 10^{-6}$)	2.54	3.70	0.38	−3.05	0.000	1.000

Panel B: Agent-matched construction benchmark

Period	Obj.	Measure	Observed	Null mean	Null SD	Std. diff.	Percentile	Upper-tail p
Validation	A	H	0.596	0.330	0.047	5.71	1.000	0.001
Validation	A	SD	0.439	0.268	0.033	5.28	1.000	0.001
Validation	A	P ($\times 10^{-6}$)	8.94	4.13	1.95	2.46	0.976	0.025
Validation	B	H	0.571	0.317	0.048	5.33	1.000	0.001
Validation	B	SD	0.477	0.290	0.038	4.99	0.999	0.002
Validation	B	P ($\times 10^{-6}$)	7.10	3.65	1.72	2.01	0.965	0.036
Locked test	A	H	0.588	0.347	0.050	4.78	1.000	0.001
Locked test	A	SD	0.432	0.274	0.035	4.45	1.000	0.001
Locked test	A	P ($\times 10^{-6}$)	3.31	1.42	0.60	3.16	0.997	0.005
Locked test	B	H	0.561	0.333	0.050	4.55	1.000	0.001
Locked test	B	SD	0.465	0.297	0.041	4.14	0.997	0.004
Locked test	B	P ($\times 10^{-6}$)	2.54	1.28	0.53	2.38	0.984	0.017

Panel C: Classification by benchmark construction

Objective	Measure	Validation	Locked test	Stability
A	H	Elevated under both	Elevated under both	Unchanged
A	SD	Elevated under one	Elevated under one	Unchanged
A	P	Elevated under one	Elevated under one	Unchanged
B	H	Elevated under both	Elevated under one	Changed
B	SD	Elevated under one	Elevated under one	Unchanged
B	P	Elevated under one	Elevated under one	Unchanged

Table VII

Return Commonality Across Periods

This table reports mean pairwise correlations of monthly strategy returns across the six agent pairs, for raw returns and FF6-residual returns, gross and net of costs, in the validation and locked test periods. FF6 denotes the Fama–French five factors plus momentum; FF6 residuals are from the regressions of Table III, and the locked-period residual sample is 142 months.

Objective	Returns	Correlation type	Validation	Locked test	Difference
A	Gross	Raw	0.798	0.869	+0.071
A	Gross	FF6 residual	0.766	0.819	+0.052
A	Net	Raw	0.798	0.867	+0.069
A	Net	FF6 residual	0.765	0.817	+0.052
B	Gross	Raw	0.795	0.867	+0.073
B	Gross	FF6 residual	0.758	0.815	+0.057
B	Net	Raw	0.794	0.867	+0.072
B	Net	FF6 residual	0.757	0.815	+0.058

Appendix A. Notation and Proofs

This appendix collects the model notation and proofs of Lemmas 1–2 and Propositions 1–3.

A.1. Notation

Symbol	Meaning
$x \in [0, 1]$; A ; E_j	AI share of active equity; total active equity; manager j 's equity ($\sum_j E_j = xA$)
$\eta \in [0, 1]$; M ; w_m ; HHI	Common-direction risk share; number of AI systems; system equity shares; $\sum_m w_m^2$
μ ; z_c ; ζ_m ; ϵ	True alpha; common error $N(0, \sigma_c^2)$; composite system-specific error $N(0, \sigma_\zeta^2)$ (absorbing sleeve alphas exactly); payoff noise, variance σ^2
$\rho_e = \sigma_c^2 / (\sigma_c^2 + \sigma_\zeta^2)$	Forecast-error correlation across systems
L ; γ ; $\theta \equiv \gamma\sigma^2$	Leverage cap; risk-aversion scale; risk-bearing coefficient
$\Gamma_j = \gamma / (LE_j)$ (Assumption 2, A-RC)	State-dependent absolute risk aversion
$u_j = \hat{\alpha}_m / \theta$; u_{j0} ; u_0	Leverage utilization; its date-0 value; equity-weighted mean $\sum_j E_j u_{j0} / \sum_j E_j$
$u_E^2 = \sum_j E_j u_{j0}^2 / \sum_j E_j$	Equity-weighted second moment of utilization (wealth-drag argument in (V.7))
$\Phi = \eta x LA / \theta$; $a = A / \theta$	AI demand slope; equity-scaled capacity
$k_R(x) = k_0 + (1 - x)a$; $\lambda = 1/k_R$	Residual depth (reallocation path); inverse depth
p_0 ; $\delta = p_0 - p_1$; $y = \delta / p_0$	Date-0 price; date-1 decline; normalized decline
π ; ϕ	Catalyst probability; proportional redemption fraction
τ ; τ_s	Routine and stress trade synchronization ($\tau_s = 1$ fixed in the formal stress model)
$\bar{\delta}_j = (1 - u_{j0})\theta / \sqrt{\eta}$	Manager j 's constraint boundary (also j 's buy/sell boundary under Assumption 3)
$u_b = 1/\sqrt{\theta L}$	Buy threshold (Assumption 3 $\iff u_b \geq 1$)
$g = \lambda \eta L^2 x A$; $\kappa(\phi) = g(1 - \phi)$	Feedback-intensity statistic; surviving-sector gain
$\mathcal{A} = 1/(1 - \kappa(\phi))$	First-order amplification of interior declines
H ; SD ; P^{common}	Frozen empirical similarity measures

A.2. Proofs

Throughout, $\theta \equiv \gamma\sigma^2$. Every derivation below is additionally verified symbolically and numerically in a companion verification script that is part of the archived workflow record.

Lemma 1 (Demand aggregation). *Under Assumption 2 (A-RC), aggregate unconstrained AI demand for the composite at price p is $E_{AI}(p) = \Phi(\mu + z_c + \bar{\zeta} - p)$ with $\Phi = \eta x LA / \theta$ and $\bar{\zeta} = \sum_m w_m \zeta_m$, exactly.*

Proof. System m delivers the composite (V.1). With the sleeve alphas absorbed into ζ_m

(Section V.A), the perceived alpha of s_m at composite price p is exactly $\hat{\alpha}_m = \sqrt{\eta}(\mu + z_c + \zeta_m - p)$. Manager j 's risk-capital scale is $r_j \equiv LE_j$, an input to the program (V.2). Applying the first-order condition once, $b_j^* = r_j \hat{\alpha}_m / (\gamma \sigma^2)$, and the manager's exposure to the priced direction s_c is $\sqrt{\eta} b_j^*$ — one factor $\sqrt{\eta}$ is the composite's loading on s_c ; the second converts the composite position into s_c exposure. Summing over managers, $\sqrt{\eta} \sum_j b_j^* = \eta L (\sum_j E_j) (\mu + z_c + \bar{\zeta} - p) / \theta$ with $\sum_j E_j = xA$, which is the stated demand. The aggregation is exact — no expectation over sleeve alphas is taken. If managers could trade s_c directly, η would not appear: the “system outputs one portfolio, the manager scales it” restriction is the key institutional assumption. ■

Lemma 2 (The manager's stress program). *Under the proportional redemption and a subsequent decline δ : (i) $E_j^+ = (1 - \phi)E_j$, $b_j^+ = (1 - \phi)b_j$, and $u_j^+ = u_{j0}$; (ii) on the unconstrained branch, $u_j(\delta) = u_{j0} + \sqrt{\eta}\delta/\theta$ and $E_j(\delta) = (1 - \phi)E_j \exp\{-\sqrt{\eta}L(u_{j0}\delta + \sqrt{\eta}\delta^2/(2\theta))\}$, so the book $b_j = LE_j(\delta)u_j(\delta)$ satisfies $db_j/d\delta = \sqrt{\eta}LE_j(\delta)[1/\theta - Lu_j(\delta)^2]$: the manager buys iff $u_j(\delta) < u_b = 1/\sqrt{\theta L}$; (iii) manager j 's cap binds iff $\delta \geq \bar{\delta}_j = (1 - u_{j0})\theta/\sqrt{\eta}$, with the book continuous at $\bar{\delta}_j$; under Assumption 3 (A-MOD), $\bar{\delta}_j$ is simultaneously j 's buy/sell boundary and constraint boundary; without A-MOD the two split ($\delta_{b,j} < \bar{\delta}_j$) and an unconstrained-selling region appears; (iv) once at the cap, $E_j(\delta) = E_j e^{-\sqrt{\eta}L\delta}$ and the composite-direction sales rate per unit decline is $\eta L^2 E_j(\delta)$.*

Proof. (i) The redemption removes the fraction ϕ of equity at pre-decline marks. Both the interior book $b^* = LEu$ and the cap LE scale one-for-one with E , so the optimal response is to liquidate exactly the fraction ϕ of the book; utilization, which is equity-free, is unchanged, and an at-cap manager stays at the cap. (ii) A decline δ raises perceived alpha by $\sqrt{\eta}\delta$, giving $u_j(\delta)$; marking the book, $dE/d\delta = -\sqrt{\eta}b = -\sqrt{\eta}LEu(\delta)$, a linear ODE whose exact solution is the stated exponential. Differentiating $b = LEu$: $db/d\delta = LE'u + LEu' = \sqrt{\eta}LE[1/\theta - Lu^2]$. Thus rising utilization does not itself imply buying; the manager buys while the utilization gain outruns the equity loss, i.e., iff $u < u_b$. A-MOD ($\theta L \leq 1 \iff u_b \geq 1$) makes the buy region cover the whole unconstrained region. (iii)

Because $u_j(\delta)$ is equity-free, equity dynamics do not move the constraint boundary: $u_j(\bar{\delta}_j) = 1$ exactly, giving $\bar{\delta}_j$; continuity holds since $LE(\bar{\delta}_j)u(\bar{\delta}_j) = LE(\bar{\delta}_j)$. With heterogeneous u_{j0} , the $\bar{\delta}_j$ are ordered inversely to u_{j0} : entry is staggered from $\min_j \bar{\delta}_j$. (iv) At the cap, $b = LE$ and $dE/d\delta = -\sqrt{\eta}LE$ give the exponential path; sales projected on the composite are $\sqrt{\eta}|db/d\delta| = \eta L^2 E$. The single L in Φ (from A-RC at date 0) and the L^2 here (cap slope $\times$ book scale) measure different objects. ■

Proof of Proposition 1. (a) Differentiate (V.3): $\partial p_0/\partial \Phi = k_R(\mu + z_c + \bar{\zeta})/(k_R + \Phi)^2 > 0$ and $\partial p_0/\partial k_R < 0$. η and L enter only through Φ (positively). For x , both channels push the same way: $d\Phi/dx > 0$ and $dk_R/dx = -a < 0$, so dp_0/dx strictly exceeds its fixed-depth counterpart (where $dk_R/dx = 0$). For A : $\Phi \propto A$ and $k_R = k_0 + (1-x)A/\theta$, and substituting, $dp_0/dA > 0$ iff $k_0 > 0$, with exact invariance at $k_0 = 0$ since A then scales both sides of the market. (b) Differentiate (V.4) in Φ holding k_R : $\partial \text{MSE}/\partial \Phi = 2k_R(\Phi S - k_R\sigma_\mu^2)/(k_R + \Phi)^3$, whose sign is $\text{sign}(\Phi S - k_R\sigma_\mu^2)$; the depth channel obeys $\partial \text{MSE}/\partial k_R = -(\Phi/k_R) \partial \text{MSE}/\partial \Phi$, and combining along the path gives the total derivative displayed in the proposition, which preserves the sign. At $\eta = 0$, $\Phi = 0$ and $\text{MSE} = \sigma_\mu^2$ regardless of x . When $S \rightarrow 0$, MSE is strictly decreasing in Φ . Otherwise, along the path $\Phi(x)$ is strictly increasing and $k_R(x)$ nonincreasing, so the marginal term $\Phi S - k_R\sigma_\mu^2$ is strictly increasing in x and changes sign at most once; it is negative at $x = 0$ ($\Phi(0) = 0$), so an interior optimum exists iff it is positive at $x = 1$, i.e., iff $\Phi(1)S > k_R(1)\sigma_\mu^2$, in which case the unique crossing solves $\Phi(x^*)S = k_R(x^*)\sigma_\mu^2$; when the sign change does not occur within $[0, 1]$, MSE is strictly decreasing on the whole feasible range. Because the reallocation path lowers k_R as x rises, the crossing (when interior) occurs at smaller x than under fixed depth (benchmark illustration, where both are interior: $x^* = 0.347$ versus 0.520). With unequal system shares, $\text{Var}(\bar{\zeta}) = \sigma_\zeta^2 \text{HHI}$, giving S . ■

Proof of Proposition 2. Immediate from (V.3): conditional on $\mu + z_c + \bar{\zeta} > 0$, an increase in x (or η , or A with $k_0 > 0$) raises p_0 ; an incumbent's realized return over the entry window contains $\Delta p_0 > 0$, which is price pressure, not payoff news. Expected future return falls by the same displacement (second expression in (V.3)). Persistence: the residual sector

does not condition on prices, so displacement is not arbitrated within the window; in the temporary-impact limit the displacement decays before measurement and the term vanishes.

■

Proof of Proposition 3. (a) By Lemma 2(ii), survivor j 's demand path is $b_j(\delta) = LE_j(\delta)u_j(\delta)$ with slope at $\delta = 0$ equal to $\sqrt{\eta}LE_j[1/\theta - Lu_{j0}^2]$. Summing over managers, the aggregate slope is $\sqrt{\eta}L(\sum_j E_j)[1/\theta - L\bar{u}_E^2]$ — the equity-weighted *second* moment appears because each manager's wealth drag is quadratic in its own utilization. Clearing $k_R(p_0 - \delta_u) =$ aggregate survivor demand, linearized in δ , gives (V.7) (using $\eta LxA/\theta = \Phi$). Absorption $\delta_u < \phi p_0$ holds iff the wealth-drag factor is positive, i.e., $\theta L\bar{u}_E^2 < 1$; A-MOD implies it on the unconstrained region since every $u_{j0} < 1$ there. Jensen gives $\bar{u}_E^2 \geq \bar{u}_0^2$ with equality iff utilization is homogeneous, so dispersion weakly deepens δ_u . The homogeneous case $u_{j0} \equiv u_0$ recovers the representative-agent formula with u_0^2 .

(b) *Exact.* Let $f(y) = 1 - y - (1 - \phi)e^{-gy}$. Then $f(0) = \phi > 0$, $f(1) = -(1 - \phi)e^{-g} < 0$, and $f''(y) = -g^2(1 - \phi)e^{-gy} < 0$: f is strictly concave with a sign change, so (V.6) has a unique root $y^* \in (0, 1)$ for every finite g and $\phi \in (0, 1)$, and $y = 1$ is never an exact solution. Implicit differentiation gives $dy^*/dg > 0$ and $dy^*/d\phi > 0$, with $y^* \rightarrow 1$ only as $g \rightarrow \infty$.

Global convergence of the exact recursion. Let $T(y) = 1 - (1 - \phi)e^{-gy}$, so that $y_{n+1} = T(y_n)$ and $f(y) = T(y) - y$. T is continuous and strictly increasing ($T'(y) = g(1 - \phi)e^{-gy} > 0$). Since $f(0) = \phi > 0$, f is strictly concave, and y^* is its unique root in $(0, 1)$, we have $f(y) > 0$ for all $y \in [0, y^*)$: $T(y) > y$ below the root. Monotonicity of T gives, for any $y < y^*$, $T(y) < T(y^*) = y^*$. Hence, starting from the economically relevant initial condition $y_0 = 0$ (no decline before the catalyst), induction yields $y_n < y_{n+1} < y^*$ for every n : the sequence is strictly increasing and bounded above by y^* , so it converges to some limit $\bar{y} \leq y^*$. Continuity of T implies $\bar{y} = T(\bar{y})$, i.e., $\bar{y}$ is a fixed point, and uniqueness of the fixed point forces $\bar{y} = y^*$. This is global convergence from $y_0 = 0$ for every finite g and $\phi \in (0, 1)$, not merely local stability; locally, in addition, $T'(y^*) = 1 + f'(y^*) < 1$, since strict concavity and $f(0) > 0$ force $f'(y^*) < 0$. The symbolic and numerical sweeps in the verification script confirm this

convergence across the tested parameter grid as a check on, not a substitute for, the argument.

Linear approximation. Linearizing the constrained clearing condition $p_0 - \delta = (1 - \phi)p_0 e^{-\sqrt{\eta}L\delta}$ at $\delta = 0$ and using the clearing identities gives the interior fixed point $\delta^* = \phi p_0 / (1 - \kappa(\phi))$; it is admissible (weakly below the position floor p_0) iff $\kappa(\phi) \leq 1 - \phi \iff g \leq 1$, with equality throughout at $g = 1$, giving (V.8); continuity at $g = 1$ follows from $\phi p_0 / (1 - (1 - \phi)) = p_0$. The corner branch is derived, not imposed: iterating $\delta_{n+1} = \min\{\phi p_0 + \kappa(\phi)\delta_n, p_0\}$ from $\delta_0 = 0$, the recursion converges to the interior fixed point when $g < 1$ and reaches the floor in finitely many rounds when $g > 1$ (either $\kappa(\phi) \geq 1$, divergence, or $\kappa(\phi) < 1 < g$, a convergent recursion whose fixed point exceeds the floor). Remark A.1 states the $g = 1$ equality case and the iteration-divergence threshold, and fixes the economic status of the recursion. Monotonicity and convexity of $\mathcal{A}$ in η , L , x on the interior region follow by differentiation; $d\delta/d\phi = (1 - g)p_0 / (1 - g(1 - \phi))^2 > 0$ for $g < 1$. (The *total* decline is not globally convex in L : entry displacement p_0 is concave in capacity and dominates at small L ; convexity claims attach to $\mathcal{A}$, monotonicity to δ^* .)

(c) Immediate from Lemma 2(iii): the boundaries $\bar{\delta}_j$ are ordered inversely to u_{j0} , entry begins at $\min_j \bar{\delta}_j$, and each crossing adds that manager's $\eta L^2 E_j$ to the induced-sales slope, so amplification phases in. Absorption of small catalysts follows from part (a) applied on $\delta < \min_j \bar{\delta}_j$. ■

Remark A.1 (Economic status of the recursion; admissibility versus iteration divergence; the $g = 1$ equality case). The equilibrium object throughout is the *static* date-1 clearing condition (V.6) (or its linearization); the round-by-round recursion is a *constructive fixed-point device* for computing it, not a model of real-time trading dynamics. Its rounds do admit a stylized within-round reading — mark positions at the current price, collect the constraint-induced sales, let the residual sector absorb them into a new price — but the model attaches no calendar time or intra-date equilibrium concept to a round, so we do not describe the recursion's convergence as economic “dynamic stability.” With that status fixed, two thresholds of the linear-feedback approximation must not be conflated. $g = 1$

is the *position-floor (admissibility) boundary*: the point at which the interior fixed point $\phi p_0/(1 - \kappa(\phi))$ reaches the floor p_0 . $\kappa(\phi) = 1$, i.e., $g = 1/(1 - \phi)$, is the *iteration-divergence threshold*: the gain at which the unconstrained linear iteration for a realized catalyst ϕ ceases to contract — a property of the algorithm, not an asserted real-time instability. It lies strictly inside the supercritical region; the two coincide only in the small-catalyst limit $\phi \rightarrow 0$. In particular, at $g = 1$ the iteration gain is $\kappa = 1 - \phi < 1$ — still a contraction — and its fixed point is exactly the floor: the iterates approach full liquidation at $g = 1$ only *asymptotically*, never in finitely many rounds, since each round contracts the remaining gap by $(1 - \phi)$: $1 - y_N = (1 - \phi)^N > 0$. The complete taxonomy is: $g < 1$, interior fixed point, floor never binds; $g = 1$, asymptotic approach to the floor; $g > 1$, floor reached in finitely many rounds. $\kappa(\phi) = 1$ separates the two routes by which the supercritical iteration reaches the floor (divergence versus an inadmissible fixed point) without changing the equilibrium outcome.

Remark A.2 (Attribution of leverage effects). The factor L in Φ (hence in Propositions 1 and 2) is generated by Assumption 2 (A-RC) and disappears under $\Gamma_j = \gamma/E_j$; the factor L^2 in g (hence the amplification analysis of Proposition 3(b)) is generated by the binding leverage constraint and survives any specification of Γ_j (Section V.D).

A.3. Model Objects and Their Empirical Counterparts

Table A.I — Model Objects and Their Empirical Counterparts

This table maps the model’s objects to the experiment’s frozen measures. The rightmost column records the epistemic status of each mapping: **empirical counterpart or scenario discipline** (the statistic is measured; the parameter is not structurally identified), **scenario input**, **external calibration**, or **supplementary consistency check**. Model symbols are defined in the notation glossary of Appendix A.1; the empirical measures H (signed holdings cosine), SD (same-direction desired-trade overlap), and liquidity-scaled common trading per dollar are defined in Table V. FF6 denotes the Fama–French five factors plus momentum (Table III); AUM is assets under management.

Model object	Empirical counterpart	Status
η (similarity)	Signed holdings cosine $H \approx 0.58$; scenario values $\{0.33, 0.58, 0.81\}$ from the agent-matched null, observed, and standardized construction	Counterpart / scenario anchor; no exact identification
τ (routine sync.)	SD ≈ 0.44 – 0.48 relative to H ; construction alone moves SD by ≈ 0.36 – 0.38 at fixed signals	Empirical counterpart / scenario discipline for $\tau < 1$; τ itself not structurally identified
τ_s (stress sync.)	—	Fixed at $\tau_s = 1$ in the formal stress model; imperfect synchronization discussed qualitatively
ρ_e (error corr.)	Not identified by the frozen measures (score correlation 0.82 disciplines upstream signal similarity, not the z_c/ζ_m split)	Sensitivity grid $\rho_e \in \{0, \frac{1}{2}, 1\}$
Common-flow scale	Liquidity-scaled common trading per dollar, objectives A/B: validation 8.9×10^{-6} / 7.1×10^{-6} ; locked test 3.3×10^{-6} / 2.5×10^{-6}	Empirical discipline for common-flow scenarios; no structural capacity-use parameter identified
Return commonality	Average pairwise return correlation ≈ 0.80 (validation) $\rightarrow 0.87$ (locked test); FF6-residual correlation $\approx 0.76 \rightarrow 0.82$	Supplementary consistency check: model-implied manager correlation = $\eta + (1 - \eta) \times$ sleeve correlation
λ (inverse depth)	$\lambda = 1/(\zeta_{\perp} W)$; $\zeta_{\perp} \approx 1$ (range 0.5–2)	External; scenario grid
x, A, k_0	AI share of active AUM; anomaly-space active capital; market-maker depth	Counterfactual axis; external
$L; u_{j0}$	Fund-level leverage ranges (≈ 2 – $6\times$); utilization distribution	Scenario grids; not identified
π, ϕ	Historical deleveraging episodes as reference points	Scenario parameters; no frequency claims
HHI	AI system/vendor market shares	External; comparative static

Appendix B. Offsetting AI Strategies (Negative Average Similarity)

The baseline factor representation (V.1) requires $\eta \geq 0$: a common component loaded with weight $\sqrt{\eta}$ exists only for nonnegative average similarity. Negative average pairwise similarity — offsetting AI strategies — is handled by working directly with the pairwise-cosine structure rather than a factor model. Let the M AI systems' unit-norm strategy weight vectors have equicorrelated pairwise cosines $\bar{\eta}$. Positive semidefiniteness of the resulting Gram matrix requires $\bar{\eta} \geq -1/(M-1)$, so average similarity can be only mildly negative for large M , and no single-factor representation of the form (V.1) is claimed or needed. The relevant aggregate object is the squared norm of summed demand: with each system contributing a unit-norm demand direction q_m ,

$$\left\| \sum_m q_m \right\|^2 = M(1 + (M-1)\bar{\eta}) < M \quad \text{for } \bar{\eta} < 0,$$

so offsetting strategies partially internalize one another's order flow before it reaches the residual sector. In the aggregate clearing condition this acts as *added effective depth*: common price pressure per unit of AI capital is scaled down relative to the independent case ($\bar{\eta} = 0$), the entry displacement analogue of (V.3) shrinks, and the feedback statistic analogue of (V.5) falls. The qualitative conclusion is one-directional and unsurprising — negative similarity reduces common-flow pressure on both the routine and stress margins — and no baseline result (Propositions 1–3) relies on this extension; the main text (Section V) accordingly maintains $\eta \geq 0$ throughout.

Appendix C. Crowding Measure Definitions and Benchmark Construction

This appendix collects the formal definitions of the cross-agent similarity measures (Section IV.A) and the construction details of the randomized benchmarks (Section II.E). All measures and both portfolio-construction benchmarks were specified, implemented, simulated, and frozen on validation-period data before any locked-period data was accessible in the environment; locked-period values apply the identical frozen code to January 2013–December 2024.

C.1. Similarity Measures

Measures operate at four layers. Throughout, m and n index strategies under the same objective (A or B), t indexes formation months, and i indexes securities in the common eligible universe U of 126 signals’ investable cross-section.

Signal choice. For strategies m and n with selected signal sets $S_m, S_n \subseteq U$, signal overlap is the Jaccard index

$$J_{mn} = \frac{|S_m \cap S_n|}{|S_m \cup S_n|},$$

computed both on signal identities and on the signals’ OSAP economic-family classifications (the family-level index applies the same formula to multisets of family labels). The system-level statistic is the mean of J_{mn} over the six agent pairs.

Security rankings. Ranking similarity is the cross-sectional Spearman rank correlation, each formation month, between two strategies’ composite scores over the common set of eligible securities with valid scores for both; the reported statistic averages over pairs and months.

Holdings. The principal holdings measure is the signed cosine similarity between full portfolio weight vectors,

$$H_{mn,t} = \frac{\sum_i w_{i,m,t} w_{i,n,t}}{\sqrt{\sum_i w_{i,m,t}^2} \sqrt{\sum_i w_{i,n,t}^2}} \in [-1, 1],$$

with opposite-side positions in the same security scored as offsetting rather than crowded exposure. Aggregation is month-first: the cross-pair mean is computed each month, then averaged over months within the evaluation period. Supplementary leg overlaps normalize each leg to unit gross weight and report the fraction held in common at the smaller of the two weights, separately for long and short legs.

Desired trades. Using the drifted-weight convention of Section II.C, strategy m 's desired trade in security i at rebalance t is

$$\Delta_{i,m,t} = w_{i,m,t} - \tilde{w}_{i,m,t},$$

where $\tilde{w}_{i,m,t}$ is the return-drifted weight carried from $t - 1$. The principal trade measure is same-direction overlap,

$$SD_{mn,t} = \frac{\sum_i \min(\Delta_{i,m,t}^+, \Delta_{i,n,t}^+) + \sum_i \min(\Delta_{i,m,t}^-, \Delta_{i,n,t}^-)}{\min(\sum_i |\Delta_{i,m,t}|, \sum_i |\Delta_{i,n,t}|)} \in [0, 1],$$

where Δ^+ and Δ^- denote the positive and negative parts (desired buys and sells in absolute value). Opposite-direction overlap OD applies the same construction to buy-against-sell coincidences, and the signed trade cosine T applies the cosine formula to the vectors $\Delta_{.,m,t}$ and $\Delta_{.,n,t}$.

Liquidity-scaled common trading. Hypothetical capital is allocated equally across the four agents. For each security and month, the *common* component of desired buying is the portion of aggregate desired buying contributed by two or more strategies (zero when all buying in the security comes from a single strategy), and symmetrically for selling. The measure divides each security's common trade dollars by the security's formation-month dollar volume and sums across securities:

$$P_t = \frac{1}{K} \sum_i \frac{\text{CommonBuy}_{i,t} + \text{CommonSell}_{i,t}}{\text{DollarVolume}_{i,t}},$$

expressed per dollar of total system capital K ; the measure is linear in K , so results at fixed AUM levels are deterministic rescalings. The unscaled companion, common trade per dollar of capital, omits the dollar-volume denominator. P is a sum of trading-to-volume ratios under an assumed allocation—not an estimate of transaction costs, price changes, or realized market impact.

Principal versus supplementary. The three principal outcomes— H , SD , and P —were designated in advance of any benchmark comparison; leg overlaps, OD , T , common trade per dollar, and the return-correlation measures of Section IV.F are supplementary.

C.2. Signal-Selection Null Distributions

Raw signal-selection overlap (Section IV.B) is evaluated against three randomization models. Each uses 100,000 Monte Carlo repetitions per objective with seed 20260827, draws without replacement within each pseudo-strategy and independently across pseudo-agents, and matches each pseudo-agent’s number of selected signals to the corresponding actual agent. The three models differ in the pool from which signals are drawn:

1. *Uniform*: independent draws from the full 126-signal eligible universe U .
2. *Family-matched*: draws preserve each agent’s economic-family composition—each pseudo-agent draws the same number of signals from each OSAP family as the agent it shadows—so the null asks only whether agents picked the same signals within families.
3. *Performance-screened*: draws are restricted to the top development-period quartile of standalone gross Sharpe ratios, capturing one important selection pressure—a preference for signals that backtested well standalone—but not the agents’ full search over combinations, transformations, weighting rules, and net objectives (Section IV.B).

For each model the reported p-value is the Monte Carlo upper-tail probability of the observed mean pairwise Jaccard index under the null. Full lossless draws are retained in the archived Parquet files.

C.3. Portfolio-Construction Benchmarks

The two portfolio-level benchmarks share one signal-randomization design and differ only in how drawn signals become portfolios. For each objective, a pseudo-experiment replaces each agent’s selected signals with random family-matched draws from U (without replacement within a pseudo-agent, independent across pseudo-agents), and exactly 1,000 feasible four-agent pseudo-systems are retained per objective. The frozen crowding measurement code of Section C.1 is then applied to each pseudo-system, generating null distributions for H , SD , and P .

Standardized construction (pre-specified). Each pseudo-agent’s drawn signals are mapped into portfolios by a single common rule: each signal’s canonical standalone portfolio is retrieved from the centralized standalone library (built from raw OSAP, CRSP, Hasbrouck, and idiosyncratic-volatility inputs under uniform conventions—nonfinite signal values treated as missing, held-position missing returns contributing zero, iterative pro-rata application of the 5% position cap), the signed weight vectors are averaged with equal weights, and the legs are renormalized and capped under the experiment’s exposure rules. Each objective-specific benchmark initializes its own PCG64 generator with seed 20260827 before its first draw; 1,000 feasible groups were accepted per objective with zero discards.

Agent-matched construction (post hoc in pre-specification status; frozen before locked-sample access). Each pseudo-agent’s portfolio is built with that agent’s *actual frozen construction code*—its own signal transformation, combination rule, leg cutoffs, weighting scheme, no-trade bands, and hysteresis where applicable—so that only signal identities are randomized and construction is held at what the agents actually did. Each objective initializes an independent PCG64 generator with seed 20260828; exactly 1,000

feasible four-agent systems are retained per objective. A reconstruction diagnostic validates the engines: feeding each agent’s actual frozen signal set through its engine reproduces the observed holdings with a monthly cosine of one to numerical precision and maximum absolute weight differences below 5×10^{-9} . The paired comparison of Section IV.D evaluates the standardized construction on each identical attempted signal realization, so the two benchmarks differ draw-by-draw in exactly one design element. This benchmark was added after the standardized results were computed—hence the post hoc label—but its draws, seed, and implementation were frozen before any locked-period access.

C.4. Locked-Period Evaluation and Classification

For the locked test period, the frozen benchmark signal identities were retained with no replacement draws, and the identical frozen measurement code was applied to 2013–2024 under a predetermined missing-observation rule. Every benchmark pseudo-system experienced at least one locked strategy-month attrition event. Standardized-benchmark H and SD remain estimable for all 1,000 draws per objective; the agent-matched liquidity-scaled common trading measure has 869 effective draws per objective (approximately 31,000 strategy-month feasibility failures per objective under the predetermined rule, with zero replacement draws), and agent-matched locked-period p-values for P are computed over those 869 draws.

A post-unlock diagnostic, computed for presentation after the reviewed draft and disclosed as such, characterizes the 131 excluded draws per objective using validation-period quantities only. The excluded draws are systematically narrow, mechanical portfolios: their mean monthly breadth (distinct securities held across the four pseudo-agents) is roughly 210 securities for Strategy A and 185 for Strategy B, against roughly 684 and 612 for the 869 retained draws—a standardized difference above 4—and their unscaled common trade per dollar is correspondingly lower (0.29 versus 0.35 for A; 0.26 versus 0.32 for B). Their signal composition is statistically indistinguishable from the retained draws: both groups average 8.2 distinct signals across exactly five economic families (Mann–Whitney p-values of 0.61

and 0.32 for the signal counts). Most importantly for the direction of any selection effect, the retained draws have *higher* validation-period P than the excluded draws (means of 4.24 versus 3.39×10^{-6} for A and 3.76 versus 2.92×10^{-6} for B), so conditioning on locked-period feasibility retains the higher- P portion of the null distribution and, if anything, biases the locked-period comparison against finding the observed system elevated. This diagnostic informs interpretation; the reported p-values remain those of the frozen rule.

All benchmark tests follow the pre-specified testing convention—upper-tail tests for excess similarity, with the p-value the fraction of null draws weakly exceeding the observed value—but only the standardized benchmark is itself pre-specified in the measurement protocol; the agent-matched benchmark was specified afterward and frozen before locked-sample access (Section II.E). Following the protocol-defined classification, a measure-objective outcome is classified as *elevated under both benchmark constructions* if it lies in the 5% upper tail of both null distributions and *elevated under only one* if in the upper tail of exactly one; the thresholds are nominal and unadjusted for the multiplicity of measures, objectives, periods, and benchmarks.

Sources (frozen): measurement protocol and CROWDING_REPORT.md (measure definitions, signal-null design, seed 20260827, 100,000 repetitions, standardized benchmark acceptance); AGENT_MATCHED_BENCHMARK_REPORT.md (seed 20260828, PCG64 streams, 1,000 feasible systems, reconstruction diagnostic, paired design); VALIDATION_REPRODUCTION_REPORT.md (seed confirmation); LOCKED_SAMPLE_REPORT.md and data_quality_and_benchmark_feasibility.csv (locked feasibility, 869 effective draws for the liquidity-scaled common trading measure, zero replacement draws).

Appendix D. Supplementary Tables

This appendix collects supplementary exhibits referenced in Sections III and IV. All values are taken directly from the frozen locked-run tables in the archived workflow record; no value is recomputed.

Table D.I — Descriptive Statistics for the Combined 2003–2024 Sample

This table reports annualized performance of the eight frozen strategies over the combined January 2003–December 2024 sample. Because this sample mixes the validation period, which informed final model selection, with the locked test period, the figures are descriptive only; no inference in the paper is based on them (Section III.C). Gross returns are before transaction costs; net returns subtract effective-cost estimates from the Hasbrouck (2009) trading-cost model, applied to turnover measured against prior-month weights drifted by realized returns (Section II.C). Volatility (vol.) is the annualized standard deviation of monthly returns. Sharpe ratios are monthly mean over monthly standard deviation, annualized by $\sqrt{12}$, with no risk-free adjustment (dollar-neutral portfolios). Maximum drawdown (max. DD) is the largest peak-to-trough decline of the cumulative arithmetic return path. Turnover is annualized one-sided turnover; Cost is the annualized average transaction-cost drag. Months are 264 (263 for Agent 2, whose one infeasible locked month is disclosed under the frozen attrition rule). Agents 1 and 3 hold identical A and B strategies; their rows are repeated for completeness.

Agent	Strategy	Gross mean	Net mean	Gross vol.	Gross Sharpe	Net Sharpe	Gross max. DD	Net max. DD	Turnover	Cost
1	A	7.37%	4.51%	9.16%	0.80	0.49	−26.37%	−29.95%	7.53	2.86%
1	B	7.37%	4.51%	9.16%	0.80	0.49	−26.37%	−29.95%	7.53	2.86%
2	A	8.72%	5.71%	10.13%	0.86	0.56	−23.18%	−26.06%	7.84	3.01%
2	B	6.35%	4.54%	8.75%	0.73	0.52	−25.08%	−27.55%	4.65	1.81%
3	A	5.26%	2.07%	9.90%	0.53	0.21	−31.69%	−38.63%	8.90	3.20%
3	B	5.26%	2.07%	9.90%	0.53	0.21	−31.69%	−38.63%	8.90	3.20%
4	A	7.63%	5.04%	8.30%	0.92	0.61	−16.46%	−19.60%	6.77	2.60%
4	B	9.32%	7.03%	9.90%	0.94	0.71	−21.05%	−22.34%	5.67	2.29%

Table D.II — Turnover, Breadth, and Liquidity Context

This table reports strategy-level turnover, cost, and breadth context (Panel A) and the liquidity of commonly traded securities (Panel B) in the validation and locked test periods. Turnover is annualized one-sided turnover; cost is the annualized transaction-cost drag under the Hasbrouck (2009) cost model (Section II.C); breadth is the average number of names (securities) per leg; Max $|w|$ is the largest absolute position weight observed, as a share of its leg. Panel B reports the mean and median dollar volume of securities in which at least two agents desire same-direction trades in a month. Dollar volume is a liquidity proxy, not a measure of structural market depth (Section IV.E).

Panel A: Strategy context

Agent	Strategy	Period	Turnover	Cost	Avg. long names	Avg. short names	Max $ w $
1	A/B	Validation	7.41	2.99%	224	224	0.62%
1	A/B	Locked test	7.63	2.76%	194	194	0.76%
2	A	Validation	7.76	3.14%	223	223	0.63%
2	A	Locked test	7.91	2.90%	193	193	0.61%
2	B	Validation	4.63	1.91%	331	332	0.46%
2	B	Locked test	4.66	1.73%	294	286	0.42%
3	A/B	Validation	8.74	3.37%	214	214	0.66%
3	A/B	Locked test	9.03	3.05%	182	182	0.78%
4	A	Validation	6.72	2.70%	466	466	0.37%
4	A	Locked test	6.82	2.52%	410	410	0.46%
4	B	Validation	5.60	2.34%	315	308	0.47%
4	B	Locked test	5.73	2.26%	277	269	0.59%

Panel B: Dollar volume of common-trade securities (\$ per security-month)

Objective	Period	Mean	Median
A	Validation	8,289,969	1,274,904
A	Locked test	20,300,159	2,590,910
B	Validation	8,451,553	1,232,839
B	Locked test	20,226,638	2,519,145