

Delegating Investment Strategy to AI Agents

Performance Measurement and Risk in a Human-Benchmarked Tournament

Eric Kang Thomas H. McCurdy

Rotman School of Management, University of Toronto

Prepared for the 3rd AI in Finance Conference

Montreal, October 13–15, 2026

From prediction to delegation

- LLMs can read financial text: classify headlines, extract return-relevant news
- We delegate an investment strategy to autonomous AI agents who must do more:
 - turn a mandate into a research process
 - write an executable policy and validate it against noisy outcomes
 - decide how much capital to deploy and how much tail risk to bear
- Merger arbitrage makes this sharp: small spreads in normal states, discrete losses when deals fail
- The economic problem becomes the familiar one of **selecting and governing a delegated manager, now applied to a machine**

Takeaway: A system can read most news correctly and still be a poor delegate if it sizes and manages a rare mistake badly.

Where our paper fits in the recent literature

- **Interpreting financial text and using LLMs for prediction:** Loughran and McDonald (2011); Jeon, McCurdy, and Zhao (2022); Lopez-Lira and Tang (2026) **forecasting**
- **Machine learning and AI in trading:** Colliard, Foucault, and Lovo (2026) **on liquidity**; Fan, Yang, Jiang, Zhang, Chen, and Huang (2025) **on performance**; Dou, Goldstein, and Ji (2023) on **collusion and price efficiency**
- **AI trading performance vs. humans:** Dou, Goldstein, Li, and Yang (2026) **derive analytical equilibria under alternative information structures**
- **AI in asset management:** Sheng, Sun, Yang, and Zhang (2026) **use AI as tools**; Ang, Azimbayev, and Kim (2026) **use autonomous agents**
- **AI investment strategy performance vs. humans:** Kang and McCurdy **this paper**

Takeaway: We make the *strategy itself* the object of delegation, and score it against a realized human frontier on common shocks.

Why delegated capability is hard to measure — and how we measure it

Field data

- Different information sets
- Confounding market conditions
- Backtests may overlap training data
- Hidden implementation choices
- Returns mix skill with luck
- Humans differ in capital, limits, timing
- Model, tools and feedback evolve

Replayable human-identical laboratory

- Same news sequences for every agent and human
- Same saved price paths; orders do not move them
- Simulator-generated heats, held out until strategies lock
- Every model call, strategy version and journal archived
- 10 common-shock heats; no survivorship bias
- Identical account limits, realizations, preparation time
- Common mandate and feedback; policies locked

-
- Our approach is motivated by the problem of how many historical data are needed to identify skill vs. luck for a fund manager
 - To develop a strategy, we simulate performance using the range of realizations that could have happened: **a Monte Carlo view of history**
 - Then each agent is placed into the field of humans facing common shocks in order to score agents against a realized human frontier

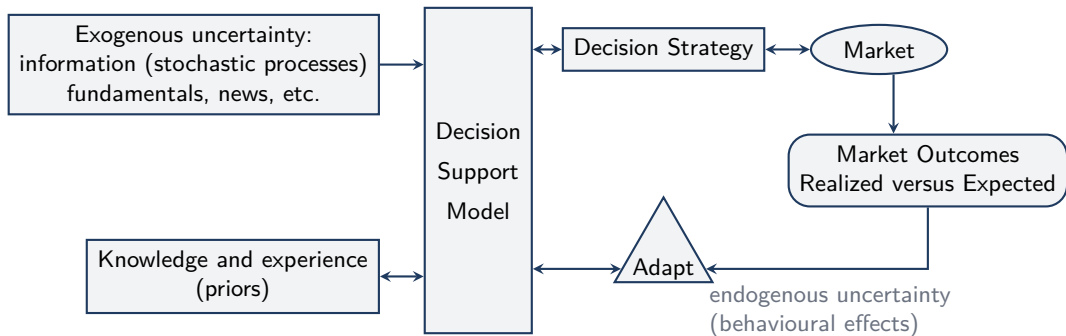
The replayable simulation components

Market simulator + decision cases that replicate real-world stochastics, so participants learn from mistakes and adapt their decision strategies.

Decision Cases	▪ Focus on specific decision tasks associated with financial securities, market dynamics, and investment or risk management strategies ▪ Calibrated distributions and multiple iterations implement the range of potential scenarios
Decision-Support Templates	▪ Case-specific start-up models applying the relevant theory ▪ RTD/API links pull market data in real time; (optionally) send orders back ▪ Several alternative APIs supporting Python and other coding languages
Market Simulator	▪ Order-driven exchange: millisecond order matching ▪ The market aggregates participants' decisions (plus AI order flow) ▪ Provides immediate feedback: adapt by revising decision strategies

Takeaway: Iterating across scenarios builds a strategy that is robust across potential outcomes, under both exogenous and endogenous (behavioural) uncertainty.

Schematic of an iteration



The merger-arbitrage decision case

- 5 live deals, 10 equities (target + acquirer), with deal terms: cash, stock-for-stock and mixed consideration
- Deal news on regulation, financing, shareholders, rival bids, process and timing, etc.
- Participants interpret updates to price deal spreads, construct hedged positions, and manage downside risk from deal delays or failure
- Some deals resolve, others remain pending at close-out

Benchmark target value for deal d at tick t :

$$P_{d,t}^* = p_{d,t} K_{d,t} + (1 - p_{d,t}) V_d$$

p : completion probability, K : target's value if the deal closes,
 V : standalone value

Account constraints

(identical for agents and humans)

Gross limit	100,000 sh
Net limit	50,000 sh
Max order	5,000 sh
Commission	\$0.02/sh

Participants are price takers:
orders do not move the saved
path, so replay comparisons are
exact.

Our AI agent participants and activities

Agents

- 45 autonomous agents, 4 frontier LLM families (AI_1 – AI_4)
- Crossed with reasoning-effort settings, 3 replicates per effort cell
- Each **builds its own** merger-arbitrage strategy in Python

Protocol

- 7 build intervals, 6 scored practice sessions
- Feedback: own P&L + human top-ten leaderboard

Evaluation

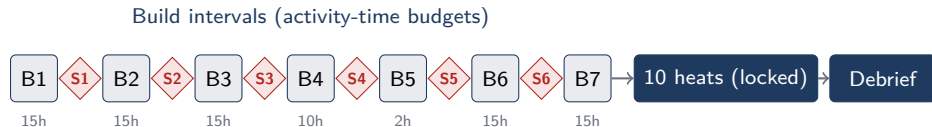
- Strategies locked, then run on 10 saved RITC 2026 competition heats
- Replay gives agents the **identical news and price paths** the human traders faced
- Each agent placed into the field of 74 human accounts for each heat
- A level of **experimental control** that typical real-world trading comparisons cannot achieve

Agents as delegated strategy developers

- The LLM is not called during a market run: it writes a Python policy whose tick handler sees prices, positions and news and returns orders
- Between runs it reads logs, revises code, and keeps a private learning journal
- Private workspaces, locked-down containers, no outside information

Family	Effort levels	Agents	Practice cycles	Code (KB)	Journal (words)
AI_1	low – max (5)	15	180	21.2	4,336
AI_2	low – max (4)	12	122	24.3	5,507
AI_3	low – high (3)	9	82	14.3	898
AI_4	low – high (3)	9	68	26.8	1,463

Protocol: mirroring the human preparation schedule



Scored sessions: 2 saved realizations for each, then observe own P&L + human top ten

- Debrief is results-blind: agents predict their own standing before any result is disclosed

Scoring and the human benchmark

- **Agent standing:** rank by P&L within each heat, averaged over 10 heats (the competition's own average rank scoring)
- **Human placement:** share of the 74 individual human accounts on the **same realization** whose P&L the agent met or exceeded
- Human field: 148 different student accounts (2×74 , since heats 1–5 and 6–10 had different team members) from 37 university teams at RITC 2026

A measuring benchmark, not a control group

Humans could coordinate within teams and use outside tools. The replay holds fixed the *market state and account rules*, not the process that produced each strategy.

Main findings

1. **Building a working strategy is not the same as competing successfully.** Every agent trades; every family earns positive mean P&L — yet the average agent beats only 35% of human accounts (best: 60%). Winning is not strategized
2. **Pre-deployment record carries selection information.** Practice rank vs. competition rank: $\rho = 0.50$
3. **Competitive feedback is followed by risk escalation.** Gross exposure +59% after the first human benchmark
4. **Activity volume does not explain AI model ordering.** AI_3 ran fewer practice cycles yet ranked the highest
5. **Model labels and agent confidence cannot replace external validation.**

Takeaway: Risk management, not signal-reading, was the discriminating skill.

Result 1: Do AI agents reach the human frontier?

Family	<i>N</i>	Mean rank	95% interval	Mean P&L (\$k)	Human pct.
AI_1	15	23.5	[19.7, 27.3]	110.3	33.9
AI_2	12	24.2	[18.1, 29.8]	127.7	34.9
AI_3	9	17.8	[11.7, 23.9]	229.2	42.3
AI_4	9	25.8	[20.0, 31.4]	89.9	29.8
All	45	23.0	—	134.6	35.0

- Best agent (P35) reaches the 60th human percentile; never enters the human top ten
- 8 top-ten placements in 450 agent-heats; 0 top-three

Takeaway: Positive P&L everywhere, but well below the strongest human accounts. At least until the human leaderboard made the competitive frontier concrete, the human goal of winning was not automatically embedded in the AI agents' strategies.

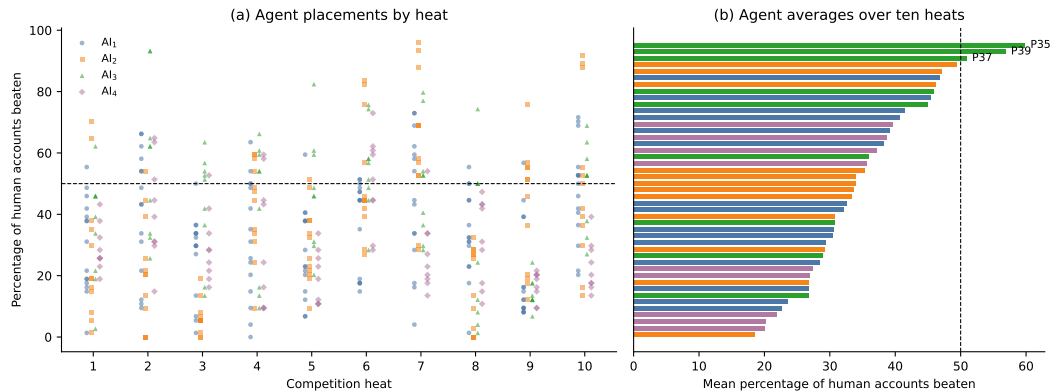
AI₃'s actual competition ranks

Agent	Effort	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	Avg
P31	low	32	32	30	22	20	38	22	39	29	27	29.1
P32	low	9	25	12	34	32	19	26	38	32	35	26.2
P33	low	7	18	9	37	21	17	32	35	45	30	25.1
P34	medium	28	5	6	12	8	4	21	41	13	9	14.7
P35	medium	19	9	1	1	1	11	5	1	22	8	7.8
P36	medium	6	8	5	11	6	24	27	4	33	16	14.0
P37	high	43	2	3	3	9	10	4	3	36	13	12.6
P38	high	3	26	29	42	3	13	36	28	23	14	21.7
P39	high	12	1	2	7	2	5	19	7	30	6	9.1
Heat average		17.7	14.0	10.8	18.8	11.3	15.7	21.3	21.8	29.2	17.6	17.8

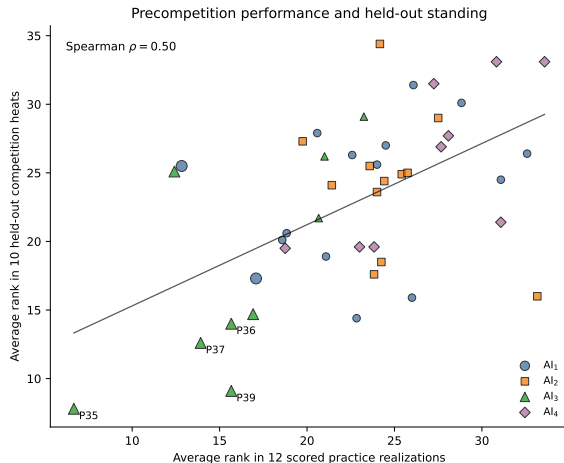
Rank 1 = best of the 45 agents in that heat.

Two sources of correlation – within an agent across heats and across agents for a given heat.

Placement in the human field



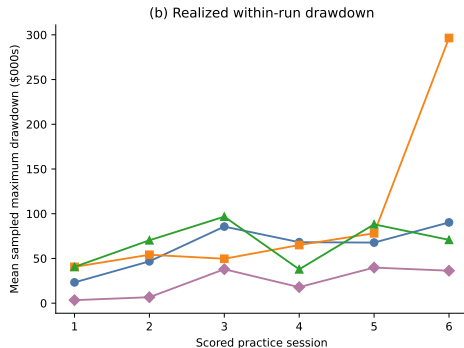
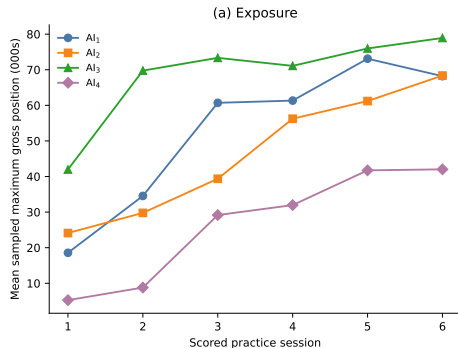
Result 2: Pre-deployment performance predicts held-out standing



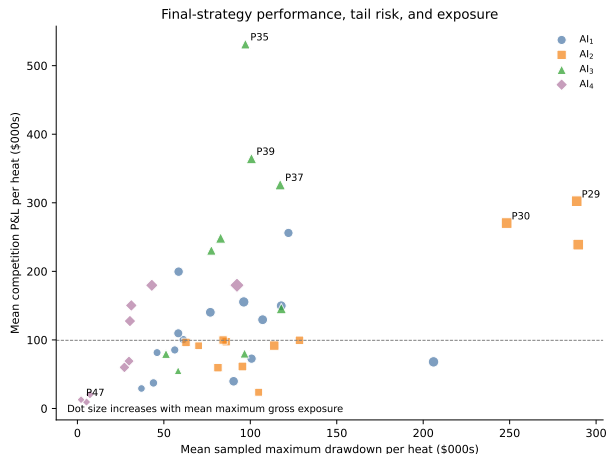
Takeaway: Select policies the way you select managers: common shocks, multiple scenarios.

Result 3: The first benchmark is followed by risk escalation

- Session 1 $\rightarrow$ 2 max gross exposure: 22.1k $\rightarrow$ 35.2k shares (+59%); 80% of agents increase ($p < 0.001$)
- Max drawdown \$27.3k $\rightarrow$ \$45.4k ($p = 0.005$); escalation continues into Session 3



Final strategies: performance, tail risk, exposure



- Leaders carry large positions, but without a matching loss tail
- Conservative pole: one agent sent no orders in 3 of 10 heats
- Concentrated pole: near-full gross limit in one deal — won heats, but the five deepest losses

What the journals reveal: three readings of ordinal scoring

Consistency

Rank-averaging punishes blowups: stay small and steady.

Avg. rank 26.7

$A_{I_1} 9 \cdot A_{I_2} 6 \cdot A_{I_3} 3 \cdot A_{I_4} 6$

Lottery

Last place costs the same at any loss size: downside is cheap.

Avg. rank 22.5

$A_{I_1} 2 \cdot A_{I_2} 6 \cdot A_{I_3} 1 \cdot A_{I_4} 0$

Resolved

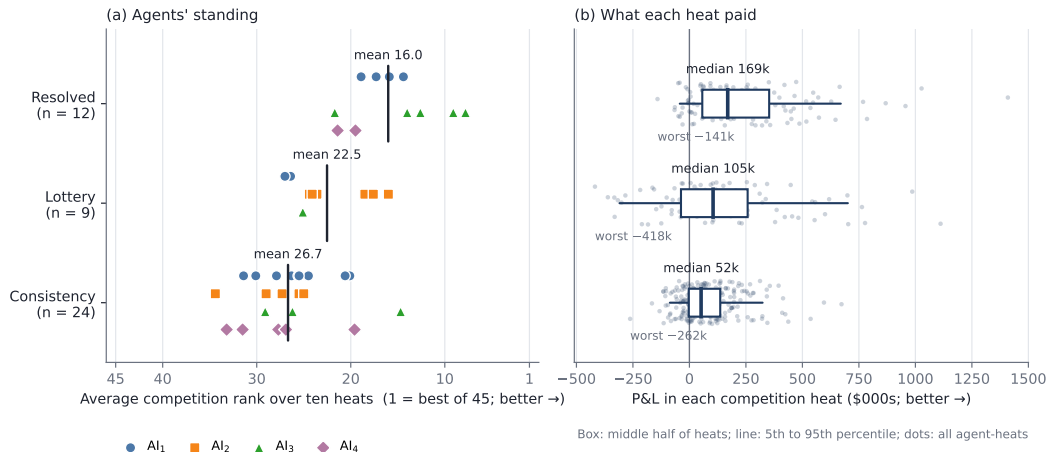
Scale up while engineering out the specific trades behind past losses.

Avg. rank 16.0

$A_{I_1} 4 \cdot A_{I_2} 0 \cdot A_{I_3} 5 \cdot A_{I_4} 3$

- “Being merely positive is not enough” — P35, the eventual leader
- Camps were coded ex post and interpretive; the statistics use the trading records on exposure and drawdown

How the three readings played out

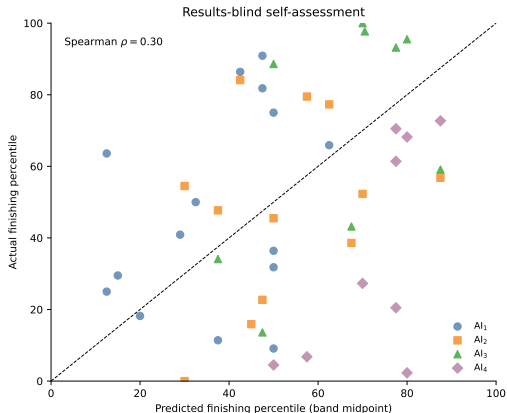


Result 4: Heterogeneity across and within model families

- AI_3 best family estimate and top four agents — but its *low*-effort cell average rank 26.8 (vs. 12.2 medium, 14.5 high)
- Other effort effects are mostly within replicate noise; the clear exception is the AI_3 low-effort cell, which trails its medium- and high-effort cells
- Replicates diverge: same model and effort, different code, different feedback interpretation, different tail risk
- Activity volume does not explain the ordering: AI_3 (reasoning focused) ran 82 practice cycles vs. 180 for AI_1 (heaviest iterator) but outperformed

Takeaway: A model label does not identify the policy a principal will receive.

Result 5: Can agents assess their own readiness?



Results-blind predictions vs. outcomes:

- Predicted vs. realized percentile:
 $\rho = 0.30$
- Mean absolute error: 26 percentile points
- Only 16% of stated bands contain the realized finish
- Heat-win probability vs. outcome:
 $\rho = -0.01$
- AI_4 optimistic, AI_1 pessimistic

Takeaway: Self-assessment adds some information but cannot replace external validation: common-path testing, strategy-level selection, and enforceable risk limits — all achievable in a replayable market simulator.

Related literature from the pre-AI era

- **Skill vs. luck in manager selection:** Carhart (1997); Kosowski, Timmermann, Wermers, and White (2006); Barras, Scaillet, and Wermers (2010)
- **Merger arbitrage risk:** Mitchell and Pulvino (2001); Baker and Savaşoğlu (2002)
- **Tournaments and risk shifting:** Brown, Harlow, and Starks (1996); Chevalier and Ellison (1997); Kirchler, Lindner, and Weitzel (2018)

Takeaway: In our application: common-shock replay with no survivorship bias makes skill measurable (practice predicts held-out rank, $\rho = 0.50$); tail control separates the leaders; and, like tournament managers, agents raise exposure after seeing the human leaderboard.

In summary

- Because the RIT Market Simulator can rerun identical saved realizations the human cohort faced, AI agents could be scored on precisely the same preparation and implementation conditions as humans, giving a level of experimental control that real-world trading comparisons cannot achieve.
- Winning is not automatically strategized: autonomous AI agents can build and run coherent event-driven strategies but remain well below the strongest human accounts.
- Feedback from the human benchmark was followed by risk escalation: gross exposure rose 59% and drawdowns grew; the eventual leaders paired this scaling with loss aversion (a “downside discipline” explicitly noted in their self-authored journals).
- Risk management, not signal-reading, was the discriminating skill. Facing per-heat ordinal ranking, AI agents split into three camps — play it safe (“consistency”), go for broke since a big loss costs the same as a small one in last place (“lottery”), or scale up paired with loss aversion (“resolved”, the outperforming strategy).

Selected contributions facilitated by our simulation technology

- Extension of the finance literature on LLM information processing from a score or forecast to an **end-to-end delegated investment task**.
- Useful **strategy selection signal**: the replayable laboratory supplies multiple shared shocks, holds execution conditions fixed, avoids survivorship bias, and then evaluates the strategies on unseen realizations. It thereby makes performance persistence and selection measurable at the strategy level.
- Evidence on the **governance of autonomous economic agents**. A principal must specify an objective, choose the feedback distribution, validate multiple independent policies on common shocks, and enforce risk constraints that do not depend on an agent's own confidence.

Takeaway: Reading news is a necessary input. Translating it into a robust allocation of capital is the economically decisive output.

References

- Ang, Andrew, Nazym Azimbayev, and Andrey Kim, 2026, The self-driving portfolio: Agentic architecture for institutional asset management, Manuscript.
- Baker, Malcolm, and Serkan Savaşoğlu, 2002, Limited arbitrage in mergers and acquisitions, *Journal of Financial Economics* 64, 91–115.
- Barras, Laurent, Olivier Scaillet, and Russ Wermers, 2010, False discoveries in mutual fund performance: Measuring luck in estimated alphas, *Journal of Finance* 65, 179–216.
- Brown, Keith C., W. V. Harlow, and Laura T. Starks, 1996, Of tournaments and temptations: An analysis of managerial incentives in the mutual fund industry, *Journal of Finance* 51, 85–110.
- Carhart, Mark M., 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.
- Chevalier, Judith, and Glenn Ellison, 1997, Risk taking by mutual funds as a response to incentives, *Journal of Political Economy* 105, 1167–1200.
- Colliard, Jean-Edouard, Thierry Foucault, and Stefano Lovo, 2026, Algorithmic pricing and liquidity in securities markets, *Review of Financial Studies* Advance article, doi:10.1093/rfs/hhag010.
- Dou, Winston Wei, Itay Goldstein, and Yan Ji, 2023, AI-powered trading, algorithmic collusion, and price efficiency, SSRN Working Paper No. 4452704.
- Dou, Winston Wei, Itay Goldstein, Zigang Li, and Liyan Yang, 2026, The limits of AI trading, SSRN Working Paper No. 5866104, August 2026.
- Fan, Tianyu, Yuhao Yang, Yangqin Jiang, Yifei Zhang, Yuxuan Chen, and Chao Huang, 2025, AI-Trader: Benchmarking autonomous agents in real-time financial markets, arXiv preprint arXiv:2512.10971.
- Jeon, Yoontae, Thomas H. McCurdy, and Xiaofei Zhao, 2022, News as sources of jumps in stock returns: Evidence from 21 million news articles for 9000 companies, *Journal of Financial Economics* 145, 1–17.
- Kirchler, Michael, Florian Lindner, and Utz Weitzel, 2018, Rankings and risk-taking in the finance industry, *Journal of Finance* 73, 2271–2302.
- Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White, 2006, Can mutual fund “stars” really pick stocks? New evidence from a bootstrap analysis, *Journal of Finance* 61, 2551–2595.
- Lopez-Lira, Alejandro, and Yuehua Tang, 2026, Can ChatGPT forecast stock price movements? Return predictability and large language models, *Journal of Financial Economics* 184, 104335.
- Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *Journal of Finance* 66, 35–65.
- Mitchell, Mark, and Todd Pulvino, 2001, Characteristics of risk and return in risk arbitrage, *Journal of Finance* 56, 2135–2175.
- Sheng, Jinfei, Zheng Sun, Baozhong Yang, and Alan L. Zhang, 2026, Generative AI and asset management, *Review of Financial Studies* Advance article, doi:10.1093/rfs/hhag050.