AI "Errors"

Wenqian Huang[1] Albert J. Menkveld[2] Shihao Yu[3]

[1]Bank for International Settlements

[2]Vrije Universiteit Amsterdam

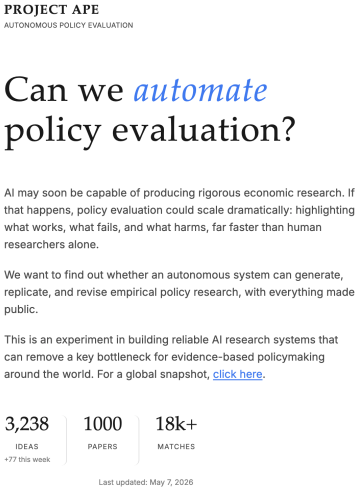
[3]Singapore Management University

UMD / SMU / UBS Quant Investment Forum: AI and Finance
2026

Outline

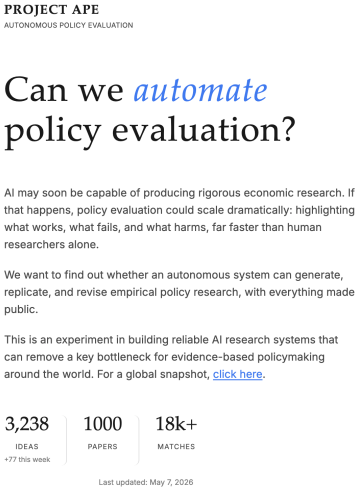
Motivation

AI is moving from drafting prose to running empirical research



APE (Autonomous Policy Evaluation): runs empirical economics at scale

AI is moving from drafting prose to running empirical research

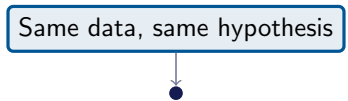


APE (Autonomous Policy Evaluation): runs empirical economics at scale

Human × AI Finance (UCLA, April 2026): AI writes the papers; AI also referees them

Small implementation choices can move estimates materially

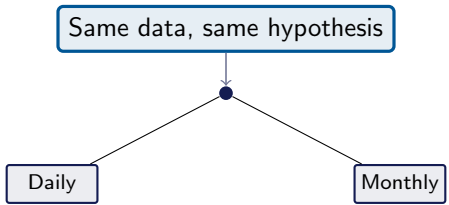
- Pivotal choices are often implicit in code and preprocessing, and largely invisible in the final write-up
- Modest differences can move estimates materially and shift conclusions



Small implementation choices can move estimates materially

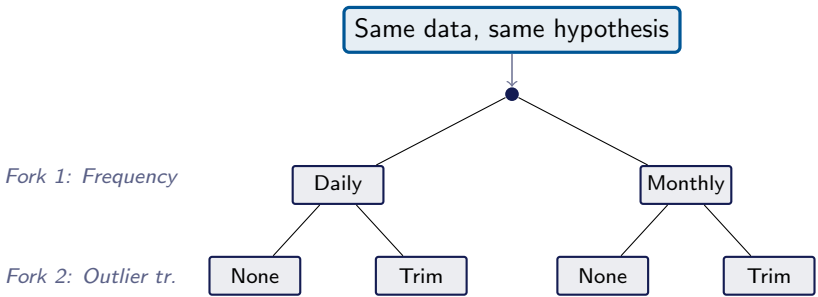
- Pivotal choices are often implicit in code and preprocessing, and largely invisible in the final write-up
- Modest differences can move estimates materially and shift conclusions

Fork 1: Frequency



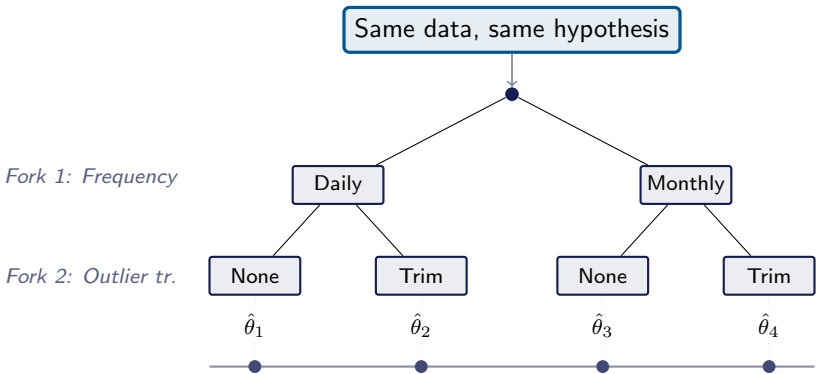
Small implementation choices can move estimates materially

- Pivotal choices are often implicit in code and preprocessing, and largely invisible in the final write-up
- Modest differences can move estimates materially and shift conclusions



Small implementation choices can move estimates materially

- Pivotal choices are often implicit in code and preprocessing, and largely invisible in the final write-up
- Modest differences can move estimates materially and shift conclusions

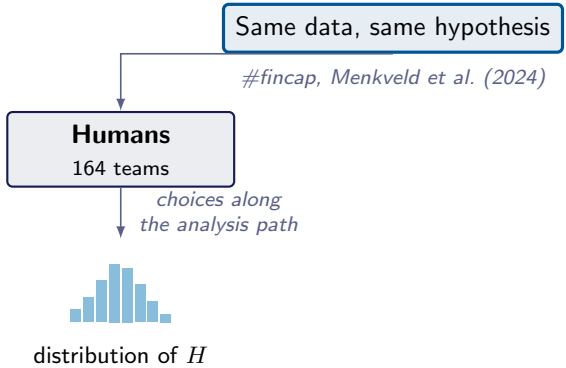


This paper asks how AI and human empirical analyses differ, and why

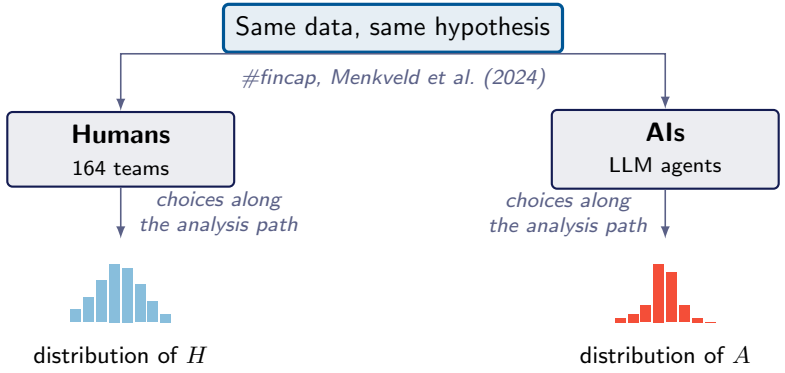
Same data, same hypothesis

#fincap, Menkveld et al. (2024)

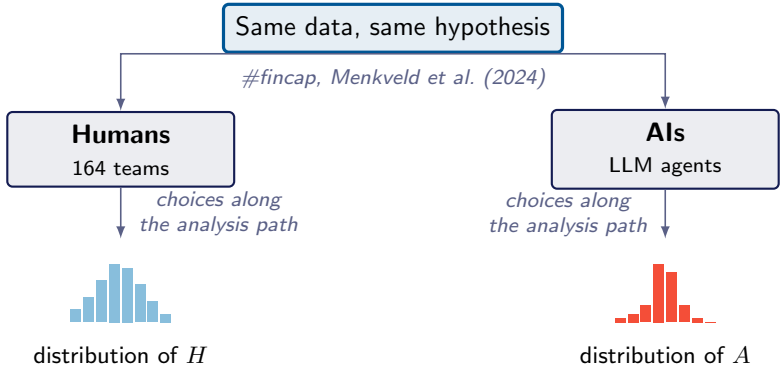
This paper asks how AI and human empirical analyses differ, and why



This paper asks how AI and human empirical analyses differ, and why



This paper asks how AI and human empirical analyses differ, and why



Q1. How do they differ in location and dispersion?
Q2. Which decision forks on the analysis path drive the gap?

Takeaways

1. **Distributions differ.** For all six hypotheses, AI and human outcome distributions are statistically different.
   - AI estimates are tighter than the human benchmark
   - For complex hypotheses, they are also shifted in location

Takeaways

1. **Distributions differ.** For all six hypotheses, AI and human outcome distributions are statistically different.
   - AI estimates are tighter than the human benchmark
   - For complex hypotheses, they are also shifted in location

2. **The drivers are identifiable.**
   - AIs concentrate on a narrow set of paths
   - They choose different options at key forks. AIs
     - never treat outliers
     - choose lower analysis frequency
     - choose more robust statistical models

Experiment

Human benchmark: 164 teams test 6 hypotheses on identical data (#fincap)

- #fincap (Finance Crowd Analysis Project): 164 independent research teams
- **Same proprietary dataset**: 720 million trades in EURO STOXX 50 Index Futures
- **Same six pre-specified hypotheses**, each phrased as "has not changed over time":
  - **H1: Market efficiency**
  - H2: Realized bid-ask spread on market orders
  - **H3: Client share of volume**
  - H4: Client realized bid-ask spread
  - H5: Fraction of client trades via market/marketable limit orders
  - H6: Relative gross trading revenue (GTR) for clients
- Each team submits a short paper reporting estimate, standard error, and t-statistic
- Each team fills out a detailed survey on their implementation choices against a pre-specified multiverse

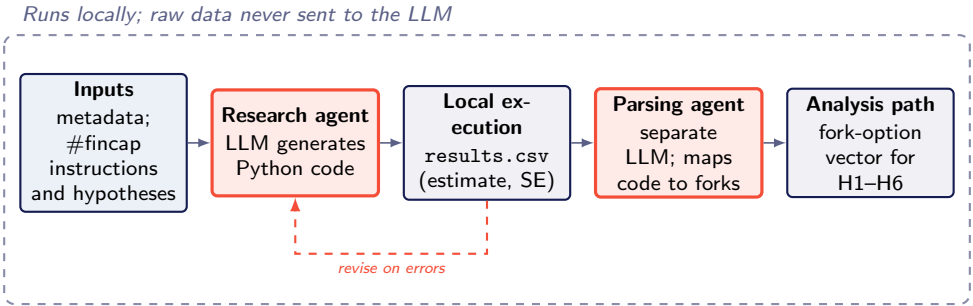
Mapping to a common multiverse: every estimate becomes one analysis path

- Multiverse from Menkveld et al. (2024, Table V) for H1 and H3

Example: H1 (Efficiency, complex) vs. H3 (Client Volume, simple)

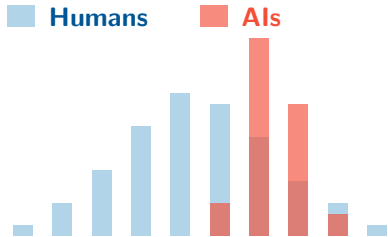
Fork	Description	Options	#	H1	H3
`RmvOpnCls`	Remove open/close obs	N, Y	2	✓	✓
`DysExcldd`	Days excluded	N, Y	2	✓	✓
`OtlrTrtmnt`	Outlier treatment	N, Trm, Wns	3	✓	✓
`FrqncyAnlyss`	Analysis frequency	D, W, M, A	4	✓	✓
`Mdl`	Statistical model	LgDff, RltvChg, TrndSttnry	3	✓	✓
`Msr`	Efficiency measure	VrncRt, AtCrrltns	2	✓	–
`Frqncy`	Return horizon	SM, 15, 530, DW, DM	5	✓	–
`Units`	Volume units	Euro, NOfCncts	2	–	✓
		Multiverse paths		1,440	288

We replicate the experiment with AI research agents



Methodology

Task 1: Whether the human and AI outcome distributions are equal

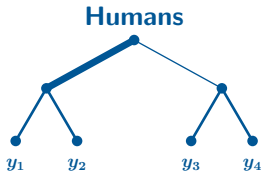


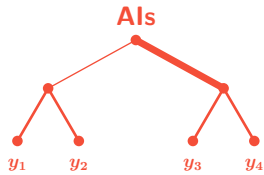
Two-sample Anderson-Darling test. For each hypothesis, $AD2(\mathrm{H}, \mathrm{A})$ tests

$$H_0 : F_{\mathrm{H}} = F_{\mathrm{A}} \quad \text{vs.} \quad H_1 : F_{\mathrm{H}} \neq F_{\mathrm{A}},$$

where F_{H}, F_{A} are the human and AI outcome distributions.

Task 2: Where the distributions differ: mapping to a multiverse



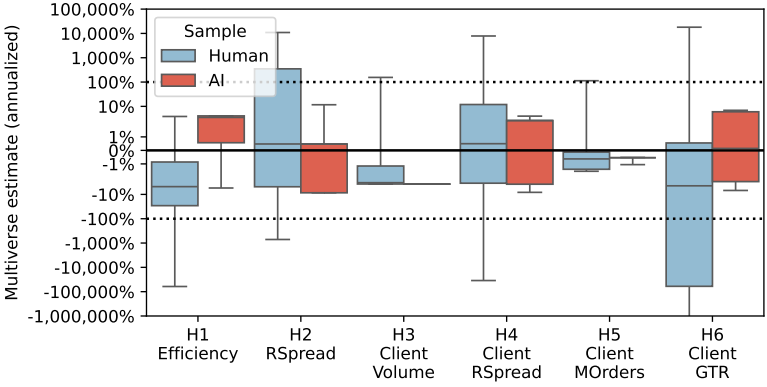


A fork-option importance measure:

$$\underbrace{\beta_{i,j}(\tau)}_{\substack{\text{intrinsic importance} \\ \text{quantile regressions}}} \times \underbrace{\Delta_{\mathsf{HA}}(i,j)}_{\text{choice difference}}$$

Results

In the multiverse, $AD2$ rejects equality for all six hypotheses



	$AD2$
H1	107.07^{***}
H2	26.21^{***}
H3	93.48^{***}
H4	15.26^{***}
H5	30.52^{***}
H6	35.32^{***}

- **Dispersion.** AI IQR and IDR are much lower than humans; for H3 and H5 the AI IQR collapses to *zero*
- **Location.** Medians align for the simpler hypotheses but diverge for the complex ones

Three forks drive the gap: AI plays one analytical chord on H1 and H3

		H1		H3	
Fork	Option	w_{H}	w_{A}	w_{H}	w_{A}
RmvOpnCls	N	0.76	1.00	0.84	1.00
	Y	0.24	0.00	0.16	0.00
DysExcldd	N	0.84	0.99	0.84	0.99
	Y	0.16	0.01	0.16	0.01
OtlrTrtmnt	N	0.66	1.00	0.68	1.00
	Trm	0.14	0.00	0.13	0.00
	Wns	0.20	0.00	0.20	0.00
FrqncyAnlyss	D	0.32	0.01	0.37	0.03
	M	0.23	0.24	0.21	0.82
	A	0.45	0.75	0.41	0.15
Mdl	LgDff	0.06	0.00	0.04	0.00
	RltvChg	0.58	0.00	0.60	0.00
	TrndSttnry	0.36	1.00	0.35	1.00

1. **Model (“Mdl”).** AIs pick one spec; humans split

Three forks drive the gap: AI plays one analytical chord on H1 and H3

		H1		H3	
Fork	Option	w_{H}	w_{A}	w_{H}	w_{A}
RmvOpnCls	N	0.76	1.00	0.84	1.00
	Y	0.24	0.00	0.16	0.00
DysExcldd	N	0.84	0.99	0.84	0.99
	Y	0.16	0.01	0.16	0.01
OtlrTrtmnt	N	0.66	1.00	0.68	1.00
	Trm	0.14	0.00	0.13	0.00
	Wns	0.20	0.00	0.20	0.00
FrqncyAnlyss	D	0.32	0.01	0.37	0.03
	M	0.23	0.24	0.21	0.82
	A	0.45	0.75	0.41	0.15
Mdl	LgDff	0.06	0.00	0.04	0.00
	RltvChg	0.58	0.00	0.60	0.00
	TrndSttnry	0.36	1.00	0.35	1.00

1. **Model ("Mdl").** AIs pick one spec; humans split
2. **Frequency.** AIs avoid daily; many humans use it

Three forks drive the gap: AI plays one analytical chord on H1 and H3

		H1		H3	
Fork	Option	w_H	w_A	w_H	w_A
RmvOpnCls	N	0.76	1.00	0.84	1.00
	Y	0.24	0.00	0.16	0.00
DysExcldd	N	0.84	0.99	0.84	0.99
	Y	0.16	0.01	0.16	0.01
OtlrTrtmnt	N	0.66	1.00	0.68	1.00
	Trm	0.14	0.00	0.13	0.00
	Wns	0.20	0.00	0.20	0.00
FrqncyAnlyss	D	0.32	0.01	0.37	0.03
	M	0.23	0.24	0.21	0.82
	A	0.45	0.75	0.41	0.15
Mdl	LgDff	0.06	0.00	0.04	0.00
	RltvChg	0.58	0.00	0.60	0.00
	TrndSttnry	0.36	1.00	0.35	1.00

1. **Model ("Mdl").** AIs pick one spec; humans split
2. **Frequency.** AIs avoid daily; many humans use it
3. **Data cleaning.** AIs skip filters; humans apply them selectively

Three forks drive the gap: AI plays one analytical chord on H1 and H3

		H1		H3	
Fork	Option	w_{H}	w_{A}	w_{H}	w_{A}
RmvOpnCls	N	0.76	1.00	0.84	1.00
	Y	0.24	0.00	0.16	0.00
DysExcldd	N	0.84	0.99	0.84	0.99
	Y	0.16	0.01	0.16	0.01
OtlrTrtmnt	N	0.66	1.00	0.68	1.00
	Trm	0.14	0.00	0.13	0.00
	Wns	0.20	0.00	0.20	0.00
FrqncyAnlyss	D	0.32	0.01	0.37	0.03
	M	0.23	0.24	0.21	0.82
	A	0.45	0.75	0.41	0.15
Mdl	LgDff	0.06	0.00	0.04	0.00
	RltvChg	0.58	0.00	0.60	0.00
	TrndSttnry	0.36	1.00	0.35	1.00

1. **Model ("Mdl").** AIs pick one spec; humans split
2. **Frequency.** AIs avoid daily; many humans use it
3. **Data cleaning.** AIs skip filters; humans apply them selectively

Three forks drive the gap: AI plays one analytical chord on H1 and H3

		H1		H3	
Fork	Option	w_H	w_A	w_H	w_A
RmvOpnCls	N	0.76	1.00	0.84	1.00
	Y	0.24	0.00	0.16	0.00
DysExcldd	N	0.84	0.99	0.84	0.99
	Y	0.16	0.01	0.16	0.01
OtlrTrtmnt	N	0.66	1.00	0.68	1.00
	Trm	0.14	0.00	0.13	0.00
	Wns	0.20	0.00	0.20	0.00
FrqncyAnlyss	D	0.32	0.01	0.37	0.03
	M	0.23	0.24	0.21	0.82
	A	0.45	0.75	0.41	0.15
Mdl	LgDff	0.06	0.00	0.04	0.00
	RltvChg	0.58	0.00	0.60	0.00
	TrndSttnry	0.36	1.00	0.35	1.00

1. **Model ("Mdl").** AIs pick one spec; humans split
2. **Frequency.** AIs avoid daily; many humans use it
3. **Data cleaning.** AIs skip filters; humans apply them selectively

Model choice is the dominant driver for H1 locational difference

- Intrinsic importance (Q50): "RltvChg" lowers the H1 median by 2.83

	Opt	Q50	Δ_{HA}	Q50 $\times\Delta_{\text{HA}}$
Mdl	LgDff	0.03	-0.06	0.00
	RltvChg	-2.83***	-0.58	1.64***
	TrndSttnry	2.22^{***}	0.64	1.42^{***}
FrqncyAnlyss	D	-0.83^{***}	-0.30	0.25^{***}
	M	-0.31^{***}	0.01	0.00
	A	0.58^{***}	0.30	0.17^{***}
RmvOpnCls	N	0.54^{***}	0.24	0.13^{***}
DysExcldd	N	0.48^{**}	0.16	0.08^{**}
OtlrTrtmnt	N	0.59^{***}	0.34	0.20^{***}

Model choice is the dominant driver for H1 locational difference

- Intrinsic importance (Q50): "RltvChg" lowers the H1 median by 2.83
- Choice difference (Δ_{HA}): AIs pick "RltvChg" 0.58 less often than humans

	Opt	Q50	Δ_{HA}	Q50 $\times\Delta_{HA}$
Mdl	LgDff	0.03	-0.06	0.00
	RltvChg	-2.83***	-0.58	1.64***
	TrndSttnry	2.22^{***}	0.64	1.42^{***}
FrqncyAnlyss	D	-0.83^{***}	-0.30	0.25^{***}
	M	-0.31^{***}	0.01	0.00
	A	0.58^{***}	0.30	0.17^{***}
RmvOpnCls	N	0.54^{***}	0.24	0.13^{***}
DysExcldd	N	0.48^{**}	0.16	0.08^{**}
OtlrTrtmnt	N	0.59^{***}	0.34	0.20^{***}

Model choice is the dominant driver for H1 locational difference

- Intrinsic importance (Q50): "RltvChg" lowers the H1 median by 2.83
- Choice difference (Δ_{HA}): AIs pick "RltvChg" 0.58 less often than humans
- Importance = Q50 $\times\Delta_{HA} \approx$ 1.6: "Mdl" is the dominant driver; data-cleaning forks add smaller, consistent contributions

	Opt	Q50	Δ_{HA}	Q50 $\times\Delta_{HA}$
Mdl	LgDff	0.03	-0.06	0.00
	RltvChg	-2.83***	-0.58	1.64***
	TrndSttnry	2.22^{***}	0.64	1.42^{***}
FrqncyAnlyss	D	-0.83^{***}	-0.30	0.25^{***}
	M	-0.31^{***}	0.01	0.00
	A	0.58^{***}	0.30	0.17^{***}
RmvOpnCls	N	0.54^{***}	0.24	0.13^{***}
DysExcldd	N	0.48^{**}	0.16	0.08^{**}
OtlrTrtmnt	N	0.59^{***}	0.34	0.20^{***}

Why H1 dispersion collapses: avoided Jensen bias on daily × relative-change

- Intrinsic importance: "FrqncyAnlyss-D" and "Mdl-RltvChg" collapse Q25 ($\sim -1{,}560$, ~ -182) while Q75 barely moves: a Jensen left-tail bias

	Opt	Q25	Q75	Δ_{HA}	IQR$\times\Delta_{\text{HA}}$
FrqncyAnlyss	D	-1559.83***	-0.66^{***}	-0.30	-474.63***
	A	4.72^{***}	0.53^{***}	0.30	-1.26^{***}
Mdl	RltvChg	-181.73***	-1.00^{***}	-0.58	-104.69***
	TrndSttnry	131.10^{***}	0.76^{***}	0.64	-83.45^{***}
RmvOpnCls	N	44.97^{***}	0.68^{***}	0.24	-10.53^{***}
DysExcldd	N	26.59^{***}	0.56^{***}	0.16	-4.12^{***}

Why H1 dispersion collapses: avoided Jensen bias on daily × relative-change

- Intrinsic importance: "FrqncyAnlyss-D" and "Mdl-RltvChg" collapse Q25 ($\sim -1{,}560$, ~ -182) while Q75 barely moves: a Jensen left-tail bias
- Choice difference (Δ_{HA}): AIs avoid both (-0.30, -0.58); humans pick them more often

	Opt	Q25	Q75	Δ_{HA}	IQR$\times\Delta_{\text{HA}}$
FrqncyAnlyss	D	-1559.83***	-0.66^{***}	-0.30	-474.63***
	A	4.72^{***}	0.53^{***}	0.30	-1.26^{***}
Mdl	RltvChg	-181.73***	-1.00^{***}	-0.58	-104.69***
	TrndSttnry	131.10^{***}	0.76^{***}	0.64	-83.45^{***}
RmvOpnCls	N	44.97^{***}	0.68^{***}	0.24	-10.53^{***}
DysExcldd	N	26.59^{***}	0.56^{***}	0.16	-4.12^{***}

Why H1 dispersion collapses: avoided Jensen bias on daily × relative-change

- Intrinsic importance: "FrqncyAnlyss-D" and "Mdl-RltvChg" collapse Q25 ($\sim -1{,}560$, ~ -182) while Q75 barely moves: a Jensen left-tail bias
- Choice difference (Δ_{HA}): AIs avoid both (-0.30, -0.58); humans pick them more often
- Importance = IQR $\times \Delta_{\text{HA}}$: the product ($\sim -475$, -105) means AI dispersion mechanically compresses

	Opt	Q25	Q75	Δ_{HA}	IQR$\times\Delta_{\text{HA}}$
FrqncyAnlyss	D	-1559.83***	-0.66^{***}	-0.30	-474.63***
	A	4.72^{***}	0.53^{***}	0.30	-1.26^{***}
Mdl	RltvChg	-181.73***	-1.00^{***}	-0.58	-104.69***
	TrndSttnry	131.10^{***}	0.76^{***}	0.64	-83.45^{***}
RmvOpnCls	N	44.97^{***}	0.68^{***}	0.24	-10.53^{***}
DysExcldd	N	26.59^{***}	0.56^{***}	0.16	-4.12^{***}

Conclusion

Conclusion

1. AIs and humans produce statistically different outcome distributions
2. AIs concentrate on a narrow set of analysis paths, yielding lower dispersion; for complex hypotheses, AI estimates are also shifted in location
3. Three forks consistently drive the gap: "Mdl" (model), "FrqncyAnlyss" (frequency), and discretionary data cleaning.
4. AI delivers *precision without alignment* when discretion is high; some deviations are even "benevolent" (e.g., the daily-frequency Jensen bias)

We should audit the forks the AI silently picks!

Appendix

References I

Cao, Sean et al. (Oct. 2024). "From Man vs. Machine to Man + Machine: The art and AI of stock analyses". In: *Journal of Financial Economics* 160, p. 103910.

Dou, Winston Wei, Itay Goldstein, and Yan Ji (Jan. 27, 2024). *AI-Powered Trading, Algorithmic Collusion, and Price Efficiency*. URL: https://papers.ssrn.com/abstract=4452704 (visited on 04/03/2024). Pre-published.

Fedyk, Anastassia et al. (Apr. 8, 2024). *AI and Perception Biases in Investments: An Experimental Study*. URL: https://papers.ssrn.com/abstract=4787249 (visited on 01/30/2026). Pre-published.

Horton, John J. (Apr. 2023). *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* URL: https://www.nber.org/papers/w31122 (visited on 01/30/2026). Pre-published.

References II

Lopez-Lira, Alejandro and Yuehua Tang (Oct. 28, 2025). *Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models*. URL: http://arxiv.org/abs/2304.07619 (visited on 01/30/2026). Pre-published.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan (Jan. 2025). *Large Language Models: An Applied Econometric Framework*. URL: https://www.nber.org/papers/w33344 (visited on 01/10/2026). Pre-published.

Manning, Benjamin S., Kehang Zhu, and John J. Horton (Apr. 2024). *Automated Social Science: Language Models as Scientist and Subjects*. URL: https://www.nber.org/papers/w32381 (visited on 08/26/2025). Pre-published.

Menkveld, Albert J. et al. (2024). "Nonstandard Errors". In: *The Journal of Finance* 79.3, pp. 2339–2390.

Novy-Marx, Robert and Mihail Z. Velikov (Jan. 2025). *AI-Powered (Finance) Scholarship*. URL: https://www.nber.org/papers/w33363 (visited on 08/26/2025). Pre-published.

References III

Ouyang, Shumiao, Hayong Yun, and Xingjian Zheng (May 1, 2024). *AI as Decision-Maker: Ethics and Risk Preferences of LLMs*. URL: https://papers.ssrn.com/abstract=4851711 (visited on 01/30/2026). Pre-published.

Pérignon, Christophe et al. (Nov. 1, 2024). "Computational Reproducibility in Finance: Evidence from 1,000 Tests". In: *The Review of Financial Studies* 37.11, pp. 3558–3593.

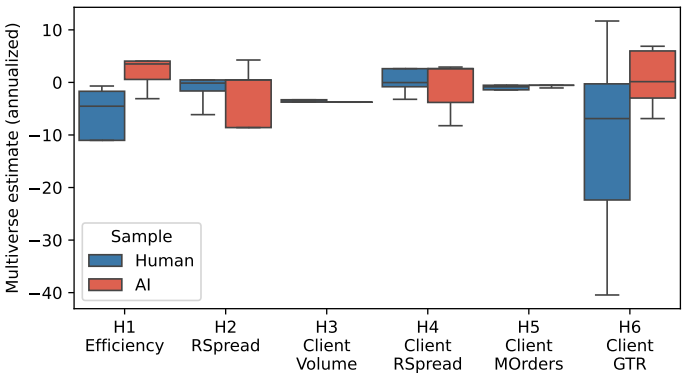
Sarkar, Suproteem K. and Keyon Vafa (June 28, 2024). *Lookahead Bias in Pretrained Language Models*. URL: https://papers.ssrn.com/abstract=4754678 (visited on 01/13/2026). Pre-published.

Multiverse estimates are tighter and shifted, especially for complex tasks

- **Dispersion.** AI's IQR and IDR are substantially lower than humans; for H3 and H5, AI IQR collapses to *zero* (few forks plus AI consistency)
- **Location.** Medians are aligned for simpler hypotheses and more pronounced for complex ones

	H1		H2		H3		H4		H5		H6	
	A	H	A	H	A	H	A	H	A	H	A	H
Median	3.5	−4.8	0.5	0.5	−3.8	−3.3	2.6	0.5	−0.5	−0.6	0.1	−4.4
IQR	3.5	28.2	9.1	354	0.0	2.6	6.4	15.4	0.0	1.3	9.0	60k
IDR	7.1	14k	12.9	5.0k	0.0	148	11.2	21k	0.5	105	13.8	5.0M

Top-quality humans look like the full sample: AI gap is not a quality artifact



- 8 humans pass all three top-quartile filters: researcher quality, reproducibility, peer rating (Menkveld et al., 2024)
- Top-quality medians stay close to the full sample
- AI-human gap persists, largely unchanged

LLM choice and look-ahead bias

- **Concern.** The LLM may have absorbed humans' analytical choices from prior #fincap drafts (Ludwig, Mullainathan, and Rambachan, 2025; Sarkar and Vafa, 2024)
- **Choice.** GPT-5.2 (OpenAI), fixed snapshot dated December 11, 2025; robustness across o3, GPT-5, GPT-5.1
- **Why not chronological or open-source.** Open weights still post-date the early #fincap SSRN release; ChronoGPT-style models cannot handle the task complexity
- **Direction of bias.** Residual leakage, if any, would attenuate (not generate) the divergence we document

Diagnostic: GPT-5.2 does not reproduce #fincap verbatim

Prompt (first half of the #fincap abstract): "In statistics, samples are drawn from a population in a data-generating process (DGP). . . . In science, evidence is generated to test hypotheses in an evidence-generating process (EGP). We claim that EGP variation across researchers adds uncertainty: nonstandard errors (NSEs)."

Actual continuation (Menkveld et al. 2024): "We study NSEs by letting **164 teams test the same hypotheses on the same data**. NSEs turn out to be sizable, but smaller for more reproducible or higher rated research. **Adding peer-review stages reduces NSEs**."

GPT-5.2 completion: "These NSEs reflect uncertainty induced by researchers' choices about data, methods, and specifications, which can materially affect reported results. . . . We propose a framework for quantifying NSEs by modeling researcher-specific variation in the EGP . . ."

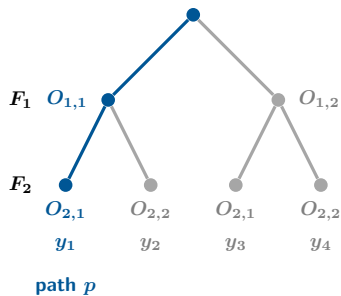
No mention of "164 teams", peer review, or quality ratings $\implies$ *no evidence of verbatim memorization*

Related literature: AI as agent, in research, and in markets

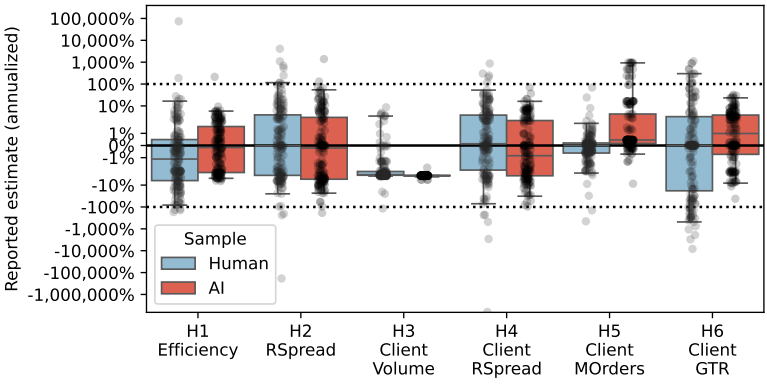
- **AI in empirical research.** Generating papers, hypotheses, and simulated agents (Novy-Marx and Velikov, 2025; Horton, 2023; Manning, Zhu, and Horton, 2024)
- **AI as economic agent.** Collusion, preferences, and risk-taking among AI-generated agents (Dou, Goldstein, and Ji, 2024; Fedyk et al., 2024; Ouyang, Yun, and Zheng, 2024)
- **AI in markets and forecasting.** AI vs. analysts; LLM-based return prediction (Cao et al., 2024; Lopez-Lira and Tang, 2025)
- **Reproducibility and researcher discretion.** Computational reproducibility and nonstandard errors (Menkveld et al., 2024; Pérignon et al., 2024)
- **This paper.** *Unique advantage: a human golden standard*
  - Fully comparable human benchmark with *identified* analysis paths, making AI "errors" measurable and interpretable

Decision forks span the analysis path

- Each hypothesis has M decision forks $F_1, \ldots, F_M$; fork F_i has N_i mutually exclusive options $O_{i,1}, \ldots, O_{i,N_i}$
- A complete analysis path p picks one option at each fork (example at right), leading to a specific outcome y_p
- For H1, seven forks generate $2{\times}2{\times}3{\times}4{\times}3{\times}2{\times}5 = 1{,}440$ paths
- The decision tree is mostly hidden in empirical research



AI and human outcome distributions differ for 5 of 6 hypotheses



	$AD2$
H1	11.61^{***}
H2	0.87
H3	73.47^{***}
H4	5.41^{***}
H5	52.00^{***}
H6	14.16^{***}

Actual estimates are tighter and shifted, especially for complex tasks

- **Dispersion.** AI IQR/IDR smaller than human for most hypotheses
- **Location.** Median diverges more for the more complex H1, H4, H6; medians are similar for the simpler H3, H5

	H1		H2		H3		H4		H5		H6	
	A	H	A	H	A	H	A	H	A	H	A	H
Median	−0.2	−1.1	−0.2	−0.0	−3.7	−3.3	−0.8	0.1	0.5	−0.0	1.0	0.0
IQR	4.2	6.7	8.4	7.5	0.1	1.2	5.9	5.9	4.1	0.8	4.5	21.4
IDR	7.5	27.3	25.7	28.4	0.1	3.7	16.2	27.1	120.2	2.5	11.5	248.5

The mechanism: averaging high-frequency relatives then compounding is biased

Estimate a T-period change by compounding T high-frequency relatives $M_t = X_t/X_{t-1}$. Since $f(x) = x^T$ is convex, by Jensen's inequality,

$$\prod_{t=1}^{T} \underbrace{E(M_t)}_{\substack{\text{Expected} \\ \text{high frequency} \\ \text{relative}}} \quad < \quad \underbrace{E\left[\prod_{t=1}^{T} M_t\right]}_{\substack{\text{Expected low} \\ \text{frequency relative}}} .$$

First-order Taylor expanding the left side around one and subtracting one,

$$T\big(E(M_t) - 1\big) \lesssim E\left[\prod_{t=1}^{T} M_t\right] - 1.$$

- $T\times$ the average high-frequency return falls short of the low-frequency return
- The gap compounds with T: longer horizon $\Rightarrow$ larger downward bias

Only the relative-change model is nonlinear; AIs avoid it

- *Trend-stationary* ($X_t = \alpha + \beta t + \epsilon_t$): linear, no compounding
- *Log-difference*: $\log(\prod_t M_t) = \sum_t \log M_t$, linear in logs, no compounding
- *Relative change*: aggregates multiplicatively as $\prod_t M_t$: **nonlinear** $\Rightarrow$ **Jensen bias**
- Daily $\times$ relative-change is the worst combination; it produces the $\sim -1{,}560$ Q25 effect for H1
- $\sim 60\%$ of humans pick "RltvChg" but 0% of AIs do, so the AI multiverse skips this Jensen-biased path entirely
- The AI's left tail is therefore absent: dispersion compresses, location shifts up

Quantile regressions: frequency and model drive the right tail for H3

- Median gap is modest; right tail is where AIs and humans diverge
- "Mdl-RltvChg" and "Mdl-TrndSttnry" jointly suppress the AI right tail; "FrqncyAnlyss-D" adds another large negative term; AI IQR collapses to zero

	Opt	Q75	Δ_{HA}	Q75 $\times\Delta_{HA}$
Mdl	LgDff	1.26^{**}	-0.04	-0.05^{**}
	RltvChg	264.62^{***}	-0.60	-159.74^{***}
	TrndSttnry	-251.38^{***}	0.65	-162.48^{***}
FrqncyAnlyss	D	323.88^{***}	-0.34	-110.22^{***}
	M	-1.19^{***}	0.61	-0.73^{***}
	A	0.84^{***}	-0.26	-0.22^{***}
OtlrTrtmnt	Wns	228.89^{***}	-0.20	-44.66^{***}
RmvOpnCls	N	-323.61^{***}	0.16	-53.28^{***}

Robustness across LLMs (o3, GPT-5, GPT-5.1, GPT-5.2)

- Main text uses GPT-5.2 (snapshot dated December 11, 2025) for reproducibility
- Robustness across alternative OpenAI reasoning models: o3, GPT-5, and GPT-5.1
- For H1, the choice distributions across forks are qualitatively similar across all four models:
  - All concentrate on "TrndSttnry" for the "Mdl" fork
  - All sample primarily at the annual frequency
  - None apply discretionary data cleaning
- This stability across LLMs suggests that the patterns we document reflect a shared empirical style, not a quirk of any single model
- Reported in Table A of the paper (chc_diff_ha_results_h1_estmt_mltvrs across LLMs)

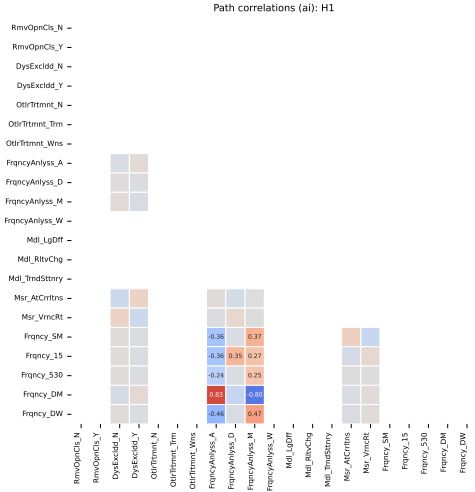
Alternative $AD2$-based fork-importance measure

- Quantile regressions focus on specific quantiles of the outcome distribution
- Alternative: use $AD2$ *within* the multiverse, fork-by-fork
- For each fork, compare conditional outcome distributions $F(Y \mid \text{option } O_{i,j})$ across humans and AIs:

$$AD2(\mathsf{H}, \mathsf{A} \,|\, \mathsf{FO}^c) = \text{contribution of fork-option } c$$

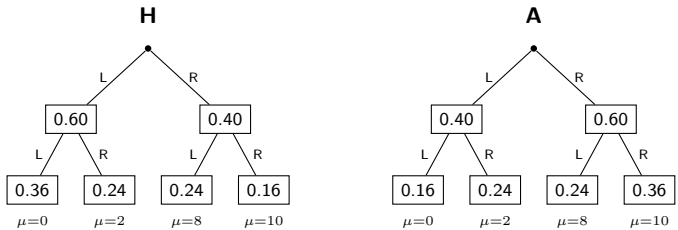
- Limitation: informative *only* when both humans and AIs exhibit nonzero variation at a fork (i.e., both groups select more than one option)
- This is why we prefer quantile regressions; reassuringly, both approaches yield qualitatively similar conclusions

Fork-option correlations: AI choices are bundled across forks



- Many off-diagonal cells are missing for AIs: certain option pairs are never jointly observed
- Conditional on availability, cross-fork correlations are more often statistically significant for AIs
- Human choices display weaker dependence across forks
- AI choices are bundled across forks
- Effective set of paths is even narrower than marginal counts suggest

A lab experiment validates the design



- Two forks, two options each; humans pick L more often, AIs pick R more often
- Fork 1 is much more consequential ($\Delta\mu = 8$) than Fork 2 ($\Delta\mu = 2$)
- $AD2 = 13.13$ ($p < 0.01$): distributions are statistically different

Quantile regression recovers the consequential fork

Fork	Q Opt	Q25	Q50	Q75	Δ_{HA}	Q50 $\times\Delta_{\text{HA}}$
F1	L	-6.39^{***}	-8.06^{***}	-9.73^{***}	-0.20	1.61^{***}
	R	6.39^{***}	8.06^{***}	9.73^{***}	0.20	1.61^{***}
F2	L	-2.37^{***}	-2.48^{***}	-2.95^{***}	-0.20	0.50^{***}
	R	2.37^{***}	2.48^{***}	2.95^{***}	0.20	0.50^{***}

- Q50 effect at F1 is ± 8.06; at F2 it is only ± 2.48
- $\Delta_{\text{HA}} = 0.20$ is identical at both forks
- Importance for the median: 1.61 at F1 vs. 0.50 at F2

$\Rightarrow$ Methodology correctly identifies F1 as the consequential fork