

AI “Errors”¹

Wenqian Huang

Albert J. Menkveld

Shihao Yu²

March 13, 2026

¹Wenqian Huang, Bank for International Settlements, wenqian.huang@bis.org. Albert J. Menkveld, Vrije Universiteit Amsterdam, albertjmenkveld@gmail.com. Shihao Yu, Singapore Management University, Lee Kong Chian School of Business, shihaoyu@smu.edu.sg. We thank Wei Jiang, Sean Cao and Fernando Perez-Cruz for their helpful comments. We are grateful to Deutsche Börse for all data support.

²Corresponding author.

AI “Errors”

Abstract

When AIs are tasked with empirical research, how do their outcomes compare to those of humans? Are the distributions similar? We run an experiment where we let AI models repeat an experiment that was run with 164 human teams. Not surprisingly, distributions differ. The deeper question is: Why? We develop an approach that identifies which decisions on the analysis path drive these differences, which we define as “AI errors.” The results show that AI concentrates on a narrow set of analysis paths, yielding markedly lower dispersion. For complex tasks, their estimates are systematically shifted relative to the human benchmark. Fork-level diagnostics and quantile regressions trace these shifts primarily to the choice of the statistical model, e.g., identifying a time trend by adding a stationary trend to a model in levels, or by computing relative changes and taking the average.

1 Introduction

Artificial intelligence (AI), powered by large language models (LLMs), is increasingly being adopted in scientific research. It is not merely used for drafting, but for research design and implementation. Evidence of growing AI involvement in scientific manuscripts is already visible in large-scale text-based measurement across major venues and preprint servers (Liang et al. 2024). In economics, this shift has been conceptualized as “cognitive automation,” where AI can automate micro-tasks across ideation, background research, coding, and data analysis (Korinek 2023) or even the entire empirical pipeline end-to-end.¹ This can increase productivity, but also raise the stakes for evaluating reliability in the parts of the pipeline where small choices can have large consequences. In finance, in particular, end-to-end generation of plausible, convention-following research output is already feasible at scale, illustrating both efficiency gains and risks, e.g., industrialized post-hoc storytelling and occasional fabricated citations (Novy-Marx and Velikov 2025).

The key open question raised by these developments is whether AI can conduct research in a manner consistent with accepted practices. This matters in finance, because pivotal choices are often implicit in coding and preprocessing, such as sample construction and outlier treatment. Modest differences along these choices can move estimates materially while remaining largely invisible in the final write-up.

Against this backdrop, we study whether AI models acting as empirical researchers produce results that align with those generated by human researchers, and, when they do not, what drives the gap. A unique feature of our setting is that the same datasets and hypotheses have already been analyzed by 164 independent human research teams, yielding a benchmark distribution that reflects how human researchers implement a shared empirical task (Menkveld et al. 2024). This human outcome distribution provides a natural yardstick for assessing AI outcomes; systematic deviations of AI outcomes can be viewed as “AI errors.”

¹For example, see the Autonomous Policy Evaluation project (APE): <https://ape.socialcatalystlab.org>.

This benchmark is particularly valuable, because the most consequential failures in empirical workflow are rarely written down precisely in the research report; they arise from latent methodological defaults embedded in code, or even implementation errors. Evaluations anchored on a single outcome can blur the line between genuine mistakes and reasonable, but different choices. By contrast, comparing the AI outcome distribution to a human outcome *distribution* makes it possible to distinguish plausible human-like implementations from systematically different implementations, and then trace deviations back to specific decisions along the analysis path.

We ask two specific questions in this paper. First, how do AI and human outcome distributions differ in location, dispersion, and tails when testing the same hypotheses using identical datasets? Second, which decision forks on the analysis-path account for these differences?

To answer these two questions, we run an experiment to compare human research teams and AI models, henceforth referred to as humans and AIs, using #fincap (Finance Crowd Analysis Project) to obtain the human outcomes. In the original #fincap study, these teams tested the same six pre-specified hypotheses on a shared proprietary dataset of EURO STOXX 50 Index Futures trades. We replicate this setting with AIs by building AI research agents that receive the same project instructions and dataset metadata as humans obtained.² Menkveld et al. (2024) analyze human outcomes in the #fincap experiment and refer to outcome variation as nonstandard error. Importantly, Menkveld et al. (2024) further analyze the key decisions that generate nonstandard errors. They do so by projecting human analysis paths onto a multiverse, which is a set of analysis paths that are constructed by identifying key forks and, for each fork, a finite set of common alternatives to pick from. We use the exact same multiverse to analyze what the source of AI errors is.

The experiment yields several insights. First, humans and AIs produce meaningfully different outcome distributions, both in location (e.g., medians) and dispersion (e.g., interquartile and

²An important concern is that #fincap outcomes have served as AI input. In Section 2.2.1, we discuss this concern, and after some analysis, we conclude that its effect, most likely, is extremely small, and would bias us in underestimating AI errors.

interdecile ranges). The gap in medians is larger for hypotheses that require more discretionary implementation choices, where AI outcomes are typically less dispersed than human outcomes. Consistent with these patterns, two-sample Anderson-Darling (*AD2*) tests reject equality of the AI and human outcome distributions for most hypotheses, underscoring that using AI for empirical analysis does change the outcome distribution.

Second, these distributional differences map directly to differences in analysis-path choices. The multiverse analysis shows that the human-AI differences can be explained by three factors. The first factor is that relative to humans, AIs concentrate on a narrower set of decisions at many decision forks along the analysis path. For example, for the decision on which statistical model to use to estimate a trend, AIs mostly add a linear trend to a model in levels, whereas human choice is more evenly spread among such linear trend, but also computing the average relative change, the average log difference. The second factor relates to AIs' concentrated frequency choices, with near-zero use of sampling at the daily frequency, which is a frequency that humans often use. The third factor is the systematic omission of discretionary data handling such as removing open/close periods, excluding settlement weeks, and applying outlier treatments, which AIs rarely implement, while a nontrivial share of humans do. Importantly, this concentration is not specific to a single LLM: We show that the same pattern holds across alternative LLMs, suggesting that these models share a stable empirical style. The fewer paths traveled in the multiverse provide a natural explanation for the lower dispersion of AI estimates.

Third, even when measurement seems straightforward (e.g., for hypotheses where metrics can be readily computed from the data), the *AD2* tests show that outcome distributions are significantly different for humans and AIs. The multiverse results suggest this can happen, because, when flexibility is limited, any systematic AI convention (e.g., their model and frequency choices) can dominate outcomes and collapse dispersion. To some extent, "simple" does not mean "immune." AIs can still induce a strong pattern of results through their default modeling conventions.

Fourth, in settings with richer design space (i.e., hypotheses with more decision forks), AI

outcomes are still less dispersed; but they are different in location relative to the human benchmark. This difference is not driven by “low-quality” human outcomes, where low-quality is identified in the same way as in [Menkveld et al. \(2024\)](#). Using a small subset of top-quality human outcomes, we find that the differences between AI and top-quality humans remain largely unchanged.

Taken together, the evidence points to a trade-off faced by researchers using AI: LLMs deliver less dispersion in outcomes, but they may do so by defaulting to a smaller set of analytical choices when benchmarked against humans.³ This reinforces the need to make discretionary steps explicit when deploying AI in empirical work, including those that may appear to be common sense to most researchers. This echoes the message that human oversight, “human in the loop”, remains essential for ensuring that AI-assisted empirical analyses reflect accepted research practice, shifting the relevant comparison from “man versus machine” to “man plus machine” ([Cao et al. 2024](#)).

Our paper is most closely related to the growing literature on the use of AI in empirical research. Existing work has primarily focused on whether large language models (LLMs) can assist with or automate specific parts of the research process. For example, [Novy-Marx and Velikov \(2025\)](#) show that LLMs can be used to automatically generate academic-style finance papers, including naming predictors and constructing narratives around empirical findings. [Horton \(2023\)](#) uses LLMs as simulated economic agents in experimental settings, while [Manning, Zhu, and Horton \(2024\)](#) illustrates how LLMs can automatically generate and test social-scientific hypotheses. Our paper differs from this line of work by shifting the focus from idea generation or narrative construction to empirical implementation: We study how AIs execute end-to-end empirical analyses and how their resulting estimates compare to those produced by human researchers when facing the same data and hypotheses.

³In this paper, AI errors are defined as AIs deviating from a human gold standard. The motivation for this standard is that human outcomes reflect “accepted practice.” However, we like to caution that some human decisions are debatable and arguably lead to biases in outcomes. In these special cases, one could say that AIs avoiding such decisions, and thus generating AI error, actually undo “human errors.” We, therefore, do not take a stance on whether the AI errors we report are right or wrong. We believe there is value in reporting them and identifying their root causes. We discuss an example of a “benevolent” AI error in [Section 4.2.2](#) (see the paragraph that contains footnote 7).

There is also a growing literature that uses AI as a tool for investment decisions. Cao et al. (2024) find that AI systems outperform human analysts in certain stock return prediction tasks, particularly when institutional knowledge is less important, while Lopez-Lira and Tang (2025) show that LLMs can predict stock returns using news headlines without domain-specific training. Our paper does not evaluate AIs’ predictive performance; instead, we treat AIs as empirical researchers and evaluate them against the methodological choices and distributional outcomes of human researchers.

Finally, our paper contributes to the literature on reproducibility and researcher discretion in empirical finance. Using the same dataset as Menkveld et al. (2024), Pérignon et al. (2024) study computational reproducibility and show that many published results cannot be mechanically reproduced even with access to code and data. Our approach is complementary: Rather than asking whether a given result can be reproduced (by default, AI results are fully reproducible), we ask how AI results differ from those of human researchers. In addition, by mapping both humans and AIs onto a common multiverse, we provide evidence on how analytical choices can drive significant differences in empirical conclusions.

The paper is organized as follows. In Section 2, we describe our experimental setup. In Section 3, we describe the methodology used to analyze the results. We discuss our results in Section 4. Section 5 concludes the paper.

2 Experiment

2.1 Human benchmark: The #fincap project

Our benchmark is the set of results from the #fincap project, which produced Menkveld et al. (2024). In this section we provide a brief overview of the #fincap design and refer readers to that paper for full details.

The #fincap (Finance Crowd Analysis Project) is a crowd-sourced research initiative in empirical finance. It was designed to study how scientific conclusions vary when multiple independent research teams test the same hypotheses on the same data. The project involved 164 independent research teams from around the world. Each team received access to the same proprietary dataset comprising 720 million trades in EURO STOXX 50 Index Futures and the same pre-specified hypotheses, and was asked to conduct its own analysis and submit a short academic paper reporting estimated effects and statistical uncertainty (e.g., standard errors and t-statistics).

For completeness, we list the six hypotheses in the #fincap project below.

H1: Market efficiency has not changed over time.

H2: Realized bid-ask spread on market orders has not changed over time.

H3: Client share of volume has not changed over time.

H4: Client realized bid-ask spreads have not changed over time.

H5: Fraction of client trades executed via market/marketable limit orders has not changed over time.

H6: Relative gross trading revenue (GTR) for clients has not changed over time.

The hypotheses differ in the amount of measurement and modeling required. Some are relatively straightforward to test (e.g., H3 and H5): The dataset includes fields indicating whether a trade is a client trade and whether it is executed via a market or marketable-limit order. Others require additional choices (e.g., H1, H2, H4, and H6). For example, testing H1 first requires selecting a measure of market efficiency, such as a variance-ratio statistic or the autocorrelation of returns.

Even for relatively straightforward hypotheses, there are many decisions to be made at different decision forks along the analysis path. For example, at what frequency should one compute client

volume share? To estimate changes over time, should one use average relative changes or estimate a regression with a time trend? More generally, should one exclude opening/closing periods or settlement weeks, and how should one treat outliers? All these decisions are nontrivial and can potentially lead to different results.

2.2 Building the AI research agent

Due to confidentiality restrictions in our agreement with Deutsche Börse, we cannot provide the proprietary dataset to large language models (LLMs). We therefore build an AI research agent that executes the project locally.

Specifically, the AI research agent receives (i) the dataset metadata and (ii) the project instructions and requirements, and uses these inputs to generate analysis code. The generated code is executed locally to produce results. If execution fails (e.g., syntax errors) or produces incomplete outputs (e.g., missing values in results files), the agent revises the code using the error messages until it runs successfully and produces the required deliverables. We detail the agent workflow in Section A in the Appendix.

2.2.1 Selecting the LLM

A central identification concern is training-data leakage: An LLM may have been trained (or subsequently updated) on information that post-dates the intended information set, creating temporal look-ahead bias (Ludwig, Mullainathan, and Rambachan 2025). While this concern is most salient in forecasting settings (Sarkar and Vafa 2024; Levy 2024), in our setting the specific risk is that the model may have been exposed to Menkveld et al. (2024) (or earlier public versions), including descriptions of the analysis choices made by human teams, thereby influencing its implementation choices.

The LLM literature proposes two approaches to mitigate look-ahead bias: (i) using open-source

models with published and fixed weights (e.g., Meta’s Llama family), or (ii) using models trained by the researcher on timestamped corpora (e.g., ChronoGPT) (Ludwig, Mullainathan, and Rambachan 2025; He et al. 2025). Neither approach is well-suited to our application. First, widely used open-source LLMs have training cutoffs that post-date early public versions of Menkveld et al. (2024) (an earlier version was posted on SSRN on November 23, 2021), so open weights do not by themselves rule out exposure. Second, chronologically trained models such as ChronoGPT are trained on a limited amount of textual data. While they are capable of semantic analysis, they are not capable of more complex tasks such as the ones required for the #fincap project (e.g., the model fails to provide correct definitions of standard concepts such as bid–ask spreads).

Given these constraints, we use the state-of-the-art reasoning model GPT-5.2 (OpenAI). To ensure reproducibility, we use a fixed snapshot dated December 11, 2025. As robustness checks, we also evaluate other OpenAI reasoning models (o3, GPT-5, and GPT-5.1). As Table D.1 shows, these models select qualitatively similar analysis paths to GPT-5.2; accordingly, the main text reports results for GPT-5.2.

To alleviate the look-ahead bias concern, we implement a text-completion diagnostic following Sarkar and Vafa (2024) and Levy (2024). We select several paragraphs from Menkveld et al. (2024) (the abstract, the first paragraph of the introduction, and one paragraph discussing results) and provide the first half of each paragraph to GPT-5.2, asking it to complete the remainder.

Results are reported in Section B in the Appendix. Across all selected passages, the model fails to reproduce the original continuations and instead generates completions that are materially different from the source text, suggesting limited verbatim memorization of these paragraphs. Also, we note that we use API calls and do not allow for web searching, so the results by GPT-5.2 should not be directly influenced by the “Nonstandard Errors” paper, or, more generally, by the #fincap project.

We therefore view direct textual leakage from the most salient #fincap materials as unlikely to be a first-order driver of the AI-human differences that we document. We note, however, that this

diagnostic is not definitive evidence of non-exposure: A model may have trained on a document without reliably generating verbatim continuations. Our inference is correspondingly narrow: We find no evidence of direct textual memorization in the tested passages. Moreover, to the extent that residual leakage exists, it would most plausibly bias the AI toward human-like implementations, attenuating rather than generating the divergences that we observe.

2.2.2 Prompt engineering

Interacting with an LLM requires specifying system and user prompts. In our setting, the system prompt instructs the model to act as a finance researcher and to follow the user prompt to produce the required deliverables.

To align the AI and human information sets, we provide the AI agent with the same project materials that were available to human teams in #fincap. In particular, human teams received three sources of information:

- Instruction sheet (context, assignment, and hypotheses);
- The #fincap website (project overview and detailed data description); and
- Email communications from the project coordinators (updates and clarifications).

We merge the relevant text from these sources and include it in the user prompt. The exact prompts are provided in Section A.3 in the Appendix.

2.2.3 Identifying the analysis path

After code generation and successful execution, we use the same model (GPT-5.2) in a separate parsing step to identify the analysis path implemented in each AI run. The parsing agent takes as input: (i) a reference CSV that enumerates analysis forks (by hypothesis) and the set of admissible alternatives at each fork from the #fincap project, and (ii) the generated Python code. It then

performs a structured code audit to classify which alternatives were implemented. Details are provided in Section A in the Appendix.

Two design features help ensure that the parsing step does not influence the code-generation step. First, the reference fork table is never provided to the code-generating agent; it is used only for auditing. Second, parsing and code generation are executed as separate processes, with independent LLM calls. The parsing agent does not observe the code-generation dialogue and cannot pass information back to it, even though both stages use the same underlying model.

3 Methodology

We have two primary goals. First, we want to see whether AIs will produce estimates that are different from humans. Second, we want to understand what drives the differences, if there are any. To arrive at an estimate for a given hypothesis, one has to make many choices at various decision forks. Take H1 as an example. To test the hypothesis that market efficiency has not changed over time, a research team will need to first decide on the choice of efficiency measure. Should variance ratio or autocorrelation or some other measure be used? At which frequency should the analysis be conducted? And, should outliers be excluded? Essentially, our goal is to identify the key fork-option combinations that drive the difference in outcome distributions between AIs and humans.

The first goal is descriptive and inferential. After running the experiment, we report summary statistics of the outcome distributions for humans and AIs. To assess whether the distributions differ statistically, we apply the two-sample Anderson-Darling test, denoted $AD2(H, A)$. The asymptotic distribution for $AD2$ is well-known under the null of the distribution being the same. Appendix F in Menkveld et al. (2024) provides a short discussion of the Anderson-Darling test.

For the second goal, we use fork-by-fork quantile regressions to quantify how outcomes differ across analysis paths that do versus do not select a given option. This is because we are interested

in the distributional differences (e.g., location, dispersion, and tails) between AIs and humans. Quantile effects therefore provide a direct and robust way to link fork choices to where the outcome distributions diverge. We then combine these distributional shifts with AI-human differences in option selection frequencies to identify candidate drivers of the aggregate divergence. We will detail it below.

3.1 Decision forks and quantile effects

Assume there are M decision forks, denoted F_1, F_2, \dots, F_M . Each fork F_i has N_i mutually exclusive options, denoted $O_{i,1}, O_{i,2}, \dots, O_{i,N_i}$, where $O_{i,j}$ is the j -th option of the i -th fork. A complete analysis path is a specific combination selecting exactly one option at each fork.

Let p index analysis paths, and let Y_p denote the outcome associated with analysis path p . For each fork-option pair (i, j) , define the indicator variable

$$D_{i,j}^{(p)} \equiv \mathbb{1}\{\text{analysis path } p \text{ selects option } O_{i,j}\}.$$

To characterize how a given option relates to different parts of the outcome distribution, we estimate univariate quantile regressions. For a given quantile level $\tau \in (0, 1)$, we estimate the conditional quantile function

$$Q_Y(\tau \mid D_{i,j}^{(p)}) = \alpha(\tau) + \beta_{i,j}(\tau) D_{i,j}^{(p)},$$

where $\beta_{i,j}(\tau)$ captures the difference in the τ -th quantile of the outcome distribution between analysis paths that select option $O_{i,j}$ and those that do not.

The quantile regression coefficient $\beta_{i,j}(\tau)$ measures the importance of a fork option by quantifying how strongly outcomes differ in that part of the distribution across paths that do versus do not implement $O_{i,j}$. Because fork choices can be correlated along the analysis pipeline, this univariate

estimate should be interpreted as a descriptive (non-causal) association that may reflect bundled, correlated choices rather than an isolated effect of the single fork.⁴ However, for a fork option to contribute to systematic differences between human and AI outcome distributions, humans and AIs must also differ in how frequently they select that option. We therefore additionally consider the human-AI choice difference

$$\Delta_{\text{HA}}(i, j) \equiv w_{\text{A}}(i, j) - w_{\text{H}}(i, j),$$

where $w_I(i, j)$, $I \in \{\text{H}, \text{A}\}$ are the fractions of human or AI analysis paths that select option $O_{i,j}$ at fork F_i . This captures differences between humans and AIs in selection probabilities at fork F_i , option $O_{i,j}$.

3.2 Lab experiment

We illustrate the methodology in a simple lab experiment. For simplicity, suppose there are only two forks, 1 and 2, and only two options, L and R, for each fork. So in total, there are four different analysis paths: LL, LR, RL, and RR. Let the different analysis paths lead to four normal distributions with means $\mu_{\text{LL}} = 0$, $\mu_{\text{LR}} = 2$, $\mu_{\text{RL}} = 8$, $\mu_{\text{RR}} = 10$, and standard deviations $\sigma_{\text{LL}} = \sigma_{\text{LR}} = 2$ and $\sigma_{\text{RL}} = \sigma_{\text{RR}} = 4$.

Consider the following scenario where humans and AIs differ in their choices at both Fork 1 and 2.

- H: $w_{\text{H}}(\text{LL}) = 0.36$, $w_{\text{H}}(\text{LR}) = 0.24$, $w_{\text{H}}(\text{RL}) = 0.24$, $w_{\text{H}}(\text{RR}) = 0.16$.
- A: $w_{\text{A}}(\text{LL}) = 0.16$, $w_{\text{A}}(\text{LR}) = 0.24$, $w_{\text{A}}(\text{RL}) = 0.24$, $w_{\text{A}}(\text{RR}) = 0.36$.

So, 36%, 24%, 24%, and 16% of humans pick LL, LR, RL, and RR, respectively, while 16%,

⁴The univariate specification provides a transparent decomposition of where outcome variation concentrates across the analysis pipeline. It is not designed to identify a causal effect of a single fork when forks co-move.

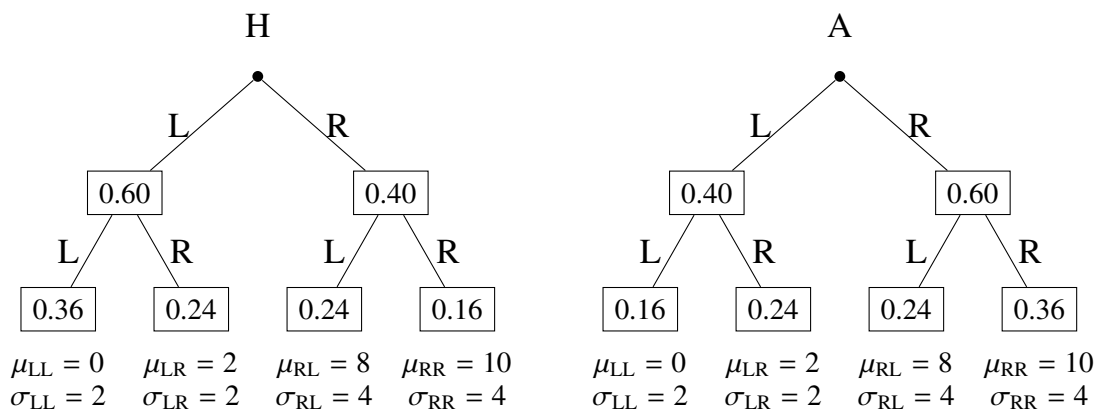


Figure 1. Analysis paths chosen by humans and AIs for the lab experiment. This figure illustrates the decision trees for human (H, left) and AI (A, right) research teams in a toy scenario with two forks. Numbers inside the boxes represent the fractions of analysis paths that choose each specific option at each fork. Numbers below the terminal boxes are the means and standard deviations of the outcome distributions for the corresponding analysis paths.

24%, 24%, and 36% of AIs pick LL, LR, RL, and RR, respectively. Figure 1 illustrates these choices.

By construction, Fork 1 is more consequential for outcomes than Fork 2: Switching from *L* to *R* at Fork 1 produces a large upward shift in both the mean and the dispersion (moving from the LL/LR states to the RL/RR states), whereas switching at Fork 2 produces only a modest shift (LL vs. LR and RL vs. RR). Humans and AIs differ in their choice at both Fork 1 and 2: AIs tend to pick R while humans tend to pick L at both forks. If our methodology works as intended, we should find that the difference in outcome distributions between AIs and humans is primarily driven by their choice at Fork 1.

To implement the lab experiment, we first simulate the outcome distributions for both the human and AI samples according to their own choices. We set the total number of analysis paths to 250. Out of the 250 human analysis paths, 90 outcomes are simulated from a normal distribution with mean 0 and standard deviation 2, that is, the outcome distribution for LL. Similarly, 60 outcomes are simulated from a normal distribution with mean 2 and standard deviation 2, that is, the outcome distribution for LR. The same exercise is done for RL and RR. Essentially, the ultimate outcome distributions are mixed normals.

Table 1. Univariate quantile estimates for the lab experiment. This table reports univariate quantile regression results for the lab experiment. For each fork-option combination, the table reports: Q25, Q50 (median), and Q75 (quantile effects), showing how selecting a given fork-option affects the 25th, 50th, and 75th percentiles of the outcome distribution relative to selecting alternative options at the same fork; Δ_{HA} , the difference in selection probabilities between AIs and humans; and the product terms $Q25 \times \Delta_{HA}$, $Q50 \times \Delta_{HA}$, and $Q75 \times \Delta_{HA}$ (combined importance measures). Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Fork	Quantile Option	Q25	Q50	Q75	IQR	Δ_{HA}	$Q25 \times \Delta_{HA}$	$Q50 \times \Delta_{HA}$	$Q75 \times \Delta_{HA}$	$IQR \times \Delta_{HA}$
F1	L	-6.39***	-8.06***	-9.73***	-3.34	-0.20	1.28***	1.61***	1.95***	0.67
	R	6.39***	8.06***	9.73***	3.34	0.20	1.28***	1.61***	1.95***	0.67
F2	L	-2.37***	-2.48***	-2.95***	-0.58	-0.20	0.47***	0.50***	0.59***	0.12
	R	2.37***	2.48***	2.95***	0.58	0.20	0.47***	0.50***	0.59***	0.12

Figure C.1 in the appendix illustrates the empirical outcome distributions for the lab experiment. The results are as intended. Compared with humans, AIs choose more often the option R at both Fork 1 and Fork 2, leading to higher means and standard deviations. The *AD2* statistic is 13.13 and statistically significant at the 1% level, confirming that human and AI outcome distributions are statistically different. We now turn to the quantile regressions to quantify the impact of the fork options on the outcome distributions.

Table 1 reports the univariate quantile regression results for the lab experiment. As intended, the results show that the difference in outcome distributions between AIs and humans is primarily driven by their choice at Fork 1. Focusing on the median, $Q50 \times \Delta_{HA}$ is 1.61 for Fork 1-Option R and 0.50 for Fork 2-Option R⁵. These results arise because the median shift associated with choosing *R* at Fork 1 is large (median difference 8.06), whereas the median shift associated with choosing *R* at Fork 2 is modest (median difference 2.48), even though the human-AI choice difference Δ_{HA} is the same at both forks.

In Section C of the appendix, we develop another fork-option importance measure based on the two-sample AD test to identify the fork-option combinations that drive the difference in outcome distributions between AIs and humans. Instead of focusing on specific quantiles, this measure focuses on the overall outcome distribution. Nonetheless, this *AD2*-based measure is informative

⁵Fork 1-Option L and Fork 1-Option R are mirror images of each other.

only when both humans and AIs exhibit nonzero variation at a fork (i.e., both groups select more than one option). This limitation of an *AD2*-based measure is why we prefer quantile regressions. It is reassuring that both approaches yield qualitatively similar results.

4 Results

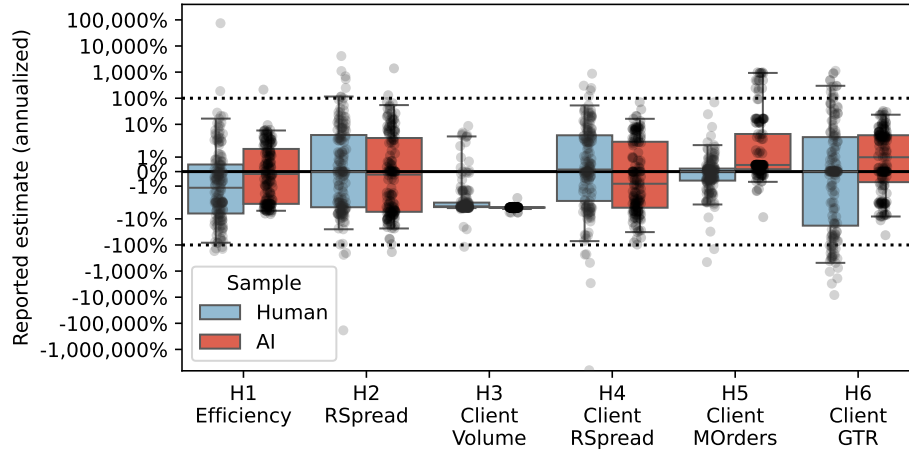
This section reports our findings. We first compare the outcome distributions between AIs and humans. The AI outcomes are generated by an AI research agent that reads the #fincap project instructions and dataset metadata, generates analysis code, and executes the code to produce estimates. The human outcomes are obtained from the #fincap project. We then use a multiverse framework to interpret the differences between AIs and humans by mapping each estimate to an explicit sequence of fork choices and comparing how AIs and humans traverse the admissible analysis space.

4.1 Do human and AI estimates differ?

Figure 2 plots the outcome distributions for AIs and humans, and reports the two-sample Anderson-Darling test statistics (*AD2*). These results lead to a couple insights. First, the AI estimates show nontrivial dispersion. This might seem odd, since we used a single AI model to generate the outcomes (see Section 2.2.1). The reason that a single model produces variation is that LLM output is probabilistic. Each time it processes a request, even if it is the same request, it will sample from a distribution of answers. This is why there is AI error to begin with.

Second, the *AD2* tests show that for all hypotheses except for one, the human and AI outcome distributions are significantly different. The exception is hypothesis H2 (RSpread). Notably, *AD2* is largest for the relatively direct hypotheses H3 (Client Volume) and H5 (Client MOrders), and smaller for more discretion-intensive hypotheses such as H1 (Efficiency). In other words, the

Figure 2. Dispersion of estimates: Human vs. AI. This figure presents box plots comparing the outcome distributions between humans (in blue) and AIs (in red) across six hypotheses (H1: Market Efficiency, H2: Realized Spread, H3: Client Volume, H4: Client Realized Spread, H5: Client Market Orders, H6: Client Gross Trading Revenue). Each box plot shows the median (center line), interquartile range (box), and whiskers extending to the 2.5% and 97.5% quantiles. The plots further depict all estimates as gray dots.



Hypothesis	<i>AD2</i>
H1 (Efficiency)	11.61 ^{***}
H2 (RSpread)	0.87
H3 (Client Volume)	73.47 ^{***}
H4 (Client RSpread)	5.41 ^{***}
H5 (Client MOrders)	52.00 ^{***}
H6 (Client GTR)	14.16 ^{***}

Anderson-Darling tests show that the outcome distributions are, statistically speaking, most different between AIs and humans for H3 and H5. Interestingly, this pattern does not seem to be driven by the number of forks. For H3 and H5 the underlying quantities are directly identified in the raw data (client-initiated trades and market/marketable-limit orders are flagged), while for hypotheses such as H1, additional discretionary choices are required (e.g., defining an efficiency measure and selecting the frequency at which it is computed).

Table 2. Summary statistics of estimates: Human vs. AI. This table reports summary statistics comparing the outcome distributions between humans and AIs across six hypotheses. For each hypothesis and sample, the table reports: N (sample size), Mean, SD (standard deviation), Min, Q(0.10) through Q(0.90) (deciles), Median, IQR (interquartile range, Q0.75 - Q0.25), IDR (inter-decile range, Q0.90 - Q0.10), and Max. This table complements Figure 2.

Hypothesis	H1 Efficiency		H2 RSpread		H3 Client Volume		H4 Client RSpread		H5 Client MOrders		H6 Client GTR	
	AI	Human	AI	Human	AI	Human	AI	Human	AI	Human	AI	Human
Sample	158	164	158	164	158	164	158	164	158	164	158	164
N	158	164	158	164	158	164	158	164	158	164	158	164
Mean	1.1	446.3	8.3	-1093.4	-3.7	-3.5	-2.2	-38276.1	65.8	-3.5	2.2	-87.1
SD	17.2	5817.5	113.3	14537.2	0.4	9.4	15.1	490024.2	209.1	37.6	7.6	728.5
Min	-6.6	-171.1	-187.1	-186074.5	-5.8	-117.5	-96.3	-6275383.0	-8.6	-452.9	-42.2	-8254.5
Q(0.10)	-4.1	-23.7	-10.4	-6.9	-3.7	-3.8	-9.3	-6.7	-0.2	-1.6	-2.6	-192.1
Q(0.25)	-2.7	-6.2	-5.4	-3.6	-3.7	-3.5	-3.7	-2.1	0.2	-0.6	-0.7	-18.2
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
Q(0.75)	1.6	0.5	3.0	3.9	-3.7	-2.4	2.1	3.8	4.3	0.2	3.8	3.2
Q(0.90)	3.4	3.7	15.3	21.5	-3.7	-0.1	6.8	20.4	120.0	1.0	8.9	56.5
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
Max	213.3	74491.1	1389.8	4124.0	-1.8	8.7	70.4	870.2	981.1	69.5	32.6	1119.0

While the Anderson-Darling tests can tell us whether the human and AI outcome distributions are statistically different or not, the summary statistics reveal more details about the differences in location and dispersion. Table 2 reports common summary statistics, in particular the different quantiles, interquartile range (IQR), and inter-decile range (IDR). Several observations emerge from the comparison.

First, the wedge in the *location* of the outcome distributions of humans and AIs, as measured by the differential in their medians, differs notably across hypotheses. The median outcome of AIs and humans diverges more for more complex hypotheses including H1 (Efficiency), H4 (Client RSpread), and H6 (Client GTR). In contrast, medians are more similar for simpler hypotheses such as H3 (Client Volume) and H5 (Client MOrders). H2 (RSpread) is an exception, because, while rather complex, the median outcomes of AIs and humans are quite similar (consistent with the insignificant *AD2* statistic reported in Figure 2).

Second, AI outcomes are generally less dispersed than human outcomes. Across most hypotheses, the IQR of the AI outcome distribution is smaller than that of the human outcome distribution. The main exception is H5 (Client MOrders): For H5, AIs exhibit both a larger IQR and a substantially larger IDR. In addition, AI distributions tend to exhibit fewer extreme observations. Minimums, maximums, and tail quantiles show that human outcomes are more prone to extreme values in most hypotheses. The only exception is H5 (Client MOrders), where AIs exhibit much larger right-tail extremes, as reflected in both the $Q(0.90)$ decile and the maximum.

In summary, the results reveal that AI and human outcomes differ, even for simple hypotheses. The AI outcome distribution is less dispersed, and its location measured by the median is different from the human outcome distribution, in particular for more complex hypotheses. In the following section, we conduct a multiverse analysis to understand why the AI and human estimates differ. In other words, which decision forks drive these outcome differences?

4.2 Why do human and AI estimates differ?

To generate insight into why AI and human estimates differ, we open the black box of the analysis process by setting up the multiverse analysis proposed by [Menkveld et al. \(2024\)](#) (Table V). For each hypothesis, they identify several key decision forks with, for each fork, a set of common alternatives to pick from. Some forks are common to all hypotheses, while others are hypothesis-specific. For example, whether or not to remove the opening or closing observations (“RmvOpnCls”) is a common fork, while the choice of the market efficiency measure (“Msr”) is specific to H1 (Efficiency). As for the options, they are specific to the forks. For the “RmvOpnCls” fork, there are only two options: “Yes” and “No”. For the “Msr” fork, the options are variance ratio (“VrncRt”) and autocorrelations (“AtCrrltns”). For H1 (Efficiency), there are seven forks, and they generate $2 \times 2 \times 3 \times 4 \times 3 \times 2 \times 5 = 1,440$ distinct analysis paths. Each analysis path leads to a corresponding estimate.

For humans, the analysis paths are identified from self-reported survey responses, while AI paths are recovered using an automated code-parsing agent (see Section 2.2.3 for details). Note that the same path can be chosen by multiple humans or AIs. As a result, the observed outcome samples form a weighted multiverse, where weights reflect the frequency with which each path is chosen.

The motivation for the multiverse analysis is interpretability: Once projected onto the multiverse, each outcome estimate can be traced back to a specific analysis path, fully reflecting all the choices made by humans and AIs in the analysis process. Therefore, differences in AI and human estimates, if any, can be explained by the differences in their choices at various decision forks along the analysis path.

4.2.1 Do human and AI outcome distributions differ in the multiverse?

In the previous section, we showed that the *actual* estimates of AIs and humans differ. How about the corresponding estimates that come out of the multiverse analysis? These are the estimates of the multiverse analysis path that is closest to the *actual* path that humans and AIs used.

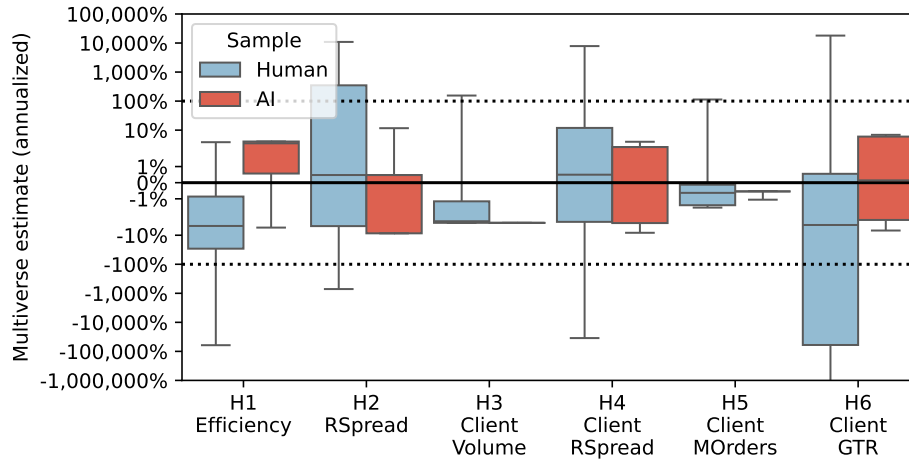
As before, Figure 3 plots the multiverse outcome distributions of humans and AIs and reports the two-sample Anderson-Darling test statistics. It shows that the AI outcome distribution in the multiverse is statistically different from the human outcome distribution, for all hypotheses, with H1 (Market Efficiency) showing the largest difference, followed by H3 (Client Volume). This is interesting, because these two hypotheses are very different in nature: H1 has much more flexibility in the analysis paths, while H3 is much more straightforward. Yet, AI and human estimates differ in both cases. Interestingly, for H2, the multiverse produces a significant *AD2* statistic this time, which, eyeballing the boxplots of Figures 2 and 3, seems driven by the larger dispersion in human estimates in the multiverse.⁶

Table 3 reports summary statistics. Perhaps the most striking pattern is that AI estimates exhibit substantially lower dispersion than human estimates, as measured by the IQR and IDR. We observe the same patterns as we did for the actual estimates. But this pattern is much more pronounced in the multiverse: For both H3 (Client Volume) and H5 (Client MOrders), the IQR is zero.

Again, these results reflect differences in the complexity of the hypotheses. H3 (Client Volume) and H5 (Client MOrders) are based on relatively straightforward calculations: The underlying data directly identify client-initiated trades and market orders, leaving limited scope for alternative analytical choices. Consequently, the multiverse for these hypotheses contains relatively few forks, and, given that all AIs consistently select a small set of options, the resulting outcome distributions collapse or nearly collapse. In this sense, the low dispersion of AI estimates for these hypotheses is mechanically linked to the limited flexibility they experience when executing the underlying

⁶The actual outlier analysis humans implemented might have been tuned to remove extremely high or low estimates.

Figure 3. Dispersion of multiverse estimates: Human vs. AI. This figure presents box plots comparing the distribution of multiverse estimates between humans and AIs across six hypotheses. Multiverse estimates are computed by mapping each team’s analysis path choices onto a common multiverse of admissible analysis paths. Each box plot shows the median (center line), interquartile range (box), and whiskers extending to 2.5% and 97.5% quantiles. Outliers beyond the whiskers are shown as individual points.



<i>AD2</i>	
Hypothesis	
H1 (Efficiency)	107.07***
H2 (RSpread)	26.21***
H3 (Client Volume)	93.48***
H4 (Client RSpread)	15.26***
H5 (Client MOrders)	30.52***
H6 (Client GTR)	35.32***

empirical task.

In contrast, the remaining hypotheses involve more analytical choices, leading to greater dispersion in multiverse outcomes. For example, H1 requires researchers to pick an efficiency measure, and if they pick a particular measure, they might need to make additional decisions. For instance, if they pick the variance ratio as an efficiency measure, then they have to choose the sampling frequencies used in the numerator and in the denominator. These additional degrees of freedom expand the multiverse and allow both AIs and humans to generate a wider range of outcomes.

In addition to differences in dispersion as measured by the IQR and IDR, the location of the outcome distributions also differs between human and AI estimates, as reflected in the median.

Similar to the results based on actual estimates, the medians of the AI and human distributions are closely aligned for simpler hypotheses. For more complex hypotheses, however, differences in distributional location become more pronounced.

Combining AI-human differences in dispersion and location, an important observation emerges. For simple hypotheses, AIs have similar distributional locations as humans while exhibiting substantially lower dispersion. Interpreted conservatively, this suggests that AIs achieve comparable outcomes with less variability. In contrast, for more complex hypotheses, although AIs continue to display lower dispersion, their locations differ from those of humans. This indicates that the tighter AI distributions are not simply less dispersed versions of the human distributions, but they exhibit different locations.

This observation highlights an important trade-off when using or involving AI in empirical research. Lower dispersion alone does not guarantee closer alignment with human benchmarks when hypotheses involve greater analytical complexity. If human estimates are considered to be the gold standard that machines are required to emulate, then more human guidance is needed.

Table 3. Summary statistics of estimates: Human vs. AI. Multiverse estimates. This table reports summary statistics comparing the distribution of multiverse estimates between humans and AIs across six hypotheses. Multiverse estimates are computed by mapping the analysis path choices of each human and AI team onto a common multiverse of admissible analysis paths. For each hypothesis and sample, the table reports: N (sample size), Mean, SD (standard deviation), Min, Q(0.10) through Q(0.90) (deciles), Median, IQR (interquartile range), IDR (inter-decile range), and Max.

Hypothesis	H1 Efficiency		H2 RSpread		H3 Client Volume		H4 Client RSpread		H5 Client MOrders		H6 Client GTR	
	AI	Human	AI	Human	AI	Human	AI	Human	AI	Human	AI	Human
Sample	158	164	158	164	158	164	158	164	158	164	158	164
Mean	1.9	-5383.3	-2.4	557.3	-3.8	23.9	-0.8	2698.5	-0.6	19.8	0.1	-3661807776.3
SD	3.0	18901.3	5.4	7930.5	0.0	56.9	4.6	43712.4	0.2	43.2	5.4	36351053006.3
Min	-7.0	-157963.0	-9.2	-60759.7	-3.8	-9.4	-8.2	-54498.9	-1.1	-2.1	-24.7	-452048467844.7
Q(0.10)	-3.1	-13893.5	-8.6	-8.6	-3.8	-3.8	-8.2	-17448.4	-1.1	-1.5	-6.9	-5025932.8
Q(0.25)	0.6	-29.0	-8.6	-4.8	-3.8	-3.7	-3.8	-3.5	-0.5	-1.4	-3.0	-60227.8
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
Q(0.75)	4.1	-0.9	0.5	349.2	-3.8	-1.2	2.6	11.9	-0.5	-0.1	6.0	0.5
Q(0.90)	4.1	0.4	4.3	5031.2	-3.8	143.9	2.9	3516.4	-0.5	103.4	6.9	38.7
IQR	3.5	28.2	9.1	354.1	0.0	2.6	6.4	15.4	0.0	1.3	9.0	60228.4
IDR	7.1	13893.9	12.9	5039.8	0.0	147.6	11.2	20964.8	0.5	104.8	13.8	5025971.5
Max	4.1	11.6	11.7	32766.7	-3.7	164.8	4.0	380143.6	-0.1	151.9	6.9	60868435124.5

One potential concern is that human outcomes differ in quality, making it inappropriate to treat the full human sample as a benchmark. To address this, we restrict the human sample to a subset of humans with the highest-quality outcomes, and compare their multiverse estimates to those of AIs. Specifically, we retain only human outcomes that satisfy three criteria: (i) the first principal component of several quality indicators of the human researchers (including academic rank and publication record) is in the top 25% of the distribution; (ii) code reproducibility is in the top quartile (corresponding to full reproducibility); and (iii) the average peer-review rating across hypotheses is in the top quartile. Only eight human outcomes satisfy all three criteria and are included in this subsample.

The resulting multiverse estimates are reported in Table E.1 and Figure E.2 in the appendix. The results show that the medians of the highest-quality humans are very close to those of the full human sample across hypotheses. This suggests that the earlier distributional differences between AI and human estimates are unlikely to be driven by lower-quality human outcomes.

4.2.2 Which fork options drive human-AI differences in the multiverse?

Having shown that AI and human estimates differ in the multiverse, we now turn to the next question: Which forks drive this difference? We first discuss the choices that humans and AIs make at the different forks, then implement univariate quantile regressions to quantify the impact of fork-options on the outcome differences.

It would be tedious to report the choice distribution for all hypotheses. Instead, we focus on H1 (Efficiency) and H3 (Client Volume). This choice is motivated by the fact that H1 (Efficiency) can serve as an example of a relatively complex hypothesis, while H3 (Client Volume) serves as a relatively simple hypothesis. In addition, the $AD2$ statistic is largest for H1 (Efficiency) among the four relatively more complex hypotheses, and larger for H3 (Client Volume) than for H5 (Client MOrders), which is another relatively simple hypothesis.

Human and AI analysis path choices. Tables 4a and 4b report the choice distributions for H1 (Market Efficiency) and H3 (Client Volume), respectively. A key feature of both tables is the much higher concentration of AI choices relative to human choices: For most forks, AIs select a narrow subset of options compared to humans who select a wider set of options.

Several forks are shared across both hypotheses and reveal consistent patterns. First, at the three discretionary data-handling forks, removing open/close observations (“RmvOpnCls”), excluding days (“DysExcldd”), and treating outliers (“OtlrTrtmnt”), no AI applies any of these filters, while a nontrivial share of humans does (up to 34% for outlier treatment in H1). This pattern holds virtually unchanged between H1 and H3, suggesting that AI agents systematically forgo data cleaning steps that a substantial fraction of human researchers do implement.

Second, the model (“Mdl”) fork shows perhaps the sharpest and most consistent divergence. All AIs select the trend-stationary specification in both hypotheses, whereas humans distribute mainly between relative changes (58–60%) and trend-stationary (35–36%), with very few opting for log differences. This uniform AI preference for trend stationarity is a robust feature that does not depend on hypothesis complexity.

Third, AI choice of the frequency at which the data is sampled for the analysis (“FrqncyAnlyss” fork) is concentrated, but it differs across hypotheses, which is interesting. For H1, AIs sample mostly at the annual frequency (75%), whereas for H3, AIs shift to a monthly frequency (82%), potentially reflecting the higher-frequency nature of volume dynamics. In both cases, human choices are spread more broadly, including a substantial fraction who sample at a daily frequency, which is what virtually no AI does. For H1, two additional forks are present. At the efficiency measure (“Msr”) fork, the AI–human gap is modest: Both prefer the variance ratio (71% vs. 63%, respectively), with AIs slightly more concentrated on this measure. For the frequencies picked for the variance ratio at the “Frqncy” fork, however, divergence is sharp: AIs concentrate on the daily-to-monthly sampling interval (DM, 67%), while humans spread across all five options with no single dominant choice. For H3, the hypothesis-specific fork concerns the unit of measurement (“Units”):

Humans predominantly use number of contracts (70%), whereas AIs are split more evenly between euros and contracts (51% vs. 49%, respectively).

More broadly, the concentration of AI choices implies that differences in AI and human estimates can arise not only from how outcomes are weighted within the multiverse, but also from which regions of the analysis space are explored. While this does not, by itself, indicate that one approach is superior, it provides a parsimonious explanation for the tighter outcome distributions observed for AI in earlier results: After projecting the analysis paths onto the multiverse, the results show that AIs place greater mass on a smaller set of analysis paths, whereas human choices span a larger set of paths.

Human and AI analysis path correlations. Beyond differences in choice frequencies, AIs and humans also differ in the joint structure of their choices across decision forks. Figure 4 reports correlation heatmaps for H1 and H3, and two patterns emerge consistently across both hypotheses. First, many off-diagonal cells are missing for AIs because certain option pairs are never jointly observed: When an option is chosen by no AI at a given fork, all its pairwise correlations with other forks are undefined. This is an immediate result of the concentration in the marginal distributions of AIs as discussed in Section 4.2.2. This implies that AIs effectively traverse a smaller region of the multiverse.

Second, conditional on availability, cross-fork correlations are more often statistically significant for AIs than for humans. Human choices display weaker dependence across forks, suggesting that human researchers make their decisions at each fork more independently. For AIs, choices tend to be bundled: Selecting one option at a fork is more predictive of choices at other forks.

Taken together, both the higher concentration in marginal distributions, and the higher correlation in options across forks, reflect the tighter set of paths traveled by AIs relative to humans.

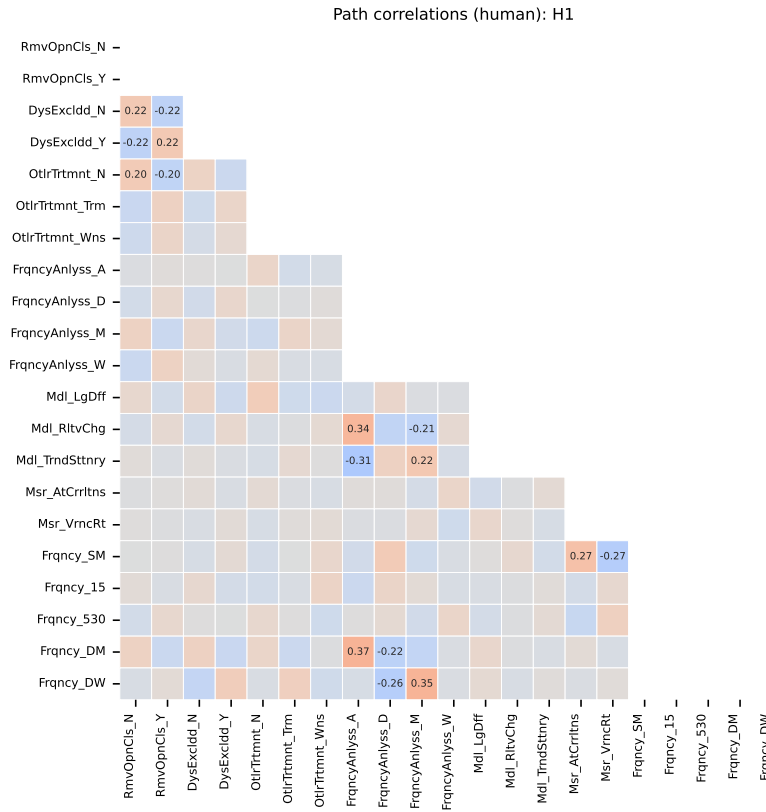
Table 4. Analysis path choices: AI vs. human. This table reports the choice distribution between humans and AIs across different decision forks in the multiverse analysis. Each row represents a fork-option combination, showing the percentage of paths that choose each option. Specifically, # H and # A are the number of humans and AIs that choose the fork-option combination, and w_H and w_A are the weights of the fork-option combination for humans and AIs, respectively. Δ_{HA} is the difference in weights between AIs and humans.

(a) H1 (Market Efficiency).

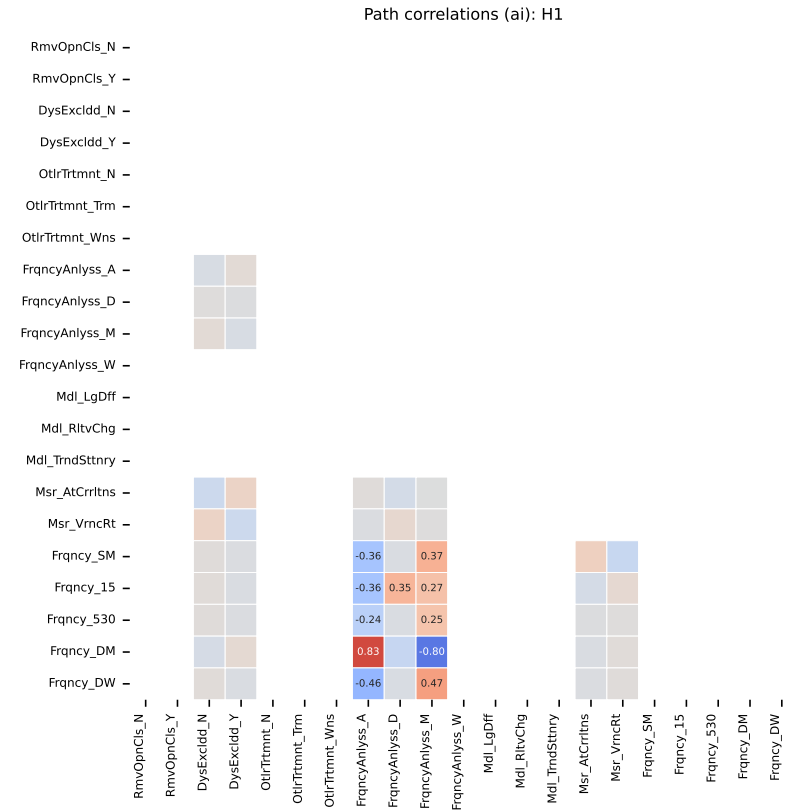
Fork	Option	# H	# A	w_H	w_A	Δ_{HA}
		164	158			
RmvOpnCIs	N	125	158	0.76	1.00	0.24
	Y	39	0	0.24	0.00	-0.24
DysExcldd	N	137	157	0.84	0.99	0.16
	Y	27	1	0.16	0.01	-0.16
OtlrTrtmnt	N	108	158	0.66	1.00	0.34
	Trm	23	0	0.14	0.00	-0.14
	Wns	33	0	0.20	0.00	-0.20
FrqncyAnlyss	D	52	2	0.32	0.01	-0.30
	W	1	0	0.01	0.00	-0.01
	M	38	38	0.23	0.24	0.01
	A	73	118	0.45	0.75	0.30
Mdl	LgDff	10	0	0.06	0.00	-0.06
	RltvChg	95	0	0.58	0.00	-0.58
	TrndSttnry	59	158	0.36	1.00	0.64
Msr	AtCrrltns	61	46	0.37	0.29	-0.08
	VrncRt	103	112	0.63	0.71	0.08
Frqncy	SM	29	11	0.18	0.07	-0.11
	15	42	15	0.26	0.09	-0.16
	530	56	11	0.34	0.07	-0.27
	DW	21	15	0.13	0.09	-0.03
	DM	16	106	0.10	0.67	0.57

(b) H3 (Client Volume).

Fork	Option	# H	# A	w_H	w_A	Δ_{HA}
		164	158			
RmvOpnCls	N	137	158	0.84	1.00	0.16
	Y	27	0	0.16	0.00	-0.16
DysExcldd	N	137	157	0.84	0.99	0.16
	Y	27	1	0.16	0.01	-0.16
OtlrTrtmnt	N	111	158	0.68	1.00	0.32
	Trm	21	0	0.13	0.00	-0.13
	Wns	32	0	0.20	0.00	-0.20
FrqncyAnlyss	D	61	5	0.37	0.03	-0.34
	W	1	0	0.01	0.00	-0.01
	M	34	129	0.21	0.82	0.61
	A	68	24	0.41	0.15	-0.26
Mdl	LgDff	7	0	0.04	0.00	-0.04
	RltvChg	99	0	0.60	0.00	-0.60
	TrndSttnry	58	158	0.35	1.00	0.65
Units	Euro	50	80	0.30	0.51	0.20
	NOfCncts	114	78	0.70	0.49	-0.20



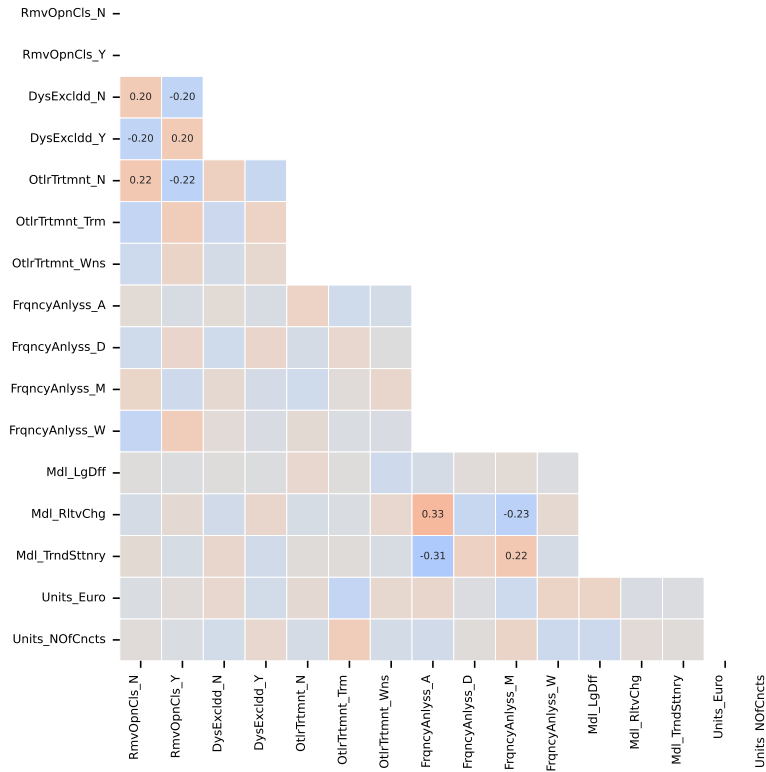
(a) Human decisions, H1 (Efficiency)



(b) AI decisions, H1 (Efficiency)

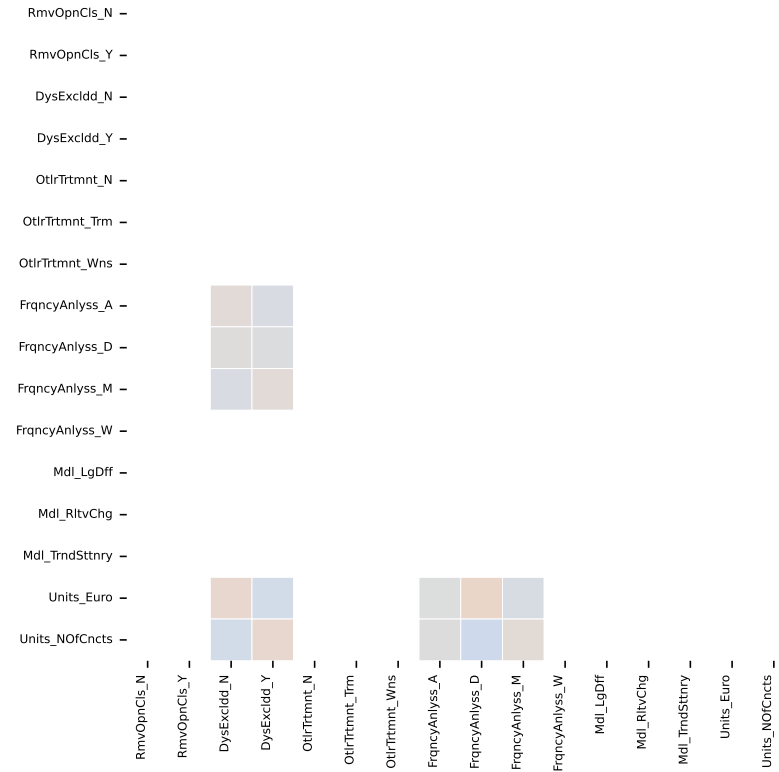
Figure 4. Fork-option choice correlations: Human vs. AI. This figure presents correlation heatmaps showing pairwise correlations between fork-option choices for humans and AIs in the multiverse analysis. Each cell represents the correlation coefficient between two fork-option choices. Red indicates positive correlation, blue indicates negative correlation, and darker colors indicate stronger correlations. Coefficients larger than 0.2 in absolute value are labeled.

Path correlations (human): H3



(c) Human decisions, H3 (Client Volume)

Path correlations (ai): H3



(d) AI decisions, H3 (Client Volume)

Fork option importance. To further understand how specific fork options contribute to differences in outcome distributions between AIs and humans, we run univariate quantile regressions to estimate the distributional effects of the fork options.

Specifically, for each fork-option, we report three pieces of information. First, we estimate how a given quantile (e.g., the median) of the outcome distribution changes when a particular fork-option is selected, relative to selecting alternative options at the same fork. This estimate captures the intrinsic effect of this fork option on the location of the estimates. Second, we report the choice difference between AIs and humans at that fork-option. Third, we compute the product of these two quantities, which serves as a summary measure of importance. A larger value indicates that a fork-option both significantly shifts the outcome distribution and is selected differently by AIs and humans. Tables 5a and 5b report the results for H1 and H3, respectively.

This decomposition allows us to distinguish fork-options that matter because they significantly affect estimates from those that matter because they are differentially chosen. Fork-options with large values of the combined metric therefore play a more prominent role in generating the observed distributional differences between AI and human outcomes within the multiverse.

Table 5. Univariate quantile estimates. Multiverse estimates. This table reports univariate quantile regression results within the multiverse analysis. For each fork-option combination, the table reports: Q25, Q50 (median), and Q75 (quantile effects), showing how selecting a given fork-option affects the 25th, 50th, 75th percentiles, and the interquartile range (IQR) of the outcome distribution relative to selecting alternative options at the same fork; Δ_{HA} , the difference in selection probabilities between AIs and humans; and the product terms $Q25 \times \Delta_{HA}$, $Q50 \times \Delta_{HA}$, $Q75 \times \Delta_{HA}$, and $IQR \times \Delta_{HA}$ (combined importance measures). Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

(a) H1 (Efficiency).

Fork	Quantile Option	Q25	Q50	Q75	IQR	Δ_{HA}	$Q25 \times \Delta_{HA}$	$Q50 \times \Delta_{HA}$	$Q75 \times \Delta_{HA}$	$IQR \times \Delta_{HA}$
RmvOpnCls	N	44.97***	0.54***	0.68***	-44.30	0.24	10.69***	0.13***	0.16***	-10.53
	Y	-44.97***	-0.54***	-0.68***	44.30	-0.24	10.69***	0.13***	0.16***	-10.53
DysExcldd	N	26.59***	0.48**	0.56***	-26.04	0.16	4.21***	0.08**	0.09***	-4.12
	Y	-26.59***	-0.48**	-0.56***	26.04	-0.16	4.21***	0.08**	0.09***	-4.12
OtlrTrtmnt	N	6.53***	0.59***	0.61***	-5.91	0.34	2.23***	0.20***	0.21***	-2.02
	Trm	-6.61***	-0.28	-0.52***	6.10	-0.14	0.93***	0.04	0.07***	-0.86
	Wns	-4.06***	-0.66***	-0.82***	3.24	-0.20	0.82***	0.13***	0.17***	-0.65
FrqncyAnlyss	D	-1559.83***	-0.83***	-0.66***	1559.17	-0.30	474.84***	0.25***	0.20***	-474.63
	W	-59.13***	-59.61***	-60.13***	-1.01	-0.01	0.36***	0.36***	0.37***	0.01
	M	0.00	-0.31***	-0.42***	-0.42	0.01	0.00	0.00***	0.00***	0.00
	A	4.72***	0.58***	0.53***	-4.19	0.30	1.42***	0.17***	0.16***	-1.26
Mdl	LgDff	0.26	0.03	0.39*	0.13	-0.06	-0.02	0.00	-0.02*	-0.01
	RltvChg	-181.73***	-2.83***	-1.00***	180.73	-0.58	105.27***	1.64***	0.58***	-104.69
	TrndSttnry	131.10***	2.22***	0.76***	-130.34	0.64	83.94***	1.42***	0.49***	-83.45
Msr	AtCrrltns	-0.32**	-0.02	-0.25***	0.07	-0.08	0.03**	0.00	0.02***	-0.01
	VrncRt	0.32**	0.02	0.25***	-0.07	0.08	0.03**	0.00	0.02***	-0.01
Frqncy	SM	0.01	-0.39***	-0.76***	-0.78	-0.11	0.00	0.04***	0.08***	0.08
	15	-2.31***	-0.84***	-0.94***	1.37	-0.16	0.37***	0.14***	0.15***	-0.22
	530	0.38***	-0.30***	-0.53***	-0.91	-0.27	-0.10***	0.08***	0.14***	0.25
	DW	-26.53***	0.14	-0.06	26.47	-0.03	0.88***	0.00	0.00	-0.88
	DM	1.51***	0.82***	0.53***	-0.98	0.57	0.87***	0.47***	0.30***	-0.56

(b) H3 (Client Volume).

Fork	Quantile Option	Q25	Q50	Q75	IQR	Δ_{HA}	$Q25 \times \Delta_{HA}$	$Q50 \times \Delta_{HA}$	$Q75 \times \Delta_{HA}$	$IQR \times \Delta_{HA}$
RmvOpnCls	N	-0.30***	-1.11***	-323.61***	-323.31	0.16	-0.05***	-0.18***	-53.28***	-53.23
	Y	0.30***	1.11***	323.61***	323.31	-0.16	-0.05***	-0.18***	-53.28***	-53.23
DysExcldd	N	-0.35***	-1.19***	-4.97***	-4.61	0.16	-0.06***	-0.19***	-0.79***	-0.73
	Y	0.35***	1.19***	4.97***	4.61	-0.16	-0.06***	-0.19***	-0.79***	-0.73
OtlrTrtmnt	N	-0.38***	-1.00***	-4.82***	-4.44	0.32	-0.12***	-0.32***	-1.56***	-1.44
	Trm	0.55***	1.00***	3.12***	2.57	-0.13	-0.07***	-0.13***	-0.40***	-0.33
	Wns	0.16*	1.00***	228.89***	228.74	-0.20	-0.03*	-0.19***	-44.66***	-44.63
FrqncyAnlyss	D	-0.04	2.22***	323.88***	323.91	-0.34	0.01	-0.76***	-110.22***	-110.23
	W	38.86***	36.06***	38.87***	0.01	-0.01	-0.24***	-0.22***	-0.24***	0.00
	M	0.02	-1.00***	-1.19***	-1.21	0.61	0.01	-0.61***	-0.73***	-0.74
	A	-0.02	1.00***	0.84***	0.86	-0.26	0.01	-0.26***	-0.22***	-0.23
Mdl	LgDff	0.21	0.99***	1.26**	1.05	-0.04	-0.01	-0.04***	-0.05**	-0.04
	RltvChg	1.02***	1.19***	264.62***	263.60	-0.60	-0.62***	-0.72***	-159.74***	-159.13
	TrndSttnry	-1.02***	-1.18***	-251.38***	-250.36	0.65	-0.66***	-0.76***	-162.48***	-161.82
Units	Euro	0.00	0.00	-0.70***	-0.70	0.20	0.00	0.00	-0.14***	-0.14
	NOfCncts	0.00	0.00	0.70***	0.70	-0.20	0.00	0.00	-0.14***	-0.14

The results are broadly consistent with the patterns documented in Table 3 and Figure 3. For H1, the combined importance measures for all quantiles ($Q25 \times \Delta_{HA}$, $Q50 \times \Delta_{HA}$, and $Q75 \times \Delta_{HA}$) are predominantly positive, confirming that AI choices shift the entire outcome distribution upward. Focusing on the median, the model fork (“Mdl”) is the dominant driver. Choosing the trend-stationary option (“TrndSttnry”) raises it by 2.22%, while the relative change option (“RltvChg”) reduces the median estimate by 2.83%; 64% more AIs choose “TrndSttnry” and 58% fewer choose “RltvChg” relative to humans. The combination of large quantile effects and large AI-human selection gaps makes this fork the primary source of the upward shift in AI median outcomes. Smaller contributions are from the data-cleaning forks: Not removing open/close observations, not excluding days, and not applying outlier treatments are each associated with higher estimates. AIs choose none of these three options.

The quantile regressions also shed light on why AI outcomes are less dispersed. The analysis frequency fork (“FrqncyAnlyss”) is the largest contributor to IQR differences. Using daily frequency (“D”) is associated with a dramatically heavier left tail: Q25 drops by roughly 1,560 while Q50 and Q75 are barely affected, producing a very large impact on IQR. Intuitively, by Jensen’s inequality, compounding daily relative changes over a year generates a downward bias in the estimate.⁷ So for all the analysis paths that sample at a daily frequency (“FrqncyAnlyss-D”), those that choose relative change for the model (“Mdl-RltvChg”) result in annual-trend estimates with a substantial negative bias, thus creating a fat left tail in the distribution. Because virtually no AI chooses relative change for the model (see Table 4a), this source of left-tail risk is absent from the AI multiverse, which substantially compresses AI dispersion. This is an interesting special case. The left-tail effect is a source of AI error, because it creates a wedge between its outcome distribution and the one of humans. It is special, because this is a type of error, i.e., the undoing of a bias, that researchers guiding AIs will be glad to make.

⁷The point boils down to $E(X_t/X_{t-1}) \times E(X_{t-1}/X_{t-2}) \leq E[(X_t/X_{t-1}) \times (X_{t-1}/X_{t-2})] = E(X_t/X_{t-2})$, so that, in this case, the two-period trend is underestimated by the squared average one-period trend (Menkveld et al. 2024, eqn. (7)).

For H3, the earlier finding that AI and human distributions differ mainly in dispersion and the right tail, with a modest median difference, is confirmed by the quantile regression. The combined importance measures for the median, $Q50 \times \Delta_{HA}$, are negative across virtually all fork-option combinations, indicating that AI choices collectively pull the median down. The estimates are normalized by the IQR, and since the IQR of H3 is small, the absolute magnitudes of these median effects are modest. The right tail shows a markedly different picture. The $Q75 \times \Delta_{HA}$ terms are large and negative at several forks, most notably “FrqncyAnlyss-D” (−110), “Mdl-RltvChg” (−160), and “Mdl-TrndSttnry” (−162), driven by both large upper-tail quantile effects and large AI-human selection gaps. These forks jointly suppress the AI right tail and compress the IQR, mirroring the dispersion patterns observed for H1.

5 Conclusion

In this paper, we study how large language models behave when deployed as *empirical researchers*. We leverage the #fincap project as a uniquely credible benchmark in finance: 164 independent human research teams analyzed the same proprietary dataset and tested the same pre-specified hypotheses, yielding a human *distribution* of outcomes rather than a single canonical estimate. We replicate this setting with AI research agents that receive the same project instructions and dataset metadata. Crucially, we classify the analysis-path choices made by both humans and AIs and project all estimates onto a common multiverse of admissible analysis paths. This mapping makes the comparison legible in finance terms: It lets us compare outcome distributions and, using quantile regressions, attribute AI-human differences to specific decision-fork choices rather than to opaque AI behavior.

We document economically and statistically meaningful AI-human differences. A dominant pattern is concentration: Relative to humans, AIs traverse a much narrower region of the multiverse and, therefore, generate outcome distributions that are typically less dispersed and exhibit

fewer extreme realizations. Importantly, this reduction in dispersion is not the product of an explicit design choice to restrict researcher degrees of freedom; it emerges endogenously from the behavioral regularity of the AI agents. For hypotheses with richer design space, the AI-human gap is not merely variance compression. AI distributions are both tighter and shifted relative to the human benchmark, and these shifts persist when the benchmark is restricted to the humans with the highest-quality outcomes. This highlights a trade-off: AI can deliver precision without alignment, i.e., a system can be “confidently wrong” relative to the distribution of plausible human implementations when discretion is high.

Methodologically, the paper offers a practical audit framework for AI-generated empirical work. The multiverse projection and fork-level diagnostics translate AI outputs into the same degrees-of-freedom objects that empirical finance already uses to reason about researcher discretion, robustness, and specification sensitivity. Quantile regressions are a key part of the analysis, because the divergences are distributional, i.e., there are differences in medians, in dispersion, and in tails. This combination yields interpretable mechanisms: We can identify which fork-option differences are most closely associated with distributional shifts and which matter primarily because AIs and humans choose them at different rates.

These contributions are timely given the rapid emergence of end-to-end autonomous research systems. The Autonomous Policy Evaluation project (APE), for example, explicitly aims to run empirical research pipelines at scale, while emphasizing that human evaluation capacity is limited and that some form of triage and auditing will be necessary. Our results suggest what such triage should prioritize in empirical finance: Not whether AI can produce a polished narrative, but whether its outcomes fall within the human benchmark distribution and, when they do not, which implicit defaults in code and preprocessing are responsible for the divergence in outcome distributions. The practical implication is not that AI should be excluded from empirical work, but that AI assistance should be paired with explicit disclosure and scrutiny of discretionary steps: a “man plus machine” approach that preserves productivity gains while maintaining methodological

accountability.

References

Cao, Sean, Wei Jiang, Junbo Wang, and Baozhong Yang (2024). “From Man vs. Machine to Man + Machine: The art and AI of stock analyses”. *Journal of Financial Economics* 160, p. 103910.

He, Songrun, Linying Lv, Asaf Manela, and Jimmy Wu (2025). “Chronologically Consistent Large Language Models”. *SSRN Electronic Journal*.

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

Korinek, Anton (2023). *Language Models and Cognitive Automation for Economic Research*. Working Paper 30957. National Bureau of Economic Research.

Levy, Bradford (Lynch) (2024). *Caution Ahead: Numerical Reasoning and Look-ahead Bias in AI Models*. URL: <https://papers.ssrn.com/abstract=5082861>. Pre-published.

Liang, Weixin, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou (2024). *Mapping the Increasing Use of LLMs in Scientific Papers*. arXiv preprint.

Lopez-Lira, Alejandro and Yuehua Tang (2025). *Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models*. URL: <http://arxiv.org/abs/2304.07619>. Pre-published.

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

- Manning, Benjamin S., Kehang Zhu, and John J. Horton (2024). *Automated Social Science: Language Models as Scientist and Subjects*. URL: <https://www.nber.org/papers/w32381>. Pre-published.
- Menkveld, Albert J. et al. (2024). “Nonstandard Errors”. *The Journal of Finance* 79.3, pp. 2339–2390.
- Novy-Marx, Robert and Mihail Z. Velikov (2025). *AI-Powered (Finance) Scholarship*. URL: <https://www.nber.org/papers/w33363>. Pre-published.
- Pérignon, Christophe, Olivier Akmansoy, Christophe Hurlin, Anna Dreber, Felix Holzmeister, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Albert J Menkveld, Michael Razen, and Utz Weitzel (2024). “Computational Reproducibility in Finance: Evidence from 1,000 Tests”. *The Review of Financial Studies* 37.11, pp. 3558–3593.
- Sarkar, Suproteem K. and Keyon Vafa (2024). *Lookahead Bias in Pretrained Language Models*. URL: <https://papers.ssrn.com/abstract=4754678>. Pre-published.

Appendix

A AI agent implementation details

A.1 Analysis code generation

The agent implements a state machine with five sequential stages, executed in a loop until either success or maximum retry attempts are reached. The workflow proceeds as follows:

1. **Code generation.** The agent receives the research prompt (containing hypotheses, data description, and requirements) and generates an initial code bundle. The LLM is instructed to return only valid JSON containing two files: `main.py` (the analysis script) and `requirements.txt` (Python package dependencies). The system prompt explicitly requires JSON-only output without markdown formatting or explanatory text.
2. **File writing.** The generated files are written to a dedicated working directory. Each independent run receives its own isolated directory to prevent interference between parallel executions.
3. **Dependency installation.** A Python virtual environment is created in the working directory, and packages listed in `requirements.txt` are installed. This ensures each run has a clean, isolated execution environment.
4. **Code execution and validation.** The agent executes `main.py` within the virtual environment, subject to resource constraints (2-hour timeout, 4GB memory limit). Upon successful execution, the agent validates the output file `results.csv` to ensure:
 - The file exists and is properly formatted
 - All six hypotheses (H1–H6) have valid estimates and standard errors

- No infinite or missing values are present in the Estimate or SE columns
- No suspicious zero values that might indicate masking rather than proper calculation

If validation passes, the run is marked as successful and terminates.

5. **Error repair.** If execution fails or validation detects issues, the agent enters a repair cycle. The error message (including specific hypotheses with problems, if applicable) is provided to the LLM along with the current code state. The LLM is instructed to fix root causes rather than symptoms, with particular emphasis on addressing the specific hypotheses mentioned in the error. The repaired code bundle is then returned to step 2, and the cycle repeats.

The agent tracks the number of repair attempts and terminates if the maximum number of retries (8 by default) is reached. Each repair cycle accumulates token usage statistics, and the full history of code changes is preserved for analysis.

A.2 Analysis path identification

The parsing process proceeds as follows:

1. **Code audit.** The LLM is instructed to act as a meticulous code auditor. For each fork in the reference table (identified by hypothesis and fork ID), the agent examines the Python code to detect evidence of implemented alternatives. Detection relies on code-level signals including imports, function calls, parameters, keyword arguments, constants, identifiers, and explicit comments. The agent does not execute the code, but performs static analysis only.
2. **Alternative matching.** For each fork, the agent identifies which alternatives from the reference table are present in the code. The matching process allows for partial matches—for example, if the reference table lists “R-squared of an autoregressive model for returns” but the code computes autocorrelation directly without computing R-squared, this is considered

a partial match. The agent is constrained to only consider alternatives explicitly listed in the reference table and cannot invent new alternatives.

3. **Primary alternative selection.** When multiple alternatives are detected for a single fork, the agent selects exactly one as the “primary” alternative and provides a brief note explaining the choice. If no strict evidence-based match is found, the agent performs a conceptual audit and selects the closest matching alternative with an appropriate confidence level.
4. **Evidence and confidence scoring.** For each detected alternative, the agent provides 1–5 evidence items, each consisting of a line number and a code snippet (up to 120 characters). The agent also assigns a confidence score on a 1–5 scale, where 5 indicates explicit/direct implementation and 1 indicates speculative/weak signals. This granular scoring allows us to distinguish between clear implementations and ambiguous cases.
5. **Structured output.** The agent returns a structured JSON object containing, for each fork: the detected alternatives, the primary alternative, confidence scores, evidence items, and (if applicable) a reason why no match was found. This JSON is validated against a Pydantic schema to ensure consistency and completeness.
6. **Data transformation.** The JSON output is converted into a tidy DataFrame format with one row per (hypothesis, fork ID, alternative) combination. This format facilitates downstream analysis and comparison with human research team choices.

The parsing agent handles edge cases such as JSON extraction from markdown-wrapped responses, escape sequence errors, and validation failures. If the initial parsing fails validation, the agent attempts a repair cycle using a secondary LLM call to fix JSON formatting issues.

A.3 System and user prompts

We feed the following system prompt to the LLM.

```

"You are a finance researcher. Follow the given prompt and complete the task."
"CRITICAL: You must respond ONLY with valid JSON in this exact format:\n"
'{"files": {\n    "main.py": "file content here",\n    "requirements.txt": "file
content here"\n  }\n}'
"Do not include any explanatory text, comments, markdown formatting, or other content
outside the JSON. "
"The response must be parseable by Python's json.loads() function."

```

Listing 1. System prompt used to generate the analysis code.

We feed the following user prompt to the LLM to generate the analysis code. Note that we explicitly ask the AI not to consult the “Nonstandard Errors” paper or any of its results, descriptions, tables, or conclusions.

```

You need to do the following for all hypotheses below:

1. Propose a statistical measure for each hypothesis
2. For this measure, estimate the average per-year change in percentage terms, based on
the full sample (or at least the longest possible period because some series are not
available yet at the beginning of the sample). Test it against the null of no change.
3. Report this estimate along with its standard error in four decimals (e.g., ‘‘measure Z
declined by 1.251\% with a standard error 0.241\%’’)

Note: If you pick measure M and find its value to be 1 in year 1, and 2 in year 2, then
its per-year change is +100\%. This is true irrespective of what the units are that
you picked for M (e.g., euro, percentage, basis points).

# Hypotheses

1. Assuming that informationally-efficient prices follow a random walk, did market
efficiency change over time?
* Null hypothesis 1: Market efficiency has not changed over time.

2. Did the (realized) bid-ask spread paid on market orders change over time? The realized
spread could be thought of as the gross-profit component of the spread as earned by
the limit-order submitter.
* Null hypothesis 2: The realized spread on market orders has not changed over time.

The remaining hypotheses focus on client trades only (i.e., trades implemented by exchange
members on behalf of their clients).

3. Did the share of client volume in total volume change over time?
* Null hypothesis 3: Client share volume as a fraction of total volume has not changed
over time.

4. On their market orders and marketable limit orders, did the realized bid-ask spread
that clients paid, change over time?
* Null hypothesis 4: Client realized spreads have not changed over time.

5. Realized spread is a standard cost measure for market orders, but to what extent do
investors continue to use market and marketable limit orders (as opposed to
non-marketable limit orders)?
* Null hypothesis 5: The fraction of client trades executed via market orders and
marketable limit orders has not changed over time.

6. A measure that does not rely on the classic limit- or market-order distinction is gross
trading revenue (GTR). Investor GTR for a particular trading day can be computed by
assuming a zero position at the start of the day and evaluating an end-of-day position
at an appropriate reference price. Relative investor GTR can then be defined as this

```

GTR divided by the investor's total (euro) volume for that trading day. This relative GTR is, in a sense, a realized spread. It reveals what various groups of market participants pay in aggregate for (or earn on) their trading. It transcends market structure as it can be meaningfully computed for any type of trading in any type of market (be it trading through limit-orders only, through market-orders only, through a mix of both, or in a completely different market structure).

* Null hypothesis 6: Relative gross trading revenue (GTR) for clients has not changed over time.

Data

The data consist of 720 million trade records for EuroStoxx 50 futures and will be made available in monthly gzipped semicolon separated text files (''csv''). Each zipped monthly file is no larger than 50 MB. The data is clean in the sense that for all files the format is identical. Please find below the first ten lines of the December 2018 file as an example.

```
''  
DATETIME; EXPIRATION; BUY_SELL_ID; TRADE_SIZE; MATCH_PRICE; AGGRESSOR_FLAG;ACCOUNT_ROLE;  
EXEC_TYPE_ID  
2018-12-03 08:00:06.400; 201812; S; 2; 3229; N; A; F  
2018-12-03 08:00:06.410; 201812; S; 1; 3229; N; A; F  
2018-12-03 08:00:06.410; 201812; S; 1; 3229; N; A; F  
2018-12-03 08:00:06.410; 201812; B; 4; 3229; Y; A; F  
2018-12-03 08:00:06.540; 201812; S; 1; 3229; N; A; F  
2018-12-03 08:00:06.550; 201812; B; 2; 3229; Y; A; F  
2018-12-03 08:00:06.550; 201812; S; 1; 3229; N; A; F  
2018-12-03 08:00:06.630; 201812; B; 1; 3229; Y; A; F  
2018-12-03 08:00:06.630; 201812; S; 1; 3229; N; A; F  
''
```

The variables are defined as follows (the characterizations are short and therefore imprecise, please refer to any standard textbook on futures to get a detailed description of what futures are and how they are traded):

- * DATETIME: Time stamp of the trade denoted as YYYY-MM-DD hh:mm:ss.sss where ss.sss denotes seconds up to a third decimal (i.e., the precision is tens of milliseconds as the last digit is always zero).
- * EXPIRATION: The expiration date of the futures contract being traded. All data pertain to Eurex trading in EuroStoxx 50 (SX5E) futures contracts. Expiration months are: March, June, September, and December. Contracts expire on the third Friday of the expiration month. The notation of expiration is YYYYMM (where MM is in [03, 06, 09, 12]).
- * BUY_SELL: This indicator shows if the trade record is for a buyer 'B' (who goes long the index) or for a seller 'S' (who goes short the index).
- * TRADE_SIZE: This is the size of the trade expressed in number of contracts. The contract value per index point is EUR 10 (e.g., per contract traded, the long side is entitled to receive 10 euro from the short side of the trade each time the index increases by one point).
- * MATCH_PRICE: The price at which the trade between buyer and the seller is concluded (i.e., the long and the short side of the trade, respectively).
- * AGGRESSOR_FLAG: If the trade record pertains to a market order (or marketable limit order) that is executed against a standing limit order, this flag takes the value 'Y'. If the record pertains to a limit order, resting in the book before being matched with an incoming market order, or to an order in an auction (e.g., the opening and closing auction), then this flag takes the value 'N'. This flag became available as of November 2009.
- * ACCOUNT_ROLE: This variable is either:
 - * A: Agency trade (i.e., a trade an exchange member does for a client).
 - * M: Market-maker principal trade (i.e., a trade an exchange member does for his own account in his role as market maker).
 - * P: Non-market-maker principal trade (i.e., a trade an exchange member does for his own account).
 - * P.S.: The distinction between M and P is not an economically meaningful one for the purpose of this project.

```

* EXEC_TYPE_ID: This variable is:
  * F if the full order was executed in the trade.
  * P if the order was only partially executed in the trade.
  * N if not assigned.

# Deliverables

Please output exactly two files:

1. A 'main.py' file, a complete Python script to perform the assignment below.
  * The path to the data files is
    '..../input/raw-data-from-deutsche-boerse-for-fincap/'.
  * Data files: monthly 'csv.gz' named '2009-01.csv.gz', '2009-02.csv.gz', etc.
  * Save results to results.csv with exactly three columns:
    1. Hypothesis - string labels 'H1' through 'H6'.
    2. Estimate - average per-year change in percentage terms. Include no percentage
       signs.
    3. SE - standard error of the estimate. Include no percentage signs.
    4. Do not include any row index in the file. The CSV should contain only the
       header row and the six data rows in order 'H1' to 'H6'.
2. A 'requirements.txt' file listing all Python packages needed to run 'main.py'.

Other requirements:

1. Generate complete, runnable Python code with no placeholders, pseudo-code, or 'fill in
   later' comments
2. Use only standard open-source statistical libraries such as numpy, pandas, statsmodels,
   scipy, etc.
3. Process the data files and compute the metrics month by month. Do not concatenate all
   data files.
4. Be efficient. Use vectorized operations in pandas or numpy. Do not use per-row loops
   (iterrows, itertuples, for i in range(len(df)), etc.).

When completing this task, do not consult, rely on, or draw from the 'Nonstandard
Errors' paper or any of its results, descriptions, tables, or conclusions. Your
response should be based solely on general domain knowledge and the information
explicitly provided in this prompt. Do not attempt to recall or reconstruct the
content of the paper, even if you are aware of it.

```

Listing 2. User prompt used to generate the analysis code.

Below is the prompt used to parse the code:

```

"You are a meticulous code auditor. You will receive a CSV reference table of analysis
forks and their allowed alternatives and a Python script."
"\n\n"
"Reference table CSV columns:\n"
"- RT-Hypothesis: Which hypothesis the fork applies to.\n"
"- Fork ID: A unique RT-Hypothesis-specific ID of the decision fork.\n"
"- Fork Description: short description of the fork.\n"
"- Question: Extra detail about the fork; explains the fork and the alternatives.\n"
"- Alternatives: The list of valid alternative implementations.\n\n"
"Your task: For each CSV row (i.e., each (RT-Hypothesis, Fork ID)), audit the Python
script and identify the implemented alternative(s) and the single primary alternative.
\n\n"
"Rules:\n"
"- Use 'Fork Description' and 'Question' as context about the fork and the alternatives.\n"
"- Only consider the alternatives given. Do not invent alternatives.\n"
"- Detect an alternative when there is evidence in code (imports, calls, parameters,
kwargs, constants, identifiers, or clear comments). Do not execute code.\n"
"- Allows for partial matching. For example, one alternative in the reference table is
'R-squared of an autoregressive model for returns.' If the code computes

```

```

autocorrelation directly, but not the R-squared, it is considered a partial match.\n"
"- If multiple alternatives are detected, mark ONLY one as 'primary' and include a short
note.\n"
"- If strict evidence yields no match, conceptually audit the code and pick the single
closest alternative as the primary with an appropriate confidence level. "
"- Use confidence, integers strictly between 1 and 5 (5 = explicit/direct implementation;
1 = speculative/weak signals).\n"
"- For each detected alternative, provide 1-5 evidence items with 1-indexed line numbers
and code snippets (<=120 chars). Undetected alternatives may have empty evidence.\n"
"- Line numbers must be within the provided code length. Notes and unmatched_reason must
be <=120 chars.\n"
"Output:\n"
"- Return ONLY a single JSON object matching the schema provided in the user message. No
Markdown, no extra commentary."

```

Listing 3. Prompt used to parse the analysis code.

B Text completion task results

Original paragraph: In statistics, samples are drawn from a population in a data-generating process (DGP). Standard errors measure the uncertainty in estimates of population parameters. 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).

Original paragraph: 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. We further find that this type of uncertainty is underestimated by participants.

Original paragraph: In statistics, samples are drawn from a population in a data-generating process (DGP). Standard errors measure the uncertainty in estimates of population parameters. 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).

GPT-5.2: These NSEs reflect uncertainty induced by researchers' choices about data, methods, and specifications, which can materially affect reported results even when sampling variation is small. Because the EGP is typically unobserved, conventional standard errors understate total uncertainty. We propose a framework for quantifying NSEs by modeling researcher-specific variation in the EGP and show how incorporating NSEs changes inference, often widening confidence intervals and reducing statistical significance. We illustrate the approach in common empirical settings and discuss implications for replication, robustness, and the interpretation of empirical evidence in finance.

Original paragraph: IN THEIR RECENT BOOK, Kahneman, Sibony, and Sunstein (2021), (KSS) discuss variability in human judgment in terms of noise. To illustrate their analysis, they consider the setting of judges passing sentence. They decompose total variation in sentencing into two canonical components: level noise and pattern noise (chapter 6).

Original paragraph: Level noise captures the extent to which some judges are more lenient than others, while pattern noise captures variation in the sentences of the same judge hearing similar cases. In statistical terms, this distinction can be thought of as across-judge versus within-judge variation. Variation across judges is also referred as variation in judge fixed effects.

Original paragraph: IN THEIR RECENT BOOK, Kahneman, Sibony, and Sunstein (2021), (KSS) discuss variability in human judgment in terms of noise. To illustrate their analysis, they consider the setting of judges passing sentence. They decompose total variation in sentencing into two canonical components: level noise and pattern noise (chapter 6).

GPT-5.2: Level noise captures differences in the overall harshness of judges, while pattern noise captures idiosyncratic variation in how individual judges respond to case characteristics. KSS argue that pattern noise is typically large and often overlooked, and that it can be reduced through the use of structured decision protocols, aggregation, and other debiasing tools. In our setting, we adopt this decomposition to study variation in [the relevant outcome], distinguishing between systematic differences in average levels across decision makers and heterogeneity in the mapping from information to decisions across otherwise similar cases.

Original paragraph: We first show that the group of #fincap participants is representative of the academic community in empirical finance/liquidity. About one-third of the 164 RTs have at least one member with publications in the top-three finance or top-five economics journals. For the group of PEs, this share is 85%.

Original paragraph: Similarly, 52% of RTs consist of at least one associate or full professor, and for the group of PEs, this share is 88%. On a scale from 1 (low) to 10, the average self-ranked score on experience with empirical finance is 8.1 for RTs and 8.4 for PEs. For experience with market liquidity, the average self-ranked score is 6.9 for RTs and 7.8 for PEs.

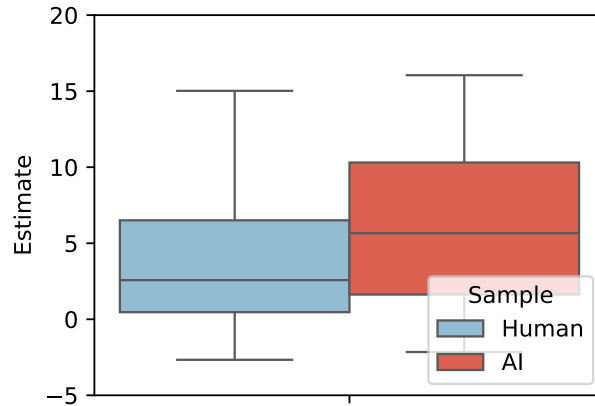


Figure C.1. Outcome distributions for the lab experiment. This figure illustrates the outcome distributions for the lab experiment. The distributions are mixed normals with the means and standard deviations as specified in Figure 1. The two-sample Anderson-Darling test statistic is 13.13 and statistically significant at the 1% level.

Original paragraph: We first show that the group of #fincap participants is representative of the academic community in empirical finance/liquidity. About one-third of the 164 RTs have at least one member with publications in the top-three finance or top-five economics journals. For the group of PEs, this share is 85%.

GPT-5.2: Moreover, participants have a wide range of experience: RTs have a median of 8 years since PhD (mean 10), while PEs have a median of 16 (mean 18). In terms of research output, the median RT has published three papers in peer-reviewed journals and the median PE has published 18. Finally, participants are geographically diverse and come from a broad set of institutions, suggesting that our results are not driven by a small number of highly connected individuals.

C An Alternative Fork-Option Importance Measure

In addition to the quantile regression approach, we also develop an alternative fork-option importance measure based on the two-sample Anderson-Darling (AD_2) test. In this section, we detail the methodology and apply it to the lab experiment and the multiverse analysis for Hypothesis H1.

To measure the importance of each fork-option combination driving the human-AI difference, we define the following $AD2$ statistic:

$$AD2(H, A|FO^c).$$

Specifically, it is the two-sample AD test statistic for the human and AI outcome distributions conditional on paths not choosing the specific fork-option combination FO. Hereafter, we drop the subscripts i and j for brevity. Essentially, it tests the null hypothesis that after discarding the paths that choose the specific fork-option combination FO, the human and AI outcome distributions are the same. Imagine that if humans and AIs primarily differ in their choice at FO while agreeing on other fork-options, then $AD2(H, A|FO^c)$ should be small and statistically insignificant. So, a lower and more statistically insignificant $AD2(H, A|FO^c)$ indicates a more important fork-option combination driving the human-AI difference.

Note that two factors jointly determine the importance of a fork-option. First, the fork-option combination is intrinsically important itself. Such importance can be measured by $AD2(FO, FO^c)$, the two-sample AD test statistic between two samples, one containing paths that choose the specific fork-option combination and the other sample containing paths that do not choose the specific fork-option combination, pooling the human and AI samples together. Second, humans and AIs differ in their choice at the fork-option as measured by $w_A(FO) - w_H(FO)$. Only when both factors are present, the fork-option combination is important. We now use the methodology developed above to identify which fork-option combinations are most important in driving the differences between human and AI outcome distributions. As before, the analysis is conducted within the multiverse, ensuring that all comparisons are made over a common and transparent set of admissible analysis paths.

The key object of interest is the statistic $AD2(H, A|FO^c)$, defined as the two-sample Anderson-Darling test statistic comparing human and AI outcome distributions after conditioning on paths

Table C.1. Anderson-Darling test results for the lab experiment. This table reports the two-sample Anderson-Darling test statistics comparing outcome distributions between human and AI samples in the toy scenario. $A^2(H, A)$ is the Anderson-Darling test statistic between the human and AI samples. $A^2(H, A|FO^c)$ is the $AD2$ statistic between the human and AI samples, using paths that do not choose the specific fork-option (FO) combination. $A^2(FO, FO^c)$ is the $AD2$ statistic between two samples, one containing paths that choose FO and the other containing paths that do not choose FO, measuring the intrinsic importance of each fork-option. # H and # A are the number of paths that choose the specific fork-option in the human and AI samples, respectively. w^H and w^A are the fractions (weights) of paths that choose the specific fork-option in the human and AI samples, respectively. Δ_{HA} is the difference between the human and AI weights ($w^A - w^H$), measuring the difference in selection probabilities. The analysis uses 250 simulated analysis paths. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Fork	Option	$AD2(H, A)$	$AD2(H, A FO^c)$	$AD2(FO, FO^c)$	# H	# A	w^H	w^A	Δ_{HA}
		13.13***							
1	L		0.55	212.02***	150	100	0.60	0.40	-0.20
1	R		-0.10	212.02***	100	150	0.40	0.60	0.20
2	L		3.68**	22.69***	150	100	0.60	0.40	-0.20
2	R		3.80***	22.69***	100	150	0.40	0.60	0.20

that do not select a given fork-option combination FO. Intuitively, this statistic measures how different the two distributions remain once the influence of a particular analytical choice has been removed.

A smaller and statistically insignificant value of $AD2(H, A|FO^c)$ indicates that excluding the fork-option combination substantially reduces the difference between human and AI outcomes. We therefore interpret such fork-option combinations as being more important contributors to the observed distributional differences, whereas combinations whose exclusion leaves the test statistic largely unchanged play a more limited role.

Based on the same simulated outcomes as in the main text, we then compute the two-sample AD test statistics. The results are shown in Table C.1.

The results are consistent with our expectation. The two options at Fork 1 are more intrinsically important than the two options at Fork 2, as measured by their intrinsic importance $AD2(FO, FO^c)$. When dropping two fork-options at Fork 1, the difference in outcome distributions between AIs and humans as measured by $AD2(H, A|FO^c)$ becomes statistically insignificant. It shows that the difference in outcome distributions between AIs and humans is indeed primarily driven by the

Table C.2. Anderson-Darling test results: Human vs. AI samples for H1. Multiverse estimates. This table reports fork-option importance analysis results for Hypothesis H1 (Market Efficiency) within the multiverse analysis. The table reports: $A^2(H, A)$, the Anderson-Darling test statistic between humans and AIs; $A^2(H, A|FO^c)$, $AD2$ statistic between humans and AIs using paths that do not choose the specific fork-option combination; $A^2(FO, FO^c)$, $AD2$ statistic between two samples, one containing paths that choose FO and the other sample containing paths that do not choose FO; # H and # A, number of paths choosing the fork-option in humans and AIs; w_H and w_A , weights of the fork-option combination for humans and AIs, respectively; Δ_{HA} , the difference in weights between AIs and humans. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. A smaller and statistically insignificant value of $AD2(H, A|FO^c)$ indicates that excluding the fork-option combination substantially reduces the difference between human and AI outcomes.

Fork	Option	$AD2(H, A)$	$AD2(H, A FO^c)$	$AD2(FO, FO^c)$	Δ_{HA}
		107.07***			
RmvOpnCls	N			20.28***	0.24
	Y		88.34***	20.28***	-0.24
DysExcldd	N		3.49**	11.80***	0.16
	Y		94.65***	11.80***	-0.16
OtlrTrtmnt	N			27.89***	0.34
	Trm		98.73***	6.99***	-0.14
	Wns		91.18***	18.65***	-0.20
FrqncyAnlyss	D		89.46***	38.57***	-0.30
	W		106.41***	0.70	-0.01
	M		108.42***	15.84***	0.01
	A		14.51***	56.37***	0.30
Mdl	LgDff		121.88***	5.27***	-0.06
	RltvChg		36.14***	119.57***	-0.58
	TrndSttnry			102.65***	0.64
Msr	AtCrrltns		79.37***	24.25***	-0.08
	VrncRt		39.11***	24.25***	0.08
Frqncy	SM		101.11***	22.55***	-0.11
	15		107.87***	47.10***	-0.16
	530		92.75***	24.03***	-0.27
	DW		96.41***	5.08***	-0.03
	DM		16.91***	134.93***	0.57

choice at Fork 1 as we expected.

Moving to the multiverse analysis, Table C.2 reports the $AD2$ statistics for Hypothesis H1. The unconditional Anderson-Darling statistic comparing the multiverse outcome distributions of AIs and humans is 107.07 and statistically significant, confirming that the distributional differences

illustrated in Figure 3 are unlikely to be driven by random variation alone.

Among all fork-option combinations, the option “DysExcldd-N” (i.e., not excluding settlement weeks) emerges as the most important contributor to the difference between AI and human outcome distributions. This is indicated by the largest reduction in the $AD2$ statistic, from 107.07 to 3.49, when this option is excluded from the analysis. The second most important contributor is “FrqncyAnlyss-A”. Removing this fork option reduces the $AD2$ statistic from 107.07 to 14.51.

For several forks, all AIs select the same option, and as a result the conditional statistic $AD2(H, A|FO^c)$ is not available. An example is the “RmvOpnCls” fork, for which all AIs choose the “N” option. In these cases, inference can still be drawn by combining two pieces of information: the intrinsic importance of the fork-option as measured by $AD2(FO, FO^c)$, and the difference in selection probabilities between AIs and humans, captured by Δ_{HA} . For example, for the “Mdl-TrndSttnry” fork, the conditional $AD2$ statistic is not available as all AIs choose this fork option. But the intrinsic importance of this fork option is 102.65, and the difference in selection probabilities is 0.64. Thus, by inference, this fork option is important.

D Analysis paths of other LLM models

We conduct the analysis using other reasoning models by OpenAI including o3, GPT-5, and GPT-5.1. The results are qualitatively similar to the results using GPT-5.2.

E Highest-quality human research teams subsample

Table D.1. Analysis path choices: Human vs. AIs. All GPT reasoning models including o3, GPT-5, GPT-5.1, and GPT-5.2. This table reports the distribution of analytical choices made by humans and AIs across different forks in the multiverse analysis for Hypothesis H1 (Market Efficiency), using multiple GPT reasoning models (o3, GPT-5, GPT-5.1, and GPT-5.2). Each row represents a fork-option combination, showing the percentage of paths that choose each option for each model.

Fork	Option	# H	# A	w_H	w_A	Δ_{HA}
		164	620			
RmvOpnCls	N	125	619	0.76	1.00	0.24
	Y	39	1	0.24	0.00	-0.24
DysExcldd	N	137	619	0.84	1.00	0.16
	Y	27	1	0.16	0.00	-0.16
OtrTrtmnt	N	108	620	0.66	1.00	0.34
	Trm	23	0	0.14	0.00	-0.14
	Wns	33	0	0.20	0.00	-0.20
FrqncyAnlyss	D	52	7	0.32	0.01	-0.31
	W	1	0	0.01	0.00	-0.01
	M	38	174	0.23	0.28	0.05
	A	73	439	0.45	0.71	0.26
Mdl	LgDff	10	8	0.06	0.01	-0.05
	RltvChg	95	169	0.58	0.27	-0.31
	TrndSttnry	59	443	0.36	0.71	0.35
Msr	AtCrrltns	61	180	0.37	0.29	-0.08
	VrncRt	103	440	0.63	0.71	0.08
Frqncy	SM	29	29	0.18	0.05	-0.13
	15	42	80	0.26	0.13	-0.13
	530	56	35	0.34	0.06	-0.29
	DW	21	68	0.13	0.11	-0.02
	DM	16	408	0.10	0.66	0.56

Table E.1. Summary statistics of multiverse estimates between human and AI samples. This table reports summary statistics comparing the distribution of multiverse estimates between the highest-quality human research teams and AIs across six hypotheses. The highest-quality human research teams are identified as those satisfying three criteria: (i) first principal component of quality indicators in top 25%, (ii) code reproducibility in the top quartile, and (iii) average peer-review rating in the top quartile. For each hypothesis and sample, the table reports: N (sample size), Mean, SD (standard deviation), Min, Q(0.10) through Q(0.90) (deciles), Median, IQR (interquartile range), IDR (inter-decile range), and Max. Multiverse estimates are computed by mapping each team’s analysis path choices onto a common multiverse of admissible analysis paths.

Hypothesis	H1 Efficiency		H2 RSpread		H3 Client Volume		H4 Client RSpread		H5 Client MOrders		H6 Client GTR	
	AI	Human	AI	Human	AI	Human	AI	Human	AI	Human	AI	Human
N	158	8	158	8	158	8	158	8	158	8	158	8
Mean	1.9	-7.5	-2.4	1.3	-3.8	-3.2	-0.8	1.3	-0.6	-0.9	0.1	-572.7
SD	3.0	8.5	5.4	10.2	0.0	1.0	4.6	4.9	0.2	0.5	5.4	1600.7
Min	-7.0	-25.9	-9.2	-9.1	-3.8	-3.8	-8.2	-3.4	-1.1	-1.4	-24.7	-4534.0
Q(0.10)	-3.1	-15.5	-8.6	-7.0	-3.8	-3.8	-8.2	-3.3	-1.1	-1.4	-6.9	-1388.5
Q(0.25)	0.6	-11.0	-8.6	-1.6	-3.8	-3.8	-3.8	-0.8	-0.5	-1.4	-3.0	-22.4
Median	3.5	-4.5	0.5	-0.1	-3.8	-3.5	2.6	-0.0	-0.5	-0.8	0.1	-6.9
Q(0.75)	4.1	-1.7	0.5	0.5	-3.8	-3.3	2.6	2.6	-0.5	-0.5	6.0	-0.3
Q(0.90)	4.1	-0.7	4.3	7.8	-3.8	-2.6	2.9	5.5	-0.5	-0.4	6.9	12.8
IQR	3.5	9.3	9.1	2.1	0.0	0.5	6.4	3.4	0.0	0.9	9.0	22.1
IDR	7.1	14.8	12.9	14.8	0.0	1.2	11.2	8.8	0.5	1.0	13.8	1401.3
Max	4.1	-0.7	11.7	24.9	-3.7	-0.9	4.0	12.2	-0.1	-0.1	6.9	15.5

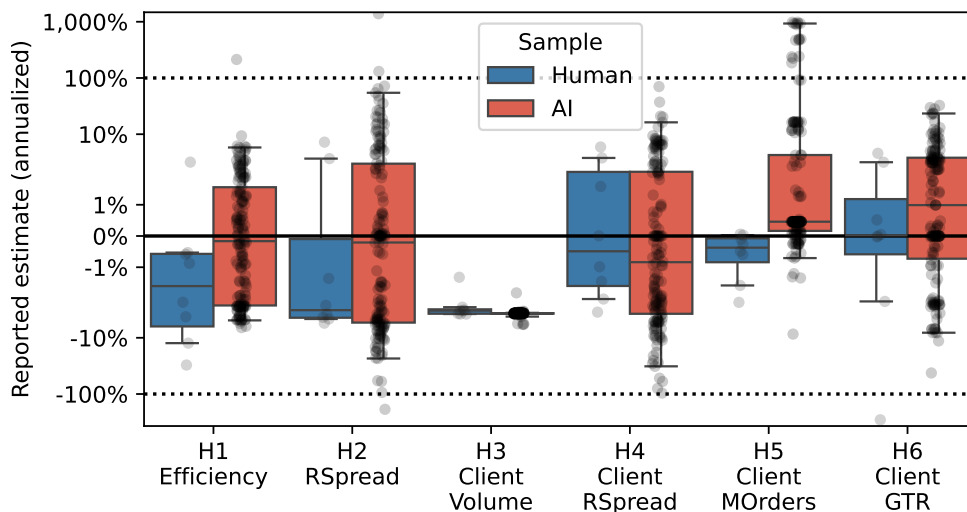


Figure E.1. Dispersion of estimates between human and AI samples. This figure presents box plots comparing the distribution of estimates between the highest-quality human research teams and AIs across six hypotheses. The highest-quality human research teams are identified as those satisfying three criteria: (i) first principal component of quality indicators in top 25%, (ii) code reproducibility in the top quartile, and (iii) average peer-review rating in the top quartile. Each box plot shows the median (center line), interquartile range (box), and whiskers extending to 2.5% and 97.5% quantiles. Outliers beyond the whiskers are shown as individual points. The figure shows that differences between AI and top-quality human estimates remain largely unchanged compared to the full sample.

F Results for other hypotheses

F.1 Anderson-Darling test results for other hypotheses

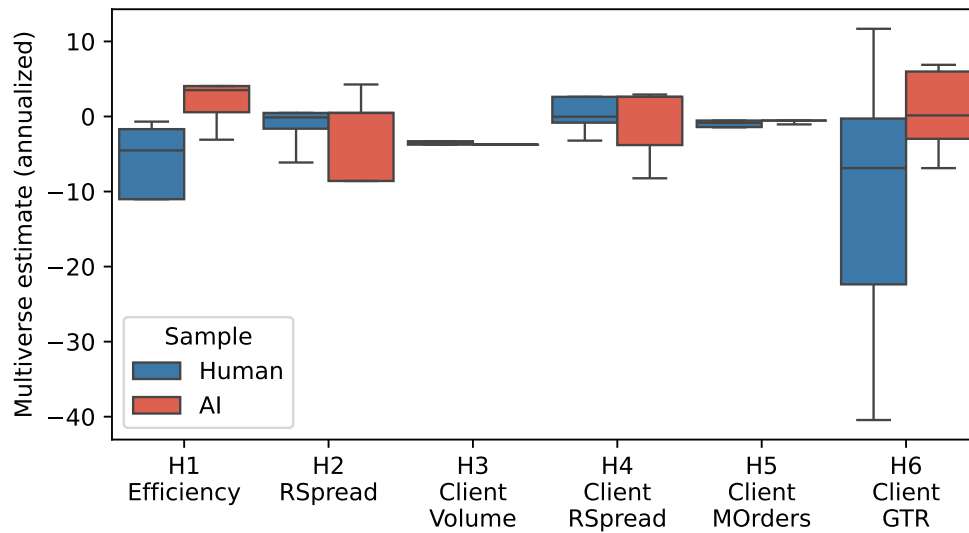


Figure E.2. Dispersion of multiverse estimates between human and AI samples. This figure presents box plots comparing the distribution of multiverse estimates between the highest-quality human research teams and AIs across six hypotheses. The highest-quality human research teams are identified as those satisfying three criteria: (i) first principal component of quality indicators in top 25%, (ii) code reproducibility in the top quartile, and (iii) average peer-review rating in the top quartile. Multiverse estimates are computed by mapping each team’s analysis path choices onto a common multiverse of admissible analysis paths. Each box plot shows the median (center line), interquartile range (box), and whiskers extending to 2.5% and 97.5% quantiles. Outliers beyond the whiskers are shown as individual points.

Table F.1. Anderson-Darling test results for H2. Multiverse estimates. This table reports fork-option importance analysis results for Hypothesis H2 (Realized bid-ask spread on market orders) within the multiverse analysis. The table reports: $A^2(H, A)$, the Anderson-Darling test statistic between humans and AIs; $A^2(H, A|FO^c)$, $AD2$ statistic between humans and AIs using paths that do not choose the specific fork-option combination; $A^2(FO, FO^c)$, $AD2$ statistic between two samples, one containing paths that choose FO and the other sample containing paths that do not choose FO; # H and # A, number of paths choosing the fork-option in humans and AIs; w_H and w_A , weights of the fork-option combination for humans and AIs, respectively; Δ_{HA} , the difference in weights between AIs and humans. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Fork	Option	$AD2(H, A)$	$AD2(H, A FO^c)$	$AD2(FO, FO^c)$	Δ_{HA}
		26.21***			
RmvOpnCls	N			2.07**	0.24
	Y		27.07***	2.07**	-0.24
DysExcldd	N		-0.66	2.33**	0.21
	Y		25.84***	2.33**	-0.21
OtlrTrtmnt	N			11.41***	0.38
	Trm		25.43***	2.02**	-0.16
	Wns		19.72***	8.74***	-0.21
FrqncyAnlyss	D		21.29***	19.41***	-0.37
	W		25.69***	0.58	-0.01
	M		4.15***	34.39***	0.62
	A		17.69***	24.65***	-0.24
Mdl	LgDff		27.12***	1.30*	-0.05
	RltvChg		2.04**	67.28***	-0.60
	TrndSttnry			63.24***	0.65
TckTstOrAgg	Agg		1.17	13.73***	0.15
	TckTst		19.52***	13.73***	-0.15
Aggrgtn	EqlWghtdAvrg		33.44***	9.40***	-0.37
	TrdSzWghtdAvrg		13.17***	9.40***	0.37
PstTrdVI	5M		6.51***	58.03***	-0.26
	10M		21.34***	3.61**	0.03
	30M		16.47***	53.69***	0.23

Table F.2. Anderson-Darling test results for H3. Multiverse estimates. This table reports fork-option importance analysis results for Hypothesis H3 (Client share of volume) within the multiverse analysis. The table reports: $A^2(H, A)$, the Anderson-Darling test statistic between humans and AIs; $A^2(H, A|FO^c)$, $AD2$ statistic between humans and AIs using paths that do not choose the specific fork-option combination; $A^2(FO, FO^c)$, $AD2$ statistic between two samples, one containing paths that choose FO and the other sample containing paths that do not choose FO; # H and # A, number of paths choosing the fork-option in humans and AIs; w_H and w_A , weights of the fork-option combination for humans and AIs, respectively; Δ_{HA} , the difference in weights between AIs and humans. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Fork	Option	$AD2(H, A)$	$AD2(H, A FO^c)$	$AD2(FO, FO^c)$	Δ_{HA}
		93.48***			
RmvOpnCls	N			16.38***	0.16
	Y		83.66***	16.38***	-0.16
DysExcldd	N		-0.17	22.94***	0.16
	Y		79.58***	22.94***	-0.16
OtlrTrtmnt	N			38.41***	0.32
	Trm		82.71***	13.13***	-0.13
	Wns		77.90***	21.28***	-0.20
FrqncyAnlyss	D		92.63***	38.08***	-0.34
	W		92.91***	0.67	-0.01
	M		26.46***	42.51***	0.61
	A		65.22***	13.82***	-0.26
Mdl	LgDff		91.53***	1.68*	-0.04
	RltvChg		22.83***	164.01***	-0.60
	TrndSttnry			167.06***	0.65
Units	Euro		50.39***	26.75***	0.20
	NOFCncts		45.15***	26.75***	-0.20

Table E.3. Anderson-Darling test results for H4. Multiverse estimates. This table reports fork-option importance analysis results for Hypothesis H4 (Client realized bid-ask spreads) within the multiverse analysis. The table reports: $A^2(H, A)$, the Anderson-Darling test statistic between humans and AIs; $A^2(H, A|FO^c)$, $AD2$ statistic between humans and AIs using paths that do not choose the specific fork-option combination; $A^2(FO, FO^c)$, $AD2$ statistic between two samples, one containing paths that choose FO and the other sample containing paths that do not choose FO; # H and # A, number of paths choosing the fork-option in humans and AIs; w_H and w_A , weights of the fork-option combination for humans and AIs, respectively; Δ_{HA} , the difference in weights between AIs and humans. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Fork	Option	$AD2(H, A)$	$AD2(H, A FO^c)$	$AD2(FO, FO^c)$	Δ_{HA}
		15.26***			
RmvOpnCls	N			2.03**	0.23
	Y		14.88***	2.03**	-0.23
DysExcldd	N	-0.70		0.78	0.19
	Y	15.01***		0.78	-0.19
OtlrTrtmnt	N			13.89***	0.37
	Trm	14.33***		2.39**	-0.16
	Wns	10.99***		10.89***	-0.21
FrqncyAnlyss	D	25.98***		24.92***	-0.36
	W	15.36***		1.29*	-0.01
	M	3.60**		12.76***	0.62
	A	8.22***		35.91***	-0.25
Mdl	LgDff	14.70***		0.47	-0.04
	RltvChg	2.67**		39.28***	-0.60
	TrndSttnry			39.71***	0.65
TckTstOrAgg	Agg	0.40		18.35***	0.14
	TckTst	10.80***		18.35***	-0.14
Aggrgtn	EqlWghtdAvrg	28.36***		14.88***	-0.37
	TrdSzWghtdAvrg	6.52***		14.88***	0.37
PstTrdVI	5M	8.58***		29.69***	-0.25
	10M	13.39***		4.35***	0.05
	30M	14.31***		28.07***	0.19

Table F.4. Anderson-Darling test results for H5. Multiverse estimates. This table reports fork-option importance analysis results for Hypothesis H5 (Fraction of client trades executed via market/marketable limit orders) within the multiverse analysis. The table reports: $A^2(H, A)$, the Anderson-Darling test statistic between humans and AIs; $A^2(H, A|FO^c)$, $AD2$ statistic between humans and AIs using paths that do not choose the specific fork-option combination; $A^2(FO, FO^c)$, $AD2$ statistic between two samples, one containing paths that choose FO and the other sample containing paths that do not choose FO; # H and # A, number of paths choosing the fork-option in humans and AIs; w_H and w_A , weights of the fork-option combination for humans and AIs, respectively; Δ_{HA} , the difference in weights between AIs and humans. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Fork	Option	$AD2(H, A)$	$AD2(H, A FO^c)$	$AD2(FO, FO^c)$	Δ_{HA}
		30.52***			
RmvOpnCls	N			9.14***	0.20
	Y		27.18***	9.14***	-0.20
DysExcldd	N		-0.67	21.52***	0.21
	Y		26.52***	21.52***	-0.21
OtlrTrtmnt	N			5.49***	0.32
	Trm		28.78***	1.89*	-0.13
	Wns		28.28***	2.02**	-0.20
FrqncyAnlyss	D		69.51***	74.64***	-0.33
	W		30.67***	0.95	-0.01
	M		7.16***	44.11***	0.61
	A		32.50***	158.51***	-0.27
Mdl	LgDff		29.30***	16.51***	-0.04
	RltvChg		7.32***	46.31***	-0.61
	TrndSttnry			52.78***	0.65

Table F.5. Anderson-Darling test results for H6. Multiverse estimates. This table reports fork-option importance analysis results for Hypothesis H6 (Relative gross trading revenue for clients) within the multiverse analysis. The table reports: $A^2(H, A)$, the Anderson-Darling test statistic between humans and AIs; $A^2(H, A|FO^c)$, $AD2$ statistic between humans and AIs using paths that do not choose the specific fork-option combination; $A^2(FO, FO^c)$, $AD2$ statistic between two samples, one containing paths that choose FO and the other sample containing paths that do not choose FO; # H and # A, number of paths choosing the fork-option in humans and AIs; w_H and w_A , weights of the fork-option combination for humans and AIs, respectively; Δ_{HA} , the difference in weights between AIs and humans. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Fork	Option	$AD2(H, A)$	$AD2(H, A FO^c)$	$AD2(FO, FO^c)$	Δ_{HA}
		35.32***			
RmvOpnCls	N			5.56***	0.18
	Y		32.95***	5.56***	-0.18
DysExcldd	N		-0.83	5.50***	0.15
	Y		31.13***	5.50***	-0.15
OtlrTrtmnt	N			8.11***	0.35
	Trm		34.26***	0.70	-0.14
	Wns		30.17***	6.09***	-0.21
FrqncyAnlyss	D		26.38***	18.13***	-0.36
	W		35.83***	2.78**	-0.01
	M		12.46***	11.01***	0.62
	A		30.11***	0.84	-0.24
Mdl	LgDff		37.62***	4.25***	-0.04
	RltvChg		6.87***	66.11***	-0.61
	TrndSttnry			61.70***	0.65
RefPrc	LstPrcTdy		12.19***	1.73*	0.12
	LstPrcDyLtr		33.75***	2.81**	0.06
	VwpL5Trds		33.31***	-0.54	-0.08
	VwpFlldy		28.73***	1.97*	-0.09
MnOrMdn	Mn		-1.06	0.52	0.04
	Mdn		34.42***	0.52	-0.04
HndlNng	SetMssng		15.85***	14.74***	-0.06
	Eps0001		26.42***	10.86***	0.00
	Eps1		28.13***	4.32***	0.06
RtnNgTrndSgn	N		22.07***	135.64***	0.31
	Y		17.12***	135.64***	-0.31

F.2 Quantile regression results for other hypotheses

Table F.6. Univariate quantile estimates for H2. Multiverse estimates. This table reports univariate quantile regression results for Hypothesis H2 (Realized bid-ask spread on market orders) within the multiverse analysis. For each fork-option combination, the table reports: Q25, Q50 (median), and Q75 (quantile effects), showing how selecting a given fork-option affects the 25th, 50th, and 75th percentiles of the outcome distribution relative to selecting alternative options at the same fork; Δ_{HA} , the difference in selection probabilities between AIs and humans; and the product terms $Q25 \times \Delta_{HA}$, $Q50 \times \Delta_{HA}$, and $Q75 \times \Delta_{HA}$. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Fork	Quantile Option	Q25	Q50	Q75	IQR	Δ_{HA}	$Q25 \times \Delta_{HA}$	$Q50 \times \Delta_{HA}$	$Q75 \times \Delta_{HA}$	$IQR \times \Delta_{HA}$
RmvOpnCls	N	-0.06	0.06	-98.53***	-98.47	0.24	-0.01	0.01	-23.43***	-23.42
	Y	0.06	-0.06	98.53***	98.47	-0.24	-0.01	0.01	-23.43***	-23.42
DysExcldd	N	-0.11	0.06	-126.79***	-126.69	0.21	-0.02	0.01	-26.26***	-26.24
	Y	0.11	-0.06	126.79***	126.69	-0.21	-0.02	0.01	-26.26***	-26.24
OtlrTrtmnt	N	-0.20	-1.64***	-98.81***	-98.61	0.38	-0.07	-0.62***	-37.35***	-37.28
	Trm	0.15	-0.06	220.60***	220.45	-0.16	-0.02	0.01	-36.32***	-36.29
	Wns	0.37**	2.38***	82.56***	82.19	-0.21	-0.08**	-0.51***	-17.62***	-17.54
FrqncyAnlyss	D	0.03	-0.06	482.08***	482.05	-0.37	-0.01	0.02	-178.75***	-178.74
	W	99.53***	98.89***	98.53***	-1.00	-0.01	-0.61***	-0.60***	-0.60***	0.01
	M	-0.47***	-0.36***	-4.96***	-4.50	0.62	-0.29***	-0.23***	-3.08***	-2.80
	A	0.81***	0.42***	2.38***	1.57	-0.24	-0.20***	-0.10***	-0.58***	-0.38
Mdl	LgDff	-4.60***	-0.24	1.44**	6.05	-0.05	0.22***	0.01	-0.07**	-0.29
	RltvChg	0.69***	2.62***	388.58***	387.89	-0.60	-0.41***	-1.57***	-232.20***	-231.79
	TrndSttnry	-0.69***	-2.62***	-260.38***	-259.69	0.65	-0.44***	-1.70***	-168.30***	-167.85
TckTstOrAgg	Agg	-0.62***	-1.15***	-226.54***	-225.92	0.15	-0.09***	-0.17***	-32.94***	-32.85
	TckTst	0.62***	1.15***	226.54***	225.92	-0.15	-0.09***	-0.17***	-32.94***	-32.85
Aggrgtn	EqlWghtdAvrg	0.28**	-0.07	1.41***	1.13	-0.37	-0.11**	0.03	-0.53***	-0.42
	TrdSzWghtdAvrg	-0.28**	0.07	-1.41***	-1.13	0.37	-0.11**	0.03	-0.53***	-0.42
PstTrdVl	5M	0.80***	0.87***	2.30***	1.50	-0.26	-0.21***	-0.23***	-0.60***	-0.39
	10M	-0.12	-0.59***	-0.40	-0.28	0.03	0.00	-0.02***	-0.01	-0.01
	30M	-0.70***	-0.87***	-1.45	-0.75	0.23	-0.16***	-0.20***	-0.33	-0.17

Table F.7. Univariate quantile estimates for H4. Multiverse estimates. This table reports univariate quantile regression results for Hypothesis H4 (Client realized bid-ask spreads) within the multiverse analysis. For each fork-option combination, the table reports: Q25, Q50 (median), and Q75 (quantile effects), showing how selecting a given fork-option affects the 25th, 50th, and 75th percentiles of the outcome distribution relative to selecting alternative options at the same fork; Δ_{HA} , the difference in selection probabilities between AIs and humans; and the product terms $Q25 \times \Delta_{HA}$, $Q50 \times \Delta_{HA}$, and $Q75 \times \Delta_{HA}$. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Fork	Quantile Option	Q25	Q50	Q75	IQR	Δ_{HA}	$Q25 \times \Delta_{HA}$	$Q50 \times \Delta_{HA}$	$Q75 \times \Delta_{HA}$	$IQR \times \Delta_{HA}$
RmvOpnCls	N	0.02	0.41***	-1.59***	-1.60	0.23	0.00	0.09***	-0.36***	-0.36
	Y	-0.02	-0.41***	1.59***	1.60	-0.23	0.00	0.09***	-0.36***	-0.36
DysExcldd	N	-0.01	0.43***	-1.79***	-1.78	0.19	0.00	0.08***	-0.35***	-0.35
	Y	0.01	-0.43***	1.79***	1.78	-0.19	0.00	0.08***	-0.35***	-0.35
OtlrTrtmnt	N	-0.30	-0.18	-17.40***	-17.10	0.37	-0.11	-0.07	-6.47***	-6.36
	Trm	0.17	-0.41**	80.25***	80.08	-0.16	-0.03	0.07**	-13.21***	-13.18
	Wns	0.28	1.43***	17.40***	17.11	-0.21	-0.06	-0.30***	-3.61***	-3.55
FrqncyAnlyss	D	-1109.39***	-0.90***	-0.14	1109.25	-0.36	397.83***	0.32***	0.05	-397.77
	W	-98.77***	-99.69***	-99.77***	-1.00	-0.01	0.60***	0.61***	0.61***	0.01
	M	-0.35*	0.41***	-1.26***	-0.92	0.62	-0.21*	0.25***	-0.78***	-0.56
	A	1.25***	0.54***	1.44***	0.19	-0.25	-0.31***	-0.13***	-0.36***	-0.05
Mdl	LgDff	-1.61*	-0.45	0.83***	2.44	-0.04	0.07*	0.02	-0.04***	-0.10
	RltvChg	0.54***	1.52***	7.22***	6.68	-0.60	-0.33***	-0.92***	-4.36***	-4.03
	TrndSttnry	-0.54***	-1.01***	-5.59***	-5.05	0.65	-0.35***	-0.65***	-3.61***	-3.26
TckTstOrAgg	Agg	-0.89**	-0.55**	-616.57***	-615.69	0.14	-0.12**	-0.08**	-85.90***	-85.78
	TckTst	0.89***	0.55**	616.57***	615.69	-0.14	-0.12***	-0.08**	-85.90***	-85.78
Aggrgtn	EqWghtdAvrg	-0.07	-0.43***	-0.30***	-0.23	-0.37	0.03	0.16***	0.11***	0.09
	TrdSzWghtdAvrg	0.07	0.43***	0.30***	0.23	0.37	0.03	0.16***	0.11***	0.09
PstTrdVl	5M	1.25***	0.97***	0.31***	-0.94	-0.25	-0.31***	-0.24***	-0.08***	0.23
	10M	0.08	-0.52***	-0.51***	-0.59	0.05	0.00	-0.03***	-0.03***	-0.03
	30M	-1.16***	-1.68***	-0.04	1.12	0.19	-0.23***	-0.33***	-0.01	0.22

Table F.8. Univariate quantile estimates for H5. Multiverse estimates. This table reports univariate quantile regression results for Hypothesis H5 (Fraction of client trades executed via market/marketable limit orders) within the multiverse analysis. For each fork-option combination, the table reports: Q25, Q50 (median), and Q75 (quantile effects), showing how selecting a given fork-option affects the 25th, 50th, and 75th percentiles of the outcome distribution relative to selecting alternative options at the same fork; Δ_{HA} , the difference in selection probabilities between AIs and humans; and the product terms $Q25 \times \Delta_{HA}$, $Q50 \times \Delta_{HA}$, and $Q75 \times \Delta_{HA}$. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Fork	Quantile Option	Q25	Q50	Q75	IQR	Δ_{HA}	$Q25 \times \Delta_{HA}$	$Q50 \times \Delta_{HA}$	$Q75 \times \Delta_{HA}$	$IQR \times \Delta_{HA}$
RmvOpnCls	N	0.76**	0.50***	-198.15***	-198.90	0.20	0.15**	0.10***	-38.66***	-38.81
	Y	-0.76**	-0.50***	198.15***	198.90	-0.20	0.15**	0.10***	-38.66***	-38.81
DysExcldd	N	0.82***	0.69***	0.18*	-0.65	0.21	0.17***	0.14***	0.04*	-0.13
	Y	-0.82***	-0.69***	-0.18*	0.65	-0.21	0.17***	0.14***	0.04*	-0.13
OtlrTrtmnt	N	0.67***	0.17*	-2.57***	-3.24	0.32	0.22***	0.05*	-0.83***	-1.05
	Trm	-0.67	-0.35**	194.04***	194.71	-0.13	0.09	0.05**	-24.85***	-24.93
	Wns	-0.32	-0.17	2.57***	2.89	-0.20	0.06	0.03	-0.50***	-0.56
FrqncyAnlyss	D	1.04***	0.73***	202.40***	201.37	-0.33	-0.35***	-0.24***	-67.64***	-67.30
	W	27.11***	26.12***	26.08***	-1.03	-0.01	-0.17***	-0.16***	-0.16***	0.01
	M	1.67***	1.00***	-0.60***	-2.26	0.61	1.02***	0.61***	-0.36***	-1.38
	A	-1.67***	-1.67***	-1.04***	0.63	-0.27	0.45***	0.45***	0.28***	-0.17
Mdl	LgDff	-1.53	-2.23***	-0.09	1.44	-0.04	0.06	0.08***	0.00	-0.05
	RltvChg	-1.58***	-1.45***	195.89***	197.47	-0.61	0.96***	0.88***	-119.45***	-120.41
	TrndSttnry	1.67***		-195.89***	-197.56	0.65	1.08***		-126.61***	-127.69

Table F.9. Univariate quantile estimates for H6. Multiverse estimates. This table reports univariate quantile regression results for Hypothesis H6 (Relative gross trading revenue for clients) within the multiverse analysis. For each fork-option combination, the table reports: Q25, Q50 (median), and Q75 (quantile effects), showing how selecting a given fork-option affects the 25th, 50th, and 75th percentiles of the outcome distribution relative to selecting alternative options at the same fork; Δ_{HA} , the difference in selection probabilities between AIs and humans; and the product terms $Q25 \times \Delta_{HA}$, $Q50 \times \Delta_{HA}$, and $Q75 \times \Delta_{HA}$. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Fork	Quantile Option	Q25	Q50	Q75	IQR	Δ_{HA}	$Q25 \times \Delta_{HA}$	$Q50 \times \Delta_{HA}$	$Q75 \times \Delta_{HA}$	$IQR \times \Delta_{HA}$
RmvOpnCls	N	18 159.18***	0.51**	-0.16	-18 159.34	0.18	3321.80***	0.09**	-0.03	-3321.83
	Y	-18 159.18***	-0.51**	0.11	18 159.29	-0.18	3321.80***	0.09**	-0.02	-3321.82
DysExcldd	N	8016.96***	0.72***	0.47**	-8016.49	0.15	1220.24***	0.11***	0.07**	-1220.17
	Y	-8016.96***	-0.72***	-0.47**	8016.49	-0.15	1220.24***	0.11***	0.07**	-1220.17
OtlrTrtmnt	N	8011.16***	0.66***	0.06	-8011.10	0.35	2784.37***	0.23***	0.02	-2784.34
	Trm	-8016.96***	-0.14	0.11	8017.07	-0.14	1124.33***	0.02	-0.02	-1124.34
	Wns	-8011.16***	-0.60***	-0.09	8011.07	-0.21	1660.85***	0.12***	0.02	-1660.83
FrqncyAnlyss	D	-11 213.75***	-0.32**	-0.65***	11 213.10	-0.36	4089.61***	0.12**	0.24***	-4089.37
	W	1911.87***	1911.36***	1910.75***	-1.12	-0.01	-11.66***	-11.65***	-11.65***	0.01
	M	10.04***	0.30**	0.55***	-9.49	0.62	6.18***	0.19**	0.34***	-5.84
	A	-3.56***	0.05	0.21	3.77	-0.24	0.87***	-0.01	-0.05	-0.92
Mdl	LgDff	0.58	3.86***	56.38***	55.80	-0.04	-0.02	-0.16***	-2.41***	-2.38
	RltvChg	-11 652.22***	-10.76***	0.21	11 652.43	-0.61	7105.01***	6.56***	-0.13	-7105.14
	TrndStnry	11 652.19***	4.27***	-0.43***	-11 652.62	0.65	7602.34***	2.79***	-0.28***	-7602.62
RefPrc	LstPrcTdy	0.14	-0.14	0.50***	0.36	0.12	0.02	-0.02	0.06***	0.04
	LstPrcDyLtr	0.50**	0.14	-0.28	-0.78	0.06	0.03**	0.01	-0.02	-0.04
	VwpL5Trds	-2.19***	0.01	0.40*	2.59	-0.08	0.18***	0.00	-0.03*	-0.22
	VwpFlldy	-29.94***	-0.01	-0.41**	29.54	-0.09	2.76***	0.00	0.04**	-2.72
MnOrMdn	Mn	10 547.92***	1.89***	-1.01***	-10 548.93	0.04	383.46***	0.07***	-0.04***	-383.49
	Mdn	-9987.92***	-1.89***	1.01**	9988.93	-0.04	363.10***	0.07***	-0.04**	-363.14
HndlNng	SetMssng	10 289.77***	-0.29**	0.71***	-10 289.05	-0.06	-604.39***	0.02**	-0.04***	604.35
	Eps0001	-46 010 101.05***	0.21	-0.43***	46 010 100.62	0.00	46 166.36***	0.00	0.00***	-46 166.36
	Eps1	0.65***	0.21	-0.43**	-1.08	0.06	0.04***	0.01	-0.03**	-0.06
RtnNgTrndSgn	N	10.99***	1.36***	0.96***	-10.04	0.31	3.43***	0.43***	0.30***	-3.13
	Y	-10.99***	-1.37***	-0.96***	10.04	-0.31	3.43***	0.43***	0.30***	-3.13