# AI "Errors":
## Discussion slides

Wenqian Huang – Albert J. Menkveld – Shihao Yu

**Discussion by James O'Donovan**

UMD/SMU/UBS Quant Investment Forum, Singapore, June 2026

# What this paper aims to do

**Big idea:** Use the #fincap experiment to audit whether AI agents, acting as empirical researchers, produce results that look like human empirical research.

## The paper's definition

An AI "error" is a systematic deviation from the human outcome distribution, where the human distribution is treated as a benchmark for accepted empirical practice.

## Two questions

- **What:** Do AI and human outcome distributions differ in location, dispersion, and tails?
- **Why:** Which analysis-path decisions explain these differences?

# Experimental setting

**Human benchmark: #fincap**

- 164 independent human research teams
- Same proprietary dataset: EURO STOXX 50 Index Futures trades
- Same six pre-specified hypotheses
- Teams choose their own measures, filters, frequencies, and models

**AI replication**

- AI agent receives the same project materials and data metadata
- Generates Python code locally
- Code is executed, repaired if needed, and validated
- Main model: GPT-5.2 fixed snapshot; other OpenAI models as robustness

## Outcome

Each human team / AI run produces an estimate and standard error for each hypothesis. The paper then compares the human and AI *distributions* of results.

# The six hypotheses

| | Hypothesis | Nature of discretion |
|---|---|---|
| H1 | Market efficiency has not changed over time | High: choose efficiency measure, frequency, model |
| H2 | Realized bid-ask spread on market orders has not changed | Medium/high: spread construction, frequency, filters |
| H3 | Client share of volume has not changed | Lower: client trade flag exists in data |
| H4 | Client realized bid-ask spreads have not changed | Medium/high: client spread construction |
| H5 | Fraction of client trades executed via market/marketable limit orders has not changed | Lower: order-type flag exists in data |
| H6 | Relative gross trading revenue for clients has not changed | High: revenue construction, scaling, filters |

**Important:** Even the "simple" hypotheses still require empirical choices: frequency, trend model, excluding periods, settlement weeks, and outliers.

# Main findings

- AI and human outcome distributions are statistically different for most hypotheses.
- AI estimates usually have **lower dispersion** than human estimates.
- For more complex hypotheses, AI estimates are often **shifted in location**, not just tighter.
- The differences are not driven away by focusing on high-quality human teams.

## My summary take

AI does not reproduce the distribution of human empirical practice. It appears more stable, but partly because it explores a narrower set of analysis paths.

This is an important result, but it should be interpreted as **alignment with human research practice**, not accuracy relative to the true estimand.

# Why do AI and human estimates differ?

## Multiverse idea

Map each implementation onto a sequence of decision forks. Then ask whether AI and humans place different weights on different paths.

**Model choice**

- AI strongly favors trend-stationary regressions
- Humans split across trend regressions, relative changes, and log differences

**Frequency choice**

- AI rarely uses daily frequency
- Humans use a broader range of sampling frequencies

**Data handling**

- AI rarely removes open/close periods
- AI rarely excludes settlement weeks
- AI rarely applies outlier treatment

# What I like about the paper

- **Timely question:** AI is entering research design, coding, and empirical implementation, not just writing.
- **Clever benchmark:** The human comparison is a distribution, not a single "correct" result.
- **Interpretability:** The multiverse framework opens the black box and points to specific forks.
- **Constructive message:** The paper does not just say AI is bad; it tells us what to watch for.

## My favorite contribution

The paper turns AI-assisted empirical research into something auditable: not just "what answer did the AI produce?", but "which empirical defaults did the AI choose?"

# Comment 1: Benchmark vs. ground truth

**The paper convincingly shows:** AI does not reproduce the human research distribution.

**But without ground truth, it is harder to know whether this is:**

- a failure of AI,
- a failure of humans,
- a failure in the original problem specification (vagueness)

## Suggestion

Separate two categories more explicitly:

- **AI errors:** clear methodological mistakes, coding errors, invalid estimands, wrong units, mechanical failures.
- **AI divergences:** different but defensible choices that depart from the modal human path.

# Comment 2: Do we care about distribution across runs within a model or distribution across models?

## Two sources of variation

| | **Human teams** | **AI runs** |
|---|---|---|
| Unit | Different researchers | Same agent, repeated runs |
| Noise | Skill, effort, taste, priors, incentives | Sampling randomness prompt, repair loop |
| Interpretation | Across-researcher dispersion | Within-agent Monte Carlo dispersion |

Lower AI dispersion may be partly mechanical: **one model GPT** versus **many humans**.

**Closer human analog:** Kahneman, Sibony, Sunstein (2021) *occasion noise* — the same judge gives different sentences on different days. Within-judge noise is the right comparator for sampling-temperature variation, not across-judge noise.

**Suggestion:** Use additional models: Claude, Gemini, DeepSeek-R1, Grok, etc

# Comment 3: Prompt (and process) sensitivity

The user prompt is long and complex, with no intermediate feedback, and likely unlike how people use AI for research. The human researchers needed to write a paper, which forced them to justify choices and reflect on them. Research is a process.

**For humans:**

- Peer review
- Draft → review → revise

**For AI, the analogues:**

- Prompts, checklists, critic agents
- Plan → critique → revise

**Natural counterfactuals:**

- Unguided AI vs. checklist-guided AI
- Prompt that explicitly asks for multiple plausible approaches
- Prompt that requires outlier, open/close, settlement-week, and frequency discussion
- Plan → critique → revise workflow

**Question for the authors:** Is this an *LLM problem* or a *prompt problem*?

# Other comments

- **Generalizability:** #fincap is one dataset, one set of hypotheses, narrowly constructed. How does this travel to asset pricing, corporate finance, or settings where the data structure is less rigid?
- **Top-quality human subsample is small (N=8).** The robustness result that AI-human gaps persist is suggestive but power-limited.
- **Does anyone deploy AI this way?** Real researchers use AI with prompts, code review, and checklists — not as autonomous agents. The interesting comparison is guided AI vs. unguided AI vs. human vs. human + AI.


- **Replication vs. research:** for a pre-registered replication with a clear target, AI can iterate until it hits the answer. For open-ended research where humans themselves disagree, the absence of a target *is* the problem.

# What excites me

**Two results, read together:**

- #fincap (Menkveld et al. 2024): random human teams give widely dispersed estimates. Four peer-review stages compress the IQR by **47%** and the IDR by **68%**.
- This paper: random AI instances give tighter distributions, but with substantial variation and shifted defaults.

**The constructive research design I would love to see - the engineering problem of making the best human+AI researcher**

**Step 1: Pick a ground truth for #fincap**
**Step 2: Engineer the AI to get there** — vary prompts, scaffolds, critic agents, and feedback loops, and measure which interventions close the gap.

# Conclusion

- This is a clever and timely paper with a genuinely useful audit framework.
- The main result is not just that AI makes "errors", but that AI has empirical defaults.
- Those defaults can reduce dispersion, but may also shift estimates away from human practice.
- Good luck with the paper!