AI $\times$ Human Session Discussion of

"Detecting Lies When Truth is Unobservable"

Wenhao Li, USC Marshall and NBER

UCLA Fink Conference, 2026

Outline

Outline

A long tradition of statistical forensics

- **Benford (1938):** first-digit distribution of clean data and how that flags manipulation in accounting, elections, scientific papers.
- **Burgstahler-Dichev (1997), Bollen-Pool (2009):** density discontinuity at zero in earnings / hedge fund returns.
- **Henderson et al. (2012), Martinez (2022):** nighttime lights as a benchmark for autocratic GDP.

Each test requires something: a reference distribution or a known threshold.

This paper asks: what if we had even less?

- **Setup:** observe only the reported series $\{Y_t\}$. Underlying truth $\{X_t\}$ is unobservable.
- **Insight:** one-sided manipulation leaves a *distribution-free* fingerprint in the lower partial moments.
- **The model.** True outcome $X \sim F$. Reported series:

$$Y \;=\; X \;+\; \varepsilon\,(\tau - X)^{+}, \qquad \varepsilon \in [0,1].$$

  Below threshold τ, shortfalls are attenuated by ε. Above, truthful.

- **Interpretation.** The agent **adds a put option on their reported outcomes**, strike τ, scale ε.

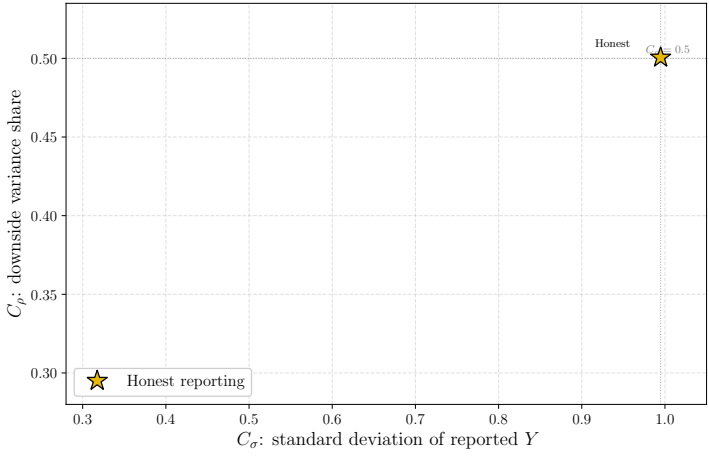
  Central identity (Prop 3). Distribution-free, almost-sure:

$$(\tau - Y)_{+} \;=\; (1-\varepsilon)\,(\tau - X)_{+} \quad \text{a.s.}$$

  Consequently $\mathrm{LPM}_{p,\tau}(Y) = (1-\varepsilon)^{p}\,\mathrm{LPM}_{p,\tau}(X)$ for every $p \geq 1$.

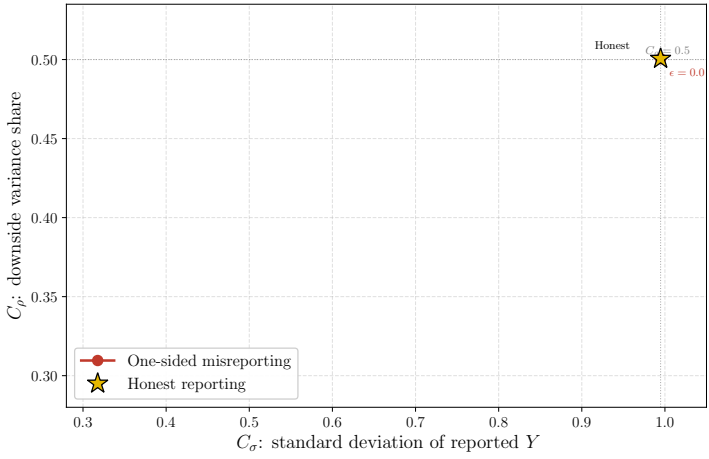
Two statistics for diagnosis

Diagnostic pair: $C_\sigma \equiv$ std of Y; $C_\rho \equiv \dfrac{\sum_{Y_t<\bar{Y}}(Y_t-\bar{Y})^2}{\sum_t (Y_t-\bar{Y})^2}$ (honest null: 0.5)



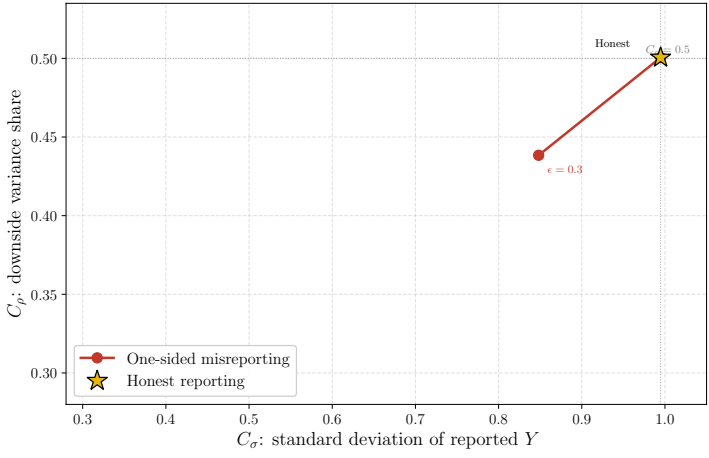
Two statistics for diagnosis

Diagnostic pair: $C_\sigma \equiv$ std of Y; $C_\rho \equiv \dfrac{\sum_{Y_t<\bar{Y}}(Y_t-\bar{Y})^2}{\sum_t (Y_t-\bar{Y})^2}$ (honest null: 0.5)



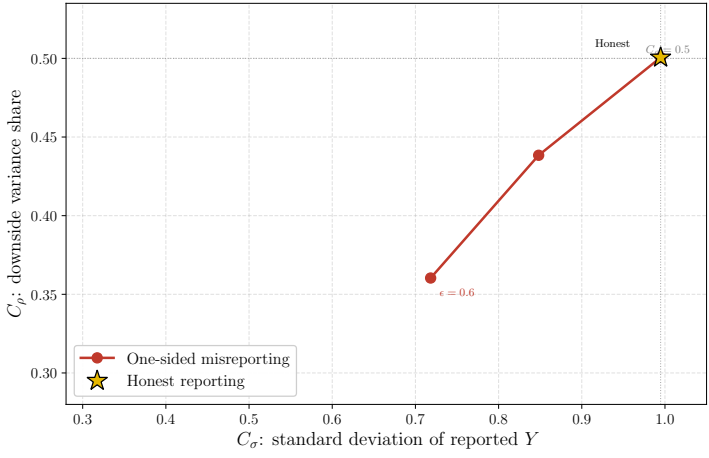
Two statistics for diagnosis

Diagnostic pair: $C_\sigma \equiv$ std of Y; $\quad C_\rho \equiv \dfrac{\sum_{Y_t<\bar{Y}}(Y_t-\bar{Y})^2}{\sum_t (Y_t-\bar{Y})^2}$ (honest null: 0.5)



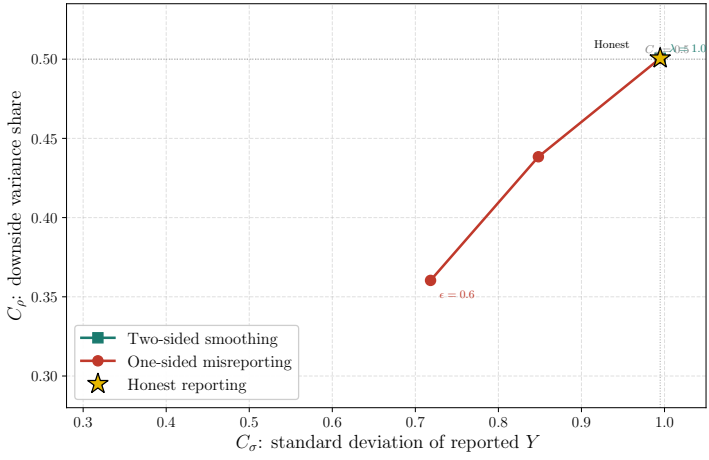
Two statistics for diagnosis

Diagnostic pair: $C_\sigma \equiv$ std of Y; $C_\rho \equiv \dfrac{\sum_{Y_t<\bar{Y}}(Y_t-\bar{Y})^2}{\sum_t (Y_t-\bar{Y})^2}$ (honest null: 0.5)



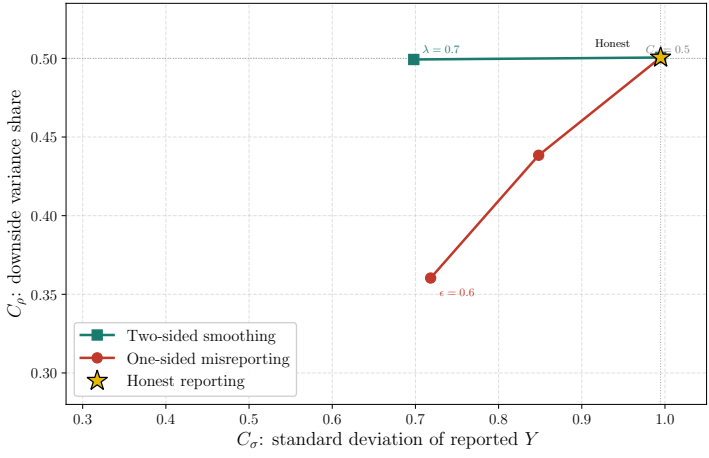
Two statistics for diagnosis

Diagnostic pair: $C_\sigma \equiv$ std of Y; $C_\rho \equiv \dfrac{\sum_{Y_t<\bar{Y}}(Y_t-\bar{Y})^2}{\sum_t (Y_t-\bar{Y})^2}$ (honest null: 0.5)



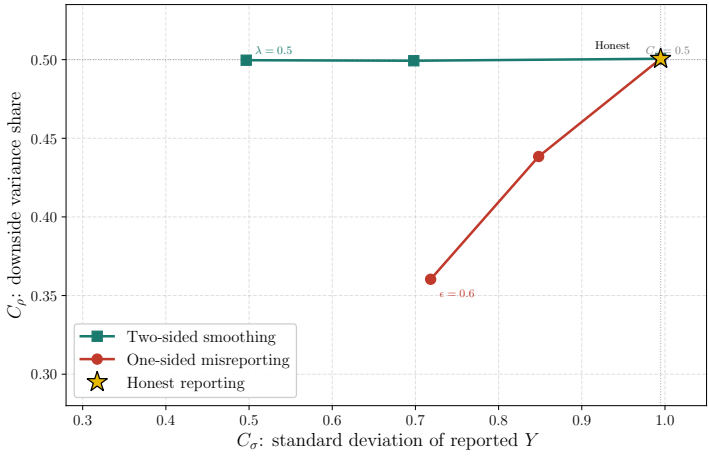
Two statistics for diagnosis

Diagnostic pair: $C_\sigma \equiv$ std of Y; $C_\rho \equiv \dfrac{\sum_{Y_t<\bar{Y}}(Y_t-\bar{Y})^2}{\sum_t (Y_t-\bar{Y})^2}$ (honest null: 0.5)



Two statistics for diagnosis

Diagnostic pair: $C_\sigma \equiv$ std of Y; $C_\rho \equiv \dfrac{\sum_{Y_t<\bar{Y}}(Y_t-\bar{Y})^2}{\sum_t (Y_t-\bar{Y})^2}$ (honest null: 0.5)



Spanning result. No strategy leaves *both* statistics unchanged.

Features of AI Writing in this paper

- **Proposition inflation, no shrinkage**
  - 9 propositions, 11 simulation subsections: all at equal depth
- **Theoretical density, empirical absence**
  - Algebra: low cost. Finding right data sources, downloading data, running regressions: costly.
- **Hedge-and-dismiss objections**
  - Every hard concern acknowledged *and* resolved in the same sentence or paragraph.

Features of AI Writing in this paper

- **Proposition inflation, no shrinkage**
  - 9 propositions, 11 simulation subsections: all at equal depth
- **Theoretical density, empirical absence**
  - Algebra: low cost. Finding right data sources, downloading data, running regressions: costly.
- **Hedge-and-dismiss objections**
  - Every hard concern acknowledged *and* resolved in the same sentence or paragraph.

Core asymmetry: humans *shrink* structure under memory pressure; AI *compresses* text while keeping every item.

Taste in research is largely a shrinkage skill built under capacity constraints guided by value judgment, but AI has no such constraints or value judgment, so no such taste.

Outline

[Me] Comment 1: Simulation-only evidence

- The paper positions itself against Bollen-Pool, Burgstahler-Dichev, and Martinez. All ran their tests *on real data at scale*.
- All exercises are simulation-based, assuming the model is correct. The toolbox is never validated on real data.
- Useful tests for the paper's claims:
  - Does it reject Madoff's returns?
  - Does it catch manipulators that existing tests *miss*?

[Me] Comment 2: Central results are not testable

- Volatility C_σ results are **not testable**
  - Detecting manipulation requires comparing the reported std to the true underlying σ, which we never observe.
- Downside variance share C_ρ results require **assuming symmetric** F
  - The honest null $C_\rho = 0.5$ only holds under a symmetric true distribution.
  - Earnings, GDP growth, and fund returns are all skewed, so the null is wrong before any manipulation occurs.
- **What is testable but never stated:**
  - Does C_ρ of the auxiliary residual drop before known fraud/restatements and recover after?
  - Does low C_ρ today predict fund failures?

[Me] Comment 3: The structure hides the hard problems

- 9 propositions and 11 simulation subsections at equal depth signal a paper that never decided what matters.
- The three hardest problems each get exactly one remark or subsection:
  - **Right skew destroys the null** (40–60% false rejection) → Section 5.10
  - **"Benchmark-free" fails in the headline GDP application** → absorbed as a remark
  - **Dynamic smoothing is a different model entirely** → Remark 7
- Equal depth is not thoroughness: A paper that buries its hardest problem in Section 5.10 has not solved it.

Suggestion: Commit to one deep economic setting

The paper covers all applications at similar depth. Pick one and go deep, trimming the rest.

Three natural candidates:

- **Central bank communication.** Inflation forecasts, forward guidance — do central banks smooth bad outcomes below target? Rich panel data, clear τ (inflation target), observable auxiliary (actual CPI).
- **Voting and electoral reporting.** Election returns in autocratic regimes — well-studied manipulation (Beber-Scacco 2012), existing forensic benchmarks to beat.
- **Financial regulation / PE fund reporting.** NAV smoothing, Preqin data, direct connection to Bollen-Pool and Getmansky-Lo-Makarov. The audience already knows the stakes.

Payoff: one deep application answers Comments 1 – 3 simultaneously: real data *and* a falsifiable prediction in the same setting.

Outline

How the AI discussant was generated

- **Training on my own past discussions.** Claude 4.7 (1M) Max efforts read ∼25 of my past discussion folders and extracted the key features from them.
- **Synthesis.** Archive patterns combined with canonical finance/econ discussion practice (NBER/AFA norms) into a skill: rules, structure template, voice anchor, etc.
- **Generation.** Ran the skill and generated discussion slides, with three key comments.

The experiment: does a skill distilled from my own archive and professional knowledge produce what I would produce myself?

[AI] Comment 1: Headline GDP application

- **The big-stakes application.** Autocratic regimes inflating growth — Martinez (2022) finds ∼35%.
- **But at Level 0:** N=200, t_5, target $\varepsilon = 0.2 \Rightarrow$ power below 20%. Effectively no test.
- **So the paper moves to Level 3:** nighttime lights, electricity, fiscal revenues as auxiliary Z (Prop 8).
- **But nighttime lights is a benchmark.** The paper ends up where Henderson et al. (2012) started, with an external, independently verified proxy.

Conclusion. In the headline application, the "benchmark-free" label does not survive contact with the power curve.

[AI] Comment 2: Reframing as *minimum-data hierarchy*, not a benchmark-free test

What the paper actually delivers:

- A sharp characterization of **what external information each level requires** to conclude what.
- The (C_σ, C_ρ) spanning result is the tool; the four-level hierarchy is the deliverable.
- Useful as a decision tree for a practitioner:
  - Have only Y? $\Rightarrow$ you can flag, not attribute.
  - Have observable transfers? $\Rightarrow$ $\tilde{Y} = Y - T$ separates mechanism.
  - Have an auxiliary series? $\Rightarrow$ projection residual gives identification.

Suggested reframing. The paper's selling point is not "no benchmark required" — it is "detection at minimum data cost, with a sharp hierarchy of what more data buys you."

[AI] Comment 3: The paper detects one manipulation strategy out of many

Framing: each manipulation strategy is like a derivatives position — it has a payoff structure, and leaves a statistical fingerprint.

Strategy	How the liar profits	Reference
Reporting put *(this paper)*	attenuate bad outcomes below τ	Chowdhry–Saxena
Stale-mark forward	delay marking losses to future periods	Getmansky-Lo-Makarov
Auditor's compound option	bet auditor won't dig at cost	Dye (1993)
Calendar spread	look good at quarter-end, reverse after	Ben-David et al.
Credibility call	spend reputation now, collect trust later	Benabou-Laroque

The problem: real manipulators use *combinations*. A fund manager may smooth returns (reporting put) *and* stale-mark *and* window-dress simultaneously.

Open question: when multiple strategies overlap, does C_ρ still identify the reporting put — or does it confound instruments?

Outline

What AI got right and what it missed

	AI discussant	My comments
Comment 1	GDP power collapse; "benchmark-free" label fails	All evidence simulation-only so the tool is unverifiable
Comment 2	Reframe as "minimum-data hierarchy"	Most theoretical predictions are not testable
Comment 3	C_ρ detects one instrument; real liars combine many	Parallel layering without actual depth

What AI got right and what it missed

	AI discussant	My comments
Comment 1	GDP power collapse; "benchmark-free" label fails	All evidence simulation-only so the tool is unverifiable
Comment 2	Reframe as "minimum-data hierarchy"	Most theoretical predictions are not testable
Comment 3	C_ρ detects one instrument; real liars combine many	Parallel layering without actual depth

Pattern: AI engages the paper on its own terms.
My comments refuse the framing and that requires sustained value judgment from a high level.

I did another round of refining the AI skills, but the gaps remain. I also tried ChatGPT and Gemini and conclusions are similar.

Why the gap is durable

Human comparative advantage in critique is protected by the economics of AI training.

- **No-forgetting ⇒ preserve over delete.** AI finds it easier to add a caveat than to cut a claim.
- **RLHF ⇒ agreeableness over refusal.** Sharp adversarial critique hurts average user satisfaction so AI training penalizes it.
- **Pattern-matching ⇒ mimicry over stance.** AI can produce critique-shaped text; it can't commit to a position. Mimicry looks critical without quite landing in depth.

Why the gap is durable

Human comparative advantage in critique is protected by the economics of AI training.

- **No-forgetting ⇒ preserve over delete.** AI finds it easier to add a caveat than to cut a claim.
- **RLHF ⇒ agreeableness over refusal.** Sharp adversarial critique hurts average user satisfaction so AI training penalizes it.
- **Pattern-matching ⇒ mimicry over stance.** AI can produce critique-shaped text; it can't commit to a position. Mimicry looks critical without quite landing in depth.

Implication. Each model generation improves coverage, algebra, and prose, which are the capabilities the broad market wants. The moves that matter for scientific discovery (refusing, cutting, committing) stay constant or regress.

Human $\times$ AI: Implications for scientific discovery

Use AI where its comparative advantage is real:

- Literature coverage, notation, algebra, simulation, polished prose
- Generating the next item in a list: brainstorming extensions, applications, robustness checks

Human $\times$ AI: Implications for scientific discovery

Use AI where its comparative advantage is real:

- Literature coverage, notation, algebra, simulation, polished prose
- Generating the next item in a list: brainstorming extensions, applications, robustness checks

Reserve human judgment for the moves AI cannot make:

- **Shrinkage:** decide which of nine propositions should be deleted entirely
- **Framing refusal:** ask whether the central theorem is scientific or definitional
- **Committed stance:** identify the one weakness that threatens the paper's identity and hold it

Human × AI: Implications for scientific discovery

Use AI where its comparative advantage is real:

- Literature coverage, notation, algebra, simulation, polished prose
- Generating the next item in a list: brainstorming extensions, applications, robustness checks

Reserve human judgment for the moves AI cannot make:

- **Shrinkage:** decide which of nine propositions should be deleted entirely
- **Framing refusal:** ask whether the central theorem is scientific or definitional
- **Committed stance:** identify the one weakness that threatens the paper's identity and hold it

Practical rule: import the constraint artificially.

- Don't ask AI "what are the weaknesses?" Ask "what would make this paper unpublishable?"
- Don't ask AI to "improve the structure." Ask "which three results survive if you cut everything else?"
- The adversarial move has to be in the *prompt*, not inherited from the model.