NEURIPS 2026 · FOUNDATION MODELS FOR TEMPORAL SYSTEMS WORKSHOP

Do LLM Forecasts Change Their Minds Coherently?

A label-free martingale audit of forecast paths

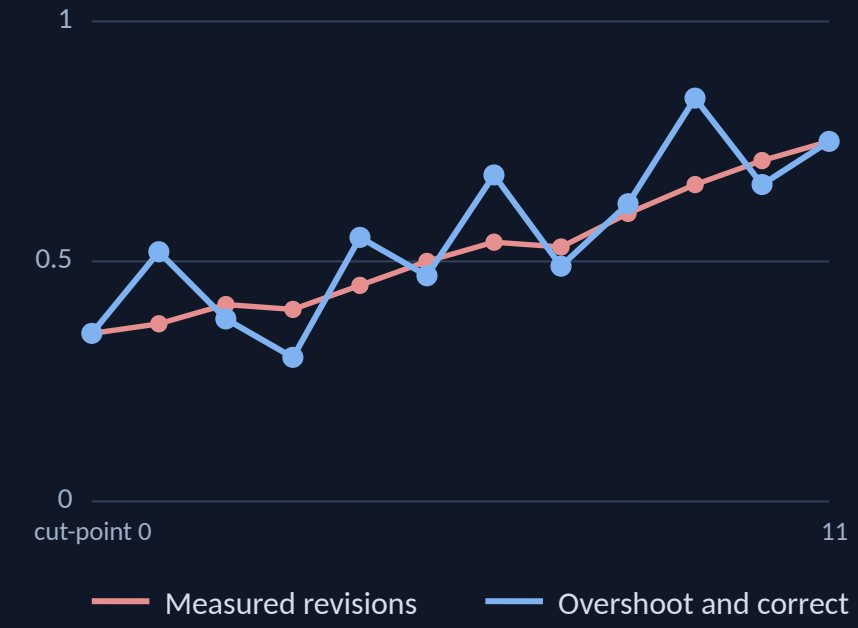
Adarsh Agrawal

Independent Researcher

openreview.net/forum?id=m9twt6TIfv

SAME DESTINATION, DIFFERENT ROUTE

Illustrative probability paths over 12 cut-points



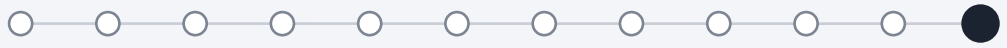
Same start and end, so endpoint scores are identical.

MOTIVATION

Endpoint scores cannot see the route

A forecast is a sequence of decisions. Revisions trigger escalation, hedging, and review before the question resolves.

Endpoint evaluation



only the final probability is scored

- Needs the resolved outcome
- Scores only where the forecast finishes
- Cannot tell large reversals from measured revisions

Path audit



every revision is scored

- Needs neither the outcome nor intermediate labels
- Scores all 11 revisions on a fixed grid
- Applies before the market resolves

Prior martingale tests probe in-context learning, reasoning, and resampling; consistency audits constrain probabilities across questions.
This audit follows one question over time, as dated evidence arrives.

METHOD

Coherent paths move as much as they resolve

$$M = \Sigma_k (p_k - p_{k-1})^2$$

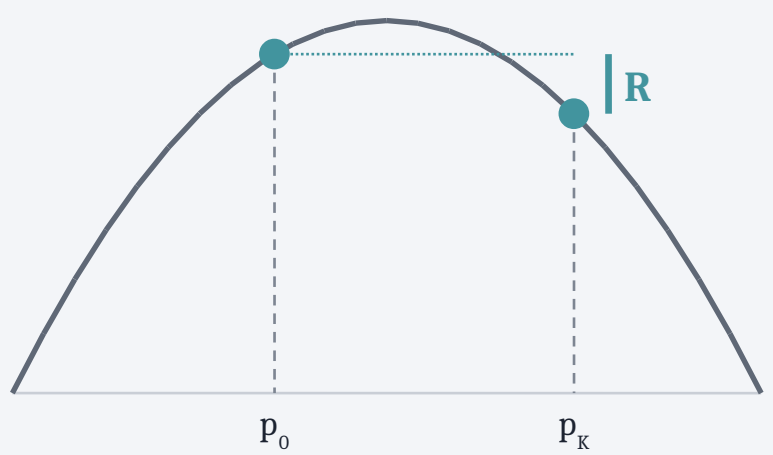
Squared movement: how far the forecast travels

$$R = p_0(1 - p_0) - p_K(1 - p_K)$$

Uncertainty reduction: how much the forecast resolves

$$C = M - R$$

Residual: movement in excess of the uncertainty resolved



Illustrative: R is the drop in uncertainty from start to end

If $p_k = \Pr(Y = 1 \mid F_k)$ on nested information sets, then $\mathbf{E[M] = E[R]}$, so $\mathbf{E[C] = 0}$.

Label-free: no outcome or intermediate labels

Moment condition: one positive C is no violation

Mean C ≈ 0 is necessary, not sufficient

STUDY DESIGN

Model and crowd paths on the same grid

73 resolved markets

Binary Polymarket markets, 21 topics, resolving Mar–Sep 2026

12 cut-points

Roughly even intervals up to a fixed horizon: 11 revisions

Nested headlines

Every dated headline available by each cut-point, appended

Stateless calls

No crowd prices or prior outputs; temperature 0

PAIRED ESTIMAND

$$\Delta C_i = \text{mean}_j\, C_{ij}(\text{model}) - C_i(\text{crowd})$$

Systems are averaged within each market, markets weigh equally, and the crowd is a comparator scored by the same identity. Inference: 10,000 two-level bootstrap draws (topics, then markets).

PANEL

8 systems

OpenAI, Anthropic, and Amazon × Efficient, Balanced, Flagship tiers, fixed before any residual was computed

COHORTS

73
primary resolved markets in 21 topics

41 of 52
prospective markets frozen Jul 20–Aug 5, resolved by Sep 15

30
later cohort, frozen Aug 12

28
nested, unresolved during elicitation

LLM paths move more than they resolve

+0.061

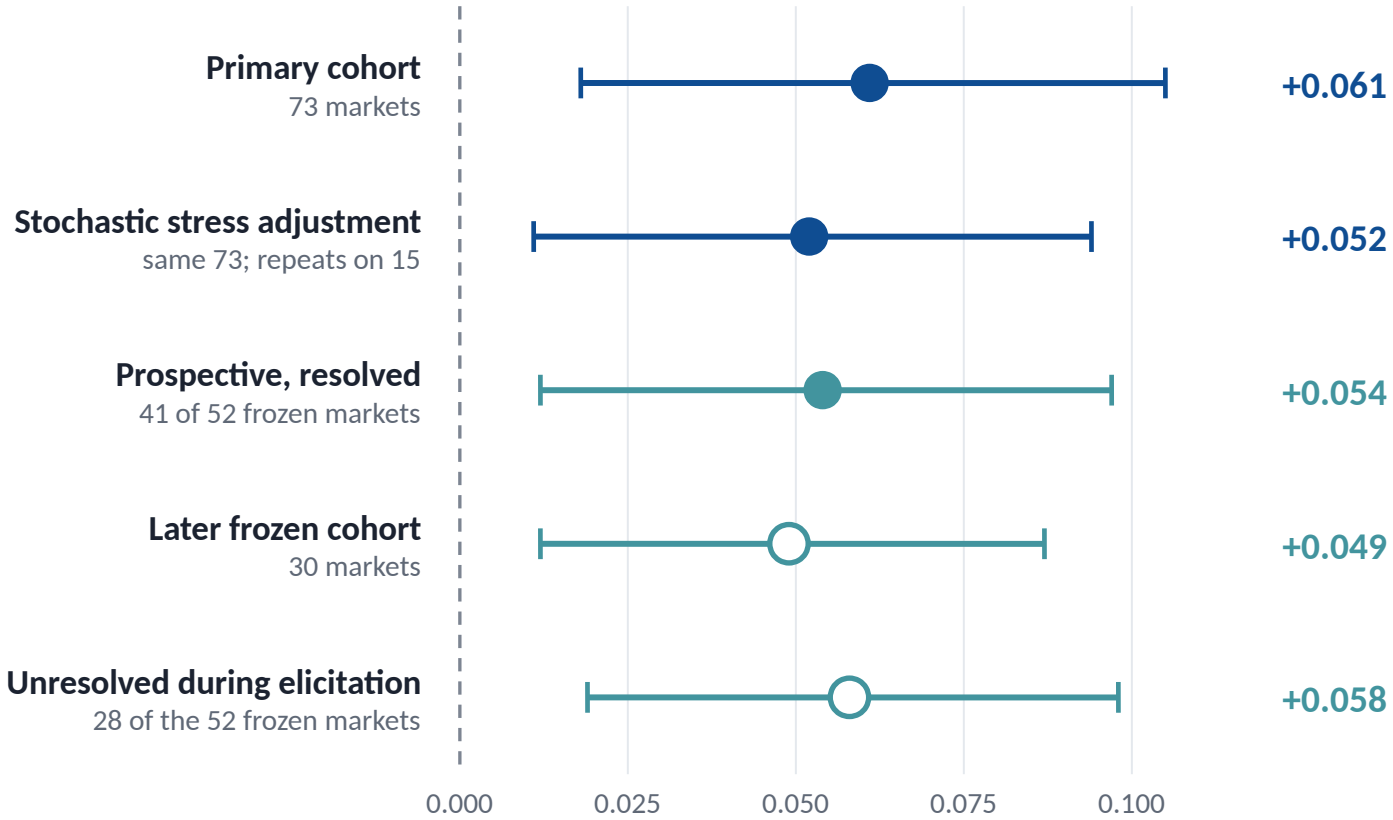
Model − crowd residual gap ΔC

95% CI [+0.018, +0.105] · Wilcoxon $p = 3.7 \times 10^{-6}$

8 of 8
systems show a positive mean gap

+0.058
averaging within providers first

+0.063
excluding all 9 markets with an extraction failure



Paired residual gap ΔC (model − crowd), 95% two-level bootstrap intervals

Reused and nested rows are complementary checks, not independent replications.

RESULT 1 · ROBUSTNESS

Not elicitation noise, not an inference artifact

ELICITATION NOISE

+0.052

Five repeats per cut-point at temperature 0.3 on 15 markets shrink the gap by about 15%

< 3%

A temperature-0 placebo with no new evidence: 0.0014 squared movement per path, under 3% of excess movement

C ≈ 0

Synthetic Bayesian agents recover mean C ≈ 0, with 94.6–95.2% interval coverage

INFERENCE AND SPECIFICATION

+0.057

Equal topic weighting [+0.012, +0.103]
Wild-cluster bootstrap [+0.014, +0.111]

p = 0.002

Jackknife cluster-robust inference over the 21 topics (SE = 0.019)

+0.053 to +0.069

Leave-one-topic-out estimates across all 21 topics, with no sign change

Also stable across **8, 12, and 24 cut-points**, prompt paraphrases, decimal vs percentage outputs, logit elicitation, and fresh vs accumulated context.

Same headlines, different updating

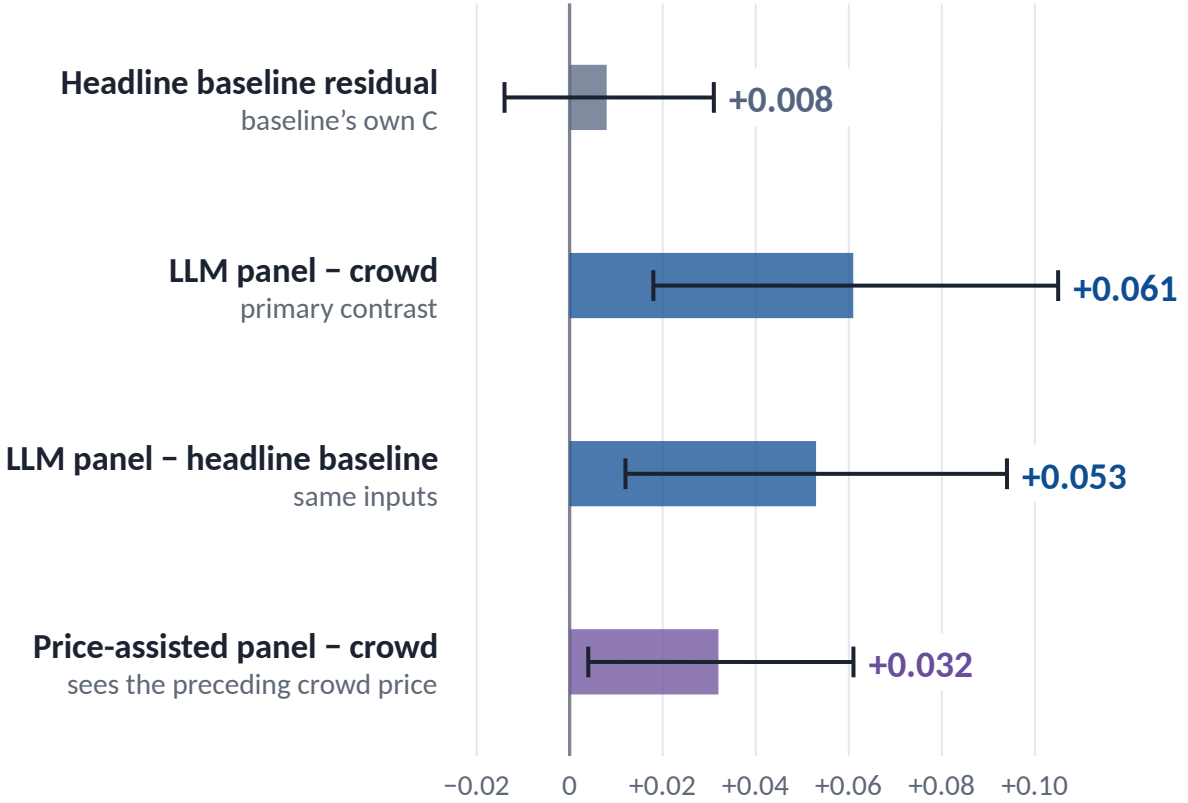
IDENTICAL EVIDENCE, TWO FORECASTERS

Nested headline snapshots
12-point grid · no prices · no prior outputs

LLM panel
eight systems

Headline baseline
TF-IDF logistic model, trained on 200 other markets, then frozen

The baseline stays near the martingale benchmark while the LLM panel exceeds it by +0.053. Revealing the preceding crowd price halves the gap (+0.032) but does not remove it.



Estimate and 95% CI on 73 markets; first row is a residual, the rest are paired contrasts

Excess movement drives 80% of the gap

EXACT DECOMPOSITION

$$\Delta C = \Delta M - \Delta R$$



Excess movement
models travel farther than the crowd

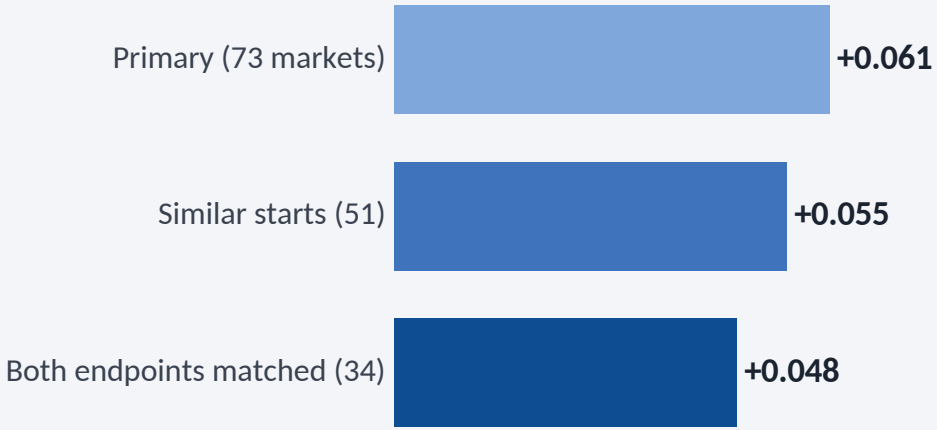
Less resolution
−ΔR = +0.012

Similar starts (51 markets)

ΔM = +0.044, ΔR = −0.011, ΔC = +0.055: still dominated by movement.

Movement, not where forecasts start or finish, carries the gap.

ENDPOINT MATCHING

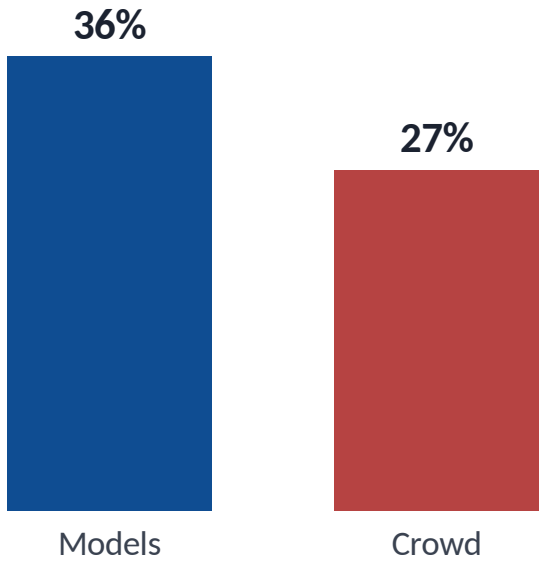


Matched: model and crowd start and finish within 0.05 of each other.
ΔC = +0.048, 95% CI [+0.010, +0.088].

Nearly the same destination, a different route.

LLMs overshoot, then correct

DIRECTION REVERSALS



Share of transitions that reverse direction; permutation p = 0.004

CONCENTRATED REVISIONS

54% vs 41%

of movement carried by the two largest revisions (models vs crowd)

−0.14 vs −0.04

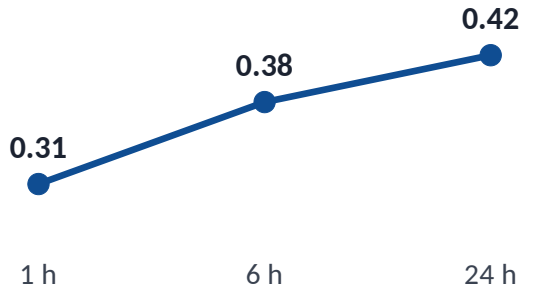
next-step slope on the current step: big LLM jumps are followed by smaller ones

+0.042

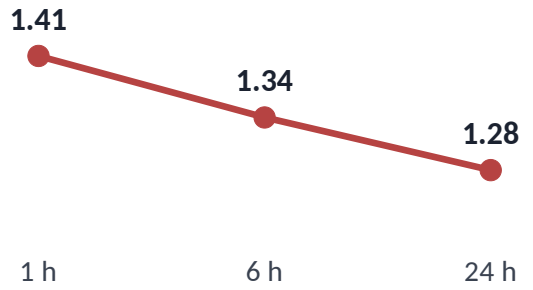
gap left after removing each path's largest squared step [+0.008, +0.077]

AFTER A HEADLINE · 784 WINDOWS

Model–crowd update correlation



Model / crowd absolute step size



Alignment improves with wider windows; LLM steps stay larger. Descriptive.

Stronger systems come closer to the crowd

MEAN ΔC BY PROVIDER AND TIER · 73 MARKETS

	Efficient < 1250 Elo	Balanced 1250–1350	Flagship > 1350
OpenAI	`gpt-5.6-luna` Elo 1198 **+0.102**	`gpt-5.6-terra` Elo 1289 **+0.055**	`gpt-5.6-sol` Elo 1407 **+0.023**
Anthropic	`claude-haiku-4.5` Elo 1165 **+0.127**	`claude-sonnet-5` Elo 1312 **+0.041**	`claude-opus-5` Elo 1421 **+0.016**
Amazon	`nova-2-lite` Elo 1142 **+0.110**	`nova-premier` Elo 1301 **+0.035**	*unavailable*

Tiers from the July 15, 2026 Chatbot Arena Elo snapshot; panel fixed before any residual was computed.

−0.024 per tier

95% CI [−0.041, −0.008]
same direction within every provider

−0.00041 per Elo point

continuous Elo, p = 0.003
two-tier and quartile splits agree

+0.112 and +0.138

held-out open-weight systems:
Llama-3.1-405B and Qwen-2.5-72B

A fixed-panel pattern, not a universal scaling law. Elo is not a forecasting benchmark.

Complements regression-based martingale tests

TWO RELATED MOMENTS

Regression test (He et al.)
a nonzero β violates a necessary implication of martingale updating

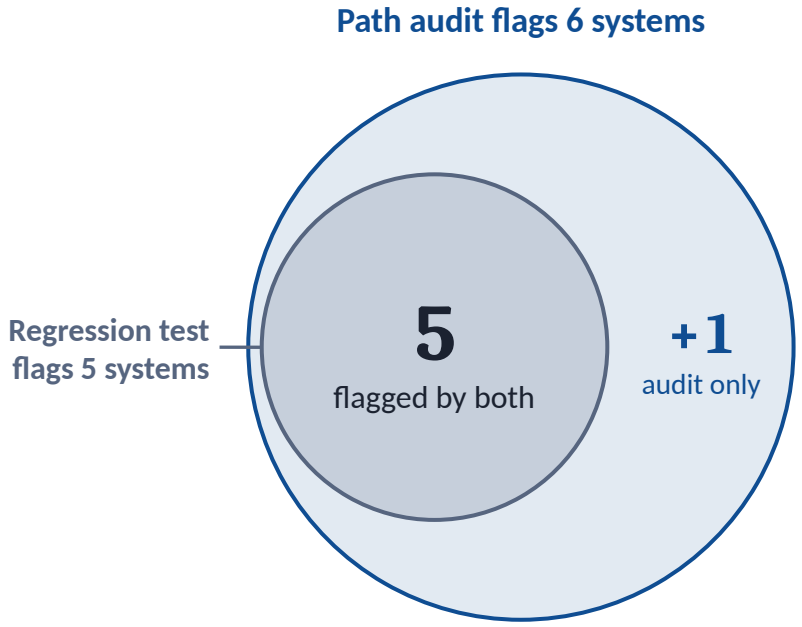
$$\Delta p_{i,k+1} = \alpha + \beta\, p_{ik} + \varepsilon$$

Path audit
the same family of moments, with different weighting and aggregation

$$C = \sum_k (2p_{k-1} - 1)(p_{k-1} - p_k)$$

r = 0.43
across markets, p < 0.001

ρ = 0.71
across systems, p = 0.047



`gpt-5.6-terra`: ΔC = +0.055, but β = −0.07 (p = 0.14). The audit catches a departure the regression misses. Report both.

Beyond binary: on 14 categorical markets (3–5 outcomes), ΔC_{vec} = +0.039 [−0.008, +0.087]: same sign as the binary result, but the interval spans zero.

A process check, not a target to optimize

HOW TO READ C

1. **A constraint, not a goal**
   A forecaster that never revises attains C = 0 while ignoring evidence.

2. **Crowd-like is not automatically better**
   Read C alongside responsiveness, calibration, accuracy, and decision-relevant stability.

3. **Comparative, not causal**
   Models and traders work from different information sets.

4. **Not a verdict on rationality**
   Neither a positive residual nor a nonsignificant slope establishes overall unreliability or rationality.

LIMITATIONS

Information sets
Traders may use private signals; the headline baseline and endpoint matching narrow, but do not eliminate, this difference.

One platform, fixed panel
Polymarket only, eight fixed systems; Elo is not a forecasting benchmark.

Decoding
Temperature 0 is operationally deterministic, not guaranteed identical across providers; the stress test is no universal noise ceiling.

Baseline design
Its training and regularization also shape how the baseline updates.

TAKEAWAY

Evaluate forecasts as processes, not endpoints

1 Audit the path

Needs no outcome or intermediate labels, so it runs before resolution.

E[C] = 0

under coherent Bayesian updating

2 LLMs overshoot

Excess movement and reversals, not endpoints, drive the gap.

+0.061

model – crowd gap, 80% of it excess movement

3 Report alongside

Pair the audit with calibration, accuracy, and regression tests.

−0.024 per tier

stronger systems sit closer to the crowd

Scope: 73 resolved Polymarket markets and a fixed eight-system panel; comparative, not causal.

Release: market manifest, prompts, headlines, all 657 attempted paths, and code.

Thank you — questions welcome

Paper and reviews on OpenReview

openreview.net/forum?id=m9twt6TIfv