NEURIPS 2026 · FOUNDATION MODELS FOR TEMPORAL SYSTEMS WORKSHOP

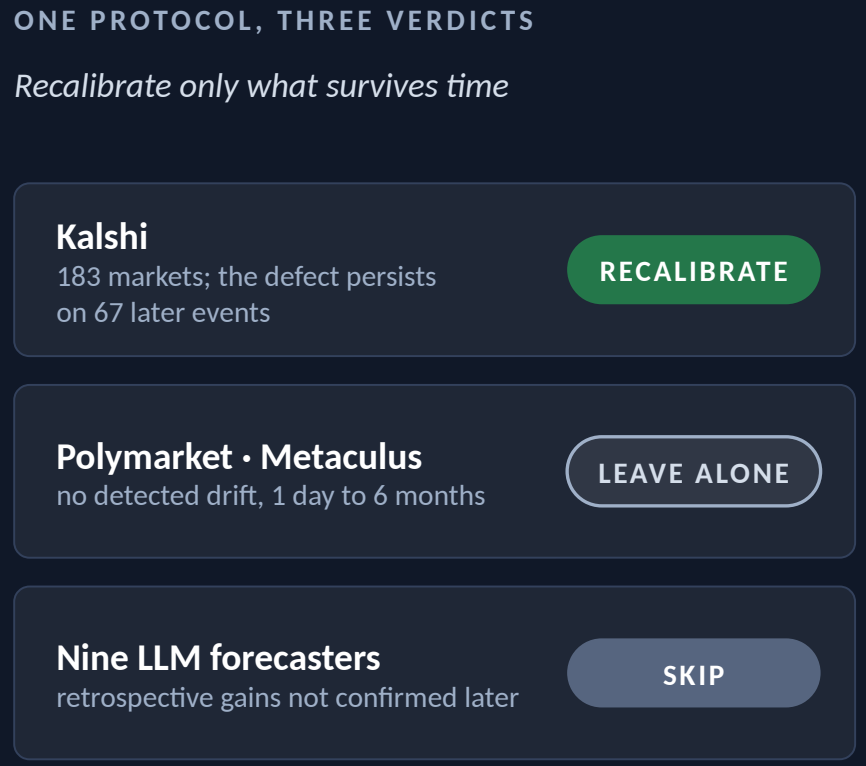
The Reliability Horizon

A Temporal Stress Test for Forecast Recalibration

Adarsh Agrawal

Independent Researcher

openreview.net/forum?id=emVhXV7PHf



MOTIVATION

A retrospective gain is not a deployment decision

A map that improves calibration on history can still worsen a proper score or fail under temporal shift.

1

Trend

Does calibration change with lead time?

Same events compared at every horizon

2

Equivalence

Is any change practically negligible?

Equivalence test at a stated margin

3

Benefit

Does a fitted correction help future forecasts without hurting utility?

A gain that persists on later events

Markets visible six months before resolution are a selected subset of those visible one day before. Fixing event identity controls that composition change; it does not make horizon differences negligible or justify a correction.

THE PROTOCOL

Three gates before any correction ships

1 Screen

Holm-controlled reliability gain across the three model tiers; a single control needs a positive 95% CI

fail **SKIP**

2 Utility

90% lower bounds above −0.08 nats for log score and −0.02 for decision loss

fail **REJECT**

pass **CANDIDATE**

3 Temporal

Gain and utility re-established on later events; frozen prospective check for model forecasts

pass **RECALIBRATE**

Precedence: unresolved reliability at any stage returns SKIP; an established gain that fails utility returns REJECT. CANDIDATE is not approval.

Δrel = Rel(p) − Rel(g(p))

binned reliability gain of the map g; positive favors recalibration

Δlog = NLL(p) − NLL(g(p))

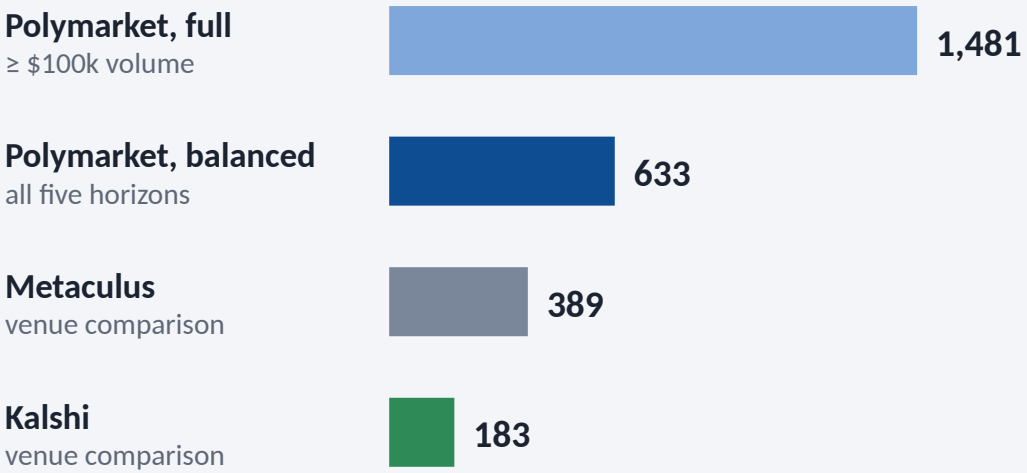
log-score gain in nats; the 0.08-nat margin is log(0.65/0.60)

0.02 decision-loss units

under 2FN + FP loss:
one extra false negative per 100 decisions

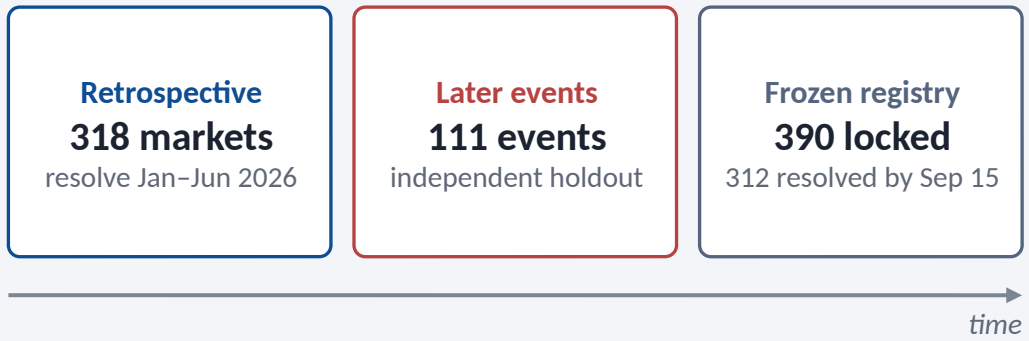
Fixed events, dated inputs, frozen predictions

CROWD CALIBRATION AUDIT



Balanced panel: 633 markets × 5 horizons (1d, 1w, 1m, 3m, 6m) = 3,165 rows, 42.7% of the full sample.

MODEL FORECASTS · NINE SYSTEMS



Efficient GPT-5.6 Luna · Claude Haiku 4.5 · Nova 2 Lite

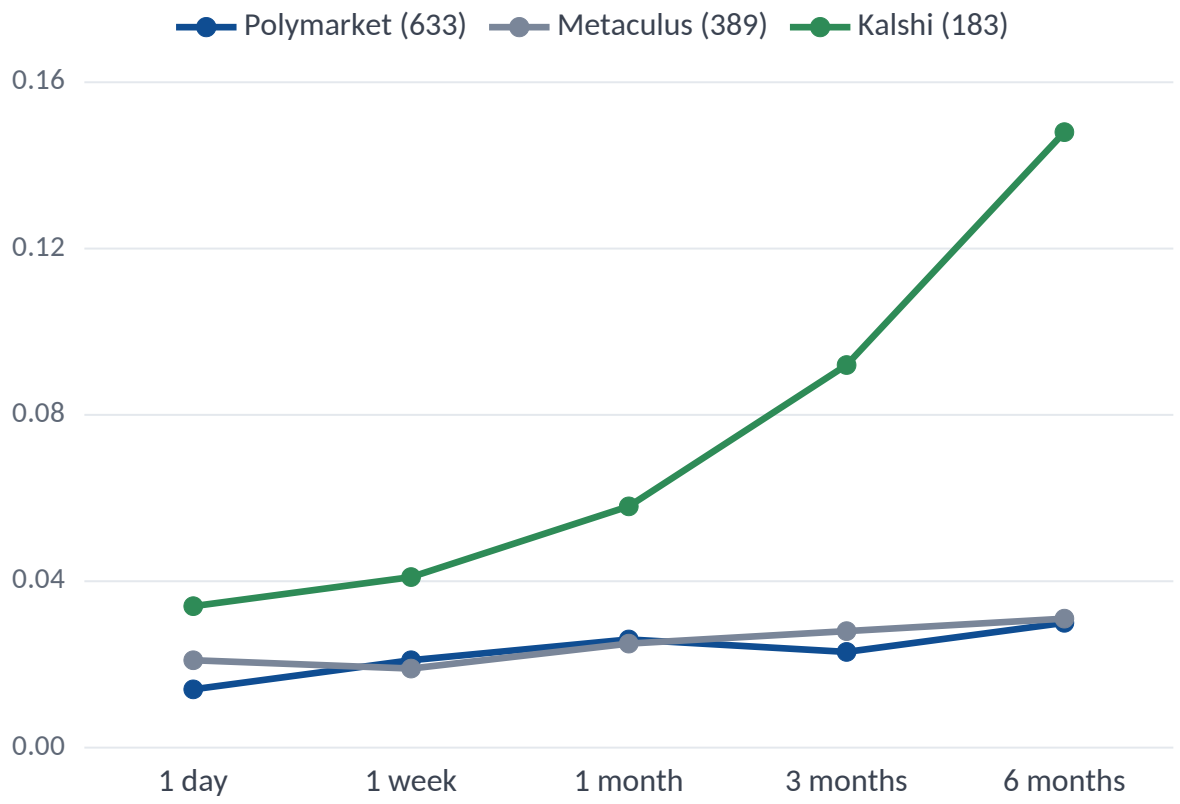
Balanced GPT-5.6 Terra · Claude Sonnet 5 · Nova Premier

Flagship GPT-5.6 Sol · Claude Opus 5 · gpt-oss-120b (anchor)

Inputs: the question plus news and reference snapshots dated no later than the forecast time, never market prices. 18–38% of retrospective events resolve after a model's training cutoff; every prospective event resolves after every cutoff.

Horizon effects are venue-specific

ECE BY HORIZON · EACH VENUE'S BALANCED PANEL



Polymarket

No detected trend (Page p = .412)

Largest contrast +0.016 [−0.004, +0.036]

Equivalence holds at ±0.04 (p = 0.021), not ±0.02 (p = 0.107): no detected trend, not equivalence

Metaculus

Same pattern (p = .351); +0.012 [−0.007, +0.032]

Kalshi

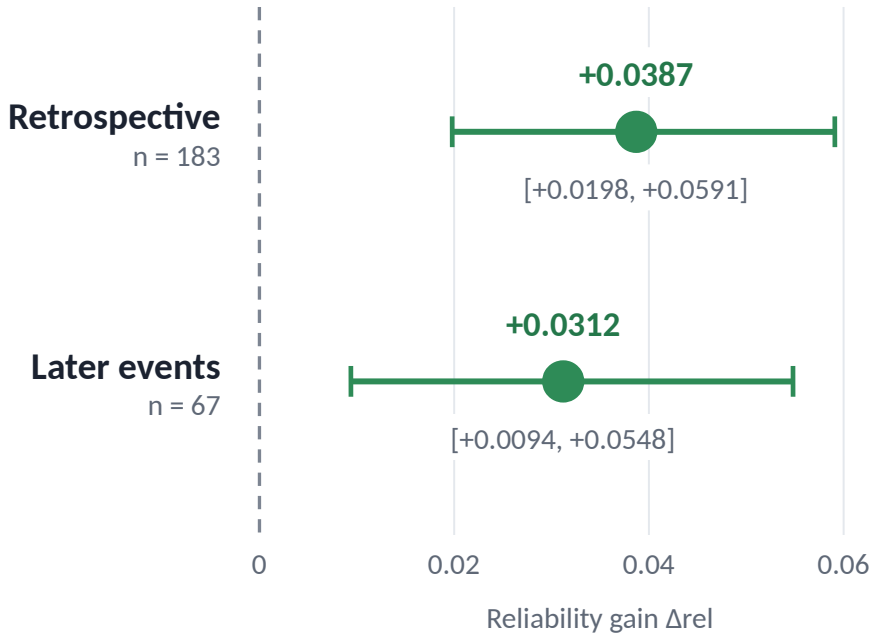
Degrades with horizon (p < .001); +0.114 [+0.062, +0.173]

Polymarket: recalibration stays within ±0.02 ECE of a no-op at every horizon, so the protocol leaves it alone.

RESULT 2 · POSITIVE CONTROL

A persistent defect passes every gate

KALSHI RELIABILITY GAIN · 95% CIs



Both intervals exclude zero; the gain persists on independent later events.

1 Screen

Retrospective Δrel +0.0387 [+0.0198, +0.0591] excludes zero

2 Utility CANDIDATE

Δlog +0.029 [−0.018, +0.077]; decision loss −0.004 [−0.019, +0.012], within the 0.02 tolerance

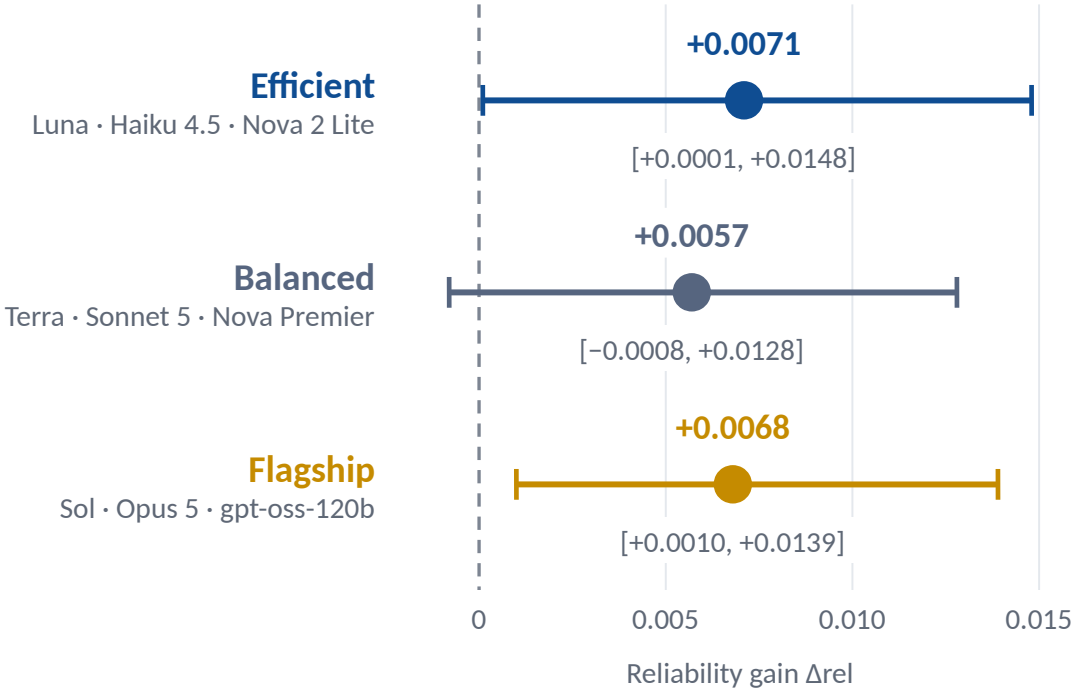
3 Temporal RECALIBRATE

On 67 later events, Δrel +0.0312 and Δlog +0.021

Venue-level temporal holdout; model forecasts also need frozen prospective checks.

Pointwise gains do not survive multiplicity

RETROSPECTIVE ENSEMBLES · POINTWISE 95% CIs



Two pointwise intervals exclude zero.

ONE HOLM FAMILY OF THREE TESTS

Tier	Raw p	Holm p
Efficient	.0498	**.1494**
Balanced	.0742	**.1494**
Flagship	.0611	**.1494**

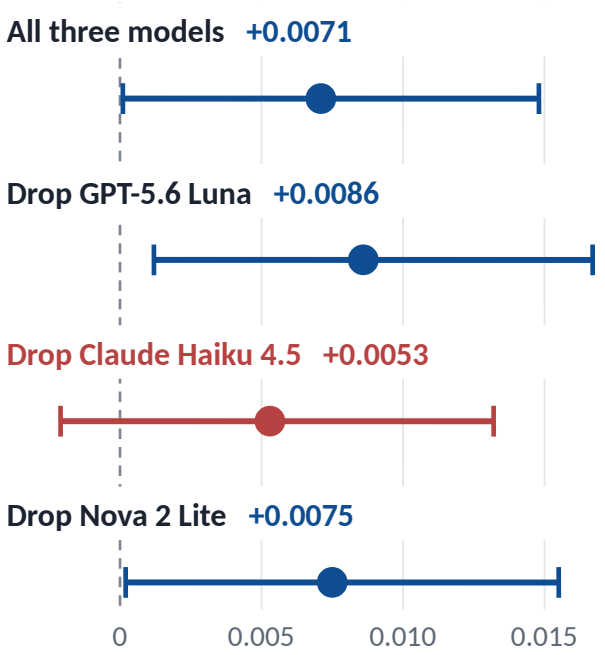
Smallest raw p = 0.0498
exceeds the first Holm threshold, 0.05 / 3 ≈ 0.0167

Screen result SKIP

Pointwise intervals are not simultaneous intervals and cannot override the familywise screen.

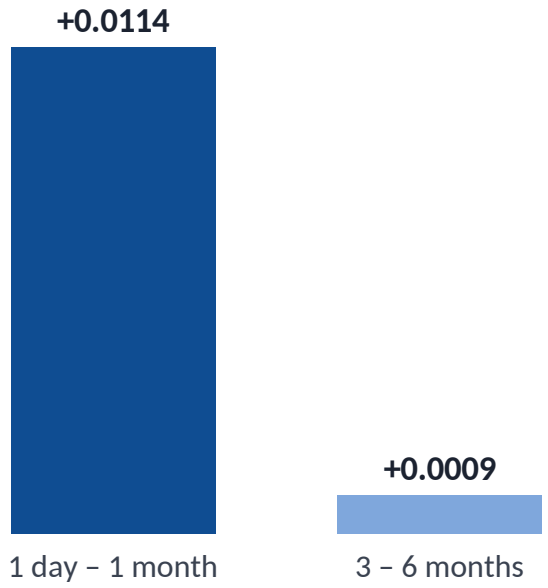
The retrospective gain does not hold up

ONE MODEL CARRIES IT



Efficient-ensemble Δrel, leave-one-model-out: only dropping Claude Haiku 4.5 pushes the interval across zero.

SHORT HORIZONS ONLY



Efficient retrospective Δrel by horizon band; interaction p = 0.021.

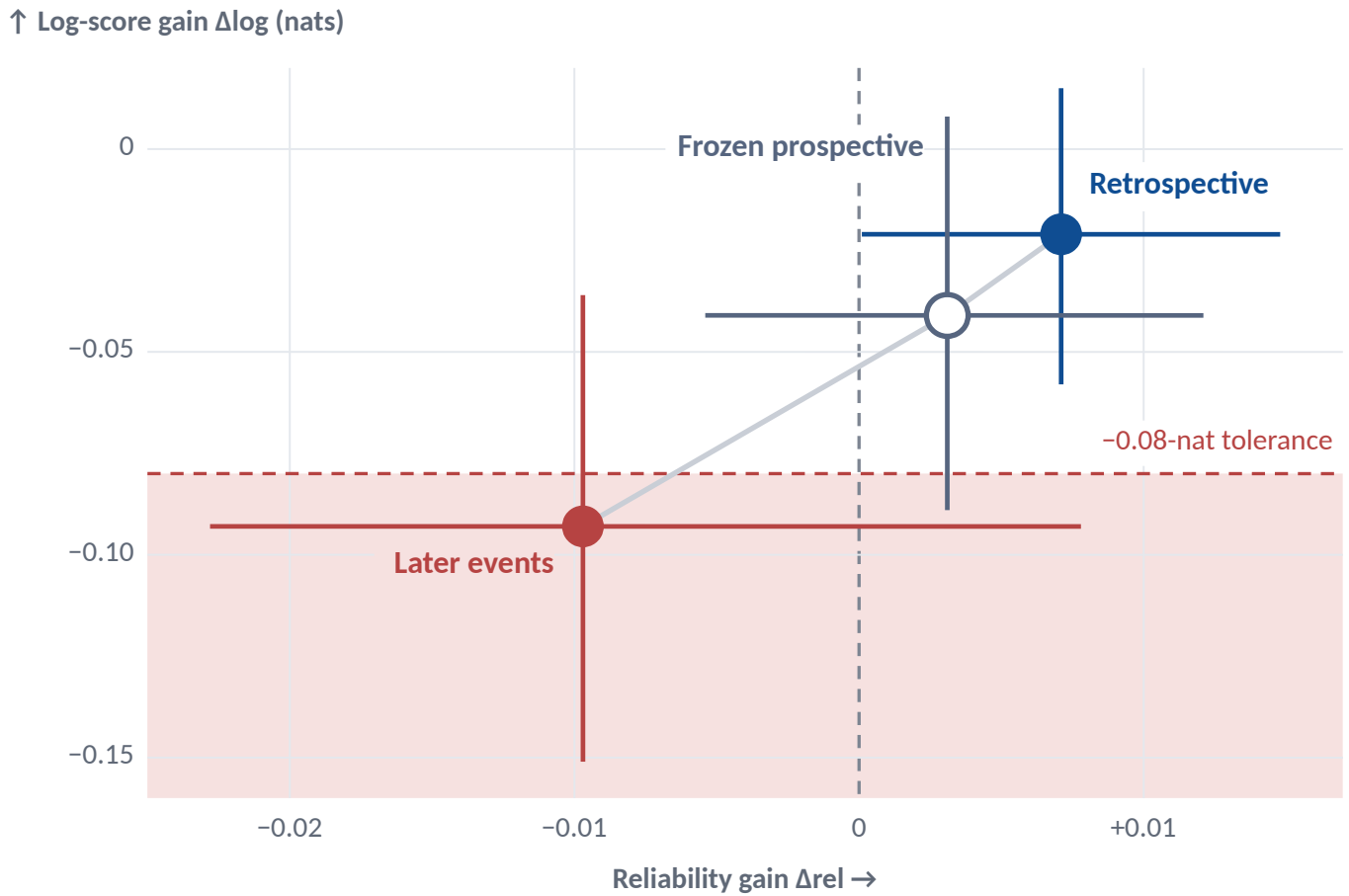
SMOOTHER MAPS

	Δrel [95% CI]	Δlog
Isotonic	**+.0071** [+.0001, +.0148]	**−.021**
Platt	**+.0034** [−.0028, +.0099]	**−.008**
Beta	**+.0048** [−.0014, +.0113]	**−.012**

Platt and beta cut the log-score penalty but lose the reliability gain: both intervals include zero. Δlog uses 90% CIs.

RESULT 3 · TEMPORAL VALIDATION

Later and frozen evaluations do not confirm it



Efficient ensemble. Reliability CIs 95%; retrospective log-score CI 90%, validation log-score CIs 95%.

−0.093 nats

later-event log-score change [−0.151, −0.036]. Trusting the retrospective interval would have shipped this loss.

+0.0031

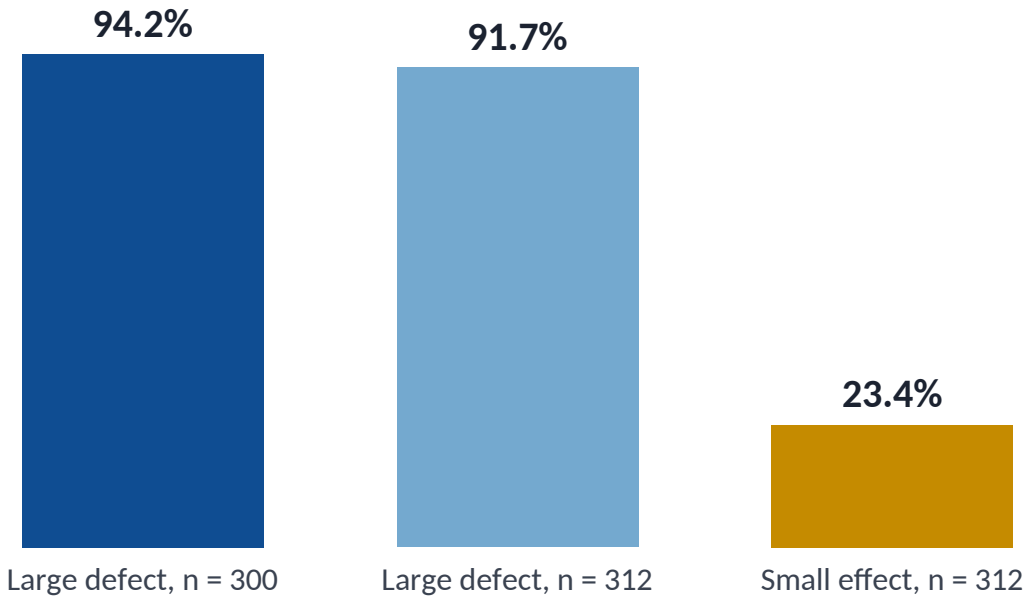
frozen prospective Δrel on 312 resolved outcomes [−0.0054, +0.0121]; every tier's interval includes zero

SKIP holds

under maturity and cohort splits, and when all 74 pending outcomes are assigned either way.

The gate still detects real defects

FULL-PROTOCOL DETECTION · 1,000 SIMULATED RUNS



Large defect: f(p) = logistic(0.15 + 0.85 logit p);
stage-one gain +0.0243 [+0.0121, +0.0379] at 318 events.
Small effect: +0.007, matching the retrospective model signal.

MINIMUM DETECTABLE EFFECTS · 80% POWER

≈ 0.015	reliability, 111 later events
≈ 0.010	reliability, 312 prospective outcomes
≈ 0.052 nats	log score, 111 later events

SKIP means unresolved, not zero: the +0.007 model signal sits below both reliability thresholds. The 0.093-nat later loss exceeds the 0.052-nat threshold and its CI excludes zero, so score deterioration is established.

DISCUSSION

Selective, not negative

Forecasts	Retrospective evidence	Later & frozen evidence	Decision
Kalshi 183 markets	Δrel +0.0387 [+0.0198, +0.0591]; utility checks pass	Δrel +0.0312 on 67 later events	RECALIBRATE
Polymarket · Metaculus 633 and 389 balanced markets	No detected horizon trend (p = .412, .351)	Recalibration within ±0.02 ECE of a no-op at every horizon	LEAVE ALONE
LLM ensembles 9 models in 3 tiers	Pointwise gains, but Holm p = .1494 for every tier	Later Δlog −0.093; every prospective interval includes zero	SKIP

Why both endpoints. A flexible map can shrink binned reliability error while assigning very low probability to realized outcomes. Better behavior on one metric does not waive the other gates.

For temporal foundation models. Post-hoc calibration is its own deployment decision. Balanced comparison, cross-fitting, leave-one-out checks, and frozen evaluation make a portable gate for it.

LIMITATIONS

What the evidence does and does not show

FOUR CAVEATS

Panel selection
The balanced panel (42.7% of markets) favors longer-lived, more liquid markets: median volume $487k vs $218k.

Sparse long horizons
Only 8 of 21 six-month prospective outcomes have resolved; most long-horizon evidence is three-month.

Knowledge in weights
As-of-date retrieval excludes future documents, not what the models learned; frozen prediction removes post-resolution querying.

Operational tolerances
Margins are application-specific, and small reliability effects stay below these panels' detection thresholds.

THE DECISION IS STABLE

- Log-score tolerances from 0.04 to 0.16 nats
- BCa and percentile intervals
- Maturity and cohort splits of the prospective set
- All 74 pending outcomes assigned either way
- ECE estimator: 15 bins, equal-mass, debiased, kernel

Model maps stay SKIP; the Polymarket pattern holds across estimators.

TAKEAWAY

Recalibrate only what survives time

1 Multiplicity

Pointwise intervals cannot override a familywise screen.

Holm p = .1494

for every LLM tier, so the screen returns SKIP

2 Utility

Reliability gains must not cost log score or decisions.

−0.093 nats

later-event loss kept out of deployment

3 Time

Confirm on later events and on frozen predictions.

+0.0312

Kalshi's gain on 67 later events: RECALIBRATE

Scope: Polymarket, Metaculus, Kalshi, and nine LLM forecasters; tolerances are operational.

Release: pinned environment, deterministic scripts, freeze timestamps, snapshots, and prompts.

Thank you — questions welcome

Paper and reviews on OpenReview

openreview.net/forum?id=emVhXV7PHf