Fluency Is Not Foresight

A Contamination-Proof Calibration Audit of LLM Forecasting

Han Kim · IOV Labs (아이오브연구소)

2026

Abstract. A language model will readily attach a probability to a future event, and it will sound like a forecaster doing so. We ask whether that number carries information. The test must be contamination-proof: we score models only on events that resolve **after** their training cutoff, where there is no answer to retrieve and a forecast must be reasoned out. Across a balanced 48-question battery of resolved world events (2024 to 2026) and the 16 races of the 2026 Korean local election, four frontier models (GPT-4o-mini, GPT-4o, Claude Haiku 4.5, Claude Sonnet 4.6) give probability forecasts that we score with the Brier rule, a reliability diagram, and an overconfidence index. Three findings. First, post-cutoff forecasts are **barely better than a coin and overconfident**: pooled Brier 0.296, worse than the 0.25 of always saying fifty percent, at 54% accuracy. Second, **remembering is not forecasting**: the same questions scored as retrieval for a model whose cutoff postdates the event yield a near-perfect Brier of 0.026, an order of magnitude better than the same model forecasting, which both validates the scoring and shows that only post-cutoff items measure foresight at all. Third, on a real election that postdates every model, a simple pre-registered statistical model (Brier 0.100) **beats every LLM** (best LLM 0.156), though handing the models the same polls closes much of the gap. We read forecasting fluency as a stylistic artifact rather than a capability, and close on the epistemics of a benchmark that scale cannot fake.

Keywords: forecasting · calibration · Brier score · contamination · knowledge cutoff · proper scoring rules · single-event probability · reproducibility

1 Introduction

A static benchmark can be memorized; a future event cannot. This single asymmetry is the spine of the present study. When a model is asked for the probability of something that has not yet happened relative to its training data, it cannot look the answer up. Whatever number it returns is a forecast, produced by whatever the model does when it reasons about an uncertain world. Forecasting is therefore the rare evaluation that scale and memorization cannot saturate [1], [2], and it is the one we use to ask a blunt question: when a large language model puts a probability on the future, is the number worth anything?

The question matters because the fluency is so convincing. A frontier model will discuss a race, weigh considerations, and produce a confident "about 70 percent" in well-formed prose. The prose is not evidence of skill. Our result is that, on events the model cannot have seen, the confident number is close to noise, and a small purpose-built statistical model out-forecasts every model we test on a real election.

1.1 The single-event problem

Underlying every probabilistic claim here is a commitment worth stating plainly. When a model says the Democratic candidate wins 서울 "with probability 73 percent," there is no long run in which 서울 votes a hundred times; the election happens once. The number is not a frequency but a degree of belief, and its only honest test is **calibration across many such claims**: if the events called seventy percent happen about seventy percent of the time, the probabilities carry information [3], [4]. A single race can never validate a probability; a battery begins to, and a track record across cycles eventually does. We therefore score many simultaneous probabilities with a proper scoring rule rather than celebrating any single call.

2 Background

LLM forecasting is an active area. Halawi et al. [2] build a retrieval-and-reasoning system that approaches human-crowd performance on a large question set; ForecastBench [1] maintains a contamination-resistant, continuously refreshed benchmark of unresolved questions; recent work compares frontier models against

human superforecasters and finds the humans still ahead but by a shrinking margin [5], trains models to forecast better from self-generated reasoning [6], and reads probabilities from token log-probabilities [7]. Our contribution is narrower and more diagnostic. We do not build a forecaster; we **audit** the raw, zero-tooling probability a model emits, and we use the **knowledge cutoff itself** as an experimental control, separating what a model retrieves from what it forecasts on the identical questions. The positive control this affords, a late-cutoff model scoring near-perfect on resolved past events, is what lets us claim the scoring is sound.

3 Method

3.1 Tasks

Election (head to head). The sixteen metropolitan-executive races of the 2026-06-03 Korean local election. Because the election resolves after every tested model's cutoff, it is a clean forecast for all of them and contains no leakage. We attach to each race the realized winner, the final pre-blackout poll two-way share, and the win probability from IOV's pre-registered poll-and-fundamentals model [8] (Brier 0.100 on these races). Two elicitation conditions: KNOWLEDGE (the model reasons from its own training) and POLLS (the same, with the final poll handed over).

General battery. A balanced set of 48 binary world events resolving 2024 to 2026, 26 that happened and 22 that did not, spanning elections, sport, macroeconomics, technology, awards, and geopolitics, each with a verified outcome and resolution month. Balance matters: with a skewed base rate, hedging toward fifty percent masquerades as calibration, so we deliberately include verified non-events (a team that lost a final, an invasion that did not occur).

3.2 Models, elicitation, scoring

Four models across two families: `gpt-4o-mini` and `gpt-4o` at a 2023-10 cutoff, and `claude-haiku-4-5` and `claude-sonnet-4-6` at a 2025 cutoff. The early pair forecasts essentially the entire battery; the late pair postdates most of it and therefore **retrieves** it, which is the contrast we exploit. Each item is tagged forecast or retrieval per model from its cutoff. We force a single integer probability, parse it robustly, and take the mean of three low-temperature samples per item (salted and cached for reproducibility). We report the Brier score and its Murphy decomposition [9], accuracy at the 0.5 threshold, an overconfidence index (mean confidence minus accuracy), and a ten-bin reliability diagram. Baselines: always-0.5 (Brier 0.25) and the category base rate. A published human-superforecaster Brier near 0.08 to 0.09 [4], [5] is cited as an anchor, not reproduced.

3.3 Confound controls

(1) **Leakage gating**: forecast claims use only post-cutoff items; the election is post-cutoff for every model. (2) **Positive control**: the late-cutoff model must score near-perfect on resolved past events, or the scoring is broken and nothing is claimed. (3) **Balanced battery** so hedging cannot pass as calibration. (4) **Neutral, future-tense wording** with no outcome cue. (5) **Three-sample averaging** at low temperature to damp sampling noise. (6) **Verified ground truth** for every outcome.

4 Results

4.1 Post-cutoff forecasts are barely better than a coin

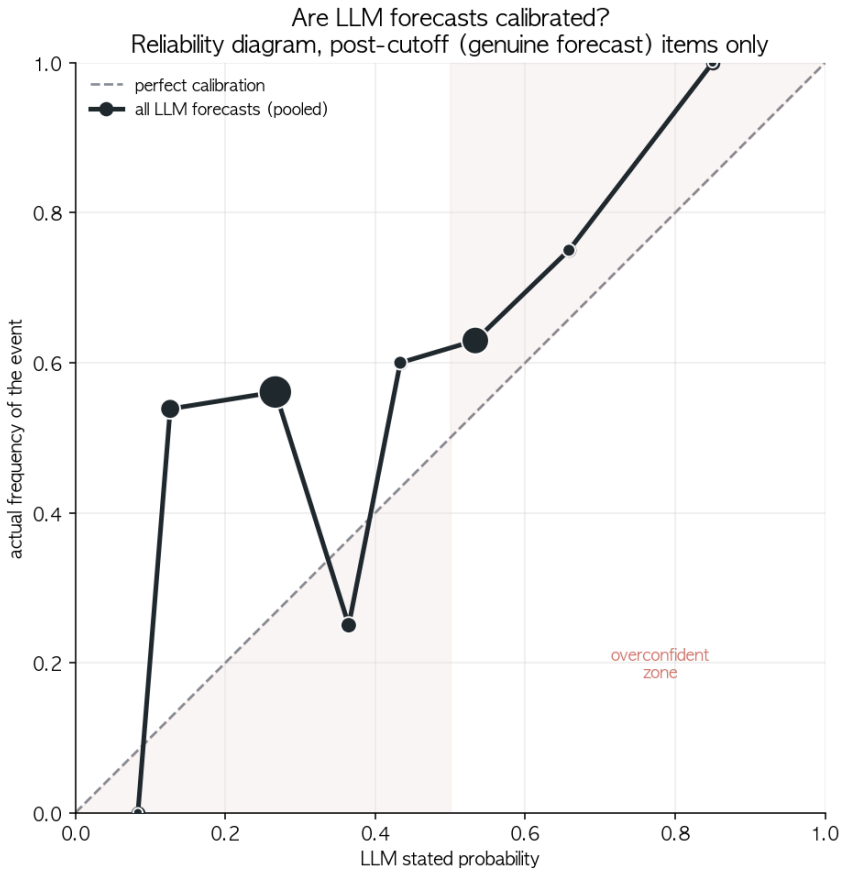

Figure 1: Reliability diagram over post-cutoff (genuine forecast) items. A calibrated forecaster lies on the diagonal; the pooled LLM curve wanders off it, and the stated probabilities barely track the realized frequencies.

Pooling 100 genuine post-cutoff forecasts over the balanced battery, the models score a Brier of **0.296**, worse than the 0.25 of always saying fifty percent, with accuracy **0.54** (a hair above a coin) and a positive overconfidence index of **+0.14**. The two early-cutoff models, the clean forecasters here, land at 0.286 (GPT-4o-mini) and 0.299 (GPT-4o). The probabilities are not merely imprecise; on a balanced set they fail to separate what happened from what did not.

4.2 Remembering is not forecasting

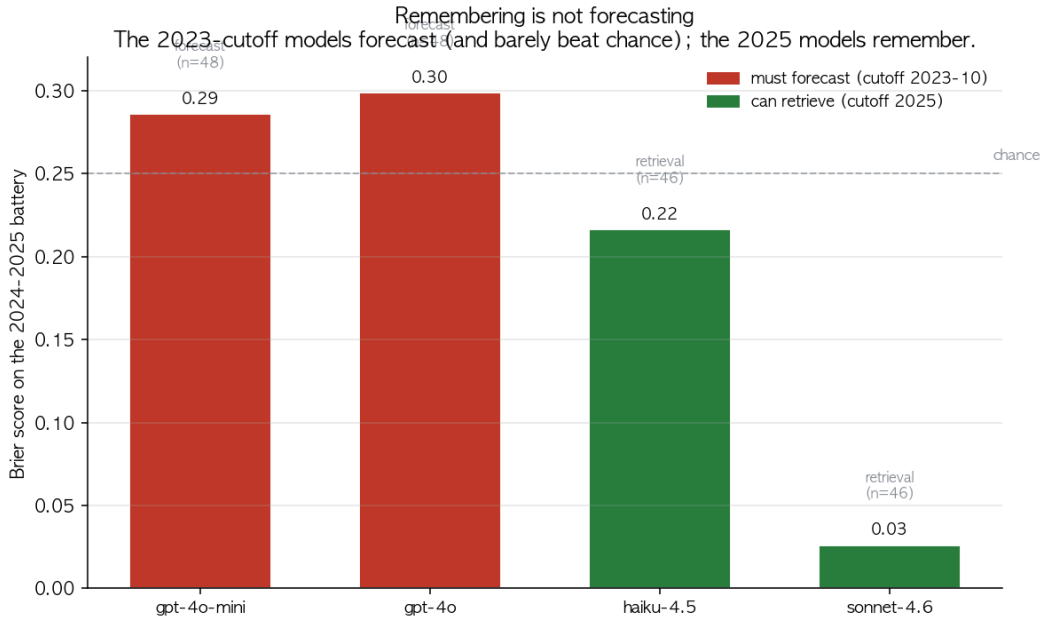

Figure 2: The natural experiment. On the identical 2024-2025 battery, the 2023-cutoff models must forecast and barely beat chance, while the 2025-cutoff models retrieve. Claude Sonnet 4.6 retrieves at Brier 0.026.

The same questions are a memory test for a model that postdates them and a forecasting test for one that predates them. Claude Sonnet 4.6 **remembers** the 2024 events at a near-perfect Brier of **0.026** (96% accuracy); on the handful of items past its own cutoff it falls toward chance. This order-of-magnitude gap does double duty. It is the positive control: a scoring pipeline that awards near-zero Brier to a model that knows the answers is working. And it is a finding: the fluent, confident probability and the well-calibrated one come from different places, retrieval and inference, and only the post-cutoff number measures forecasting. A smaller model, Haiku 4.5, is poorly calibrated **even on retrieval** (Brier 0.216), a reminder that remembering, too, is unevenly distributed across model scale.

4.3 A simple statistical model beats every LLM on a real election

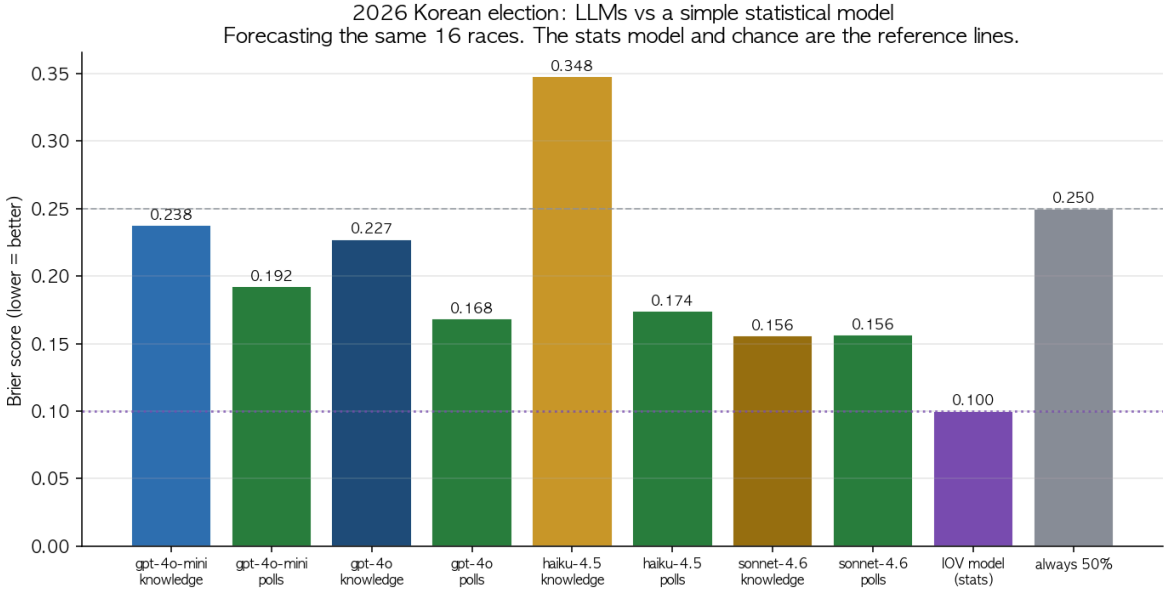

Figure 3: Forecasting the same sixteen races. IOV's pre-registered statistical model (purple) scores Brier 0.100; the best LLM from its own knowledge is Sonnet 4.6 at 0.156. Handing over the polls (green) helps the weaker models toward, but not past, the statistical model.

On the election, which is leakage-free for everyone, IOV's poll-and-fundamentals model scores **0.100**. The best LLM, Sonnet 4.6 reasoning from its own knowledge, scores 0.156; GPT-4o 0.227, GPT-4o-mini 0.238, and Haiku 4.5 a chance-beating-the-wrong-way **0.348**, having called the map backwards (25% winner accuracy) while sounding sure. Two readings follow. The frontier is not absurd: a recent large model with rich training-time context about the race lands within striking distance of a bespoke model. But the bespoke model still wins, and the cheap, confident model is actively misleading.

4.4 LLMs are weak at sourcing signal, better at using it

Handing the models the same final polls the statistical model used moves the weak forecasters sharply: Haiku 0.348 to 0.174, GPT-4o 0.227 to 0.168, GPT-4o-mini 0.238 to 0.192. Sonnet 4.6, already near the polls in its own estimate, barely moves (0.156). The models cannot reliably find the relevant signal in their own memory, but they can use it competently once it is placed in front of them. The deficit audited here is one of **sourcing and weighting evidence**, not of arithmetic.

5 Discussion

5.1 Why scale will not simply fix this

It is tempting to read the Sonnet-versus-Haiku gap as "bigger models forecast better, so wait." The natural experiment cautions against it. Most of what separates the late models from the early ones on the battery is **retrieval**, not foresight: Sonnet's 0.026 is memory. On the genuinely post-cutoff election the same model sits at 0.156, well short of a simple model that did nothing but aggregate polls and swing fundamentals. Scaling improves how much of the recent past a model has absorbed and how well it uses handed-over evidence. It does not obviously confer the thing a forecaster needs, a calibrated sense of its own ignorance about what has not happened yet.

5.2 Memory dressed as prediction

An LLM's competence about the social world is **memory**, not **measurement**: it interpolates a training-time distribution and does not observe the present. For a stable, backward-looking quantity this can suffice and even excel, as the retrieval scores show. For a contested, forward-looking one it returns a stale prior in the grammar of a forecast. The danger is not that the model is wrong, all models are wrong, but that it is **confidently** wrong in fluent prose, the overconfidence index made of words. A user cannot tell, from the answer alone, whether they are reading a recollection or a guess.

5.3 The one benchmark scale cannot fake

Because a future event cannot be in any training set, post-cutoff forecasting resists the contamination that erodes static benchmarks. This is also its discipline: the only way to keep the benchmark honest is to keep asking about the future, scoring after resolution, and reporting the calibration, not the anecdote. Our election entry is one clean, leakage-free case; it cannot validate a probability by itself, and we do not claim it does. The method, not the single number, is the contribution.

6 Limitations

Pilot scale. Forty-eight battery questions, sixteen races, four models, two families; the post-cutoff forecast sample is dominated by the two early models, by construction. **Salient questions.** The battery leans on memorable events; an obscurer set could move the numbers either way. **Elicitation.** Verbalized probabilities at low temperature, not log-probabilities; format effects are possible though mitigated by averaging. **Single election.** One election is one event. All negative and null results, including the model that beat chance backwards, are reported.

7 Conclusion

Asked to forecast what it cannot have seen, a frontier language model produces a number that sounds like a forecast and behaves like noise: pooled Brier worse than a coin, overconfident, and beaten on a real election by a model that does nothing clever. The same model, asked about the settled past, is near-perfect, which is exactly the point. Fluency about the future is not foresight; it is memory of the past, rendered in the

confident grammar of prediction. The honest use of these systems is to hand them the evidence and audit the calibration, not to mistake the prose for a probability.

Code, data, the pre-registered design, and the cached forecasts: https://github.com/hankimis/llm-forecast. Companion: IOV's seoul-2026-forecast. MIT licensed.

References

[1] E. Karger and others, "ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities." 2024.

[2] D. Halawi, F. Zhang, C. Yueh-Han, and J. Steinhardt, "Approaching Human-Level Forecasting with Language Models." 2024.

[3] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[4] P. E. Tetlock and D. Gardner, *Superforecasting: The Art and Science of Prediction*. Crown, 2015.

[5] J. Lu and others, "Evaluating LLMs on Real-World Forecasting Against Human Superforecasters." 2025.

[6] B. Turtel and others, "LLMs Can Teach Themselves to Better Predict the Future." 2025.

[7] T. Bei and others, "Leveraging Log Probabilities in Language Models to Forecast Future Events." 2025.

[8] H. Kim, "A Poll and Fundamentals Forecast of the 2026 Korean Local Elections." 2026.

[9] A. H. Murphy, "A new vector partition of the probability score," *Journal of Applied Meteorology*, vol. 12, no. 4, pp. 595–600, 1973.