An Agent on Every Open Question

Continuous agentic forecasting with an auditable update log

Drazta
apriandito@drazta.com

Abstract

Forecasting tournaments identified a procedure that outperforms unaided expert judgement: anchor on a reference class, update incrementally on identifiable evidence, and score against outcomes. The procedure is not scarce; the trained human attention it consumes is. We describe a system that executes it as a scheduled agent, so that the marginal cost of a maintained forecast falls far enough to place one on every open question an organisation carries rather than on the few that justify an analyst. The system's substantive contribution is not the reasoning, which we take unaltered from the literature, but the ledger: an append-only record in which the decision *not* to move a forecast is a first-class dated event with a stated cause. That record makes two things possible which are otherwise unavailable, namely an audit trail from any probability back to the facts that produced it, and a retrospective test of whether scheduled updating improves accuracy at all. We formalise the update rule, the scoring, and a replay operator over dated evidence; we register two evaluations in advance, with success thresholds and a power analysis; and we state the conditions under which we expect the approach to fail. No accuracy results are reported. None exist yet.

Keywords: probabilistic forecasting; calibration; LLM agents; proper scoring rules; pre-registration

1 Introduction

Let a decision-relevant proposition be a binary event $Y \in \{0,1\}$ with resolution date T and criteria fixed in advance. In most organisations, beliefs about Y are encoded as ordinal labels drawn from a small set such as $\{\text{GREEN}, \text{AMBER}, \text{RED}\}$. Three properties follow from that encoding alone.

No scoring. An ordinal label admits no proper scoring rule (Gneiting and Raftery, 2007): there is no loss function mapping AMBER and $Y = 1$ to a penalty, so no label can be wrong.

No aggregation. Labels do not compose. For n programmes at AMBER the probability that at least one fails is undefined, whereas with probabilities p_i and independence it is $1 - \prod_i (1 - p_i)$.

No attribution. Without (p, Y) pairs, per-forecaster skill is unidentifiable, and organisational trust defaults to seniority.

Asking staff to state probabilities addresses the first only partially: an unscored probability is a label with more digits. What is missing is not the number but the loop that closes it.

The method that closes it is known (Tetlock, 2005; Tetlock and Gardner, 2015; Mellers et al., 2014). It stayed rare because it consumed analyst hours per question, so it was rationed to the highest-stakes questions. Our claims concern capacity, persistence, and access rather than reasoning quality, and are stated narrowly in Section 4.

2 Method

We use the log-odds parameterisation throughout, since update arithmetic is additive in that space:

$$\lambda_t \;=\; \operatorname{logit}(p_t) \;=\; \log\frac{p_t}{1-p_t}, \qquad p_t = \sigma(\lambda_t) = \frac{1}{1+e^{-\lambda_t}}. \tag{1}$$

2.1 Anchor

Before retrieving case-specific coverage, the agent constructs a reference class $\mathscr{C}$ of resolved comparable events and estimates its rate:

$$\hat{\pi} = \frac{1}{|\mathscr{C}|}\sum_{j\in\mathscr{C}} y_j, \qquad \lambda_0 = \operatorname{logit}\big(\hat{\pi}\big). \tag{2}$$

$\mathscr{C}$, $|\mathscr{C}|$ and $\hat{\pi}$ are written as separate ledger fields so that a reader may reject the reference class without rejecting the case reasoning built on it. When $|\mathscr{C}| < 10$ the question is flagged low-reference and the anchor is shrunk toward the uninformative prior,

$$\lambda_0' = \frac{|\mathscr{C}|}{|\mathscr{C}| + \kappa}\,\lambda_0, \tag{3}$$

with κ fixed in advance. Ordering is enforced procedurally: retrieval is blocked until λ_0 is committed. An anchor chosen after reading the news is not an outside view.

2.2 Update

Let $\mathscr{E}_t$ be events observed since the previous entry, $m(e) \in \{0,1\}$ a materiality indicator, and $\delta(e) \in \mathbb{R}$ the log-odds increment attributed to e. The rule is

$$\lambda_t = \lambda_{t-1} + \gamma_t + \sum_{e\in\mathscr{E}_t} m(e)\,\delta(e), \tag{4}$$

where the case $\sum_e m(e) = 0$ is a *hold* and still appends a dated entry stating that no material evidence arrived. Figure 1 gives the control flow.

The rule exists to suppress a specific failure. A language model instructed to update a forecast emits a slightly different number on each call, because emitting a different number is what the instruction appears to request. The resulting jitter is uncorrelated with information, inflates the reliability term of (7), and conceals genuine movement.

Deadline term. For questions requiring occurrence before T, treat residual occurrence as a Poisson process with rate r estimated from $\mathscr{C}$, so that

$$p_t = 1 - e^{-r(T-t)}, \qquad \gamma_t = \operatorname{logit}\big(1-e^{-r(T-t)}\big) - \operatorname{logit}\big(1-e^{-r(T-t')}\big), \tag{5}$$

for previous entry time t'. The estimate therefore decays toward the status-quo outcome as the window closes, explicitly rather than by model intuition.

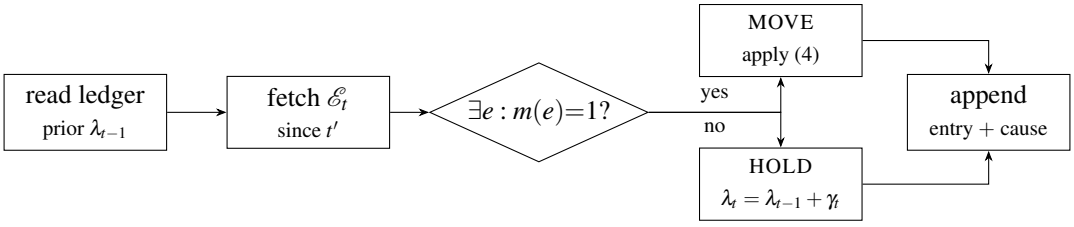

Figure 1. Control flow of one scheduled review. Both branches terminate in a written entry; a hold is a record, not a skipped step.

Algorithm 1 Scheduled review of question q

$(\lambda_{t-1}, t') \leftarrow \textsc{LastEntry}(q)$ ▷ prior comes from the ledger, not from a fresh guess
$\mathscr{E}_t \leftarrow \textsc{Events}(q, \text{after } t')$ ▷ deduplicated by content hash
$\gamma_t \leftarrow$ deadline term from (5)
$S \leftarrow \{e \in \mathscr{E}_t : m(e) = 1\}$
if $S = \emptyset$ **then**
    $\lambda_t \leftarrow \lambda_{t-1} + \gamma_t$; $\textsc{Append}(q, \textsc{hold}, \lambda_t, \text{cause})$
else
    $\lambda_t \leftarrow \lambda_{t-1} + \gamma_t + \sum_{e \in S} \delta(e)$; $\textsc{Append}(q, \textsc{move}, \lambda_t, S)$
end if
return $\sigma(\lambda_t)$

2.3 Settle

At resolution the forecast is scored with the Brier score (Brier, 1950),

$$\mathscr{B} = (p - Y)^2, \qquad \bar{\mathscr{B}} = \frac{1}{n} \sum_{i=1}^{n} (p_i - Y_i)^2. \tag{6}$$

We additionally report the Murphy decomposition over K probability bands (Murphy, 1973),

$$\bar{\mathscr{B}} = \underbrace{\frac{1}{n} \sum_{k=1}^{K} n_k (\bar{p}_k - \bar{y}_k)^2}_{\text{reliability}} - \underbrace{\frac{1}{n} \sum_{k=1}^{K} n_k (\bar{y}_k - \bar{y})^2}_{\text{resolution}} + \underbrace{\bar{y}(1 - \bar{y})}_{\text{uncertainty}}, \tag{7}$$

because the aggregate alone is not diagnostic: a system can post a respectable $\bar{\mathscr{B}}$ through low reliability error while carrying almost no resolution, which is the signature of a forecaster that hedges toward the base rate and never commits.

3 System

3.1 Ledger

One append-only plain-text file per question, version controlled. Each history entry carries a timestamp, p_t, a type in $\{\textsc{anchor}, \textsc{move}, \textsc{hold}\}$, and either the event identifier justifying δ or the cause of the hold. The format is chosen so that a reader can reconstruct a question's full reasoning history without tooling, and so the corpus is diffable with ordinary version control. Figure 2 shows the resulting trajectory shape.

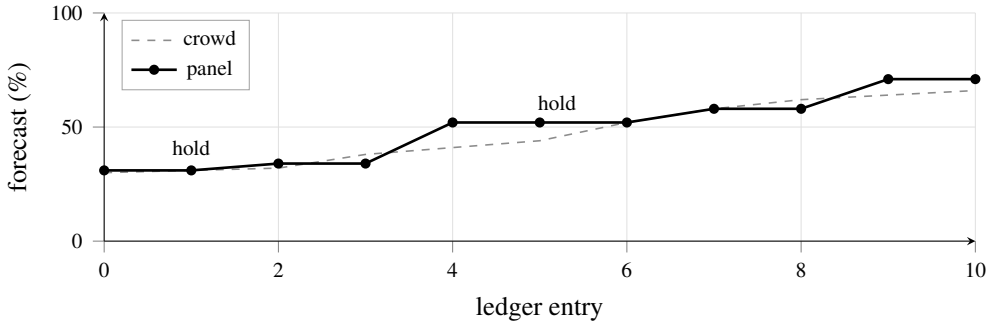

Figure 2. Illustrative trajectory. The series is piecewise constant between named events rather than continuously drifting; flat segments are holds carrying recorded causes. No result is claimed by this figure.

3.2 Pooling and calibration

Where K samples are drawn per review they are pooled in log-odds (Satopää et al., 2014),

$$\bar{\lambda} = \frac{1}{K}\sum_{k=1}^{K}\lambda_k, \qquad \lambda' = a\,\bar{\lambda}, \quad a \geq 1, \tag{8}$$

rather than by averaging probabilities, which contracts toward 0.5 and discards resolution. The extremisation coefficient a must be fitted on held-out resolved questions, never assumed. Calibration correction, when fitted, uses single-parameter Platt scaling (Platt, 1999),

$$\hat{p} = \sigma(\alpha\lambda), \qquad \hat{\alpha} = \arg\max_{\alpha}\sum_{i}\Big[Y_i \log\sigma(\alpha\lambda_i) + (1-Y_i)\log\big(1-\sigma(\alpha\lambda_i)\big)\Big]. \tag{9}$$

One parameter is deliberate. Isotonic regression is more flexible and will overfit at the sample sizes available for years.

3.3 Replay operator

The weakest component is $m(\cdot)$, which currently rests on a retrieval performed at review time plus the model's own judgement of novelty, and is unmeasured. The specified evidence layer maintains a dated event table per question with first-observation date, source, claim strength in $\{\text{REPORTED}, \text{CONFIRMED}\}$, and a content hash, so that n outlets covering one filing resolve to one event. It admits the replay operator

$$F(q,d) \;=\; \textsc{Forecast}\big(q \,\big|\, \{e : \text{date}(e) \leq d\}\big), \tag{10}$$

which reconstructs the maintained forecast as of any past date. Without (10) the hypothesis of Section 5.2 cannot be tested retrospectively at all; it can only be observed forward, one resolution horizon at a time. This component is specified but not built, and is sequenced after the first evaluation.

3.4 Portfolio coherence

With many live questions, some are logically related. For nested windows $A \subseteq B$ the forecasts must satisfy $p_A \leq p_B$. Writing $\mathscr{K}$ for the convex set induced by all detected constraints, the coherent projection is

$$p^{\star} = \arg\min_{u \in \mathscr{K}} \|u - p\|_2^2. \tag{11}$$

Because the realised outcome vector lies in $\mathscr{K}$, projection onto a convex set containing it cannot increase mean squared error. The guarantee is conditional on the constraints being correct, and they are detected by a model reading free text; Section 6 treats this as a risk rather than a feature.

4 Claims

(i) The marginal cost of a maintained forecast falls by orders of magnitude, which changes which questions are worth forecasting rather than making the same product cheaper.

(ii) A scheduled process sustains update discipline indefinitely. Update frequency correlates with accuracy in tournament data (Mellers et al., 2014), and is the first discipline human panels lose under load.

(iii) An agent operating inside the perimeter can forecast questions that external experts never see.

None is a claim about reasoning quality. A fourth property is methodological: because non-updates are recorded, the system produces a dataset in which the decision not to move is observable, which is what makes Section 5.2 testable.

5 Evaluation

Any forecasting system evaluated after the fact can be made to look good: exclude questions that went badly as unrepresentative, select the baseline that was beaten, report the flattering metric, extend the sample until the estimate settles where it is wanted. No step requires dishonesty. The defence is to fix every choice before observing results.

5.1 Protocol 01: accuracy

Question set	$n = 150$ binary questions resolving strictly after the model's training cutoff
Baseline A	constant 0.5, giving $\bar{\mathscr{B}} = 0.25$
Baseline B	community forecast at question open
Primary metric	$\bar{\mathscr{B}}$ over all 150, no exclusions, plus decomposition (7)
Uncertainty	paired bootstrap, 10^4 resamples, resampled over event clusters
Success	beat Baseline B, 90% interval excluding zero
Publication	full results and question list within 14 days of last resolution, either way

Contamination. Questions must resolve after the training cutoff. A model that met the outcome in pre-training is measured on recall, and returns an excellent, meaningless score.

Clustering. Questions sharing an underlying event are not independent. With mean cluster size $\bar{n}_c$ and intra-cluster correlation ρ, the design effect is

$$\mathrm{DEFF} = 1 + (\bar{n}_c - 1)\rho, \qquad n_{\mathrm{eff}} = \frac{n}{\mathrm{DEFF}}, \tag{12}$$

so at $\bar{n}_c = 3$, $\rho = 0.5$ the effective sample is halved.

Power. With paired per-question Brier differences of standard deviation $s_d \approx 0.12$,

$$\mathrm{SE}(\bar{d}) = \frac{s_d}{\sqrt{n_{\mathrm{eff}}}} \approx \frac{0.12}{\sqrt{75}} \approx 0.014, \qquad d_{\min} = z_{0.95}\,\mathrm{SE}(\bar{d}) \approx 0.023. \tag{13}$$

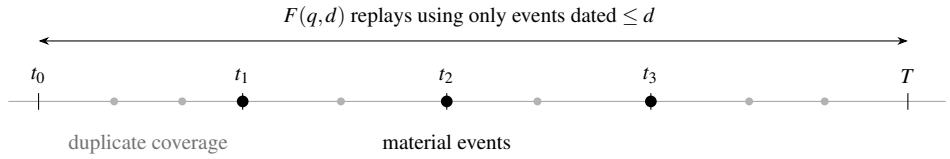

Figure 3. Dated, deduplicated evidence separates material events from repeated coverage and permits reconstruction of the forecast as of any past date.

Differences below roughly two Brier points are therefore undetectable at this sample size, and will not be reported as if they were.

5.2 Protocol 02: update policy

Registered separately and blocked until Protocol 01 closes. It asks whether scheduled updating beats freezing the day-one forecast:

$$\Delta = \frac{1}{n}\sum_{i=1}^{n}\left[\left(p_i^{\text{frozen}} - Y_i\right)^2 - \left(p_i^{\text{final}} - Y_i\right)^2\right], \tag{14}$$

with $\Delta > 0$ favouring maintenance and $p_i^{\text{frozen}} = F(q_i, d_0)$ from (10). Secondary analysis attributes Δ to individual entries and estimates the value of the hold rule by comparison against a variant permitted to move at every review. This test requires a log in which non-updates appear, which is a property of the ledger rather than of the model.

6 Failure modes

F1. The baseline may not be cleared. The structure presumes the panel is reasonably calibrated to begin with. If Protocol 01 misses, the response is publication and reconsideration, not the addition of components until something passes.

F2. Scheduled updating may add nothing. The frequency finding comes from humans whose updates reflected genuine information gathering; an agent re-reading coverage daily may be tracking publication volume. $\Delta \leq 0$ in (14) would invalidate the central operational claim, and we have committed to publishing that outcome.

F3. Transfer is assumed, not shown. Validation runs on public questions; the commercial proposition concerns internal ones. Calibration measured on the former is assumed to hold for the latter, an assumption close to untestable in advance.

F4. Internal questions resolve too slowly to calibrate on. Fitting $\hat{\alpha}$ in (9) requires resolutions that internal questions supply over years, forcing the correction back onto public data and into F3.

F5. Reference classes may be absent where they matter. Equation (2) presumes a computable $\hat{\pi}$. Semantic similarity between question texts is not evidence of exchangeability, and for highly specific internal questions no clean $\mathscr{C}$ may exist.

F6. Base rates are jurisdiction-bound. Permitting, delivery and regulatory timelines vary sharply by country; classes drawn from English-language archives may mislead elsewhere.

F7. Samples are not independent. Averaging K runs of one model reduces sampling noise but not systematic bias; where the model is confidently wrong, (8) concentrates the error and

extremisation amplifies it.

F8. Fitted layers share one statistical budget. Calibration, pooling weights and materiality thresholds fitted on the same few hundred resolutions constitute overfitting in a tidy wrapper. A held-out set must be reserved at the outset.

F9. Coherence detection is fragile. The guarantee under (11) holds only if $\mathscr{K}$ is correct. Constraints are extracted by a model from free text, and the fraction of logically related pairs in a realistic portfolio is small enough that expected gain may not justify the risk.

F10. The commercial premise is unproven. Whether organisations pay for calibrated probabilities is independent of whether the probabilities are good.

7 What is not claimed

The method is not our invention; it is taken from the tournament literature and deliberately left unaltered. We do not claim the system outperforms human superforecasters, nor that it approaches liquid prediction markets. Inspectability implies arguability, not correctness. That more updates are better is a hypothesis with a registered test attached, not a result. No figure in our public materials reflects measured performance until Protocol 01 closes.

8 Sequence

1. Backtest harness: contamination-safe selection, clustered bootstrap, single-variable comparison.
2. Protocol 01 baseline run and publication.
3. Evidence layer with dated event table and replay operator (10).
4. Protocol 02 on update policy.

Steps 3 and 4 are contingent on step 2. If the baseline is not cleared, the sequence halts and the result is published.

References

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.

Gneiting, T., and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Mellers, B., et al. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5), 1106–1115.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4), 595–600.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*.

Satopää, V. A., et al. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2), 344–356.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press.

Tetlock, P. E., and Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. Crown.