TOWARDS A GENERAL AI PREDICTION ARENA

Anonymous authors
Paper under double-blind review

ABSTRACT

Prediction, the ability to gather evidence and reason under uncertainty before an outcome is known, is a cornerstone of human intelligence and a critical frontier for artificial intelligence. Evaluating this capability requires testing models against true future events to prevent data contamination, and modern benchmarks employ live, dynamic arenas to achieve this. However, these systems suffer from three critical limitations. First, temporal asymmetry leads to unfair comparisons between early forecasts made under severe uncertainty and late predictions that benefit from accumulated evidence. Second, their question pools rely heavily on popular prediction markets while neglecting specialized, long-tail domains. Third, models cannot be scored until real-world events finally resolve, stalling iterative development. We introduce Echo, a live and general AI prediction arena that addresses these challenges. Echo implements a multi-point aligned Elo rating system that schedules predictions across an event's lifecycle and compares models only at the exact same prediction point. We propose a data acquisition pipeline combining prediction markets, automated trending question synthesis, and domain expert annotation for broad coverage. Echo further introduces narrative rubric evaluation, where rubrics are algorithmically evolved to score reasoning trajectories independently of the final outcome, providing immediate feedback for model iteration. In robustness experiments, even when half the participants miss prediction days, time-weighted Brier rankings have up to 2.18× the rank standard deviation of aligned Elo. In a revised seven-model rubric evaluation, scores computed without the recorded outcomes correlate with Elo rankings at $\rho = 0.857$.

1 INTRODUCTION

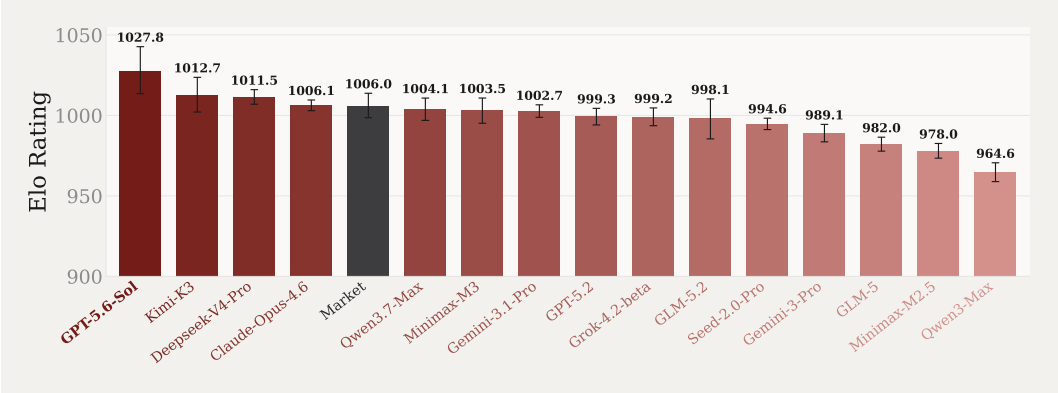

Figure 1: Echo leaderboard snapshot on 23 September 2026 with 95% confidence intervals.

General predictive reasoning is the capacity to synthesize heterogeneous information into calibrated prospective judgments across a wide range of domains. It is a fundamental cognitive attribute and a cornerstone of decision-making (Box et al., 2015; Lynch, 2006; Cowles 3rd, 1933; Jin et al., 2021; Yan et al., 2023; Hua et al., 2023). Accurate anticipation of future states is valuable in high-stakes environments, from the mitigation of public crises (Rotaru et al., 2022) to the calibration of financial risk (Xiao et al., 2024). Recent agentic systems such as DeepResearch (Team et al., 2025) and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

terminal operation (Merrill et al., 2026) have enabled AI forecasting agents that autonomously gather evidence and produce probability estimates. However, real-world predictions impose strict constraints on accuracy and reliability (Ye et al., 2024; Halawi et al., 2024; Wang et al., 2024). Without rigorous empirical validation, predictive model outputs cannot be trusted for real-world applications. Developing systematic evaluation protocols for general forecasting is therefore essential to ensure that AI predictive systems deliver calibrated and actionable foresight.

Evaluating forecasting performance accurately is highly complex. To ensure authenticity and prevent data memorization, the community has adopted live, dynamic benchmarks. These systems continuously test models on ongoing, real-world events before their outcomes occur (Karger et al., 2025; Zeng et al., 2025; Liu et al., 2026; Yang et al., 2025). Nevertheless, current frameworks remain constrained by three structural limitations. First, they lack temporal alignment. An event's informational landscape evolves continuously. Predictions made at different lead times assess distinct capabilities (Tao et al., 2025b). Early forecasts test long-horizon causal reasoning under sparse evidence. Late-stage predictions measure the capacity to aggregate dense, arriving evidence. Without synchronizing these timestamps, leaderboards conflate distinct analytical skills with timing disparities, making cross-model comparisons unreliable. Second, existing benchmarks suffer from propositional narrowness. Current live benchmarks draw heavily on highly liquid prediction markets (Yang et al., 2025) or public periodic rankings (Zeng et al., 2025), such as weekly media charts. This biases evaluation toward high-attention topics and systematically neglects high-value, long-tail uncertainties in specialized scientific or technical domains. Finally, evaluation throughput is bottlenecked by an exclusive reliance on terminal outcomes. While measuring against ground-truth resolution is essential, depending solely on post-hoc results introduces feedback latency. Standard accuracy metrics remain incalculable until an event concludes, which often takes weeks or months. This delay severely obstructs the rapid, iterative diagnostic loops required to optimize forecasting models.

To address these methodological deficiencies, we introduce Echo, a general AI prediction arena. Echo mitigates temporal confounding via a time-aligned Elo rating (Hvattum & Arntzen, 2010) framework that compares models on the same question at the same scheduled prediction point, reducing differences in prediction timing. To expand propositional breadth, our pipeline combines market-derived questions with automated synthesis and expert annotations, broadening the pool of diverse, long-tail forecasting tasks. To resolve feedback latency, Echo uses algorithmically evolved rubrics (Hashemi et al., 2024) that *serve as a fast proxy for Elo*: they score reasoning trajectories in real time, providing correlated diagnostic feedback that enables model optimization long before ground-truth resolution. We benchmark 13 leading AI systems in controlled experiments on this framework. When half the participants miss prediction days, time-weighted Brier rankings have up to $2.18\times$ Elo's rank standard deviation. Elo also reaches the rank-error threshold earlier in our new-model simulation. In the revised seven-model rubric evaluation, scores computed without outcomes correlate with Elo at $\rho = 0.857$.

Our main contributions are as follows:

1. **A live and general prediction arena.** We introduce Echo, featuring a comprehensive data acquisition pipeline for diverse domain coverage, a multipoint time-aligned Elo rating system to reduce differences in scheduled prediction time, and algorithmically evolved narrative rubrics for real-time, process-level reasoning assessment.
2. **Extensive benchmarking and validation.** We evaluate several AI systems in controlled experiments to provide a reliable benchmark of their forecasting capabilities within a shared agent environment. Our experiments show lower ranking volatility for aligned Elo than Brier under missing and sparse predictions, and a correlation of $\rho = 0.857$ between rubric scores and Elo in the revised seven-model evaluation.
3. **Actionable insights for predictive AI.** Models score lower on probability derivation and uncertainty handling than on evidence collection. These dimensions are also more closely related to Elo rankings, suggesting that better use of evidence is an important direction for improvement.

2 Related Work

Event Prediction Benchmarks. Recent forecasting benchmarks for large language models (LLMs) generally fall into two paradigms. The first employs retrospective backtesting, simulating pre-event scenarios using historical contexts. Works such as Halawi et al. (2024), OpenEP (Guan et al., 2024), EV2 (Tao et al., 2025a) and OpenForesight (Chandak et al., 2025) evaluate real-world questions via historical news search, while Tao et al. (2025b) explicitly filters datasets for predictive inferability

using Causal Intervened Likelihood estimation. However, authentically reconstructing pre-event information environments is notoriously difficult due to the pervasive risk of training data leakage. The second paradigm evaluates predictions on unresolved future events and measures accuracy after resolution. ForecastBench (Karger et al., 2025), Future-X (Zeng et al., 2025), and Future-X-Pro (Liu et al., 2026) follow this approach. Foresight Arena (Nechepurenko & Shuvalov, 2026) proposes an on-chain protocol for evaluating forecasts on prediction markets. Echo also follows the live evaluation paradigm, with comparisons aligned at scheduled prediction points.

Evaluation By Rubrics. To evaluate the complex, multi-step reasoning capabilities of agents, recent benchmarks have increasingly shifted from final-answer metrics to process-oriented, rubric-based assessments. In the medical domain, HealthBench (Arora et al., 2025) evaluates healthcare agents using a massive repository of over 48,000 rubrics meticulously crafted by expert physicians. Similarly, Researchrubrics (Sharma et al., 2025) employs human-annotated rubrics to systematically assess the performance of deep research agents on open-ended tasks.

Arena-Style and Dynamic Evaluation. Chatbot Arena (Chiang et al., 2024) popularized pairwise LLM evaluation through crowdsourced battles fitted with a Bradley-Terry model (Bradley & Terry, 1952), accumulating over 240K anonymous human votes and producing Elo-style rankings that have become a de facto industry standard. In this work, Echo adapts the pairwise Bradley-Terry framework from conversational preference to prediction accuracy, replacing human votes with soft labels derived from Brier-score differences.

3 Inside Echo: Arena Design and Methodology

Echo consists of three components: a question collection pipeline that draws from prediction markets, trend synthesis, and expert annotation, a time-aligned Elo rating that compares models only at matched prediction points, and a narrative-wise rubric that scores reasoning traces before outcomes are known. Figure 2 shows the full pipeline.

3.1 Question Collection

We define a *prediction question* as a tuple $\mathcal{Q}_i = (c_i, \mathcal{A}_i, d_i, \rho_i)$, where c_i is a natural-language claim about a future event, $\mathcal{A}_i = \{a_1, \ldots, a_J\}$ is the set of candidate answers, d_i is the resolution deadline, and ρ_i is a public resolution rule that maps observed evidence to the correct answer. Given $\mathcal{Q}_i$, a forecasting model produces at prediction time $t < d_i$ a probability vector $\hat{\mathbf{p}}_i(t) = \big(\hat{P}(a_j \mid \mathcal{Q}_i, t)\big)_{j=1}^{|\mathcal{A}_i|}$ over the candidate answers, where $\sum_j \hat{P}(a_j \mid \mathcal{Q}_i, t) = 1$. Echo evaluates the quality of $\hat{\mathbf{p}}_i(t)$ across a large and diverse set of questions from three complementary sources.

Prediction markets. We source raw contracts from prediction-market platforms such as Polymarket. To produce clean questions, we use an LLM to parse each contract, extract the event semantics, candidate answers, closing time, and resolution criteria, and synthesize related contracts into a single well-formed tuple $\mathcal{Q}_i$.

Question synthesis. Although prediction markets provide a large volume of questions, their coverage is limited to topics of public interest or events that follow periodic schedules. Many timely, rapidly evolving, and long-tail topics are absent from these platforms. To address this limitation, we run an automated pipeline that continuously tracks news, policy developments, product launches, and domain-specific events from public sources.

Expert annotation. While the above two automated pipelines can generate questions at scale, they have limited coverage of the practical uncertainties that professionals routinely encounter in fields such as scientific research, public policy, and business strategy. To bridge this gap, we recruit domain experts to propose future events whose uncertainties are directly relevant to real-world decision-making. We describe the acquisition process in Appendix A. Expert-written questions are not yet included in the resolved-question evaluation.

Narrative taxonomy. We organize all questions into a three-level hierarchy: *category* $\rightarrow$ *domain* $\rightarrow$ *narrative*. A narrative is a reusable forecasting template that captures the analytical method shared by a group of questions. For instance, all questions about whether a certain famous person would say some words under a condition share the narrative *Person · Say words · Under conditions*. This

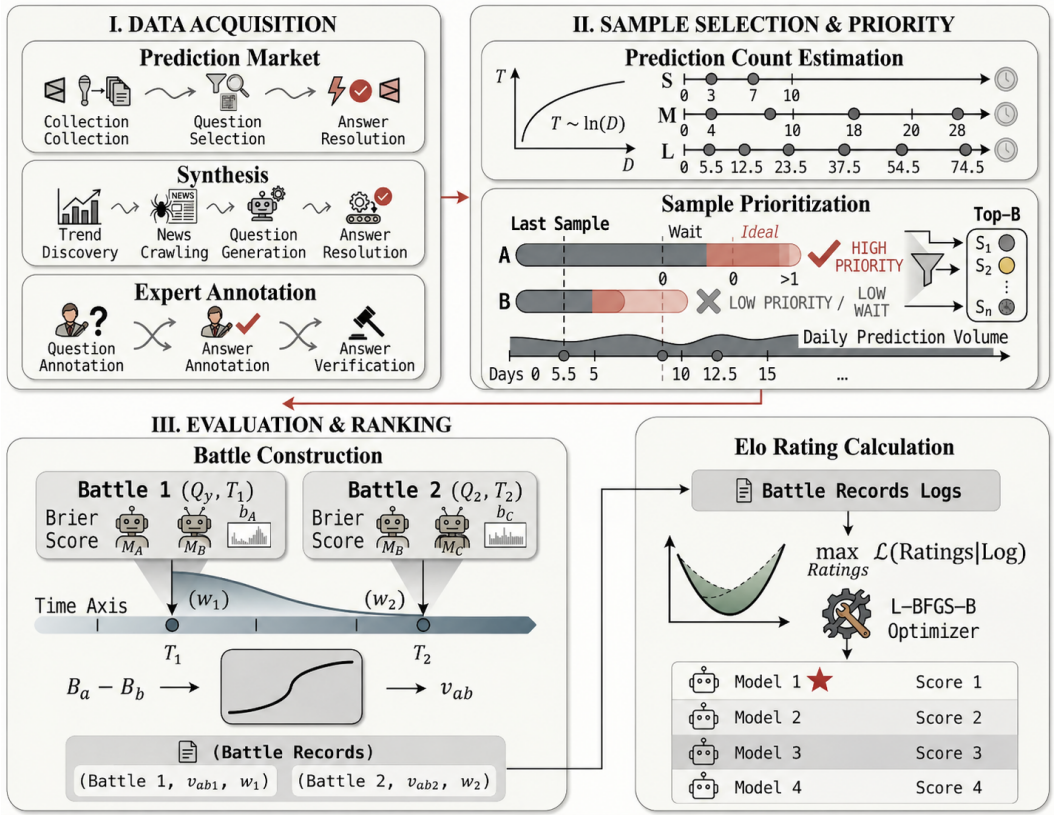

Figure 2: Echo pipeline overview. Three data sources feed into a structured question pool. Questions are scheduled across each event's lifecycle. Models are compared at matched prediction points.

fine-grained classification serves two purposes. First, it allows us to construct rubrics at the narrative level, producing evaluation criteria that are specific to each question type rather than generic across all domains. Second, it provides structured labels for balanced question sampling across categories and domains. Appendix B describes the labeling and sampling procedure in detail. As shown in Figure 7, the taxonomy currently covers 7 top-level categories with 1,075 narratives across 223 domains.

3.2 Time-Aligned Elo Rating

The information available for a prediction changes continuously as the resolution date approaches. A prediction made one month before an event relies on different evidence than one made one day before. If submissions from different time points are averaged directly, the resulting ranking conflates prediction ability with information availability. Echo addresses this by scheduling multiple prediction points across each question's lifecycle and comparing models only at the same time point. A prediction point is a designated time before a question's deadline at which all models submit probability estimates. Then we establish the ranking by adapting Elo rating (Chiang et al., 2024) mechanism into prediction scenario. The advantage of this Elo approach is that, even when prediction points do not perfectly align across models, due to differences in release times or other factors, the Elo mechanism allows for an effective comparison of model performance at shared prediction points.

Multi-point prediction scheduling. The system incorporates a scheduling mechanism based on the principle of first assigning a specific number of prediction points to each question, and then dynamically prioritizing questions for prediction each day, based on completed prediction points, until each question has undergone the number of prediction rounds initially allocated to it. One challenge lies in scheduling the prediction points for each problem. A simple schedule based on uniform time intervals would incur excessive prediction overhead. Striking a balance between prediction overhead and evaluation effectiveness is the focus of our research. Let H_i denote the forecast horizon of

question i, defined as the number of days between the question's entry into the arena and its resolution deadline. We assign

$$T_i = \text{round}(\alpha_T \ln(1 + H_i) + \beta_T) \tag{1}$$

prediction points per question, where α_T controls how quickly the number of points grows with horizon and β_T adjusts the number before rounding. The logarithmic form allocates more prediction points to longer-horizon questions without letting them dominate the daily budget. Each day, we assign each question a scheduling priority $q_i(t) = A_i(t)\, M_i(t)/L_i(t)$, where $A_i(t)$ is days since last scheduled, $M_i(t)$ is remaining prediction points, and $L_i(t)$ is days until resolution. Questions that have waited longer, still need more prediction points, or are close to resolution receive higher priority. Appendix C gives this scheduling algorithm.

Based on the fundamental principles of the Elo rating mechanism, our system compares models in a pairwise manner to derive a comprehensive ranking. When comparing two models, we construct battles based solely on the data points for which both models have generated predictions. Each pair of submissions on the same question at the same prediction point forms an *aligned battle*. In this pair, we first introduce how we measure the difference of prediction quality of both models for battle comparison. Then we show how the ranking details are based on the measures.

Aligned battle construction and prediction quality difference measurement. We measure prediction quality using the squared error of the correct-answer probability, $\text{BS}_m = (1 - p_m^c)^2$, where p_m^c is the probability that model m assigned to the correct answer. For a question with two possible answers, this is the Brier score (Glenn et al., 1950). For questions with more than two answers, we use only the probability assigned to the correct answer. Full multiclass Brier instead sums the errors across all answer options. We use "Brier" to refer to the correct-answer error in the leaderboard. For battle k between models a_k and b_k, we convert their Brier-score difference into a soft win label $v_k = \frac{1}{1+10^{(\text{BS}_{a_k} - \text{BS}_{b_k})/\sigma}}$, where σ is a temperature parameter: smaller values produce sharper transitions near 0 or 1 for even small Brier-score gaps, while larger values keep labels closer to 0.5. This soft label preserves the magnitude of the performance difference rather than reducing it to a binary win or loss. Each battle is also weighted by its lead time $w_k = 1 + \gamma \ln(1 + \Delta_k)$, where Δ_k is the number of days between the prediction point and resolution. The parameter $\gamma \geq 0$ controls the strength of this lead-time bonus, giving more credit to predictions made far in advance since early forecasts require reasoning under greater uncertainty.

Bradley-Terry rating estimation. We use the Bradley-Terry model (Bradley & Terry, 1952), a standard pairwise ranking approach that estimates the probability one model beats another from their latent ratings. Let r_m denote the rating of model m. The probability that a_k beats b_k is

$$P(a_k \succ b_k \mid \mathbf{r}) = \frac{1}{1 + 10^{(r_{b_k} - r_{a_k})/400}}. \tag{2}$$

We estimate ratings by maximizing the weighted log-likelihood over all aligned battles:

$$\mathcal{L}(\mathbf{r}) = \sum_k w_k \left[v_k \log P(a_k \succ b_k \mid \mathbf{r}) + (1 - v_k) \log\big(1 - P(a_k \succ b_k \mid \mathbf{r})\big) \right]. \tag{3}$$

The optimization $\hat{\mathbf{r}} = \arg\max_{\mathbf{r}} \mathcal{L}(\mathbf{r})$ is solved with L-BFGS-B (Zhu et al., 1997). We estimate confidence intervals by repeatedly sampling whole questions with replacement and recomputing the ratings for all participants. We keep all predictions from each sampled question together. When a model misses a prediction point, only the battles at that point are removed. All other comparisons remain valid because each retained battle uses the same scheduled prediction points.

3.3 Narrative-Wise Rubrics

So far, we describe how we build up Echo's Elo rating system. Furthermore, outcome-based metrics require waiting for question resolution, which can take weeks or months. To provide faster feedback, Echo also evaluates the reasoning trace that each model produces when making a prediction. Achieving this goal requires a method for generating rubrics, as well as the validation of the validity and robustness of LLM-based scoring using these rubrics. We employed an evolutionary approach to synthesize the rubrics and subsequently validated the effectiveness of the scoring.

Rubric generation. Each narrative in the taxonomy is associated with a set of *prediction skills* $\mathcal{S}_j = \{s_1, \ldots, s_K\}$, concrete analytical steps that characterize the reasoning required for that question

class. When a new question $\mathcal{Q}_i$ enters the arena, it is classified into a narrative. An LLM then converts the narrative's skills into a rubric $\mathcal{R}_j$ with five scoring dimensions. The first two dimensions are *domain-specific*: their criteria and checklists are drawn directly from $\mathcal{S}_j$, evaluating whether the model performed the analytical steps that the question type requires. The remaining three dimensions are *universal* and apply to all narratives: synthesis and causal structure, probability derivation, and uncertainty quantification.

Rubric evolution. The initial rubric for narrative n_j is produced by a generate-then-refine pipeline: a *generator* $\mathcal{G}$ proposes $\mathcal{R}_j^{(0)} = \mathcal{G}(\mathcal{S}_j)$ from the skill set, then a *critic* $\mathcal{C}$ audits it for coverage gaps, redundant criteria, insufficient discriminative power, weak evidence bars, and weight imbalance, producing $\mathcal{R}_j^{(1)} = \mathcal{C}(\mathcal{R}_j^{(0)}, \mathcal{S}_j)$. We provide a detailed explanation of how we generate and evolve our rubrics in Appendix D.

Rubric scoring and safeguards. Let τ denote a model's prediction trajectory, the full sequence of reasoning steps leading to a probability estimate. Each of the five rubric dimensions is scored on a 0–3 scale, and the weighted scores are normalized into a single rubric score:

$$\text{Rubric}(\tau, \mathcal{R}_j) = \sum_{d=1}^{5} \omega_d \cdot \frac{\text{score}_d(\tau)}{3}, \quad \sum_{d=1}^{5} \omega_d = 1. \tag{4}$$

An LLM scorer receives the question text, the rubric $\mathcal{R}_j$, and the trajectory τ, but not the question outcome or ground truth. The scorer evaluates each dimension independently: a trajectory that merely lists facts and asserts a probability receives at most score 1 on the synthesis and derivation dimensions, regardless of its length. When the scorer is uncertain between two adjacent scores, it assigns the lower one. To prevent circular validation, we enforce that no rubric may contain criteria referencing prediction-market prices, bookmaker odds, or other external consensus probabilities. Appendix H provides complete scoring examples.

Narrative rubrics serve as a fast proxy for Elo. The rubrics score the quality of a model's reasoning trajectory immediately after prediction, without waiting for the outcome. We compare rubric and Elo rankings in Section 4.4. Importantly, we regard Elo rating as the primary evaluation for Future prediction since it measures the real performances, while take the rubrics score as a fast proxy.

4 On Evaluating Predictive Agents in Echo

4.1 Experimental Setup

In the original controlled experiments, we evaluate 14 baselines on Echo, including 13 LLM-based prediction agents and the prediction-market consensus probability from Polymarket[1] as a non-model baseline. All LLM-based models operate through the same ReAct-style (Yao et al., 2022) agent wrapper equipped with web search, page visit, and Python sandbox tools. This shared interface controls differences in tool access or prompt format. We report three metrics: aligned Elo as the timing-corrected ranking, narrative rubric score as the process-based diagnostic, and the Brier score. For the revised rubric evaluation, we use seven models and DeepSeek-v4-Flash as the judge. We evaluate the models without Echo-specific training or fine-tuning. For the main Elo ranking, we use $\sigma = 0.10$ and $\gamma = 1.0$ and center the ratings at 1,000.

4.2 Main Results

Figure 1 presents the updated leaderboard snapshot. Among the models shown, GPT-5.6-Sol ranks first, followed by Kimi-K3 and DeepSeek-V4-Pro. Market ranks just below Claude-Opus-4.6, but their confidence intervals overlap. This live leaderboard uses predictions from each model's existing agent setup. In the controlled experiments below, all models use the same wrapper described above.

Table 1 compares Elo ratings with rubric scores. Opus ranks first under both measures. Seed ranks second by Elo and third by rubric score, while GLM-5 ranks third and second, respectively. GLM-5 scores highest on D1–D2, and Opus on D3–D5.

[1] https://polymarket.com/

Table 1: Elo ratings and rubric scores for seven models scored by DeepSeek-v4-Flash. D1–D5 are mean dimension scores, normalized to $[0, 1]$.

				Rubric Score↑					
Model	**Elo**	**Elo rank**	**Rubric rank**	**D1**	**D2**	**D3**	**D4**	**D5**	**Overall**
Claude-Opus-4.6	**1025.2**	1	1	0.446	0.523	**0.481**	**0.346**	**0.357**	**0.442**
Seed-2.0-Pro	1020.8	2	3	0.391	0.479	0.425	0.284	0.312	0.389
GLM-5	1015.0	3	2	**0.467**	**0.524**	0.467	0.295	0.323	0.429
Qwen3.5-397B-A17B	1010.0	4	4	0.401	0.463	0.395	0.291	0.316	0.383
Minimax-M2.5	998.2	5	7	0.350	0.420	0.334	0.214	0.250	0.324
Gemini-3.0-Pro	994.8	6	5	0.369	0.456	0.423	0.280	0.323	0.379
Kimi-K2.5	936.0	7	6	0.404	0.448	0.347	0.184	0.231	0.338

4.3 A Closer Look at our Elo Rating System

The ranking tests in Figures 3–5 use updated predictions from 13 AI agents and Market. To account for the limited information available early in an event, we give earlier errors less weight in the time-weighted Brier comparison. Each squared error is multiplied by $\exp(-0.3h)$, where h is the time left until resolution as a fraction of the total time from the question's first recorded prediction to resolution. We average these errors within each question and then across questions.

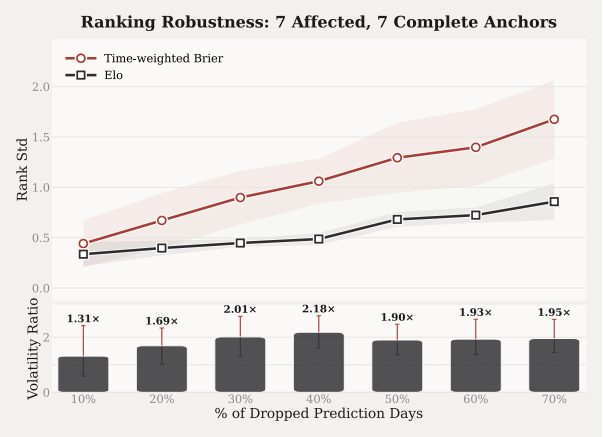

Figure 3: Ranking robustness under missing predictions. Top: mean rank standard deviation with bands based on the standard error across models. Bottom: time-weighted Brier/Elo ratio with bootstrap confidence intervals.

Ranking robustness against missing predictions. To test robustness to missing predictions, we compare changes in Elo and Brier rankings. Average losses can use different questions for different models, while Elo keeps comparisons at matched prediction points. We randomly remove 10%–70% of prediction days from seven participants, keeping the other seven complete, and repeat the test with $\gamma = 0$. Our measure is each affected model's rank standard deviation across trials, which we average across affected models. Time-weighted Brier has 1.31–2.18× Elo's mean rank standard deviation, and the ratios remain above one without time weighting. For both Brier measures, the 95% confidence intervals exclude one at 20%–70% missing days, but include one at 10%. These results indicate that Elo rankings change less under missing predictions in this test.

Additional ablations. Our next test examines ranking stability when models make fewer predictions on each question. It uses the original evaluation data and excludes predictions completed at or after resolution. Within each question, we randomly retain 25%, 50%, or 75% of prediction points, rounding up to keep at least one point. The mean absolute rank change measures how far the new ranks move from each method's ranks before removing predictions. Elo reduces the mean absolute rank change relative to average Brier by 22.7%, 32.8%, and 74.4%, respectively. These results show smaller rank changes for Elo than for average Brier in these tests with fewer predictions.

For the framework comparison, we hold the base model fixed at Opus 4.6 and evaluate ReAct and Claude Code. The comparison uses correct-answer squared error and full multiclass Brier on matched forecasts from both frameworks. Both 95% confidence intervals include zero, and we do not find a clear accuracy advantage for either framework on these measures. We report the comparison in Appendix G and test sensitivity to the Elo parameters in Appendix F.

New-model convergence. Our newcomer test replays predictions from 27 April to 13 May 2026 to measure how quickly a new model approaches its final rank. We treat each AI agent as a newcomer in turn, exclude questions resolved before its entry, and leave the other participants' histories unchanged. For each method, we measure the absolute difference between each newcomer's current and final rank, then average across newcomers. We smooth the curve with a centered three-day moving average and divide it by its initial value. In Figure 4, Elo reaches 40% of its initial rank error after 3.7 days, compared with 9.3 days for time-weighted Brier. Elo therefore approaches its final order earlier in this test. This test uses records available by the end of the replay period rather than reconstructing the data available on each day.

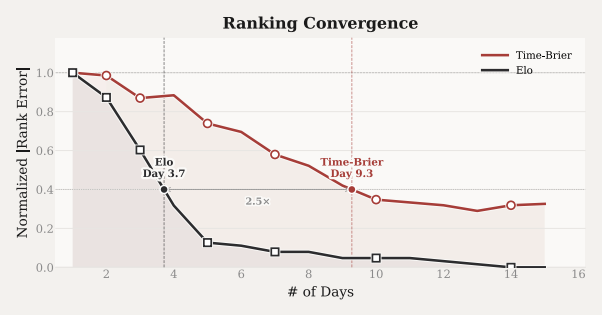

Figure 4: New-model convergence as questions resolve. Curves show smoothed mean rank error divided by its initial value. The dotted line marks 40%. Shaded areas are not confidence intervals.

Reliability under model-set changes. Since Elo ratings depend on comparisons between models, we test how removing participants changes the remaining order. We remove groups of one to six models and recompute the ratings. For each group, we compare the remaining models' order before and after removal. Kendall's τ measures agreement, with 1 indicating identical order. We average over the tested groups of the same size. In Figure 5, we obtain a mean agreement of 0.973 with one model removed and 0.947 with six removed. These results indicate that most pairs of models remain in the same relative order when other participants are removed.

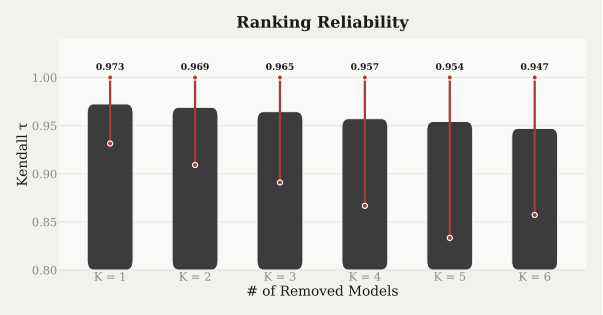

Figure 5: Ranking reliability under model-set changes. Mean Kendall's τ and the 2.5th–97.5th percentile range across removal combinations.

4.4 Narrative Rubric Evaluation

To assess agreement between rubric scores and Elo rankings, we use DeepSeek-v4-Flash to score one recorded trajectory per model on the shared evaluation set, without providing outcomes or Elo ratings. All forecasts precede recorded resolution. For each model, we average rubric scores across questions and compare the resulting ranking with Elo using Spearman's ρ. The Elo fit uses $\gamma = 0$ and $\sigma = 0.5$ within these seven models, giving each question equal weight and allowing selected prediction points to differ.

Our main comparison gives $\rho = 0.857$, with a 95% question-bootstrap interval of [0.679, 0.929]. We estimate these intervals by resampling whole questions and refitting Elo in each sample. The exact two-sided model-label permutation test gives $p = 0.02381$ for the main comparison. By restricting the comparison to questions where all selected prediction points match, we obtain a lower but still high correlation of $\rho = 0.750$, with an interval of [0.429, 0.964]. The decrease may reflect both the timing restriction and the changed question set.

When every model's scoring input must contain readable reasoning, the correlation remains $\rho = 0.857$, with an interval of [0.679, 0.929]. This restriction removes cases where Opus inputs lack explicit reasoning. The correlation remains positive under both restrictions, but its size changes when we match prediction points. These results are exploratory because we revised the model set after examining an earlier evaluation. The intervals and permutation test do not account for this model selection. Appendix D gives further controls and the scope of this revised evaluation.

Rubric analysis. Figure 6 compares each rubric dimension with Elo, using the same seven models and evaluation set as Table 1. We find the strongest correlation for probability derivation. Evidence analysis and synthesis have the next highest correlations. These results show which reasoning steps have scores that agree more closely with Elo. Since this comparison measures correlation, it does not establish that improving those steps causes better forecasts.

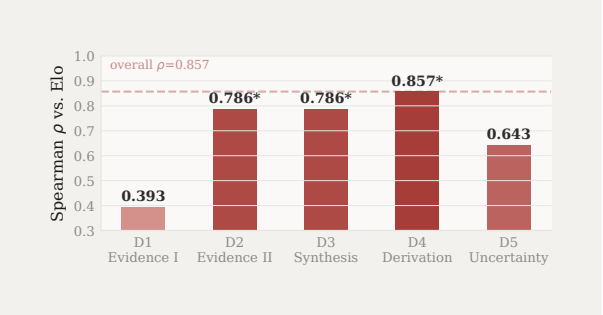

Figure 6: Per-dimension Spearman correlation with Elo for the seven-model rubric evaluation.

Agreement across judges. The original nine-model evaluation provides a separate test of whether rubric rankings depend on the judge. We use GPT-5.4, Gemini-3.1-Pro-Preview, and DeepSeek-v4-Flash to score the trajectories previously evaluated by Claude Sonnet 4.6. Every judge receives the same rubrics, without the outcomes or Elo ratings. The judges have different limits on input and response length, as described in Appendix D.

Table 2: Pairwise judge agreement with 95% question-bootstrap intervals.

Judge pair	Spearman ρ	95% confidence interval
GPT-5.4 / Claude Sonnet 4.6	0.967	[0.933, 1.000]
GPT-5.4 / DeepSeek-v4-Flash	0.933	[0.867, 0.983]
GPT-5.4 / Gemini-3.1-Pro-Preview	0.817	[0.717, 0.900]
Gemini-3.1-Pro-Preview / DeepSeek-v4-Flash	0.950	[0.900, 0.983]
Gemini-3.1-Pro-Preview / Claude Sonnet 4.6	0.817	[0.733, 0.833]
DeepSeek-v4-Flash / Claude Sonnet 4.6	0.917	[0.833, 0.950]

Our experiments show similar model rankings across judges, with a mean pairwise Spearman correlation of 0.900 in Table 2, even though the judges assign different absolute scores. The comparison with Elo uses predictions made at matching points on questions scored by all four judges. The Spearman correlations range from 0.717 to 0.800. Averaging rubric scores across the four judges gives a correlation of 0.800, matching the best individual judge in this test. A separate test removes intermediate reasoning and shows that retaining tool results alongside final answers gives closer agreement with Elo than scoring final answers alone. We report this ablation and an evaluation of models without visible reasoning in Appendix E.

Calibration and high-confidence errors. We examine whether a model's stated confidence matches how often it is correct. Using Gemini-3.1-Pro's original evaluation predictions, we compare average confidence with accuracy within groups of similar confidence. Expected Calibration Error averages the absolute differences between confidence and accuracy, weighting each group by its share of predictions. The calibration error is low at 0.017, yet 14.0% of the model's incorrect predictions have at least 80% confidence. These results indicate that low average calibration error does not rule out high-confidence errors on individual questions.

5 Conclusion

We presented Echo, a live and general AI prediction arena featuring time-aligned Elo rating, multi-source question acquisition, and narrative rubrics. Elo rankings are more stable than Brier rankings under missing and sparse predictions, and approach their final order sooner in the new-model test. Rubric scores correlate with Elo at $\rho = 0.857$ in the revised seven-model evaluation. The closest agreement is for probability derivation, followed by evidence analysis and synthesis. Further evaluation is needed to establish how well these relationships hold across other model sets. Echo serves as a fundamental and efficient benchmark for developing frontier forecasting systems.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, provide feedback on research methodology and experiments, clean and reformat datasets, and support qualitative and thematic data analysis. We have not used generative AI tools to develop theoretical models or conceptual frameworks, propose or refine hypotheses, implement research methods, assist with translation, or interpret results. Formulating mathematical claims, providing critical ingredients for proofs, and writing proofs are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code, improve the readability of the paper, and propose a title or keywords. We have reviewed all AI-assisted work. We checked that the code was correct, followed our intended design, and kept experimental comparisons fair. We also carefully reviewed AI-assisted text for hallucinations and checked that the experiments did not use shortcuts to obtain unfairly better results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Scaling open-ended reasoning to predict the future. *arXiv preprint arXiv:2512.25070*, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Alfred Cowles 3rd. Can stock market forecasters forecast? *Econometrica: Journal of the Econometric Society*, pp. 309–324, 1933.

W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. Openep: Open-ended future event prediction. *arXiv preprint arXiv:2408.06578*, 2024.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 37: 50426–50468, 2024.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, 2024.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*, 2023.

Lars Magnus Hvattum and Halvard Arntzen. Using elo ratings for match result prediction in association football. *International Journal of forecasting*, 26(3):460–470, 2010.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. Forecastqa: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4636–4650, 2021.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities. In *International Conference on Learning Representations*, volume 2025, pp. 93943–93980, 2025.

Jiashuo Liu, Siyuan Chen, Zaiyuan Wang, Zhiyuan Zeng, Jiacheng Guo, Liang Hu, Lingyue Yin, Suozhi Huang, Wenxin Hao, Yang Yang, et al. Futurex-pro: Extending future prediction to high-value vertical domains. *arXiv preprint arXiv:2601.12259*, 2026.

Peter Lynch. Weather prediction by numerical process. *The Emergence of Numerical Weather Prediction*, 11(3):1–27, 2006.

Mike A Merrill, Alexander G Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E Kelly Buchanan, et al. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *arXiv preprint arXiv:2601.11868*, 2026.

Maksym Nechepurenko and Pavel Shuvalov. Foresight arena: An on-chain benchmark for evaluating ai forecasting agents. *arXiv preprint arXiv:2605.00420*, 2026.

Victor Rotaru, Yi Huang, Timmy Li, James Evans, and Ishanu Chattopadhyay. Event-level prediction of urban crime reveals a signature of enforcement bias in us cities. *Nature human behaviour*, 6(8): 1056–1068, 2022.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, et al. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *arXiv preprint arXiv:2511.07685*, 2025.

Zhengwei Tao, Zhi Jin, Yifan Zhang, Xiancai Chen, Haiyan Zhao, Jia Li, Bin Liang, Chongyang Tao, Qun Liu, and Kam-Fai Wong. A comprehensive evaluation on event reasoning of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25273–25281, 2025a.

Zhengwei Tao, Pu Wu, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Chengfeng Dou, Xiancai Chen, Jia Li, Linyu Li, Chongyang Tao, et al. Prophet: An inferable future forecasting benchmark with causal intervened likelihood estimation. *arXiv preprint arXiv:2504.01509*, 2025b.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, et al. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*, 2025.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Iterative event reasoning in llm-based time series forecasting. In *Neural Information Processing Systems*, 2024.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024.

Qi Yan, Raihan Seraj, Jiawei He, Lili Meng, and Tristan Sylvain. Autocast++: Enhancing world event prediction with zero-shot ranking-based context retrieval. *arXiv preprint arXiv:2310.01880*, 2023.

Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. Llm-as-a-prophet: Understanding predictive intelligence with prophet arena. *arXiv preprint arXiv:2510.17638*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. Mirai: Evaluating llm agents for event forecasting. *arXiv preprint arXiv:2407.01231*, 2024.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, et al. Futurex: An advanced live benchmark for llm agents in future prediction. *arXiv preprint arXiv:2508.11987*, 2025.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, 23(4):550–560, 1997.

A Data Acquisition Pipeline

Category	Representative domains	Example narrative templates	Rubric criteria induced by narrative
Sports & Entertainment	Soccer, basketball, tennis, awards	- Match winner: team strength, injuries, venue, and form.	Score roster availability, team form, matchup context, and probability traceability.
Politics & Governance	US politics, geopolitics, policy	- Seat-count range: polling, incumbency, and turnout signals.	Score polling or institutional evidence, catalyst analysis, and conditional scenarios.
Crypto & Digital Assets	Tokens, DeFi, chains, exchanges	- Price threshold: required move under realized volatility.	Score threshold distance, volatility regime, on-chain evidence, and derivation.
Economy & Finance	Macro, equity, rates, earnings	- Inflation direction: energy, housing, and core services.	Score leading indicators, base effects, revisions, and uncertainty bands.
Science & Environment	Healthcare, AI, climate, technology	- Drug approval: trial endpoints, safety, and advisory signals.	Score technical evidence, regulatory milestones, and failure modes.
Esports & Gaming	FPS, MOBA, strategy games	- Map winner: map pool, side bias, and patch changes.	Score patch context, player form, map statistics, and matchup reasoning.
Other	Aerospace, academia, culture	- Launch success: readiness milestones and weather windows.	Score milestone readiness, expert signals, and historical base rates.

Figure 7: Three-level taxonomy of Echo. The figure shows 7 top-level categories, representative domains, and example narratives. The rightmost column indicates the narrative-wise rubric criteria.

Prediction markets. We source raw contracts from prediction-market platforms such as Polymarket. Raw contracts are not directly usable as prediction questions because a single real-world event often corresponds to multiple contracts with varying levels of specificity. To produce clean questions, we use an LLM to parse each contract, extract the event semantics, candidate answers, closing time, and resolution criteria, and synthesize related contracts into a single well-formed tuple $\mathcal{Q}_i$. This source covers events with clear resolution rules but is biased toward topics that attract sufficient trading volume.

Question synthesis. Although prediction markets provide a large volume of questions, their coverage is limited to topics that attract sufficient public interest or follow periodic schedules. Many timely and rapidly evolving topics are absent from these platforms. To address this limitation, we run an automated pipeline that continuously tracks trending news, policy developments, product launches, and domain-specific events from public sources. For each trending topic, an LLM-based agent searches news articles of that topic and identifies the primary unresolved question and converts it into a well-formed prediction question $\mathcal{Q}_i$. The agent formulates the uncertainty as a time-bounded claim c_i with a resolution deadline d_i, then constructs the candidate answer set $\mathcal{A}_i$ and a verifiable resolution rule ρ_i. A synthesized question is retained only if a publicly available source can resolve the outcome. The pipeline processes new topics daily, generating hundreds of questions per week.

Expert annotation. While the above two automated pipelines can generate questions at scale, they have limited coverage of the practical uncertainties that professionals routinely encounter in fields such as scientific research, public policy, and business strategy. To bridge this gap, we recruit domain experts to propose future events whose uncertainties are directly relevant to real-world decision-making. Each expert directly specifies the claim c_i, the candidate answer set $\mathcal{A}_i$, a resolution deadline d_i, and the evidence that should settle the outcome under ρ_i. The resulting questions undergo a consistency review before entering the question pool.

We apply the same curation pattern across data sources: candidate discovery, question synthesis, validation, and resolution. For prediction-market questions, candidate markets are first filtered to remove inactive, zero-volume, or otherwise stale markets. The remaining markets are passed to an LLM synthesis step that decides whether related markets should be merged into one multi-choice question or split into independent questions.

Each synthesized question must pass hard checks for deadline validity, option consistency, and time ordering. A second quality check verifies that the question is self-contained and that the options are mutually exclusive when the task is multi-choice. Resolution is automated whenever possible: closed

market prices or observed facts are extracted by the answer pipeline and then verified before the question is marked resolved.

B Question Labeling and Sampling

Question labels are used to keep the daily prediction set balanced. The micro–macro label captures entity scope, causal complexity, controllability, and impact breadth. The expert–public label captures knowledge barrier, decision utility, information cost, and topic longevity. Each question is also assigned a narrative from the taxonomy and a difficulty level. For each prediction round, we sample questions to balance question type, difficulty, time horizon, and category coverage. A domination-score filter removes questions where one option is already near-certain, since such questions provide little signal for comparison. The resolved-question distribution across categories is: Sports & Entertainment, Politics & Governance, Crypto & Digital Assets, Economy & Finance, Esports & Gaming, Science & Environment, and Other.

C Scheduling Algorithm

The logarithmic schedule in Eq. (1) was chosen to increase coverage for longer-horizon questions without letting them dominate the daily budget. A linear schedule would allocate too many prediction points to long-horizon questions, while a constant schedule would undersample early and middle stages. We set $\alpha_T = 1.35$ and $\beta_T = 0.5$, which yields approximately 4 prediction points for a 10-day question and 7 for a 90-day question. The scheduling priority $q_i(t) = A_i(t)M_i(t)/L_i(t)$ combines days since last scheduled, remaining prediction points, and days until resolution so that skipped questions recover while questions close to resolution are still emphasized.

D Rubric Generation and Scoring

Rubric generation. Each narrative in the taxonomy is associated with a set of *prediction skills* $\mathcal{S}_j = \{s_1, \ldots, s_K\}$, concrete analytical steps that characterize the reasoning required for that question class. For example, the narrative *Person · Say words · Under conditions* includes skills such as "identify the speaker's historical statements and stance on the topic" and "assess the likelihood of the triggering condition occurring within the specified timeframe." When a new question $\mathcal{Q}_i$ enters the arena, it is classified into a narrative. An LLM then converts the narrative's skills into a rubric $\mathcal{R}_j$ with five scoring dimensions. The first two dimensions are *domain-specific*: their criteria and checklists are drawn directly from $\mathcal{S}_j$, evaluating whether the model performed the analytical steps that the question type requires. The remaining three dimensions are *universal* and apply to all narratives: synthesis and causal structure, probability derivation, and uncertainty quantification.

Rubric evolution. The initial rubric for narrative n_j is produced by a generate-then-refine pipeline: a *generator* $\mathcal{G}$ proposes $\mathcal{R}_j^{(0)} = \mathcal{G}(\mathcal{S}_j)$ from the skill set, then a *critic* $\mathcal{C}$ audits it for coverage gaps, redundant criteria, insufficient discriminative power, weak evidence bars, and weight imbalance, producing $\mathcal{R}_j^{(1)} = \mathcal{C}(\mathcal{R}_j^{(0)}, \mathcal{S}_j)$. We also propose improving rubrics with examples where their scores fail to distinguish good and poor forecasts. The generator would use these examples to produce an updated rubric, $\mathcal{R}_j^{(t+1)} = \mathcal{G}(\mathcal{S}_j, \{\tau_{\text{fail}}\})$, which we would review and evaluate on separate questions. We provide a synthetic example in the supplementary code to illustrate this procedure. We have not yet tested whether it improves evaluation quality.

Before validating the rubrics against forecasting accuracy, we removed a D5 criterion that rewarded agreement with market predictions. We instead check whether the stated uncertainty and probabilities are supported by the analysis.

Scorer instructions. For the revised seven-model evaluation, we use DeepSeek-v4-Flash without providing outcomes or Elo ratings. We limit the question text to 500 characters, the trajectory to 25,000 characters, and each tool excerpt to 1,500 characters, while retaining the full rubric. The input includes readable reasoning and reasoning summaries but excludes encrypted fields and signatures. Long trajectories can lose their final answer under the input limit. We use a 1,000-token output limit and retry malformed or truncated responses with up to 4,096 tokens. A score is reused only when its

prediction, rubric, and complete scoring request match. We instruct the scorer to use scores 1–2 for most responses and reserve 3 for exceptional analysis.

Additional models are selected by their coverage of shared forecasts made before resolution and whether most scoring inputs contain readable reasoning. The evaluation uses earlier recorded forecasts, completed from March to June 2026.

To assess the effect of model membership, we repeat the rubric–Elo comparison with a smaller model set. Without Gemini-3.0-Pro, the correlation is $\rho = 0.886$ both on the main evaluation set and on the larger set available to the remaining models. The correlation therefore changes with model membership even when we keep the evaluation set fixed.

For the original nine-model cross-judge comparison in Table 2, the three added judges receive the full question and have a 4,096-token output limit. Claude Sonnet 4.6 uses a 500-character question limit, a 25,000-character trajectory limit, and a 1,000-token output limit. In Figure 6, stars denote unadjusted asymptotic Spearman tests across models: $*$ for $p < 0.05$, $**$ for $p < 0.01$, and $***$ for $p < 0.001$.

E On Evaluating Rubrics Without Intermediate Reasoning

To test whether rubric rankings depend on visible intermediate reasoning, we use the original nine-model evaluation to compare full trajectories with tool results plus the final answer, and with the final answer alone. Using Gemini-3.1-Pro-Preview as the judge, we find that both reduced views preserve much of the full-trajectory ranking. Scores for synthesis and probability derivation are lower when intermediate reasoning is removed. The reduced inputs no longer show the reasoning behind these steps.

Table 3 compares Elo with the rankings obtained by scoring the three views with Claude Sonnet 4.6. With tool results and final answers, the correlation is $\rho = 0.900$, compared with 0.567 for final answers alone. The same pattern holds when prediction points are matched. These results indicate that tool results improve the agreement with Elo when intermediate reasoning is unavailable.

Table 3: Sonnet rubric–Elo correlations on questions with all three scoring views available. Each row uses the same questions and models across views.

Matching rule	Full trajectory	Visible evidence	Final answer
Shared questions ($\gamma = 0, \sigma = 0.5$)	0.917	0.900	0.567
Exact selected points ($\gamma = 1, \sigma = 0.1$)	0.850	0.817	0.417

A broader evaluation covers 14 models, including models without visible reasoning. We use Sonnet to score their tool results and final answers on the same questions at matching prediction points. We first give each question source equal weight in the average rubric scores and calculate Elo with $\gamma = 1$, $\sigma = 0.5$. This comparison gives a correlation of 0.824. With a broader question set and equal weight for each question, we obtain 0.662. Because both the questions and their weights change, we cannot separate their effects on the correlation. Both comparisons show agreement with Elo without access to intermediate reasoning, although the strength of that agreement differs.

F A Closer Look at Elo Rating Stability

Our parameter tests examine how the leaderboard changes with the soft-label parameter σ and the lead-time weight γ. These tests use the same data as the current leaderboard, fitting all participants and displaying the same models as Figure 1. With $\gamma = 1$, we vary σ from 0.01 to 0.50. In Figure 8, GPT-5.6-Sol remains first, while Claude-Opus-4.6 ranges from fourth to sixth, Gemini-3.1-Pro from sixth to tenth, and GPT-5.2 from second to thirteenth. With $\sigma = 0.1$ held fixed, we then vary γ from 0 to 3. Kendall's τ relative to the default ranking remains at least 0.956 across all participants and 0.933 across the displayed models. These results show that the tested lead-time weights preserve most of the ordering, while the soft-label parameter can substantially change individual ranks.

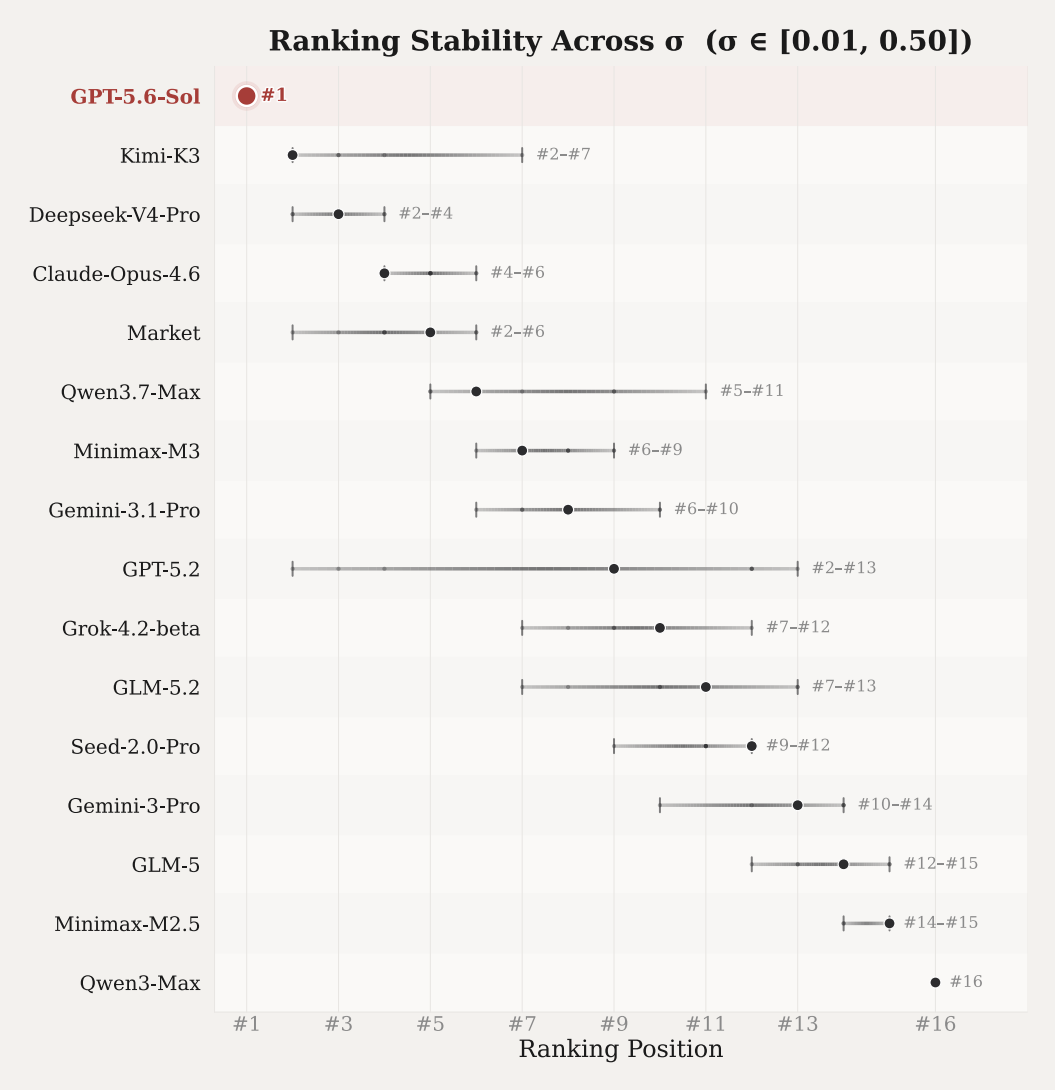

Figure 8: Leaderboard sensitivity to σ with $\gamma = 1$. Ranges show ranks among the displayed models as σ varies from 0.01 to 0.50. Dots mark $\sigma = 0.10$.

G On Evaluating the Impact of the Agent Framework

To assess how the agent framework affects forecast accuracy, we compare ReAct and Claude Code with Claude Opus 4.6 as the base model. ReAct is the wrapper used in the original controlled experiments. We match questions and prediction points and resample whole questions to estimate the intervals in Table 4.

The two errors in Eq. (5) compare a forecast with the resolved outcome. Let p_j be the predicted probability of option j, let y be the correct option, and let J be the number of options:

$$L_{\text{correct}}(\mathbf{p}, y) = (1 - p_y)^2, \qquad L_{\text{Brier}}(\mathbf{p}, y) = \sum_{j=1}^{J} (p_j - \mathbf{1}[j = y])^2. \tag{5}$$

Here, $\mathbf{1}[j = y]$ is 1 for the correct option and 0 otherwise. Correct-answer squared error uses only the probability assigned to the correct option. Full multiclass Brier sums the squared errors over all options, including the probability assigned to incorrect options. Lower values are better for both. For two options, full multiclass Brier is twice the correct-answer squared error under this sum convention.

For each error, we compute Claude Code minus ReAct, average over matched prediction points within each question, and then average equally across questions. Thus negative differences in Table 4 favor Claude Code.

Table 4: Error differences in the earlier Claude Code versus ReAct comparison with Opus 4.6. Negative values favor Claude Code.

Metric	Mean difference	95% confidence interval
L_{correct}	-0.00154	$[-0.00873, 0.00582]$
L_{Brier}	-0.00085	$[-0.01466, 0.01263]$

H Rubric Scoring Examples

The following two GLM-5 predictions illustrate how the rubrics assign scores to different reasoning traces. All thinking blocks are shown in full. Tool observations are truncated where marked with "…".

The contrast between these two cases shows the rubric's ability to distinguish analytical depth. In the high-score trajectory, the agent retrieves price data, computes volatility, runs Python simulations, and derives probabilities from a statistical model. In the low-score trajectory, the agent quickly finds the factual answer but performs no analytical work: no base rates, no individual risk assessment, no scenario analysis, and no derived probabilities.

High Rubric Score: Stock Price Threshold

GLM-5

Question Will Apple (AAPL) finish week of March 23 above $220 at 4:00 PM ET on March 27, 2026? A. Yes B. No

---

Turn 1

<thinking>
The user is asking whether Apple (AAPL) will finish the week of March 23 above $220 at 4:00 PM ET on March 27, 2026.
Today is March 25, 2026, so this is asking about the closing price on March 27, 2026 (which is 2 days from now).
I need to gather information about: 1. Current AAPL stock price 2. Recent price trends 3. Any news or factors that might affect the stock price 4. Technical analysis or analyst predictions
Let me start by searching for current AAPL stock price and recent news.
</thinking>

I'll help you predict whether Apple (AAPL) will close above $220 on March 27, 2026. Let me gather current market data and relevant information.

Search

```
{"query": ["AAPL stock price March 25 2026", "Apple stock current price today", "AAPL
    stock news March 2026", "Apple stock forecast March 27 2026"]}
```

Observation

```
A Google search for 'AAPL stock price March 25 2026' found 10 results:

## Web Results
1. [Apple Inc. (AAPL) Stock Historical Prices & Data - Yahoo Finance](https://finance.
    yahoo.com/quote/AAPL/history/)

Volume. Mar 24, 2026, 250.49, 254.82, 249.55, 251.64, 251.64, 27,882,961. Mar 24, 2026,
    250.35, 254.83, 249.55, 251.64, 251.64, 45,111,500.

2. [AAPL Mar 2026 262.500 call (AAPL260318C00262500)](https://finance.yahoo.com/quote/
    AAPL260318C00262500/history/)

Discover historical prices for AAPL260318C00262500 stock on Yahoo Finance. View daily,
    weekly or monthly format back to when AAPL Mar 202...
```

Turn 2

<thinking>

The search results provide very clear and consistent data about AAPL's current stock price. Let me analyze the key findings:

1. Current price data (from multiple sources): - March 24, 2026: Close at $251.64-$252.91 (various sources show slightly different closing prices, but all around $251-252) - March 23, 2026: Close around $253.97-$254.60 range - The stock is currently trading well above $220

2. Price trend in March 2026: - March 13, 2026: $255.40 - March 12, 2026: $258.66 - March 11, 2026: $261.09 - March 10, 2026: $257.65 - Late February 2026: around $265 - March 24, 2026: ~$251-252

3. Recent performance: - The stock has been declining slightly in March 2026 - It's been trading in the $249-$265 range - All prices are well above $220

4. The question asks: Will AAPL finish the week of March 23 above $220 at 4:00 PM ET on March 27, 2026?

Given that: - Current price on March 24 is approximately $251-252 - This is already $31-32 above the $220 threshold - The stock would need to drop approximately 12.4-12.8% in 3 trading days to fall below $220 - There's no news suggesting any catastrophic drop is imminent - Analyst forecasts remain bullish with price targets around $297

The probability of AAPL falling below $220 by March 27 is extremely low. The stock is currently trading at $251-252 and would need a massive drop of over 12% in just 3 trading days.

Let me check for any recent news or events that might affect the stock price this week.

</thinking>

I found current price data for AAPL. Let me verify the exact current price and recent trading data by visiting a financial data source.

Visit

```
{"urls": ["https://finance.yahoo.com/quote/AAPL/", "https://stockanalysis.com/stocks/aapl/
    history/"], "goal": "Extract the current stock price of AAPL, recent closing prices
    from March 23-24, 2026, and any relevant price trends or news"}
```

Observation

```
1. The current stock price of AAPL is $251.64, with an after-hours/overnight price of $252
    .70. The closing price on March 24, 2026, was $251.64. The previous day's close (
    March 23) was $251.49. The stock is down 7.44% year-to-date and 7.53% over the past
    month. Key news includes analysis that Apple is positioned to gain smartphone market
    share in 2026 despite a declining overall market, driven by strong iPhone sales and
    upcoming AI enhancements. Recent product launches and financial results indicate
    robust growth.

2. The current stock price of AAPL is $251.64, as of the close on March 24, 202...
```

Turn 3

<thinking>

I have solid data now. The current price as of March 24, 2026 close is $251.64. The question is whether AAPL will finish the week above $220 on March 27, 2026.

Key facts: - Current price (March 24 close): $251.64 - The threshold is $220, which is $31.64 below current price - That's about a 12.6% decline needed to fall below $220 - The stock is already well above $220 - The stock has 3 trading days (March 25, 26, 27) before the target date
For AAPL to NOT close above $220 on March 27, it would need to drop more than 12.6% in 3 trading days. That would be an extreme move. A 12.6% drop in 3 days would be a massive crash - something like a market-wide panic or major negative news specific to Apple.
Let me search for any upcoming events, earnings, or potential catalysts that could affect AAPL between now and March 27, 2026.
</thinking>
Good data collected. Now let me search for any upcoming events or news that could affect Apple stock between now and March 27, 2026.
Search

```
{"query": ["AAPL upcoming events March 25-27 2026", "Apple stock news today March 2026", "
    AAPL earnings March 2026", "Apple stock volatility risk March 2026"]}
```

Observation

```
A Google search for 'AAPL upcoming events March 25-27 2026' found 10 results:

## Web Results
1. [Apple's Biggest Week of 2026: Details on Every New Product ...](https://www.macrumors.
    com/2026/03/04/apple-march-2026-product-releases/)
Date published: Mar 4, 2026

Everything Apple announced this week will be launching on Wednesday, March 11, and we'll
    have more in-depth coverage and reviews of what's new.

2. [Apple March 2026 Event - 8 Things to Expect! - YouTube](https://www.youtube.com/watch?
    v=TiNm3qMLdYA)
Date published: Feb 22, 2026

Apple March 2026 Event - 8 NEW Products to Expect! Wallp...
```

Turn 4

<thinking>
Now I have comprehensive data. Let me analyze what I found:
Current Price and Recent Trends: - March 24, 2026 close: $251.64 - March 23, 2026 close: $251.49 - The stock is well ABOVE $220 currently - trading around $251-252 - Year-to-date: down 7.44% - Past month: down 7.53% - 52-week range: $169.21 - $288.62
Question: Will AAPL finish the week (March 23-27, 2026) above $220 at 4:00 PM ET on March 27, 2026?
Analysis: - Current price is ~$251.64, which is $31.64 above the $220 threshold - The stock would need to drop approximately 12.6% (from $251.64 to $220) in just 3 trading days (March 25, 26, 27) - Recent daily movements have been small (less than 1% daily) - No major earnings or events scheduled between now and March 27 - Volatility is around 21-22%, which is moderate - The stock has been in a downtrend but gradual - Recent product launches (iPhone 17e, iPad Air M4, M5 MacBook Pro, MacBook Neo) were announced in early March and already factored in
Risk factors: - General downtrend over the past month (7.53% decline) - Market volatility could cause sudden drops - No specific major negative news, but stock has been sliding
Key point: Even with the recent downtrend, AAPL would need to drop nearly 12.6% in 3 trading days to fall below $220. This would be an extremely rare and severe decline for a large-cap stock like Apple without major news.
Let me check if there are any specific risks or news that could cause such a drop.
</thinking>
Excellent data collected! Let me verify if there are any specific market-moving events or risks scheduled for March 25-27, 2026.
Search

```
{"query": ["market news March 25 2026", "stock market outlook this week March 2026", "AAPL
     analyst downgrades March 2026", "Fed meeting March 2026"]}
```

Observation

```
A Google search for 'market news March 25 2026' found 9 results:

## Web Results
1. [S&P 500 closes higher as oil price gains temper rebound from Iran ...](https://www.
    cnbc.com/2026/03/16/stock-market-today-live-updates.html)
```

```
The S&P 500 rose on Tuesday as Wall Street built on the momentum seen in the previous
     session amid developments in the Iran war.

2. [Stock Market News, March 20, 2026: Stocks Suffer Fourth ... - WSJ](https://www.wsj.com
     /livecoverage/stock-market-today-dow-sp-500-nasdaq-03-20-2026?gaa_at=eafs&gaa_n=
     AWEtsqeVWvgE8F7Ggk0-JUwg8KiQqduUCoXHyoF6MyiTtAf3k1TL15I1f_wF&gaa_ts=69c36...
```

Turn 5

<thinking>
Excellent! I've gathered a lot of relevant information. Let me synthesize what I know:
Current Price Situation (as of March 24, 2026): - AAPL closed at $251.64 on March 24, 2026 - The stock is already significantly ABOVE $220 - The target price is $220, which is $31.64 BELOW the current price - This is a margin of about 14.4% cushion above the target
Recent Price Trend: - March 23, 2026: closed at $251.49 - March 24, 2026: closed at $251.64 - The stock has been trading in the $250s recently - It's down about 7.5% YTD and 7.5% over the past month (from highs around $275 in mid-February) - Recent range has been roughly $249-$255
Market Context: - Fed meeting just concluded (March 17-18, 2026) - rates held steady at 3.5%-3.75% - Geopolitical concerns with Iran war affecting markets - Oil prices surging, creating market headwinds - Stocks have been volatile but relatively calm since start of Iran conflict
Apple-specific factors: - Recent product launches (March 4, 2026 event with iPhone 17e, MacBook Neo, etc.) - Strong Q1 2026 earnings (revenue $143.8B, up 16% YoY) - Apple projecting 13-16% revenue growth for March quarter - Tim Cook quelled retirement rumors - Some analyst downgrades in early 2026 citing AI competition concerns
Volatility Assessment: - 30-day historical volatility around 21% (moderate) - Expected move for options was about ±$3 (1.2%) for March 23 expiry
Key Question: Will AAPL close above $220 on March 27, 2026?
Given that: 1. Current price is $251.64 (already $31.64 above target) 2. Only 3 trading days remain (March 25, 26, 27) 3. For AAPL to fall below $220, it would need to drop ~12.6% in 3 days 4. Even with recent volatility, AAPL hasn't had such extreme drops 5. The stock has been relatively stable in the $249-$255 range recently 6. No major earnings or catastrophic news expected in the next 3 days
The probability seems strongly in favor of "Yes" - Apple will finish above $220 on March 27, 2026. The margin of safety is substantial (~$31 or 12.6%).
However, I should consider: - Market volatility due to Iran war situation - Any unexpected news could cause market-wide selloff - But even a significant drop of 5% would only bring it to ~$239, still well above $220
Let me check if there's any specific risk or news that could cause a 12%+ drop in 3 days.
</thinking>
I have comprehensive data. Let me check the specific market outlook for this week to understand any near-term risks.
Visit

```
{"goal": "Extract the market outlook for the week of March 23-27, 2026, including any
     expected volatility, risks, or events that could affect stock prices, especially
     Apple", "urls": ["https://www.cnbc.com/2026/03/20/stock-market-next-week-outlook-for-
     march-23-27-2026-.html"]}
```

Observation

```
1. The stock market outlook for March 23-27, 2026 is dominated by the ongoing war in Iran,
      which is entering its fourth week. The primary risk is the potential for an extended
      conflict, with reports of additional troops and warships being deployed and plans to
      occupy Kharg Island. Analysts warn this could trigger a recession and a potential
     20% market drop. Key volatility triggers include oil prices, which have spiked
     roughly 50%, and the S&P 500 closing below its 200-day moving average. The market's
     complacency is a concern; a sharp decline could force political de-escalation. The "
     biggest un...
```

Turn 6

<thinking>
I now have comprehensive information about the market situation and AAPL stock. Let me synthesize this:
Key findings:
1. **Current Price**: AAPL closed at $251.64 on March 24, 2026, and $251.49 on March 23, 2026
2. **Price relative to target**: The current price (~$251.64) is well above the $220 target. It would need to drop by more than $31.64 (about 12.6%) in 3 trading days to fall below $220.
3. **Recent trend**: The stock has been in a downtrend, falling from highs above $284 in mid-February to the $250s now, down about 7.44% YTD and 7.53% over the past month.

4. **Volatility**: 30-day historical volatility is around 20.8%, and expected move for weekly options is around 1.22% (about $3)
5. **Market context**: - Fed just held rates steady at 3.50%-3.75% on March 18 - Iran war is ongoing (entering fourth week), creating geopolitical uncertainty - Oil prices have spiked ~50% - S&P 500 closed below its 200-day moving average - Analysts warn of potential 20% market drop if war extends
6. **Company-specific news**: - Apple had strong Q1 2026 earnings in January (revenue up 16% YoY) - March 2026 product launches (iPhone 17e, iPad Air M4, M5 MacBook Pro, MacBook Neo) already announced - Analysts have a consensus price target of ~$297, suggesting upside - Recent downgrade from Raymond James but generally positive outlook
Risk Assessment:
The question is whether AAPL will close above $220 on March 27, 2026. Given: - Current price: ~$251.64 - Target: $220 - Buffer: $31.64 (12.6% cushion) - Time frame: 3 trading days (March 25, 26, 27)
For AAPL to close below $220 on March 27, it would need to drop 12.6% in 3 days. This would be an extremely rare and severe move. Even in a market crash scenario (which the Iran war situation could potentially trigger), a 12.6% drop in 3 days would be extraordinary.
Apple's stock has shown some volatility (21.54% annualized) but the expected move for weekly options is only about ±$3 (1.22%), not the $31+ needed.
Key risks that could cause a major drop: - Iran war escalation leading to broader market crash - Apple-specific negative news - Broader tech sector sell-off
Supporting factors: - Strong earnings and fundamentals - Product launches completed - Analysts generally bullish with price targets around $297 - Large buffer ($31+)
Given that the stock would need to drop more than 12% in 3 trading days to fall below $220, and there's no specific Apple-related catalyst that would cause such a drop, the probability of AAPL closing above $220 on March 27 is very high, even accounting for market uncertainty from the Iran situation.
Let me calculate this more precisely. With 21.54% annualized volatility, what's the expected 3-day move?
</thinking>
Let me calculate the probability using the current volatility data I've gathered.

Python

```
import math

# Current AAPL data from search results
current_price = 251.64 # Close on March 24, 2026
target_price = 220.00

# Days remaining: March 25, 26, 27 (3 trading days)
days_remaining = 3

# From search results: 30-day historical volatility is about 20.8%
# Also saw 21.54% volatility mentioned
annualized_volatility = 0.2154 # Using 21.54%

# Calculate daily volatility
daily_volatility = annualized_volatility / math.sqrt(252) # 252 trading days in a year
print(f"Daily volatility: {daily_volatility:.4f} ({daily_volatility*100:.2f}%)")

# Calculate expected move over 3 days
three_day_volatility = daily_volatility * math.sqrt(days_remaining)
print(f"3-day volatility: {three_day_volatility:.4f} ({three_day_volatility*100:.2f}%)")

# Expected 1-sigma move in dollars
expected_move_1sigma = current_price * three_day_volatility
print(f"Expected 1-sigma move (3 days): ${expected_move_1sigma:.2f}")

# Expected 2-sigma move (covers ~95% of outcomes)
expected_move_2sigma = expected_move_1sigma * 2
print(f"Expected 2-sigma move (3 days): ${expected_move_2sigma:.2f}")

# Price needed to drop to reach target
price_drop_needed = current_price - target_price
print(f"\nPrice drop needed to
# ... (truncated)
```

Observation

```
stdout:
Daily volatility: 0.0136 (1.36%)
3-day volatility: 0.0235 (2.35%)
Expected 1-sigma move (3 days): $5.91
Expected 2-sigma move (3 days): $11.83

Price drop needed to reach $220: $31.64
Percentage drop needed: 12.57%
```

```
This is a 5.35 sigma move (downside)

Probability of closing above $220: 100.00%
Probability of closing below $220: 0.0000%

Using options market expected move:
Options expected move (weekly): +/-$3.04

Stressed scenario (2x volatility):
Stressed 3-day volatility: 4.70%
Stressed expected move: $11.83
Sigma move in stressed scenario: 2.67
Probability above $220 (...
```

Turn 7

<thinking>
The probability calculation shows: - Even under normal volatility conditions, the probability of AAPL staying above $220 is essentially 100% - Even if volatility doubles (stressed scenario due to Iran war), the probability is still 99.63%
The $220 target would require a 12.57% drop in just 3 trading days. A 5+ sigma move is extremely rare. Even a 2.67 sigma move has very low probability.
Looking at all the evidence: 1. Current price: $251.64 (as of March 24, 2026 close) 2. Target: $220 - requires a $31.64 drop (12.57% decline) 3. Recent volatility: 21.54% annualized (about 1.36% daily) 4. Expected 3-day move: about $5.91 (2.35%) 5. Even with 2x volatility stress scenario, expected move is $11.83 6. The stock would need to drop nearly 3x the 2-sigma expected range
The Iran war is a concern, but: - Markets have already priced in much of the risk - Apple is down 7.44% YTD and 7.53% over the past month - some bad news is already priced in - A 12.57% drop in 3 days would be a crash scenario - Apple has strong fundamentals (record Q1 earnings, strong iPhone demand) - Analyst price target average is $297.58 - significant upside from current levels
Based on this analysis, the probability of AAPL closing above $220 on March 27, 2026 is extremely high - likely over 95%. Even in a stress scenario with double volatility, the probability is around 99.6%.
However, I should note that: - Black swan events can happen (major geopolitical escalation, company-specific bad news) - The Iran war situation is unpredictable - But these would need to be catastrophic to cause a 12.5%+ drop
My prediction should be strongly weighted toward "Yes" - AAPL will close above $220 on March 27, 2026.
</thinking>
Summary of Evidence Collected:
1. **Current AAPL Price (March 24, 2026 close): $251.64** - March 23 close: $251.49 - After-hours: $252.70
2. **Target Analysis:** - Target price: $220 - Drop needed: $31.64 (12.57% decline) - This is an extremely large move for just 3 trading days
3. **Volatility Data:** - 30-day historical volatility: 20.8-21.54% - Expected 3-day move: ~...
Visit

```
{"goal": "Get the most recent closing price for AAPL and confirm current trading levels",
    "urls": ["https://www.marketwatch.com/investing/stock/aapl"]}
```

Observation

```
1. AAPL's most recent closing price was $251.64. The current after-hours trading level is
    $253.47, up 0.73% from the close. The stock's day range was $249.55 to $254.83.
```

Turn 8

<thinking>
This confirms the current price. The stock is actually up in after-hours to $253.47, which makes it even less likely to drop to $220.
Let me now provide my final prediction. The evidence overwhelmingly supports AAPL finishing above $220 on March 27, 2026.
</thinking>
Final Evidence Summary:
- **Current Price (March 24, 2026): $251.64** (after-hours: $253.47) - **Target: $220** ? Requires 12.57% decline in 3 trading days - **Statistical probability: ~99.9%+ chance of staying above $220** - **No major negative catalysts** for Apple in the next 3 days - **Strong fundamentals:** Apple Q1 2026 revenue +16% YoY, iPhone demand strong - **Geopolitical risk** exists (Iran war) but already largely priced in - **Analyst price targets** average $297.58 - suggesting upside, not downside

> The question asks if AAPL will finish above $220 on March 27, 2026. Based on all evidence, the probability is extremely high that it will remain well above $220. The stock would need an unprecedented 12.5% crash in just 3 days with no major negative catalyst - a scenario with near-zero probability based on historical volatility patterns.{"Yes": 0.98, "No": 0.02}

Narrative rubric: Equity · Single Stock · Closing Price Threshold at Specified Timepoint

Dimension	Wt	Criteria & Scoring Guide
D1: Price Distance & Volatility Analysis	0.25	Does the trajectory compute the gap between current price and threshold, then compare it against the stock's historical daily move distribution (realized volatility over recent trading days)? *Scoring:* 0=No mention of current price or volatility. 1=States current price and gap but no volatility context. 2=Computes gap and uses historical volatility to assess feasibility. 3=Rigorous distributional analysis with explicit move probabilities derived from realized vol and days remaining.
D2: Catalyst & Event Identification	0.2	Does the trajectory identify relevant catalysts: earnings, macro releases, analyst actions, sector trends, and technical levels near the threshold? *Scoring:* 0=No event or catalyst check. 1=Mentions one category superficially. 2=Covers multiple catalyst types with relevance to the specific date and stock. 3=Comprehensive scan across company events, macro calendar, technical levels, and sector context with assessed directional impact.
D3: Synthesis & Causal Structure	0.2	Does the trajectory integrate multiple factors into a coherent predictive argument with scenario reasoning or conditional logic? *Scoring:* 0=No integration, just isolated facts. 1=Lists factors without connecting them. 2=Builds a structured argument weighing bull/bear scenarios. 3=Explicit scenario enumeration with conditional probabilities and clear causal chains linking factors to threshold outcome.
D4: Probability Derivation	0.2	Is the final probability traceable to the analysis–anchored on base rates, vol-implied move distributions, or explicit quantitative reasoning? *Scoring:* 0=Probability asserted with no justification. 1=Vague qualitative mapping. 2=Probability linked to specific analytical inputs (e.g., % of days stock moves X%). 3=Explicit calculation from distributional assumptions with clear parameter choices.
D5: Uncertainty & Self-Consistency	0.15	Does the trajectory acknowledge key uncertainties, test sensitivity to assumptions, and ensure internal consistency? *Scoring:* 0=No uncertainty discussion. 1=Brief hedging language. 2=Identifies key assumptions and notes how changes would shift the estimate. 3=Formal sensitivity analysis on vol or catalyst assumptions, checks consistency of implied conditionals.

Rubric evaluation (weighted score: **0.817**)

Dimension	Score	Reason
D1: Price Distance & Volatility A…	3/3	The trajectory explicitly computes the gap ($31.64, 12.57%), calculates daily and 3-day realized volatility (1.36% daily, 2.35% 3-day), derives sigma move (5.35σ), and generates explicit probability figures including a stressed 2x volatility scenario. This is rigorous distributional analysis with explicit move probabilities.
D2: Catalyst & Event Identificati…	2/3	Covers multiple catalyst categories: geopolitical risk (Iran war), macro context (Fed meeting), Apple fundamentals/earnings, analyst targets, and product launches. Doesn't deeply assess directional impact of each or technical levels near $220, but covers enough categories with relevant specifics to merit a 2.
D3: Synthesis & Causal Structure	2/3	Builds a structured argument weighing the extreme statistical improbability against geopolitical risks, with a stressed-scenario analysis. However, it doesn't enumerate explicit conditional scenarios with assigned probabilities–it mainly concludes the outcome is near-certain without detailed bull/bear conditional chains.
D4: Probability Derivation	3/3	Probability is explicitly derived from sigma calculations (5.35σ normal, 2.67σ stressed), with clear parameter choices (realized vol, 3-day horizon), yielding traceable quantitative outputs (99.6% stressed, ~100% normal). This is explicit calculation from distributional assumptions.

Dimension	Score	Reason
D5: Uncertainty & Self-Consistenc…	2/3	Identifies key assumptions (volatility regime) and tests sensitivity via stressed 2x volatility scenario. Acknowledges geopolitical uncertainty. However, no formal sensitivity analysis varying parameters systematically, and doesn't check consistency of implied conditionals. Solid but not formal.

LOW RUBRIC SCORE: CABINET DEPARTURE PREDICTION

In this example, the model treats the event as already having occurred. We use it to show why stating an answer without analysis receives a low rubric score, not as evidence of forecasting ability.

GLM-5

Question Will at least one cabinet-level Donald Trump appointee from his second term leave their position between February 4, 2026, and March 31, 2026, at 11:59 PM ET? A. Yes B. No

---

Turn 1

<thinking>
This is a prediction question about whether any cabinet-level Donald Trump appointee from his second term will leave their position between February 4, 2026 and March 31, 2026.
Today is March 24, 2026, so we're already within that time window. I need to: 1. First, identify who are the cabinet-level appointees in Trump's second term 2. Check if any have left their positions during the specified period (February 4, 2026 to March 31, 2026)
Let me search for information about Trump's second term cabinet and any departures.
</thinking>
I need to investigate whether any cabinet-level Trump appointees from his second term have left their positions during the specified period (February 4, 2026 - March 31, 2026). Today is March 24, 2026, so we're within that window.
Let me search for information about Trump's second term cabinet and any recent departures.
Search

```
{"query": ["Trump second term cabinet members 2025 2026", "Trump cabinet resignation
    2026", "Trump cabinet departures February March 2026", "Trump administration cabinet
    turnover 2026"]}
```

Observation

```
A Google search for 'Trump second term cabinet members 2025 2026' found 10 results:

## Web Results
1. [Trump's cabinet picks: Here's who will join his second administration](https://www.
    firstcoastnews.com/article/news/nation-world/trump-cabinet-picks/507-c7695a82-5ecc-4
    d10-994b-d3c916ef06a9)
Date published: Nov 13, 2024

Cabinet nominees . TREASURY SECRETARY: Scott Bessent . SECRETARY OF STATE: Marco Rubio .
    DIRECTOR OF NATIONAL INTELLIGENCE: Tulsi Gabbard . DEFENSE ...

2. [Trump's Bold Cabinet Picks: New Faces, Big Changes for 2025](https://www.youtube.com/
    watch?v=uGTj83L0zfw)
Date publishe...
```

Turn 2

<thinking>
This is very interesting. I found a key piece of evidence. The Ballotpedia result mentions:
"March 6, 2026: Donald Trump removed Kristi Noem as secretary of Homeland Security and nominated Markwayne Mullin to take the position."
This is exactly within the timeframe I need to check (February 4, 2026 - March 31, 2026). Kristi Noem was removed as Secretary of Homeland Security on March 6, 2026.
The Columbia Missourian article from Feb 24, 2026 also stated: "Nearly a month into his second year in office, President Trump's Cabinet has been remarkably stable. No Cabinet member has quit, been fired..." - but that was before March 6.
The Covers.com prediction market mentions: "Kristi Noem is still live in this market, even though President Trump has announced her departure."

Let me verify this information by visiting the Ballotpedia page to confirm the details about Kristi Noem's departure.
</thinking>
I found a crucial piece of information. The Ballotpedia result mentions that on **March 6, 2026**, Donald Trump removed Kristi Noem as Secretary of Homeland Security. This is within the specified window (February 4 - March 31, 2026). Let me verify this by visiting the Ballotpedia page.
Visit

```
{"urls": ["https://ballotpedia.org/Donald_Trump%27s_Cabinet,_2025-2026"], "goal": "Verify
    details about Kristi Noem's removal as Homeland Security Secretary on March 6, 2026,
    and identify any other cabinet departures during February-March 2026"}
```

Observation

```
1. On March 6, 2026, Donald Trump removed Kristi Noem as Secretary of Homeland Security
    and nominated Markwayne Mullin to replace her. Mullin was later confirmed on March 23,
     2026. There were no other cabinet removals or departures reported during February-
    March 2026.
```

Turn 3

<thinking>
The Ballotpedia page confirms that on March 6, 2026, Donald Trump removed Kristi Noem as Secretary of Homeland Security. This is a cabinet-level position, and this removal occurred within the specified window of February 4, 2026, to March 31, 2026.
Let me also check the Brookings tracking page and other sources to verify this information and ensure there are no other departures in that time period.
</thinking>
Let me verify this with additional sources, including the Brookings tracking page and other news sources.
Visit

```
{"goal": "Verify Kristi Noem departure and find any other cabinet departures in February-
    March 2026. List all cabinet turnover events with dates.", "urls": ["https://www.
    brookings.edu/articles/tracking-turnover-in-the-second-trump-administration/"]}
```

Observation

```
1. Based on the provided content, there is no information about Kristi Noem's departure or
     any cabinet turnover events in February-March 2026. The article only tracks turnover
     up to January 20, 2026. It states the Cabinet turnover rate for the second Trump
    administration is 7% with one position turned over, but does not list any specific
    events or dates for February-March 2026.
```

Turn 4

<thinking>
The Brookings page was published on January 20, 2026, so it wouldn't have March 2026 events. Let me search for more specific news about Kristi Noem's departure in March 2026.
</thinking>
Let me search for more specific news about Kristi Noem's departure.
Search

```
{"query": ["Kristi Noem removed Homeland Security March 2026", "Kristi Noem fired Trump
    March 2026", "Trump removes Kristi Noem secretary 2026"]}
```

Observation

```
A Google search for 'Kristi Noem removed Homeland Security March 2026' found 10 results:

## Web Results
1. [Trump fires Kristi Noem as homeland security secretary - NPR](https://www.npr.org
    /2026/03/05/nx-s1-5667546/kristi-noem-homeland-security-fired)
Date published: Mar 5, 2026

President Trump has fired his homeland security secretary, Kristi Noem, and said Markwayne
     Mullin, a senator from Oklahoma, would replace ...

2. [Noem fired from DHS amid agency shutdown, Trump taps Mullin as ...](https://www.
    govexec.com/management/2026/03/trump-fire-dhs-secretary-noem-selects-mullin-as-
    successor/41...
```

Turn 5

<thinking>
Excellent! Multiple reputable news sources confirm that Kristi Noem was fired/removed as Homeland Security Secretary by President Trump on March 5, 2026. This is clearly within the specified window (February 4, 2026 to March 31, 2026 at 11:59 PM ET).
Key findings: 1. NPR: "Trump fires Kristi Noem as homeland security secretary" - March 5, 2026 2. TIME: "Trump Removes Kristi Noem as Homeland Security Secretary" - March 5, 2026 3. CNN: "Trump fires Noem as frustrations build among White House officials" - March 5, 2026 4. NBC News: "President Donald Trump fired Homeland Security Secretary Kristi Noem on Thursday, making her the first Cabinet secretary to leave her post" 5. Reuters: "Trump fires Kristi Noem as homeland secretary after storm" - March 5, 2026 6. BBC, Al Jazeera, Axios, and others also reported this
The question asks: "Will at least one cabinet-level Donald Trump appointee from his second term leave their position between February 4, 2026, and March 31, 2026, at 11:59 PM ET?"
The answer is clearly YES. Kristi Noem was removed as Secretary of Homeland Security on March 5, 2026, which is within the specified time window. She was a cabinet-level appointee from Trump's second term.
I have sufficient evidence to make a confident prediction.
</thinking>
Let me verify one of these sources directly to confirm the exact date.
Visit

```
{"goal": "Confirm the exact date and details of Kristi Noem's firing as Homeland Security
     Secretary. Was she a cabinet-level appointee? What was the exact date of removal?", "
     urls": ["https://www.npr.org/2026/03/05/nx-s1-5667546/kristi-noem-homeland-security-
     fired", "https://www.nbcnews.com/politics/trump-administration/trump-says-kristi-noem-
     stepping-homeland-security-secretary-rcna248719"]}
```

Observation

```
1. Kristi Noem was fired from her position as Homeland Security Secretary, a Cabinet-level
      role, by President Trump on March 5, 2026.

2. Kristi Noem was fired from her position as Homeland Security Secretary, a cabinet-level
      role, on Thursday, March 5, 2026. President Trump announced her removal and stated
     her successor would take over on March 31, 2026.
```

Turn 6

<thinking>
This is excellent confirmation. Let me verify:
1. Kristi Noem was Homeland Security Secretary - a cabinet-level position (yes) 2. She was fired/removed by President Trump on March 5, 2026 (yes) 3. March 5, 2026 is within the specified window (February 4, 2026 - March 31, 2026 at 11:59 PM ET) (yes)
The question asks: "Will at least one cabinet-level Donald Trump appointee from his second term leave their position between February 4, 2026, and March 31, 2026, at 11:59 PM ET?"
The answer is definitively YES. Kristi Noem, a cabinet-level appointee (Homeland Security Secretary), left her position on March 5, 2026, which falls within the specified window.
Let me verify that there are no other departures I should be aware of, and also confirm this is from Trump's second term (not first term).
Actually, I have enough evidence. The sources are clear: - NPR confirmed on March 5, 2026 that Trump fired Kristi Noem as Homeland Security Secretary - NBC News confirmed the same, noting she was "the first Cabinet secretary to leave her post" - This is clearly from Trump's second term (2025-2026) - The date March 5, 2026 is well within the window (Feb 4 - March 31, 2026)
I should provide my final answer now.
</thinking>
I now have definitive evidence. Let me verify one more detail about whether there were any other departures to ensure completeness.
Visit

```
{"goal": "Confirm that Kristi Noem was the first cabinet secretary to leave in Trump's
     second term and find the exact timeline of her departure", "urls": ["https://www.
     reuters.com/world/us/trump-taps-us-senator-mullin-replace-noem-dhs-chief
     -2026-03-05/"]}
```

Observation

```
1. Kristi Noem was fired as Homeland Security Secretary by President Donald Trump on March
      5, 2026. The article explicitly states she was "the first Senate-confirmed member of
      Trump's Cabinet to be removed this term."
```

Turn 7

<thinking>
This is very clear evidence. The Reuters article confirms that Kristi Noem was fired on March 5, 2026, and that she was "the first Senate-confirmed member of Trump's Cabinet to be removed this term."
Let me verify the date range one more time: - Question period: February 4, 2026 to March 31, 2026 at 11:59 PM ET - Kristi Noem fired: March 5, 2026 - March 5, 2026 is within the specified period
Since Kristi Noem was a cabinet-level Trump appointee (Homeland Security Secretary), and she was fired on March 5, 2026, which falls within the February 4 - March 31, 2026 window, the answer is definitively "Yes."
Let me double-check whether there might be any ambiguity about her departure date. Some sources mentioned March 6, and some mentioned that her successor would take over on March 31. Let me verify when exactly she left - was she removed effective immediately on March 5/6, or was she removed effective March 31?
Actually, the question asks if someone "leave their position" during the window. If Trump announced on March 5 that he was removing her, and she was removed effective that day (or shortly after), that counts as leaving during the window. The exact effective date might be March 31 for her successor to take over, but she would have left or been removed from her position on March 5/6.
Based on all the evidence: 1. Kristi Noem was Homeland Security Secretary - a cabinet-level position 2. She was fired by Trump on March 5, 2026 3. March 5, 2026 falls between February 4, 2026 and March 31, 2026 4. She was the first cabinet-level appointee to be removed from Trump's second term
Therefore, at least one cabinet-level Donald Trump appointee from his second term has left their position during the specified period. The answer is "Yes" with very high confidence.
</thinking>
I now have definitive evidence. Let me verify one more detail about the exact effective date of her departure to ensure it falls within the window.
Visit

```
{"goal": "Find the exact effective date when Kristi Noem left her position as DHS
     Secretary. Was she removed effective immediately on March 5, or was there a later
     effective date like March 31?", "urls": ["https://time.com/7382719/kristi-noem-
     removed-homeland-security-secretary-markwayne-mullin/", "https://www.cbsnews.com/news
     /decision-to-remove-kristi-noem-dhs-secretary/"]}
```

Observation

```
1. Kristi Noem's removal as DHS Secretary was announced on March 5, 2026, but her last day
      on the job was March 31, 2026, according to a department-wide message to staff.

2. Kristi Noem was removed as Homeland Security Secretary effective March 31. President
     Trump announced that her replacement, Markwayne Mullin, would assume the role on that
      date. Noem was reassigned to a different position as of that time.
```

Turn 8

<thinking>
This is interesting - there seems to be some ambiguity about the exact timeline. Let me think about this:
1. March 5, 2026 - Trump announced firing Kristi Noem 2. March 31, 2026 - Her replacement Markwayne Mullin takes over
The question asks if at least one cabinet-level appointee "leave their position between February 4, 2026, and March 31, 2026, at 11:59 PM ET"
There are two interpretations: - If "leave their position" means the announcement/firing, that happened on March 5, 2026 - within the window - If "leave their position" means their last day on the job, that was March 31, 2026 - also within the window (the window ends March 31, 2026 at 11:59 PM ET)
Either way, this falls within the specified period! March 5 is clearly between February 4 and March 31. And if her last day was March 31, that's still within the window (until 11:59 PM ET).
So regardless of interpretation, a cabinet-level Trump appointee (Kristi Noem as Homeland Security Secretary) left their position during the specified window.
Actually, let me re-read the sources. One source says: - "Kristi Noem's removal as DHS Secretary was announced on March 5, 2026, but her last day on the job was March 31, 2026"
So she was removed/fired on March 5, 2026 (the announcement), which is definitely within the window (February 4 - March 31, 2026). Even if her last day was March 31, 2026, that's still within the window (the window ends March 31, 2026 at 11:59 PM ET).
This clearly answers "Yes" to the question. A cabinet-level Trump appointee (Kristi Noem, Homeland Security Secretary) left her position during the specified period.
</thinking>
I now have clear and verified evidence to answer this question.
What was accomplished: - Confirmed that **Kristi Noem**, who served as **Secretary of Homeland Security** (a cabinet-level position), was fired by President Trump on **March 5, 2026** - Multiple news

sources (NPR, NBC, Reuters, Time, CNN) confirmed this event - Reuters explicitly stated she was "the first Senate-confirmed member ...

Search

```
{"query": ["Kristi Noem Homeland Security Secretary confirmation date 2025", "Trump second
     term cabinet confirmed 2025"]}
```

Observation

```
A Google search for 'Kristi Noem Homeland Security Secretary confirmation date 2025' found
     10 results:

## Web Results
1. [US Senate Confirms Kristi Noem as Secretary of Homeland Security](https://www.dhs.gov/
    news/2025/01/25/us-senate-confirms-kristi-noem-secretary-homeland-security)
Date published: Jan 25, 2025

US Senate Confirms Kristi Noem as Secretary of the Department of Homeland Security in
    Sweeping Bi-Partisan Vote. Release Date: January 25, 2025.

2. [Senate confirms Kristi Noem as Trump's secretary of homeland ...](https://www.npr.org
    /2025/01/25/g-s1-44753/immigration-senate-kristi-n...
```

Turn 9

<thinking>
I have all the evidence I need:
1. **Kristi Noem was a cabinet-level appointee**: She was confirmed as Secretary of Homeland Security on January 25, 2025, for Trump's second term.
2. **She left her position during the specified window**: - She was fired/removed on March 5, 2026 - Her last day was March 31, 2026 - Both dates are within February 4, 2026 - March 31, 2026 at 11:59 PM ET
3. **Multiple sources confirm this**: - NPR, NBC, Reuters, CNN, Time, CBS, BBC, Al Jazeera, and others all reported on March 5, 2026 that Trump fired Kristi Noem as Homeland Security Secretary - Reuters specifically noted she was "the first Senate-confirmed member of Trump's Cabinet to be removed this term"
The question asks: "Will at least one cabinet-level Donald Trump appointee from his second term leave their position between February 4, 2026, and March 31, 2026, at 11:59 PM ET?"
The answer is definitively **YES**. Kristi Noem left her position during this timeframe.
I can now provide my final prediction.
</thinking>
I have comprehensive and verified evidence to answer this question definitively.
Summary of Findings:
1. **Kristi Noem was a cabinet-level Trump appointee**: She was confirmed as Secretary of Homeland Security on January 25, 2025, for Trump's second term (confirmed by DHS.gov, Senate records, and multiple news outlets).
2. **She left her position during the specified window**: - **March 5, 2026**: Trump announced he was firing Noem as Homeland Security Secretary (reported by NPR, NBC, Reuters, CNN, Time, CBS, BBC, Al Jazeera, etc.) - Reuters explicitly stated she was "the first Senate-confirmed member of Trump's Cabinet to be removed this term" - Her last day on the job was March 31, 2026
3. **Both key dates fall within the specified timeframe** (February 4, 2026 - March 31, 2026 at 11:59 PM ET): - Announcement of removal: March 5, 2026 (yes) - Effective departure: March 31, 2026 (yes)
The question asks if **at least one** cabinet-level appointee left their position during this period. Kristi Noem definitively did.{"Yes": 0.99, "No": 0.01}

Narrative rubric: US Politics · Cabinet · Any Member Departure Within Time Window

Dimension	Wt	Criteria & Scoring Guide
D1: Historical Base Rate Analysis	0.25	Computes cabinet departure base rates using historical turnover data (e.g., Brookings tracker), administration-stage adjustments, and Poisson/survival modeling for the specific window length. *Scoring:* 0=No historical data referenced. 1=Mentions turnover happens but no quantification. 2=Cites specific historical rates and applies window-length scaling. 3=Fits arrival rate model, adjusts for administration stage, and compares across multiple prior administrations.

Dimension	Wt	Criteria & Scoring Guide
D2: Individual Risk Assessment & Aggregation	0.25	Evaluates per-member departure signals (presidential criticism, scandals, reporting on tensions, policy conflicts) and aggregates via complement product P(≥1) = 1 - Π(1-p_i). *Scoring:* 0=No individual-level analysis. 1=Vaguely names at-risk members without evidence. 2=Identifies specific members with concrete signals and attempts aggregation. 3=Systematic risk scoring across multiple members with explicit complement-product aggregation and sourced evidence.
D3: Synthesis & Causal Structure	0.2	Integrates base rates, individual risks, and catalysts (hearings, budget deadlines, recess periods) into a coherent causal argument with scenario reasoning. *Scoring:* 0=No integration. 1=Lists factors without connecting them. 2=Builds structured argument linking catalysts to departure probability with scenarios. 3=Multi-scenario framework with explicit conditional reasoning and competing hypotheses weighed.
D4: Probability Derivation Traceability	0.15	Final probability is traceable to preceding analysis–anchored on base rates or explicit calculation, not just asserted. *Scoring:* 0=Number appears without justification. 1=Loosely linked to analysis. 2=Clear derivation path from base rate adjusted by evidence. 3=Fully transparent calculation with stated assumptions mapping to final number.
D5: Uncertainty & Self-Consistency	0.15	Quantifies uncertainty, tests sensitivity to key assumptions (window length, number of at-risk members, arrival rate), and checks internal consistency. *Scoring:* 0=No uncertainty discussion. 1=Acknowledges uncertainty verbally. 2=Tests sensitivity on 1-2 key parameters. 3=Systematic sensitivity analysis with explicit confidence range and consistency checks.

Rubric evaluation (weighted score: **0.000**)

Dimension	Score	Reason
D1: Historical Base Rate Analysis	0/3	No historical base rate analysis whatsoever. The trajectory does not reference Brookings data, prior administration turnover rates, or any quantitative historical analysis. The trajectory is entirely focused on finding factual evidence of a specific event rather than probabilistic reasoning.
D2: Individual Risk Assessment & …	0/3	No individual risk assessment or aggregation. The trajectory does not evaluate per-member departure signals, identify at-risk members with concrete evidence prior to the departure, or attempt any complement-product aggregation. It simply searches for and confirms a known departure.
D3: Synthesis & Causal Structure	0/3	No integration of base rates, individual risks, or catalysts into a causal argument. The trajectory is purely fact-finding–confirming a departure that already occurred–with no scenario reasoning or causal structure.
D4: Probability Derivation Traceab…	0/3	The final probability (0.99) is not derived from any analytical work. It is simply asserted after confirming the factual occurrence of Noem's departure. There is no base rate anchor, no calculation, and no traceability.
D5: Uncertainty & Self-Consistenc…	0/3	No uncertainty discussion, no sensitivity analysis, and no consistency checks. The trajectory treats the question as purely factual (the event already occurred) and never engages with probabilistic uncertainty or alternative scenarios.