WHAT SHOULD WE FORECAST? BENCHMARKING AGENTS ON EARLY QUESTION DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting benchmarks usually assume that the question has already been written. We study the upstream problem: can an agent discover which questions should be asked before the relevant events become obvious? We introduce **AnteBench**[1], a benchmark that evaluates whether an agent formulates forecasting questions in advance about the uncertainties that later lead to consequential developments. Using only news published before a cutoff date, the agent produces a list of forecasting questions about the next 30 days; after the 30-day window closes, we construct a reference list of questions from the available information and match the two lists. We use average precision (AP) to evaluate how much of the reference list the agent covers and how highly it ranks those questions in its own list. In a 30-day backtest covering Eastern Europe and Russia in April 2026, we evaluate frontier agents and find that performance differs across agents and improves as reasoning effort increases. We also show that AnteBench serves as a testbed for experimentally studying agents' abilities to gather information and formulate relevant forecasting questions from it.

1 INTRODUCTION

> *We know there are known knowns,...things we know we know. We also know there are known unknowns; we know there are things we do not know. But there are also unknown unknowns, the ones we don't know we don't know.* Donald Rumsfeld, February 12, 2002

Forecasting systems are usually evaluated on known unknowns: given a well-specified future event, assign a calibrated probability. This has enabled rapid progress in AI forecasting (Halawi et al., 2024; Karger et al., 2025; Metaculus, 2026). But the setup assumes that someone has already identified the relevant uncertainty and written it as a question.

Many real-world foresight failures happen one step earlier: the important question is never asked until the event is already underway. A system that can assign accurate probabilities to supplied questions may still be ineffective if it fails to surface the issues that decision-makers should be tracking. This creates a benchmark gap. How should we evaluate whether an agent can identify such uncertainties before they resolve?

Evaluating an agent's ability to discover forecasting questions requires a way to identify important questions missing from its list. The questions it proposes may all be useful, yet tell us little about what it overlooked. We therefore construct a reference set after the target period ends. We identify consequential uncertainties that resolved in that period and formulate questions about them. We then compare the agent's earlier list with this reference set to determine which of these uncertainties it had identified. The comparison measures how well the list anticipated uncertainties that later proved consequential.

Constructing such a reference set also requires choosing an appropriate level of specificity. An agent should not receive credit merely for identifying a broad topic. At the same time, a question

[1] Code and data: `https://anonymous.4open.science/r/antebench`.

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.

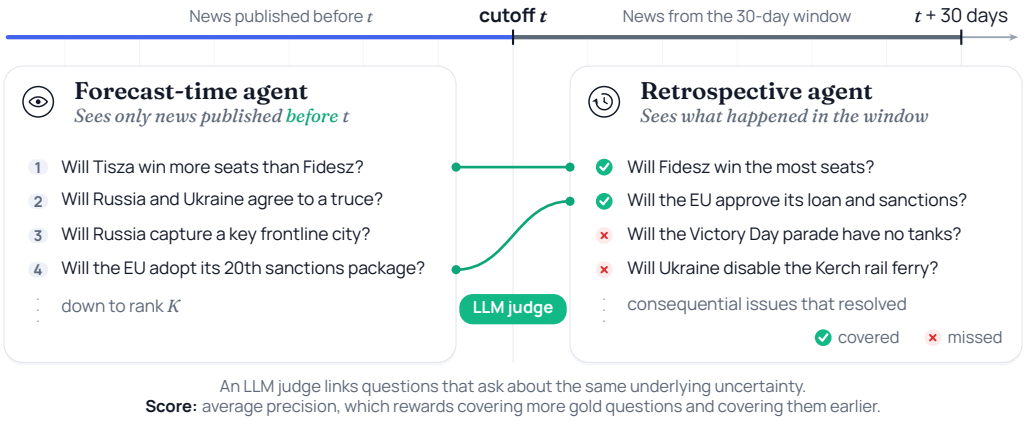


Figure 1: Overview of AnteBench. A forecast-time agent that sees only news published before the cutoff t submits a ranked list of questions. After the 30-day window closes, a retrospective agent constructs gold questions about consequential uncertainties that resolved in it, and an LLM judge matches the two lists.

constructed in hindsight should not require details that could not reasonably have been specified before the cutoff. Evaluation must therefore determine whether a question proposed before the cutoff captures the same underlying uncertainty as one constructed in hindsight.

We introduce **AnteBench** (Figure 1), a benchmark for early question discovery. At a cutoff date, the forecast-time agent sees only information available before that date and proposes a ranked list of forecasting questions about the next 30 days. After that window closes, the retrospective agent uses information available through the end of the window to construct the reference set. The forecast-time side and the retrospective side produce the same kind of object, but under different information conditions. We evaluate the forecast-time questions by matching them against the retrospective questions. AnteBench also supports forward testing, with lists frozen before the window opens.

We evaluate frontier agents on an AnteBench backtest covering Eastern Europe and Russia in April 2026. We use the retrospective questions as the reference set for evaluation and call them gold questions. Performance differs across agents. The four frontier agents occupy an overlapping band (AP@100 0.0100–0.0128) above a lower-cost control agent (0.0016), and for GPT-5.5, raising reasoning effort from none to medium increases AP@50 about five-fold. In absolute terms, the highest point estimate covers nearly 12 of the 329 gold questions at $K = 100$. This level depends on how broadly the pool is built and on the candidate budget, and by itself it does not show whether performance is high or low. Because higher reasoning effort changes both what GPT-5.5 retrieves and how it reasons over it, we then fix the news it receives, disable search, and vary only the reasoning effort used to formulate questions. More effort improves performance with news retrieved at the highest effort but barely with news retrieved without reasoning.

Our contributions are:

- We formalize early question discovery and construct AnteBench, a benchmark that makes omissions measurable. Gold questions constructed after the window show which consequential uncertainties an agent's pre-cutoff list missed, and matching questions at an appropriate level of specificity credits the same underlying uncertainty without requiring details that could not have been specified at the cutoff.
- We evaluate frontier agents, a lower-cost control agent, and five reasoning-effort settings in an Eastern Europe / Russia backtest, and find that performance differs across them.
- We show that AnteBench supports controlled experiments on what drives question discovery. When the news given to a single agent is fixed, more reasoning at question formulation improves performance with news retrieved at the highest effort but barely with news retrieved without reasoning.

2 Related Work

Early question discovery sits between forecasting benchmarks, automated question generation, open-ended event prediction, and weak-signal mining. Unlike these lines of work, AnteBench evaluates which consequential uncertainties an agent's own question list covers and which it misses, with no issue supplied in advance.

Forecasting benchmarks. Forecasting has a long evaluation tradition based on probability quality for supplied events or questions, including Brier scores and strictly proper scoring rules (Brier, 1950; Gneiting & Raftery, 2007). Recent AI forecasting benchmarks such as ForecastBench, FutureX, FutureSim, and Metaculus FutureEval evaluate increasingly agentic forecasters once questions are specified (e.g., Halawi et al., 2024; Karger et al., 2025; Zeng et al., 2026; Goel et al., 2026; Metaculus, 2026). This line of work has also surfaced limitations in consistency and benchmark design (Paleka et al., 2026; 2025). AnteBench keeps the time-bounded forecasting setup, but moves the evaluated task upstream: before assigning probabilities, a system must first decide which uncertainties are worth forecasting.

Prediction markets also evaluate forecasts on pre-specified questions, using financial incentives to aggregate beliefs about future outcomes (Wolfers & Zitzewitz, 2004), and can provide useful early signals for economic forecasting (Diercks et al., 2026; Kalshi, 2026). Their market lists are an important practical comparison point, but market availability is shaped by liquidity, settlement contracts, and platform incentives. AnteBench instead evaluates whether an agent can propose the watchlist itself.

Question generation and open-ended prediction. Automated pipelines generate forecasting questions from text, markets, or future-event databases (Bosse et al., 2026; Karger et al., 2025; Bianchi et al., 2025; Yuan et al., 2025; Tao et al., 2025; Cheng et al., 2026), often to scale training and reinforcement-learning data for forecasting agents (Turtel et al., 2025; Chandak et al., 2026; Turtel et al., 2026a;b). Open-ended prediction benchmarks similarly ask models to forecast outcomes or timelines from a topic, background, or question (Guan et al., 2026; Ye et al., 2024; Zeng et al., 2026; Yuan et al., 2025; Tao et al., 2025; Bianchi et al., 2025). These efforts expand forecasting beyond static question sets, but they usually condition on a visible topic, seed article, event background, or already-written question. Their evaluation typically covers only the questions or forecasts they produce. AnteBench instead scores a pre-cutoff list against gold questions constructed later from issues that materialized, so it also measures what the list omits.

Weak signals and automated horizon scanning. Weak-signal mining and automated horizon scanning study how precursors, anomalies, and underappreciated evidence can be extracted from large unstructured streams using text mining, NLP, retrieval, and machine learning (Ansoff, 1975; Yoon, 2012; Kim & Lee, 2017; Griol-Barres et al., 2020; Mühlroth et al., 2023; Ebadi et al., 2022; Zhang & Evans, 2025). Recent AI foresight work formulates horizon scanning as document retrieval, relevance filtering, weak-signal extraction, and explanatory comment generation, increasingly using LLMs, neural retrieval, and agentic workflows (Ishigaki et al., 2022; Schmidt et al., 2025; O'Keefe & Marshall, 2025; Hajikhani, 2026; Rozanec et al., 2024). AnteBench does not score precursor extraction itself, only whether an agent turns precursors into specific questions about uncertainties that later resolved. The gold records still keep the precursor evidence, so humans can audit whether a gold question was naturally askable at the cutoff.

3 Methodology

3.1 Evaluation Target: Early Question Discovery

An agent θ combines an underlying LLM, which we call the model, with a prompt, available tools, and an execution procedure. An evaluation instance specifies an area a, a cutoff time t, and a 30-day target window

$$T_t = [t, t + 30 \text{ days}),$$

where all boundaries are UTC start-of-day timestamps and intervals are half-open. At cutoff t, the forecast-time agent sees only a time-bounded information environment $\mathcal{D}_{<t}$ and produces a ranked

list of questions

$$Q_\theta(a, t) = (q_1, \ldots, q_k).$$

We refer to questions on the forecast-time side of the evaluation as candidate questions. Each question is intended to identify an uncertainty that may resolve during T_t. Early refers to asking before the uncertainty resolves, not before it draws attention, so questions about issues already in the news are eligible.

After T_t closes, we construct a gold question set $G(a, t) = \{g_1, \ldots, g_n\}$ retrospectively. This retrospective question-generation stage can inspect evidence from T_t and is instructed to generate questions that should have been asked at t, given what actually materialized. This does not mean recovering every possible useful question. The gold set is anchored on later-observed consequential outcomes; its role is to express those outcomes at the question granularity that would have been natural at the cutoff.

The question-level target is deliberate. Atomic news events are too fine-grained: one well-formed forecast question can cover a sequence of related reports, such as a sanctions-waiver extension, expiry, non-renewal, later extension, and pressure from other governments. Broad policy episodes avoid some over-counting, but their boundaries are subjective. AnteBench therefore defines the evaluation object as a forecast question on both sides. The two sides are meant to differ only in what they can see: past-only evidence on one side, hindsight on the other.

A judge defines a match relation $m(q, g) \in \{0, 1\}$ between candidate questions and normalized gold questions. For the top-k prefix Q_θ^k, let $c(k) = \big|\{g \in G : \exists q \in Q_\theta^k,\ m(q, g) = 1\}\big|$ be the number of gold questions covered (with $c(0) = 0$), and let $p(k)$ be the fraction of those k candidates that match at least one gold question. Our primary metric is the average precision of the ranked candidate list,

$$\mathrm{AP} = \frac{1}{|G|} \sum_{k=1}^{K} \big(c(k) - c(k-1)\big)\, p(k),$$

which credits an agent both for covering gold questions and for covering them early in the ranking. We evaluate AP at the candidate budget K, written $\mathrm{AP}@K$, and report the mean $\pm$ one standard error across runs. A full-precision list in which each candidate matches one distinct gold question attains $\mathrm{AP} = K/|G|$, about 0.30 at $K = 100$ against the 329-question pool. This is the relevant one-to-one budget ceiling, although the matcher can map one candidate to more than one related gold question; it is therefore a reference rather than a strict global bound. Its two components are reported directly in the precision–recall view: coverage (recall) $\mathrm{Coverage}@K = c(K)/|G|$ and precision $p(K)$. Because many useful but unrealized forecast-time questions are unlabeled, precision is a proxy measured against the labeled gold pool rather than full precision.

Two notes on interpretation apply throughout. First, the gold pool is designed for high recall. It retains every issue that the retrospective pipeline recovers more than once, including many low-salience ones, rather than a curated shortlist of headline events. Low salience refers to how much attention an issue drew, not to whether it was consequential. High recall is a design goal rather than a verified property, because further mining runs would still add questions (Appendix C.6). Coverage at a fixed budget is therefore recall against this pool, and its absolute level should be read together with the pool and the candidate budget, since an agent could always raise coverage by submitting more questions at the price of a longer, less precise watchlist. AP is the primary metric because it also rewards order: a gold question first matched at rank k contributes $p(k)/|G|$, so matches placed early count more than matches found deep in a long list. Adding candidates never lowers AP, which is why the candidate budget K is fixed in advance. Second, the highest performance achievable from the same pre-cutoff information is unknown, so absolute scores alone cannot show whether performance is high or low. We therefore compare how performance changes across agents, reasoning settings, and the information given to the agent, at matched budgets and against the same fixed pool, to study what drives question discovery.

3.2 PIPELINE

Forecast-time question generation. For each agent configuration, area, and cutoff, the forecast-time agent searches only pre-cutoff news through a cutoff-bounded Perigon[2] wrapper. It receives the target window T_t and submits an open-ended ranked list of consequential, concrete, verifiable questions. The harness repeats the open-generation prompt with previously submitted questions included as exclusions until a candidate budget is reached or a repetition cap is hit. The agent orders each pass from the most likely and consequential candidate downward, and later passes are appended, so ranks are positions in the merged list. Each cutoff is stateless: we do not carry forward a persistent watchlist or memory from earlier cutoffs. This makes agent comparisons simpler and avoids conflating question discovery ability with the management of accumulated state.

Retrospective question generation. The retrospective stage uses the same 30-day target window but a different information condition. After T_t closes, the agent mines later news for consequential issues that materialized during the window and expresses them as retrospective questions that would have represented those issues at cutoff t. We run this mining loop five times, each until the agent declares its list complete, then embed and cluster the pooled output to consolidate rephrasings of the same issue, and expand each recurrent issue into a structured gold record containing a normalized yes/no question, its timing and precursor signals, and the matching guidance given to the judge: positive and negative examples and a boundary note saying where the core watchlist issue ends and neighboring causal context begins. A record-level duplicate audit then merges records that still describe the same uncertainty, keeping one existing wording per issue. The result is a fixed gold pool that keeps all deduplicated records rather than a headline shortlist.

Question-to-question judge. Matching is a two-stage offline procedure with no live tools. An embedding prefilter first retains candidate–gold pairs whose questions are semantically close; each surviving pair is then judged individually by an LLM that sees the candidate question, its rationale, and selected fields of the gold record. Appendix A.1 lists the gold-record fields given to the judge. A match means the candidate question targets the same core consequential uncertainty as the gold question, including close implementation milestones of the same issue, but not merely the same actor, strategic theme, causal mechanism, warning, precondition, or downstream consequence. For example, if a gold record asks whether Russia and Ukraine will agree to an Orthodox Easter ceasefire, a candidate asking whether Putin will declare an Easter truce matches, but one asking whether the truce will hold does not (Appendix A.2). Polarity does not affect a match, because AnteBench scores whether the issue was on the watchlist, not whether its outcome was predicted. The one-pair-at-a-time design keeps each verdict independent of unrelated candidates and makes every accept or reject auditable in isolation. A separate near-miss audit channel also surfaces close rejected pairs: cases where an agent found the right neighborhood but asked about a different uncertainty from the gold question (for example, a warning rather than the event it warned about, or a downstream consequence rather than the main decision). These are not counted as coverage, but they help check whether the judge is too strict or too lenient.

3.3 EXPERIMENTAL SETUP

All results in this paper come from a single backtest instance, Eastern Europe / Russia at an April 1, 2026 00:00:00 UTC cutoff, with target window $[\text{Apr } 1, \text{May } 1)$, 2026. The retrospective stage yields a pool of 335 deduplicated gold records, of which 329 describe questions that resolve inside the target window and form the evaluation pool; the retrospective stage and the offline judge both use DeepSeek v4 Pro with xhigh reasoning. The judge prefilters candidate–gold pairs by embedding similarity (`text-embedding-3-small`, cosine ≥ 0.60) and judges each surviving pair one-to-one. Because pairs below 0.60 are not judged exhaustively, reported scores are conditional on this cutoff.

We evaluate four frontier forecasting agents (GPT-5.5, Claude Opus 4.8, Claude Fable 5, and GLM 5.2) and a lower-cost agent (GPT-5-mini) included to check that the metric separates weaker capability from the frontier band. We repeat question generation five times for each agent configuration. We

[2]Perigon is a commercial news API that supports article search over title, description, and content, with date, source, language, and metadata filters.

Agent	AP@100	Coverage@100
Claude Fable 5 (high)	0.0128 $\pm$ 0.0015	0.035 $\pm$ 0.004
GPT-5.5 (xhigh)	0.0111 $\pm$ 0.0012	0.033 $\pm$ 0.003
Claude Opus 4.8 (high)	0.0100 $\pm$ 0.0015	0.029 $\pm$ 0.002
GLM 5.2 (xhigh)	0.0100 $\pm$ 0.0013	0.026 $\pm$ 0.003
GPT-5-mini (high)	0.0016 $\pm$ 0.0003	0.016 $\pm$ 0.001

Table 1: Agent comparison at the $K = 100$ candidate budget for the Eastern Europe / Russia April 2026 main experiment. Parentheses give each model's reasoning setting. Entries are mean $\pm$ one standard error across five runs. Rows are ordered by AP point estimate (Claude Opus 4.8 and GLM 5.2 tie at the shown precision and are ordered by Coverage@100); the experiment does not resolve an ordering among the frontier agents. AP@100 is the average precision of the ranked candidate list; Coverage@100 is the fraction of the 329 gold questions covered.

refer to each complete execution as a run. A run may include multiple generation passes. We truncate all candidate lists at $K = 100$ against this fixed 329-question gold pool. Claude models use the Anthropic provider through OpenRouter with adaptive thinking at the default high effort; GPT-5.5 uses the direct OpenAI provider at xhigh reasoning effort. Because DeepSeek v4 Pro is used both to construct the gold pool and to judge matches, DeepSeek-family agents are excluded from the main comparison; their scores are reported for reference in Appendix F. Appendix B gives the full configuration.

Forward testing. AnteBench runs in two modes under one scoring rule. Our main experiment uses *backtesting*: forecast-time agents are restricted to the cutoff-bounded news corpus, and the retrospective stage reconstructs gold questions after a past window closes. In *forward testing*, question lists are submitted and frozen before the window opens, which removes any concern that submissions were shaped by later knowledge; because the cutoff is the present, live sources cannot leak future information, and our deployment provides agents with live web search through Exa. The deployment runs daily with the same 30-day evaluation window: forecast-time agents generate question lists, the identical retrospective stage and judge build gold questions for the window that has just closed, and a public leaderboard is updated. Backtesting doubles as the development and calibration harness for the live deployment. Appendix G describes how the news tools enforce the cutoff and lists each evaluated model's documented training cutoff, which the tools cannot restrict.

4 Results

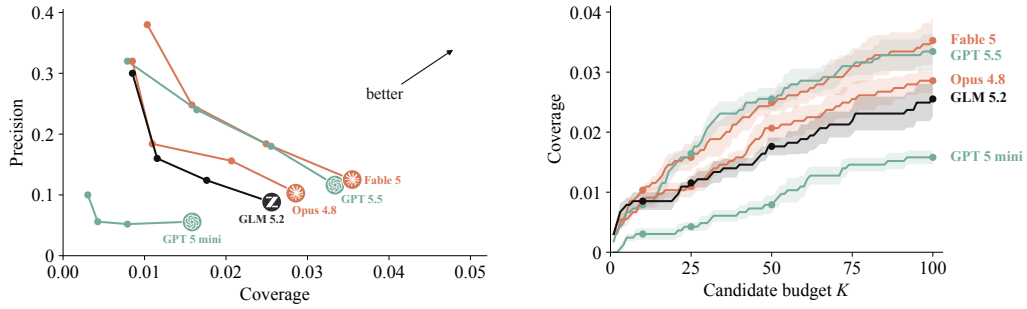
4.1 Quantitative Results

Table 1 reports agent performance at the $K = 100$ candidate budget for five agents: four frontier agents and a lower-cost control agent. The four frontier agents occupy an overlapping performance band. Claude Fable 5 has the highest point estimate, but with five runs and only about 12 covered gold questions per run, the experiment does not resolve pairwise ordering within this band. GPT-5-mini achieves substantially lower AP@100 (0.0016) than the four frontier agents (0.0100–0.0128), so the evaluation distinguishes this lower-cost control from the frontier group. AP additionally rewards agents that recover gold questions earlier in the ranking.

The question lists also differ across agents: at cosine threshold 0.80, mean Jaccard similarity is 0.340 between runs of the same agent and 0.115 between runs of different agents (Appendix C.5).

Because the gold pool is designed for high recall (Appendix A.1), absolute coverage depends on how the pool is built and on the candidate budget. For every agent, most submitted questions match no gold question, and many gold questions are first matched late in the list. Counting each run separately, 69 of the 228 covered gold questions (30.3%) are first matched beyond rank 50 (56 of 202, or 27.7%, for the four frontier agents). Figure 2a shows the precision–recall tradeoff as the top-K prefix grows from 10 to 100 candidates, and Figure 2b traces coverage at every budget up to $K = 100$.

As K increases, agents recover more gold questions while precision generally falls: later candidates are increasingly unlikely to match a gold question.



(a) Precision and coverage

(b) Coverage by candidate budget K

Figure 2: Precision and coverage of the ranked candidate lists for all five agents. Each agent runs at the reasoning setting listed in Table 1. Coverage is the fraction of the 329 gold questions matched by at least one generated question in the top-K prefix; precision is the fraction of generated questions in that prefix that match at least one gold question. (a) Points correspond to $K \in \{10, 25, 50, 100\}$, and lines connect the four operating points for each agent. (b) Coverage at every K from 1 to 100; lines show the mean over five runs, shaded bands show $\pm$ one standard error, and markers indicate the four operating points in (a).

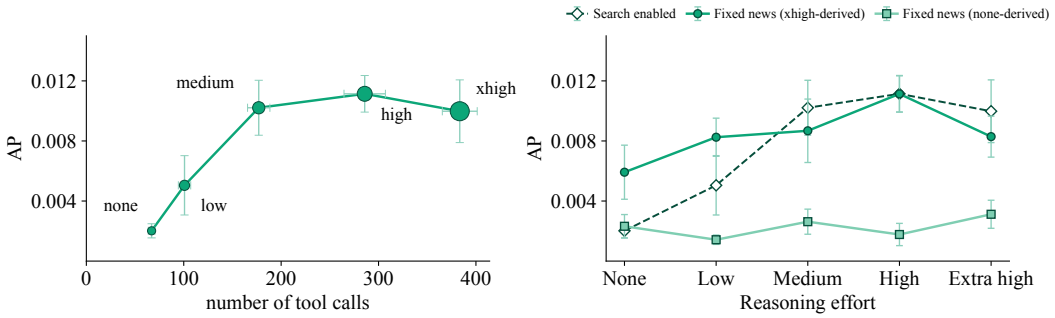


Figure 3: Reasoning effort and question discovery for GPT-5.5. Left: AP@50 against realized tool calls in the search-enabled effort sweep. Points and vertical bars show the mean $\pm$ one standard error over five runs; horizontal bars show the standard error of tool calls. Right: AP@50 against reasoning effort. The dashed line shows the search-enabled runs of the left panel. The solid lines show runs that receive a fixed news set, retrieved in the first generation pass of a none- or xhigh-effort search-enabled run, with search disabled and the effort applied at generation. Points and error bars show the mean $\pm$ one standard error over five runs or five news sets.

4.2 Inference Scaling

We vary GPT-5.5's reasoning effort over five levels, from none to xhigh, at a fixed budget of $K = 50$ questions, holding the model, area, cutoff, gold pool, and judge fixed. As reasoning effort increases, the number of tool calls (news searches and article reads) grows monotonically, and AP@50 rises about five-fold from none to medium. Beyond medium, tool calls grow by up to 2.2 times, but AP stays at about the same level (Figure 3, left).

Raising reasoning effort changes two things at once: how much news the agent searches for and reads, and how much it reasons when turning what it has read into questions. This sweep alone therefore cannot tell whether the improvement comes from gathering more relevant information or from formulating better questions. The next section separates the two experimentally.

4.3 Separating Information Gathering from Question Formulation

To separate the ability to gather relevant information from the ability to formulate relevant questions, we generate questions from fixed news. From the five none-effort and five xhigh-effort search-enabled

runs of Section 4.2, we take the news articles each agent retrieved in its first generation pass and use them as a fixed news set. Each none-derived set contains 67–176 articles, and each xhigh-derived set 695–1,323. We give each set to GPT-5.5 with search disabled and have it produce question lists at each of the five generation efforts, from none to xhigh. Generation follows the same sequential procedure as the search-enabled runs in Section 4.2, repeating open-ended passes until K questions are collected, and we score the lists at $K = 50$. This yields a comparison of two news sources and five generation efforts that varies the gathered information (the source of the news set) independently of the reasoning applied to it when forming questions (the generation effort), with 50 runs in total (Appendix C.2).

Both information gathering and question formulation affect performance (Figure 3, right). With none-derived sets, AP@50 stays between 0.0014 and 0.0031 at every generation effort, close to the search-enabled none runs (0.0020). With xhigh-derived sets, AP is already 0.0059 at none, about three times the search-enabled none runs, and it rises further with generation effort, reaching 0.0111 at high, the same as the search-enabled high runs. Raising generation effort thus clearly improves AP for xhigh-derived sets but hardly at all for none-derived sets. Moreover, at low effort or below, xhigh-derived sets yield higher AP than the search-enabled runs at the same effort, suggesting that the performance of low-effort agents is constrained mainly by the information they gather.

This comparison has two limitations. First, xhigh-derived sets are about ten times larger than none-derived sets, so the effect of the news source reflects both the amount and the content of the information. Second, because the input is fixed, the agent cannot adapt its later searches to the questions it has already formed, as the search-enabled agent can. The experiment is therefore not a full decomposition of the agent's behavior; it isolates two stages, information gathering and question formulation. More generally, AnteBench lets us study the stages of question discovery separately by controlling what information the agent receives and how it generates questions.

4.4 Hinted Recovery from Recorded Precursors

An agent may miss a gold question because the precursor evidence is sparse, because it does not turn the available evidence into the question, or both. To probe these explanations, we gave the generation agent the gold questions' own provenance. For a seeded random sample of 160 of the 329 gold questions, we re-ran the standard forecast-time setup with the same prompt and cutoff-bounded news tools. We appended to the prompt the pre-cutoff precursor signals that the retrospective pipeline had recorded for that question (signal text and date only; all outcome-describing fields are withheld). The agent (GPT-5.5, medium) produces one $\sim$12-question pass per gold question, judged against that question's record by the official judge.

The agent recovers 29 of 160 sampled questions (18.1%). Among the 21 questions found by at least one main-experiment run, it recovers 13 (61.9%); among the 139 questions found by none, it recovers 16 (11.5%). Appendix D.1 breaks recovery down by category. These rates characterize this agent configuration, not an intrinsic upper bound on what could have been asked. The signal texts were written by the hindsight-informed retrospective pipeline, so any hindsight in their wording may facilitate recovery. Pointing the agent to recorded precursors nevertheless leaves most sampled questions unrecovered.

4.5 Prediction Markets as a Question Baseline

Prediction markets and forecasting platforms maintain the closest human-curated analogue of a forecast-time watchlist: the set of questions that were live at the cutoff. We collected every Polymarket and Kalshi market and every Metaculus binary question that was listed before the cutoff and unresolved at it (excluding Kalshi's multi-leg combination markets), screened them for the benchmark's area scope with an LLM that applies the agents' own scope instructions, and matched the in-scope questions against the same 329-question gold pool with the identical judge pipeline, using each market's question text and resolution criteria. Of 90,837 platform questions, 725 are in scope: 426 on Polymarket, 138 on Kalshi, and 161 on Metaculus. Polymarket and Metaculus each cover three gold questions and Kalshi two; together, the three platforms cover four (Table 2). A single frontier-agent run of 100 questions covers 8.4–11.6 gold questions on average, about three to four times as many as any one platform's list and more than twice as many as all three combined.

Source	In scope	Gold covered	Coverage	Precision
Polymarket	426	3	0.009	0.019
Kalshi	138	2	0.006	0.058
Metaculus	161	3	0.009	0.031
All three platforms	725	4	0.012	0.029
Claude Fable 5 (mean per run, K=100)	100	∼12	0.035	0.126

Table 2: Prediction-market question lists at the cutoff versus the gold pool, under the same judge pipeline as the main results. In-scope questions were listed before the cutoff, unresolved at it, and pass an area screen that applies the agents' scope instructions. Precision is the fraction of in-scope questions that match at least one gold question. Platform lists are unranked, so AP is not reported.

This is not a claim about market accuracy, as markets forecast the questions they list quite well. Instead, it is about question supply. Platform questions cluster on scheduled events: 145 of the 185 in-scope Polymarket questions that close in April concern the Hungarian and Bulgarian parliamentary elections, and every platform match is one of these two elections or the Zaporizhzhia power-supply question. Most Metaculus questions are long-horizon (3 of 161 close in April). Listings are driven by trader and forecaster demand, not by coverage of the benchmark's reference set.

5 Discussion and Limitations

Deciding which questions to ask is a capability distinct from forecasting questions that are already specified. AnteBench makes it measurable by scoring a pre-cutoff list against uncertainties that later proved consequential, and our results show that this measurement is sensitive to the agent, to its reasoning setting, and to the information it receives.

Limitations. These results come with four caveats. First, the gold pool is constructed using LLMs: repeated mining runs, clustering, and a duplicate audit stabilize it, but it contains what the retrospective pipeline reliably recovers, not an exhaustive human-authored ground truth. Second, gold construction is hindsight-informed and matching is semantic rather than exact; each record therefore carries a normalized question, precursor signals, matching boundaries, and positive and negative examples, so the benchmark's hardest judgments can be audited rather than taken on trust, and a near-miss channel exposes the matcher's boundary behavior. Third, DeepSeek v4 Pro is used both to construct the gold pool and to judge all matches. A second-judge study finds no consistent self-favoritism on the judging side (Appendix E), but an independent cross-miner check shows that DeepSeek v4 Pro's aggregate edge does not survive on gold questions recovered by a non-DeepSeek miner (Appendix F); we therefore exclude them from the main comparison and report their scores for reference only. Fourth, the evaluation covers a single benchmark instance: one region and one 30-day window in a theater with unusually dense news coverage. Absolute levels and agent ordering may not transfer to other regions, calmer periods, or longer horizons; the protocol is cheap to re-run and the forward-testing deployment accumulates further instances, but the external validity of this first instance is untested.

The finding in Section 4.3 that information gathering affects performance suggests a natural extension. The same framework can test how the questions an agent discovers change when it has access to sources beyond mainstream news articles, such as local-language social media, sensor- and satellite-based monitoring, and domain-specific event streams. Likewise, alternative generation procedures, such as first extracting signals from the news and then formulating questions from them, can be compared under the same information and question budget.

In the April 2026 Eastern Europe / Russia backtest, question discovery separated the frontier agents from a lower-cost control and improved with GPT-5.5's reasoning effort up to a medium level, and extra reasoning at generation raised AP with richer news but barely with sparse news. Whether these findings generalize beyond this instance remains open.

AI USE STATEMENT

LLMs are both the object of study and components of AnteBench: they serve as forecast-time agents, construct the retrospective gold questions, judge candidate–gold matches, and assign gold-question categories. These uses are part of the method and are specified in Section 3 and Appendices A, H, and K; the rest of this statement concerns generative AI used as a research and writing assistant. In this work, we used generative AI tools to provide feedback on the research methodology and experiments (as a brainstorming partner, to organize discussions among the authors, and to identify potential flaws in the design), to implement methods with coding agents, and to support qualitative data analysis (selecting the examples in Appendix D.3). We have not used generative AI tools to generate datasets beyond the method components above, to develop theoretical models or conceptual frameworks, to propose or refine hypotheses, to assist with translation, to clean or reformat datasets, or to interpret results; formulating mathematical claims and writing proofs are not applicable to this work. Additionally, we used generative AI tools for brainstorming, writing analysis and plotting code, creating and modifying figures, searching for and identifying relevant literature, formatting references, and editing the paper for readability. We have reviewed all AI-assisted work: we manually inspected most of the generated code, validated it with tests, and checked all reported numbers against the stored experiment outputs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

H. Igor Ansoff. Managing strategic surprise by response to weak signals. *California Management Review*, 18(2):21–33, 1975. doi: 10.2307/41164635.

Federico Bianchi, Junlin Wang, Zain Hasan, Shang Zhu, Roy, Clémentine Fourrier, and James Zou. Back to the future: Evaluating AI agents on predicting future events. Hugging Face Blog, 2025. URL https://huggingface.co/blog/futurebench.

Nikos I. Bosse, Peter Mühlbacher, Jack Wildman, Lawrence Phillips, and Dan Schwarz. Automating forecasting question generation and resolution for AI evaluation, 2026. URL https://arxiv.org/abs/2601.22444.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. URL https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Scaling open-ended reasoning to predict the future, 2026. URL https://arxiv.org/abs/2512.25070.

Pu Cheng, Juncheng Liu, and Yunshen Long. PolyBench: Benchmarking LLM forecasting and trading capabilities on live prediction market data, 2026. URL https://arxiv.org/abs/2604.14199.

Anthony M. Diercks, Jared Dean Katz, and Jonathan H. Wright. Kalshi and the rise of macro markets. Finance and Economics Discussion Series 2026-010, Board of Governors of the Federal Reserve System, 2026. URL https://doi.org/10.17016/FEDS.2026.010.

Ashkan Ebadi, Alain Auger, and Yvan Gauthier. Detecting emerging technologies and their evolution using deep learning and weak signal analysis. *Journal of Informetrics*, 16(4):101344, 2022. doi: 10.1016/j.joi.2022.101344.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Shashwat Goel, Nikhil Chandak, Arvindh Arun, Ameya Prabhu, Steffen Staab, Moritz Hardt, Maksym Andriushchenko, and Jonas Geiping. FutureSim: Replaying world events to evaluate adaptive agents, 2026. URL https://arxiv.org/abs/2605.15188.

Israel Griol-Barres, Sergio Milla, Antonio Cebrián, Huaan Fan, and Jose Millet. Detecting weak signals of the future: A system implementation based on text mining and natural language processing. *Sustainability*, 12(19):7848, 2020. doi: 10.3390/su12197848.

Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. OpenEP: Open-ended future event prediction. *ACM Transactions on Information Systems*, 44(4):1–27, 2026. doi: 10.1145/3802922.

Arash Hajikhani. Weak signal detection in quantitative foresight: A new method for large-scale data using large language models. In Henrique da Hora, Alan L. Porter, Denise Chiavetta, and Yi Zhang (eds.), *Technology Mining: Text Analytics for Evidence-Based Foresight*, pp. 99–123. Springer Nature Switzerland, Cham, 2026. URL https://doi.org/10.1007/978-3-032-10849-4_5.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 50426–50468, 2024. doi: 10.52202/079017-1598.

Tatsuya Ishigaki, Suzuko Nishino, Sohei Washino, Hiroki Igarashi, Yukari Nagai, Yuichi Washida, and Akihiko Murai. Automating horizon scanning in future studies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 319–327, Marseille, France, 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.34/.

Kalshi. Kalshi Research partners with ProCap Financial to bring prediction market data to investment research. Kalshi News, 2026. URL https://news.kalshi.com/p/kalshi-procap-partnership-prediction-market-data-investment-research. Accessed: 2026-05-13.

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*, pp. 93943–93980, 2025.

Jieun Kim and Changyong Lee. Novelty-focused weak signal detection in futuristic data: Assessing the rarity and paradigm unrelatedness of signals. *Technological Forecasting and Social Change*, 120:59–76, 2017. doi: 10.1016/j.techfore.2017.04.006.

Metaculus. FutureEval: Measuring the forecasting accuracy of AI. Metaculus FutureEval, 2026. URL https://www.metaculus.com/futureeval/. Accessed: 2026-05-13.

Christian Mühlroth, Laura Kölbl, and Michael Grottke. Innovation signals: Leveraging machine learning to separate noise from news. *Scientometrics*, 128(5):2649–2676, 2023. doi: 10.1007/s11192-023-04672-y.

Hannah O'Keefe and Chris Marshall. Comparing large language models for horizon scanning: A feature assessment. *International Journal of Technology Assessment in Health Care*, 41(S1):S27, 2025. doi: 10.1017/S0266462325101141.

Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. Consistency checks for language model forecasters. In *International Conference on Learning Representations*, pp. 32764–32822, 2025.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language model forecasters. In *International Conference on Learning Representations*, pp. 114207–114227, 2026.

Perigon. Articles API overview. Perigon API Documentation (archived 2025-07-07), 2025. URL https://web.archive.org/web/20250707200541/https://docs.perigon.io/docs/overview.

Perigon. Perigon News API: Global media from 80,000 sources. Product page, 2026a. URL https://site.goperigon.com/data-solutions/news-api. Accessed: 2026-05-13.

Perigon. Query logic & keywords. Perigon API Documentation, 2026b. URL https://perigon.io/docs/api/search. Accessed: 2026-09-26.

Jose Rozanec, Beno Šircelj, Michael Cochez, and Gregor Leban. Back to the future: Predicting causal relationships influencing oil prices. In *The Second Tiny Papers Track at ICLR 2024*, 2024. URL https://openreview.net/forum?id=TqQlYLd52W.

Lena Schmidt, Oshin Sharma, Chris Marshall, and Sonia Garcia Gonzalez Moral. Horizon scans can be accelerated using novel information retrieval and artificial intelligence tools, 2025. URL https://arxiv.org/abs/2504.01627.

Zhengwei Tao, Pu Wu, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Chengfeng Dou, Xiancai Chen, Jia Li, Linyu Li, Chongyang Tao, and Wentao Zhang. PROPHET: An inferable future forecasting benchmark with causal intervened likelihood estimation, 2025. URL https://arxiv.org/abs/2504.01509.

Benjamin Turtel, Danny Franklin, Kris Skotheim, Luke Hewitt, and Philipp Schoenegger. Outcome-based reinforcement learning to predict the future. *Transactions on Machine Learning Research*, 2025.

Benjamin Turtel, Paul Wilczewski, Danny Franklin, and Kris Skotheim. Foresight learning for SEC risk prediction, 2026a. URL https://arxiv.org/abs/2601.19189.

Benjamin Turtel, Paul Wilczewski, and Kris Skotheim. Forecasting supply chain disruptions with foresight learning, 2026b. URL https://arxiv.org/abs/2604.01298.

Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2): 107–126, 2004. doi: 10.1257/0895330041371321.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. MIRAI: Evaluating LLM agents for event forecasting, 2024. URL https://arxiv.org/abs/2407.01231.

Janghyeok Yoon. Detecting weak signals for long-term business opportunities using text mining of web news. *Expert Systems with Applications*, 39(16):12543–12550, 2012. doi: 10.1016/j.eswa.2012.04.059.

Zhangdie Yuan, Zifeng Ding, and Andreas Vlachos. Introducing FOReCAst: The future outcome reasoning and confidence assessment benchmark. In *Advances in Neural Information Processing Systems*, volume 38, pp. 27131–27156, 2025. doi: 10.52202/085713-0808.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhenwei Zhu, Tianle Cai, Xinjie Chen, Zehui Chen, Jiecao Chen, Yantao Du, Xiang Gao, Jiacheng Guo, Liang Hu, Jianpeng Jiao, Xiangsheng Li, Jingkai Liu, Shuang Ni, Zhoufutu Wen, Ge Zhang, Kaiyuan Zhang, Xin Zhou, Jose Blanchet, Xipeng Qiu, Mengdi Wang, and Wenhao Huang. FutureX: An advanced live benchmark for LLM agents in future prediction. In *International Conference on Learning Representations*, pp. 19522–19563, 2026.

Zhen Zhang and James Evans. Language model perplexity predicts scientific surprise and transformative impact, 2025. URL https://arxiv.org/abs/2509.05591.

APPENDIX

Contents

A AnteBench 14
- A.1 Full Benchmark Pipeline 14
- A.2 Example Question Records 15

B Experimental Configuration 17

C Additional Quantitative Details 19
- C.1 Retrieval and Cost Efficiency 19
- C.2 Fixed-News Experiment Details 20
- C.3 Surprisingness 20
- C.4 Forecast-Time Tool Usage 22
- C.5 Overlap Between Generated Question Lists 23
- C.6 Retrospective Mining Behavior 23

D Additional Qualitative Details 24
- D.1 Gold-Question Categories 24
- D.2 Gold-Question Hit Frequency 25
- D.3 Matching Boundary Examples 26

E Judge Robustness 29
- E.1 Choice of the Prefilter Threshold 29
- E.2 Second-Judge Agreement 30

F DeepSeek Reference Results 31
- F.1 Reference Scores 31
- F.2 Cross-Miner Authorship Check 31

G News Backend and Cutoff Enforcement 32

H Prompt and Tool Details 33

I Runtime and Cost Details 34

J Question Lists 35

K Prompt Text 58

A ANTEBENCH

A.1 FULL BENCHMARK PIPELINE

AnteBench is organized around two question-generation processes and an offline matcher. The forecast-time process is the evaluated system: given only information available before a cutoff, it proposes a ranked watchlist of questions for the next 30 days. The retrospective process constructs the gold set: after the target window closes, it uses target-window evidence to identify the questions that would have represented the consequential uncertainties that materialized. The only intended asymmetry between the two sides is the information condition. Both sides ultimately produce forecasting questions, and evaluation asks whether the forecast-time watchlist covers the gold questions.

Evaluation instances and time windows. Each benchmark instance is defined by an area a, a cutoff time t, and a target window $T_t = [t, t + 30 \text{ days})$. All timestamps are UTC and all intervals are half-open. There is no buffer day between the cutoff and the target window: an issue resolving on the cutoff date after 00:00 UTC belongs to the target window, while forecast-time evidence is restricted to material published before t. This choice matches the operational setting in which a watchlist is frozen immediately before the next reporting period begins.

Forecast-time question generation. For each area, cutoff, agent configuration, and run, the forecast-time agent searches only pre-cutoff news through the cutoff-bounded Perigon wrapper. It receives the area description and the 30-day target window, then submits an ordered list of concrete, consequential, verifiable forecasting questions. The harness can repeat this open-generation step with previously submitted questions supplied as exclusions, so later passes extend the list rather than restating the same candidates. At evaluation time, each run is truncated to a fixed candidate budget K.

Retrospective question generation. The retrospective question-generation process runs after T_t closes and constructs the gold set in four steps: retrospective candidate mining, semantic consolidation, gold record construction, and record-level deduplication.

1. **Retrospective candidate mining.** The agent is anchored on later-observed consequential outcomes rather than being asked to recover every possible useful question. Given what actually materialized during the target window, it expresses those outcomes as forecast questions that would have been natural to ask at t. In the reported experiment, we run the retrospective mining agent five times with the same model, prompt, and reasoning effort. Each mining run continues until the agent declares its list complete, subject to a hard loop cap, rather than stopping at a fixed question count. The five runs yield 2,217 raw retrospective questions.
2. **Semantic consolidation.** We embed the 2,217 raw questions with `text-embedding-3-small` and cluster them with average-linkage agglomerative clustering at cosine threshold 0.90, producing 1,403 clusters. The 428 clusters containing at least two raw submissions form the reservoir from which gold records are built. Of these, 389 contain submissions from at least two mining runs, and the other 39 contain repeated submissions within one run.
3. **Gold record construction.** The representative question of each of the 428 reservoir clusters is expanded into a structured gold record, which pairs the normalized question with the metadata the judge needs to decide whether another question targets the same uncertainty. The record-construction agent verifies when and how the issue materialized, normalizes the question into a yes/no form, removes incidental hindsight-specific detail, and defines the boundary of the core watchlist issue. Each record includes a normalized question, summary, timing information, normalization note, matching boundary note, positive examples, negative examples, precursor signals, and causal mechanisms.
4. **Record-level deduplication.** Because normalization can make records from different clusters describe the same uncertainty, we deduplicate the constructed records. Record pairs with embedding similarity of at least 0.70 are audited by an LLM judge that applies the main judge's match semantics in both directions, and a pair counts as a duplicate if either direction

matches. The judge audits 619 of the 631 such pairs, because the other 12 already lie within one duplicate component. Each duplicate component is collapsed to a single record. A directed match graph identifies the most canonical records, and when several remain, an LLM selects exactly one *existing* record, so no new phrasing is introduced. The retained record sums the raw-submission support of the records it absorbs. This step removes 93 duplicates across 77 components and reduces the 428 records to 335.

The final Eastern Europe / Russia gold pool retains all 335 deduplicated records, ranked by merged raw-submission support with earliest-source-rank and record-id tie breaking, and no manual per-question exclusions or promotions. Applying the standard temporal eligibility rule (resolution inside $[t, t + 30 \text{ days})$) leaves 329 records as the evaluation pool. We preserve the original construction rank in the released records and assign the eligible questions consecutive evaluation ranks from 1 to 329.

Offline matching. The matcher receives forecast-time questions and the fixed pool of gold records for the same area and target window. It has no live tools and cannot search for new evidence. Matching runs in two stages. An embedding prefilter (`text-embedding-3-small`) retains candidate–gold pairs whose question texts have cosine similarity at least 0.60. This recall-oriented operating point retained 2,247 of the 2,251 positives observed during calibration (99.8%) while discarding most unrelated pairs; it does not guarantee that no true match lies below the cutoff. Each surviving pair is then judged independently. The judge receives one forecast-time question with the agent's rationale, and one gold record restricted to its resolution date, normalized question, outcome summary, matching boundary note, and positive and negative examples; precursor signals and causal mechanisms are withheld. The match criteria are stated in terms of the question, and the rationale serves as context for deciding whether the candidate engages the specific scenario rather than a broad label. The judge returns a strict-JSON verdict (match or no match) with a short rationale. A strict match means that the forecast-time question identifies the same core consequential uncertainty as the gold question, even if the wording or polarity differs. Questions that share only actors, broad topics, causal context, warnings, preconditions, or downstream consequences are rejected. A separate near-miss audit channel records close rejected pairs, while the primary score is computed from unique gold questions matched within the top-K forecast-time prefix.

A.2 Example Question Records

The forecast-time side and the retrospective side both produce forecast questions, but their records have different roles. Forecast-time candidates are lightweight watchlist items produced before the target window; gold records are richer objects, constructed after the window closes to support matching. The examples below show the contrast with a forecast-time question and the gold record that the official judge matched it to, both concerning an Orthodox Easter ceasefire.

Will Russia declare a unilateral temporary general ceasefire in Ukraine (a publicly announced halt to all offensive operations for a defined period, e.g., around Orthodox Easter on April 12), during April 2026?

Rationale.
Zelensky has proposed an Easter ceasefire and energy truce, which the Kremlin publicly rebuffed on March 31 as vague, but Peskov left the door open by demanding clarification. Putin has a track record (the April 2025 Easter truce and the May 2025 Victory Day truce) of announcing short unilateral ceasefires around symbolic dates. With peace talks paused over the Iran war, a unilateral gesture timed to Orthodox Easter (April 12) or late April is a realistic scenario that would reshape battlefield tempo and diplomatic narratives. Unlike a mutual energy-infrastructure truce, this question tracks a unilateral, all-hostilities declaration by Moscow; either polarity matters for assessing Russian intent.

Evidence.

- 2026-03-30, Kyiv Post: Zelensky proposed an Easter ceasefire and an energy truce ahead of April 12, the symbolic date this question keys on.
- 2026-03-31, The Moscow Times: the Kremlin rebuffed the proposal as vague, raising the question of whether Moscow would instead make its own unilateral move.
- 2026-03-31, Al Arabiya: the Kremlin said it had not seen concrete proposals and urged clarification, leaving the truce question open entering April.

Figure 4: Abridged example of a forecast-time question (Claude Fable 5) that the official judge matched to the gold record in Figure 5. The question text is shown as the record title, followed by the candidate rationale and evidence list.

Will Russia and Ukraine agree to a temporary ceasefire for the Orthodox Easter holiday in 2026?

Resolution and timing.
Resolved, first reported, and confirmed on 2026-04-09, when Putin declared a 32-hour ceasefire for Orthodox Easter (April 11–12) and Ukraine agreed to observe it.

Outcome summary.
At the cutoff, US-brokered peace talks were stalled and diverted by the escalating Iran conflict. Zelensky had proposed an Easter ceasefire on March 30, but the Kremlin responded dismissively, demanding clarification and insisting on a permanent settlement on Russian terms. The April 9 agreement marked the first mutually agreed pause in hostilities since the full-scale invasion began, though previous holiday truces had been extensively violated by both sides.

Matching boundary.
Include formulations asking whether Russia and Ukraine will agree to, declare, or mutually accept a ceasefire or truce for Orthodox Easter 2026, from either side's framing, using "truce" or "pause in hostilities," and with or without "temporary." The key is the agreement event, not its implementation. Exclude questions about whether the ceasefire holds (downstream consequence), a permanent ceasefire or comprehensive peace deal (different scope), an energy-infrastructure-only ceasefire (narrower domain), prisoner exchanges during Easter (different issue), or US mediation success (different actor and outcome).

Positive examples.
- Will Russia and Ukraine agree to an Orthodox Easter truce in April 2026?
- Will Putin declare a ceasefire in Ukraine for Orthodox Easter 2026?
- Will both Russia and Ukraine agree to pause hostilities for Orthodox Easter 2026?

Negative examples.
- Will the Orthodox Easter ceasefire between Russia and Ukraine hold without major violations?
- Will Russia and Ukraine agree to an energy infrastructure ceasefire in April 2026?
- Will the United States successfully broker a ceasefire agreement between Russia and Ukraine in 2026?

Pre-cutoff signals.
- 2026-03-30: Zelensky proposed a ceasefire for the Easter holidays and an energy truce, saying Ukraine was ready for compromises short of sovereignty.
- 2026-03-31: Kremlin spokesman Peskov dismissed the proposal, saying Moscow had not received a clearly formulated offer.
- 2026-03-31: Zelensky confirmed the truce proposal had been formally submitted and would be discussed with the US negotiating team.
- 2026-03-31: Russia urged Ukraine to clarify the Easter truce proposal while warning that delay would come at a higher price.

Figure 5: Abridged example of a structured gold record used for offline matching. The record defines the normalized uncertainty, the observed resolution, acceptable paraphrases, rejected neighboring questions, and pre-cutoff signals that made the issue naturally askable.

B Experimental Configuration

This appendix records the configuration of the main experiment. We separate the shared benchmark instance from the stage-specific settings used for forecast-time question generation, retrospective question generation, and offline matching.

Shared setting	Value
Area	Eastern Europe / Russia
Cutoff	2026-04-01 00:00:00 UTC
Target window	[2026-04-01, 2026-05-01) UTC

Table 3: Shared benchmark instance used for the main results. All timestamps are UTC start-of-day timestamps, all intervals are half-open, and there is no buffer day between the cutoff and target window. Only material published strictly before the cutoff is available to forecast-time agents.

Model	Provider route	Reasoning setting	Documented cutoff
DeepSeek v4 Pro	OpenRouter, DeepSeek pinned	xhigh	not documented
GPT-5.5	OpenAI direct	xhigh	2025-12
Claude Opus 4.8	OpenRouter, Anthropic pinned	high (adaptive thinking)	2026-01
Claude Fable 5	OpenRouter, Anthropic pinned	high (adaptive thinking)	2026-01
GLM 5.2	OpenRouter, Z.ai pinned	xhigh	not documented
DeepSeek v4 Flash	OpenRouter, DeepSeek pinned	xhigh	not documented
GPT-5-mini	OpenAI direct	high (highest available)	2024-05

Table 4: Forecast-time question-generation model settings. Reasoning settings are the effort levels applied by each provider. The Claude requests enabled adaptive thinking without setting an effort, so Anthropic's default (high) applied. Every Claude Fable 5 request was served by Claude Fable 5; none fell back to another model. The two DeepSeek rows configure the reference agents of Appendix F. Documented cutoff is the provider-stated training-data cutoff (Appendix G discusses parametric knowledge). Each agent using these model settings is evaluated over five runs at top-100 candidates against the same 329-question gold pool. Each run uses cutoff-bounded Perigon search/read tools and sequential generation passes with previously submitted questions supplied as exclusions; candidate ranks are the global positions in each run's merged submission list.

Retrospective stage setting	Value
Mining model	DeepSeek v4 Pro, OpenRouter route, DeepSeek provider pinned, xhigh reasoning.
Mining budget	Five runs with the same model, prompt, and reasoning effort, each continuing until the agent declares completion (loop cap 200; the first run was extended to a 100-loop cap after exhausting an initial 30-loop cap), yielding 2,217 raw retrospective questions.
Raw clustering	`text-embedding-3-small`; average-linkage agglomerative clustering at cosine threshold 0.90; 1,403 clusters.
Construction reservoir	428 clusters with at least two raw questions.
Record construction	DeepSeek v4 Pro xhigh expands each selected cluster representative into a normalized record for the judge.
Record deduplication	Record pairs with embedding similarity at least 0.70 are audited bidirectionally by an LLM judge (619 of 631; the other 12 already lay in one duplicate component) with the main-judge match semantics; 77 duplicate components collapse 428 records to 335, keeping one existing wording per component (LLM representative selection when several representatives remain).
Final gold pool	All 335 deduplicated records, ranked by merged raw-submission support with source-rank and record-id tie breaking; 329 resolve inside the target window and are reranked 1–329 for evaluation.

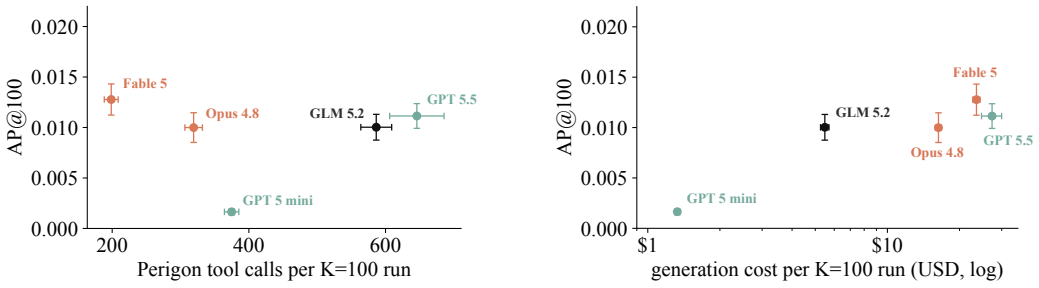
Table 5: Retrospective question generation parameters. This stage is hindsight-informed but still records precursor signals, which the construction prompt requires to predate the cutoff (29 of the 4,279 recorded signals are dated on or after it; the hinted-recovery experiment uses only earlier-dated signals, and the judge never sees signals).

Matching setting	Value
Judge model	DeepSeek v4 Pro, OpenRouter route, DeepSeek provider pinned, xhigh reasoning.
Prefilter	`text-embedding-3-small` cosine similarity between the candidate question and the gold `normalized_question`; pairs with similarity ≥ 0.60 proceed to the LLM judge.
Judge granularity	One candidate–gold pair per judge call (no batching).
Judge input	Candidate question and its rationale; gold resolution date, normalized question, summary, boundary note, and positive and negative examples. Precursor signals and causal mechanisms are withheld.
Output	Strict JSON verdict (match or no match) with a rationale; no tool call; temperature 0, at most 4,000 tokens.
Inputs	Forecast-time questions from the fixed $K = 100$ snapshots and the 329-record eligible gold pool for the same area and target window.
Tool access	None; the judge cannot search Perigon or the live web.
Audit channel	A separate near-miss audit judge is used only for qualitative examples, never for scores.
Primary output	Accepted forecast–gold matches with rationales; coverage counts unique gold questions matched within the top-K forecast-time prefix.

Table 6: Offline question-to-question matching parameters used for the main results.

C Additional Quantitative Details

C.1 Retrieval and Cost Efficiency



(a) Total news-tool invocations

(b) Estimated generation cost (log scale)

Figure 6: AP at $K = 100$ as a function of total Perigon search and article-read calls per run (left) and estimated generation cost per run (right), for all five agents. Points show mean $\pm$ one standard error over five runs on both axes. Costs are estimated from the recorded input, cache-read, and output tokens at each provider's list price around the time of the runs, including GPT-5.5's long-context surcharge; output tokens already include reasoning tokens. Judge costs are excluded.

Figure 6 shows that the candidate budget alone hides large efficiency differences. Across the overlapping frontier performance band, total news-tool use varies more than three-fold: Claude Fable 5 makes about 200 search and article-read calls per run, while GPT-5.5 makes about 650. Generation cost spans more than an order of magnitude, from about $1 for GPT-5-mini and $5 for GLM 5.2 to $16 for Claude Opus 4.8, $24 for Claude Fable 5, and $27 for GPT-5.5 (Table 19). GPT-5-mini is cheap and uses a typical number of news-tool calls, but converts far fewer of them into matched questions (5.6 matched candidates per run of 100, versus 8.8–12.6 for the frontier agents). Because reasoning settings do not reliably force proportional hidden-reasoning usage, and provider pricing, caching, and model snapshots can change, these curves should be treated as single-instance measurements of realized generation compute rather than stable agent rankings.

C.2 Fixed-News Experiment Details

Each fixed news set in Section 4.3 consists of the news articles that one search-enabled run of Section 4.2 retrieved in its first generation pass, taken from its five none-effort and five xhigh-effort runs. We keep only articles published before the cutoff and remove duplicates. Each article appears with its title, publication date, source, URL, the search queries that retrieved it, its summary (up to 2,000 characters), and, if the agent read it, its text (up to 16,000 characters). Each set is capped at 220,000 tokens; one xhigh-derived set exceeds the cap, and its remaining articles are dropped. The agent receives the forecast-time prompt with the description of the search tools replaced by the news set, and it cites articles by their identifiers in the set. Generation repeats open-ended passes, each listing the previously submitted questions as exclusions, until 50 questions are collected, with at most ten passes; every run reached 50. The lists are judged with the same judge and prompt as the main results. Table 7 reports AP@50 for each news source and generation effort, together with the search-enabled runs.

	None	Low	Medium	High	Extra high
Search enabled	0.0020 $\pm$ 0.0005	0.0050 $\pm$ 0.0020	0.0102 $\pm$ 0.0018	0.0111 $\pm$ 0.0012	0.0100 $\pm$ 0.0021
Fixed news (none-derived)	0.0023 $\pm$ 0.0008	0.0014 $\pm$ 0.0003	0.0026 $\pm$ 0.0008	0.0018 $\pm$ 0.0007	0.0031 $\pm$ 0.0009
Fixed news (xhigh-derived)	0.0059 $\pm$ 0.0018	0.0082 $\pm$ 0.0013	0.0087 $\pm$ 0.0021	0.0111 $\pm$ 0.0012	0.0083 $\pm$ 0.0014

Table 7: AP@50 of GPT-5.5 by reasoning effort. For the fixed-news rows, search is disabled and the effort applies at generation. Entries are mean $\pm$ one standard error over five runs (search enabled) or five news sets (fixed news).

C.3 Surprisingness

Coverage treats all gold questions equally. We also compute a news-attention surprisingness score for each gold question. For each question, a query-generation model writes 20 Perigon Boolean queries using 3–5 nouns per query. We count matching articles in a 7-day pre-resolution window and a 7-day post-resolution window, then score each query by the natural-log ratio

$$\ln\left(\frac{\mathrm{post}+1}{\mathrm{pre}+1}\right).$$

Here pre and post are the pre- and post-resolution article counts for that query. Queries with fewer than 5 post-event articles are excluded, and the gold-question score is the macro mean of the remaining query scores. The metric is a rough proxy for how much the issue's news attention rose when it materialized, and we do not treat it as a measure of importance. The pre window is the week before each question's resolution date, which for 82% of gold questions lies entirely after the cutoff, so the score is not computed from the news the agents could search. Issues that drew little attention just before resolving, however, plausibly also drew little coverage before the cutoff, so an association between surprisingness and discovery may partly reflect how much news an issue received before it materialized. We therefore use surprisingness only as a secondary, descriptive measure, not as evidence of a causal relationship.

Figure 7 shows the complementary per-question view. The smooth is non-monotonic, but it does not establish an inverted-U relationship. The clearer result is that highly surprising questions are rarely discovered: the 111 gold questions with surprisingness score at least 2.0 collect only 62 hits across $111 \times 25 = 2{,}775$ question–run pairs at $K = 100$, and 94 of the 111 receive no hits from any run (Table 8). Inspecting the individual questions suggests a descriptive difference. Questions about scheduled decisions, such as elections, are found more often, while many missed high-surprise questions concern wartime decisions, operational incidents, sabotage discoveries, or other fast-emerging developments whose pre-cutoff traces are sparse in ordinary news. A category breakdown in Appendix D.1 shows a similar pattern.

Figure 8 breaks down the per-question hit-frequency plot from Figure 7 by forecast-time agent. Each panel uses the same 329 gold questions, but the y-axis is now the percentage of that agent's five runs whose top-100 list matched the question. On the high-surprise set used in Table 8 (111 gold questions with surprisingness score at least 2.0), Claude Fable 5 averages 4.2 hits per run, versus 2.0–2.4 for the other three frontier agents and 1.6 for GPT-5-mini. High-surprise discovery is thus

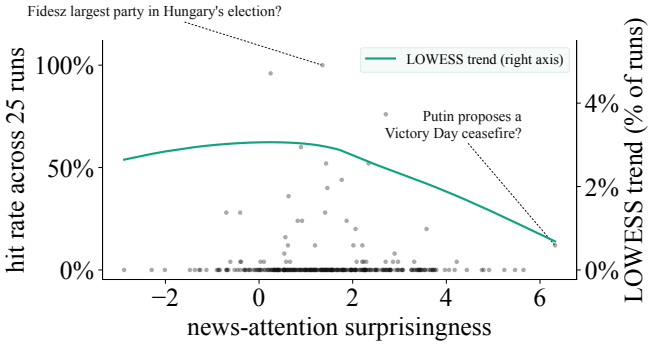


Figure 7: Gold-question hit rate as a function of news-attention surprisingness. Each semi-transparent black point is one gold question; the y-axis is the percentage of the 25 forecast-time runs (five agent configurations $\times$ five runs) whose top-100 list matched that question. The green curve is a descriptive LOWESS smooth over the question-level points; we do not fit or test an inverted-U pattern. Hit rates are low for highly surprising questions.

Agent	Coverage@100 within high-surprise questions
Claude Fable 5	0.038 $\pm$ 0.009
GLM 5.2	0.022 $\pm$ 0.007
GPT-5.5	0.020 $\pm$ 0.005
Claude Opus 4.8	0.018 $\pm$ 0.006
GPT-5-mini	0.014 $\pm$ 0.002

Table 8: Coverage@100 within the 111 high-surprise gold questions (surprisingness score at least 2.0), that is, the fraction of these questions matched by a run's top-100 list. It uses the same definition as the overall Coverage@100 in Table 1, restricted to this subset. Entries are mean $\pm$ one standard error across five runs.

not simply proportional to overall coverage: agents with similar aggregate scores differ in how much of the high-surprise tail they reach.

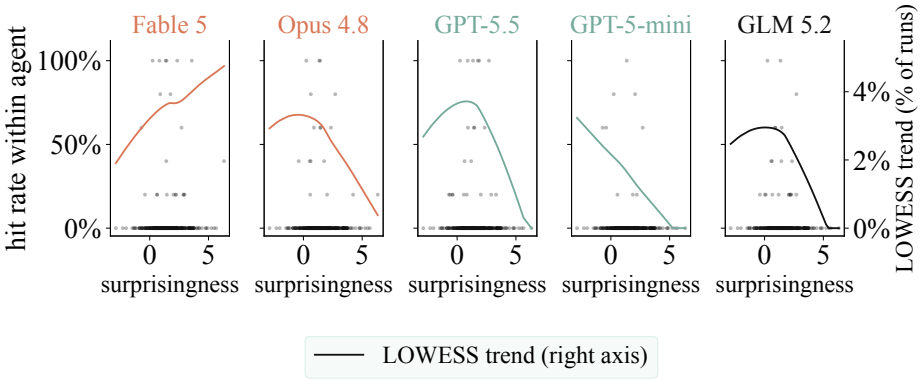


Figure 8: Agent-specific gold-question hit rate as a function of news-attention surprisingness. Each point is one gold question. Within each agent panel, the y-axis is the percentage of five runs whose top-100 forecast-time list matched that gold question, so values occur in 20-point increments. LOWESS trends (right-hand axis within each panel) are descriptive local smooths over the same question-level points, not parametric fits.

C.4 Forecast-Time Tool Usage

Figure 9 separates three measures of news retrieval through the sequential generation pass that produced the 100th candidate: search calls, compact article records returned by those searches, and full-article reads.

The retrieval pattern does not reduce to "more news access is better." Claude Fable 5 makes about 180 searches and about 20 full-article reads per run. GLM 5.2 makes the most search calls (451), whereas GPT-5.5 makes the most total news-tool calls because it also reads 259 full articles. Their performance estimates nevertheless occupy the same broad frontier band. The comparison suggests that search volume, reading volume, and the ability to turn evidence into concrete watchlist questions are distinct factors.

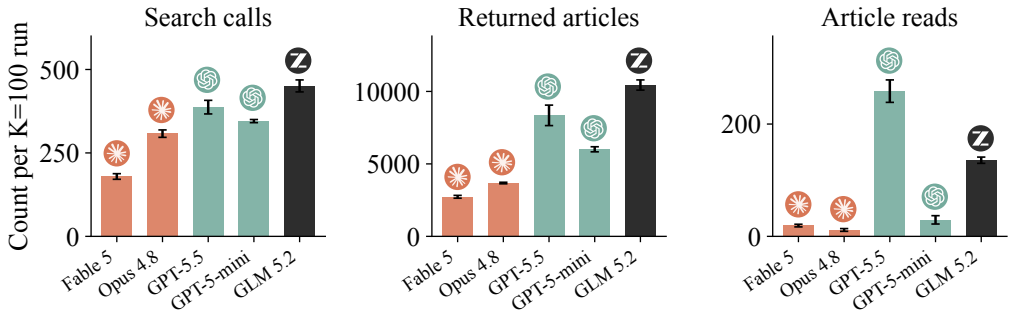


Figure 9: Forecast-time Perigon tool usage at $K = 100$ for all five agents. Bars show the mean total per run across five runs; error bars show ± 1 standard error. Search calls count `perigon_search` invocations, returned articles count compact article records returned by those searches, and article reads count `perigon_read_article` calls.

C.5 Overlap Between Generated Question Lists

We ask whether different agents generate distinct forecasting questions or largely reproduce the same lists. In particular, we compare variation across agents with variation across repeated runs of the same agent. We analyze the 25 top-100 question lists from the main experiment, comprising five separate runs of each of five agents. We embed the questions using OpenAI's `text-embedding-3-small` model. For each pair of lists, we find a maximum one-to-one matching between questions whose cosine similarity reaches a given threshold and compute the Jaccard similarity $m/(200 - m)$ from the number m of matched pairs.

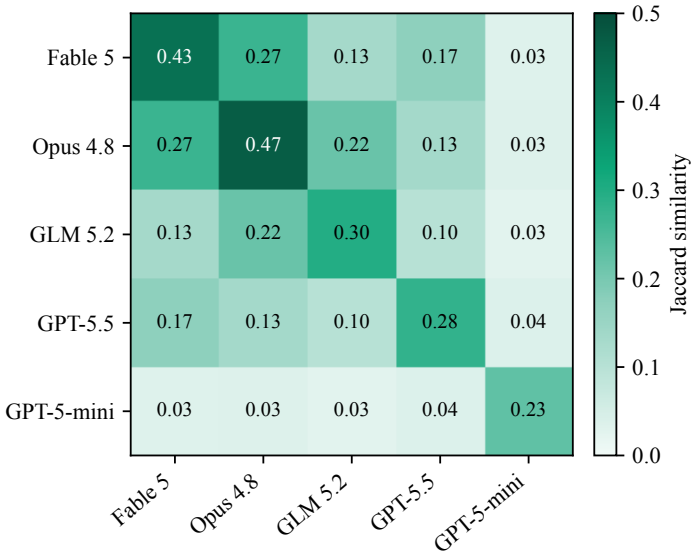


Figure 10: Jaccard similarity between question lists at cosine threshold 0.80. Diagonal cells show means over the ten pairs of runs from the same agent; off-diagonal cells show means over the 25 pairs of runs from two different agents.

At threshold 0.80, mean Jaccard similarity is 0.340 for runs of the same agent and 0.115 for runs of different agents (Figure 10). The same pattern holds at the two higher thresholds: 0.196 versus 0.048 at threshold 0.85, and 0.077 versus 0.012 at threshold 0.90. Claude Fable 5 and Claude Opus 4.8 have the highest cross-agent similarity (0.270), which may reflect their shared model family.

These results show clear agent-specific differences in the generated question lists, indicating that using multiple agents can increase question diversity. GPT-5-mini's particularly low Jaccard similarity may reflect differences in specificity or quality rather than useful diversity.

C.6 Retrospective Mining Behavior

The five mining runs used the same agent configuration, but their behavior varied widely. In each run, the agent continued until it declared its list complete rather than stopping at a fixed question count. Successful submission loops ranged from 53 to 111 per run (median 87) and submitted questions from 325 to 578 per run (median 472; 2,217 in total); loops that failed with transient provider connection errors during one batch are excluded throughout. All five runs terminated by calling `finish_question_mining` before reaching their final loop caps; the first run was extended once after exhausting an initial 30-loop cap. Exact-duplicate submissions within a run were rare (0–22 questions), so nearly all consolidation happens at the semantic-clustering stage rather than at the string level.

Figure 11 shows the cumulative number of unique questions over loop index for each run. The curves rise steadily rather than saturating: even the longest run was still producing new questions when it declared completion. This is why the pipeline combines several mining runs: no single run, and no

fixed-length run, recovers the space of consequential questions on its own. The reported ranking uses merged raw-submission support, which can include repetition within a run.

As a post-deduplication sensitivity analysis, we retain only issues with at least one source cluster represented in two or more mining runs. This removes 14 eligible records and leaves 315. It is not a reconstruction of the pipeline from the 389 cross-run clusters, because removing source clusters before record-level deduplication can change duplicate components and representative selection.

Agent	AP@100	Coverage@100
Claude Fable 5	0.0133 $\pm$ 0.0016	0.037 $\pm$ 0.004
GPT-5.5	0.0116 $\pm$ 0.0013	0.035 $\pm$ 0.003
Claude Opus 4.8	0.0104 $\pm$ 0.0015	0.030 $\pm$ 0.002
GLM 5.2	0.0105 $\pm$ 0.0013	0.027 $\pm$ 0.003
GPT-5-mini	0.0016 $\pm$ 0.0003	0.016 $\pm$ 0.001

Table 9: Post-deduplication sensitivity on the 315 eligible gold records with cross-run source support. Entries are mean $\pm$ one standard error across five runs. Existing match decisions are rescored; no new judging is introduced.

All four frontier agents retain every accepted hit under this restriction; GPT-5-mini loses one. The main-five union changes from 42 to 41 covered questions, while the conclusion that the frontier agents occupy an overlapping performance band is unchanged.

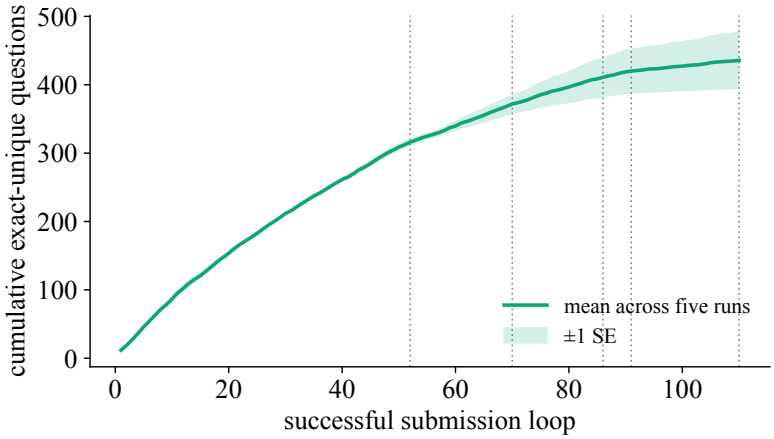


Figure 11: Cumulative unique questions submitted during five retrospective mining runs, by successful submission loop (loops lost to transient provider errors are excluded). The line and band show the mean and one standard error across the five runs (runs that declared completion earlier contribute their final counts); dotted vertical lines mark where individual runs declared completion. All five runs stopped by declaring completion rather than by the loop cap.

D Additional Qualitative Details

D.1 Gold-Question Categories

To characterize what the gold pool contains, we assigned each of the 329 eligible gold questions to one of eight fixed categories with an LLM classifier (DeepSeek v4 Pro, xhigh; one label per question from the normalized question and summary; a manual spot-check of 100 questions found the assignments largely accurate, with occasional boundary ambiguity between neighboring categories). Table 10 reports the distribution together with two discovery statistics from the main experiment: the mean hit rate (a question's matches across the 25 forecast-time runs, divided by 25, averaged within the category) and the fraction of questions matched by at least one run.

Category	n	% of pool	Mean hit rate	Hit by $\geq$1 run	Hinted recovery
Elections & domestic politics	59	17.9%	6.0%	20.3%	8/34 (23.5%)
Sanctions & economic statecraft	31	9.4%	5.7%	9.7%	6/14 (42.9%)
Energy & infrastructure	16	4.9%	4.3%	18.8%	0/8 (0.0%)
Diplomacy & negotiations	76	23.1%	2.3%	10.5%	6/44 (13.6%)
Economy, markets & society	28	8.5%	1.3%	7.1%	1/11 (9.1%)
War & military operations	40	12.2%	1.2%	10.0%	2/17 (11.8%)
Alliances & security posture	46	14.0%	0.9%	17.4%	5/18 (27.8%)
Hybrid, sabotage & espionage	33	10.0%	0.6%	6.1%	1/14 (7.1%)

Table 10: Gold-pool composition by category and per-category discovery statistics at $K = 100$. Mean hit rate averages, within each category, the fraction of the 25 forecast-time runs (five agent configurations $\times$ five runs) that matched each question. Hinted recovery gives the questions recovered out of those sampled in the hinted-recovery experiment of Section 4.4 (29 of 160 in total).

The category view shows a similar descriptive pattern. Mean hit rates reach 4.3–6.0% for categories anchored on scheduled or procedural decisions, including elections, sanctions, and infrastructure. Categories dominated by fast-emerging events, including military operations, hybrid and sabotage incidents, and shifts in alliance posture, sit at or below 1.2%. This difference is consistent with scheduled decisions being easier to formulate than fast-moving operational events, but it does not establish why the categories differ.

D.2 Gold-Question Hit Frequency

Because the gold pool contains 329 eligible questions, we do not print the full per-question matrix. Only 42 gold questions are matched by at least one of the 25 runs (five agent configurations $\times$ five runs); the remaining 287 receive no hits from any run. The matrix below covers those 42 questions, identified by their evaluation rank after removing the six out-of-window records; each cell reports how many of an agent's five runs matched that question under the official pairwise judge, and darker green indicates more runs with a match. Matched questions skew toward high-support ranks (median evaluation rank 83.5), but hits extend to rank 329, and the eight questions matched by at least 10 of the 25 runs all lie in the upper half of the ranking (ranks 2–158).

Rank	Fable 5	GPT-5.5	Opus 4.8	GLM 5.2	5-mini	Rank	Fable 5	GPT-5.5	Opus 4.8	GLM 5.2	5-mini
1	1/5	1/5	1/5	0/5	0/5	85	5/5	3/5	3/5	2/5	0/5
2	5/5	5/5	5/5	5/5	4/5	95	0/5	1/5	0/5	0/5	0/5
6	3/5	1/5	0/5	0/5	3/5	99	1/5	1/5	0/5	0/5	0/5
7	0/5	0/5	2/5	0/5	0/5	100	0/5	3/5	0/5	3/5	0/5
12	4/5	5/5	2/5	0/5	0/5	114	1/5	0/5	0/5	0/5	0/5
20	4/5	5/5	3/5	3/5	0/5	121	5/5	5/5	5/5	5/5	5/5
23	5/5	0/5	0/5	0/5	0/5	144	5/5	1/5	0/5	0/5	0/5
33	0/5	0/5	1/5	0/5	0/5	150	1/5	0/5	0/5	1/5	1/5
34	0/5	0/5	1/5	0/5	0/5	158	5/5	3/5	3/5	2/5	0/5
35	0/5	0/5	0/5	1/5	4/5	168	0/5	1/5	0/5	0/5	0/5
39	1/5	1/5	2/5	0/5	3/5	173	0/5	0/5	0/5	1/5	0/5
45	0/5	5/5	0/5	0/5	1/5	180	0/5	0/5	0/5	1/5	0/5
46	1/5	3/5	3/5	3/5	0/5	187	0/5	2/5	0/5	1/5	0/5
49	2/5	0/5	1/5	0/5	0/5	206	1/5	0/5	2/5	0/5	0/5
52	0/5	0/5	0/5	1/5	0/5	214	1/5	2/5	4/5	2/5	0/5
59	1/5	0/5	0/5	0/5	0/5	223	0/5	0/5	0/5	1/5	0/5
66	0/5	0/5	0/5	1/5	0/5	246	1/5	0/5	0/5	0/5	0/5
73	0/5	0/5	0/5	1/5	0/5	252	0/5	1/5	0/5	0/5	0/5
77	3/5	5/5	4/5	4/5	3/5	265	0/5	1/5	0/5	2/5	1/5
80	2/5	0/5	5/5	0/5	0/5	318	0/5	0/5	0/5	1/5	0/5
82	0/5	0/5	0/5	1/5	0/5	329	0/5	0/5	0/5	0/5	1/5

D.3 Matching Boundary Examples

In a separate exploratory audit of run 1 for four frontier agents and DeepSeek v4 Pro, a batched judge returned 79 strict matches and 102 near misses. After applying the official cosine cutoff of 0.60, all 79 strict matches and 98 near misses remained. Because this audit used an open-ended near-miss label and a different prompt from the official pairwise judge, we treat its outputs as a permissive diagnostic rather than an alternative benchmark score.

Agent	Strict	Strict + any audit near miss
Claude Fable 5	17/329 (5.2%)	32/329 (9.7%)
Claude Opus 4.8	14/329 (4.3%)	41/329 (12.5%)
GLM 5.2	17/329 (5.2%)	36/329 (10.9%)
GPT-5.5	14/329 (4.3%)	30/329 (9.1%)
DeepSeek v4 Pro	13/329 (4.0%)	30/329 (9.1%)

Table 11: Exploratory Coverage@100 for run 1 under the batched near-miss audit, restricted to candidate–gold pairs with cosine similarity at least 0.60. Treating every reported near miss as covered is deliberately permissive. The audit used a different prompt from the official judge, so these values are not comparable to the five-run main results.

A formal relaxed rule should distinguish strict matches, boundary near misses, episode-level near misses, and nonmatches. The audit does not make that distinction, so the combined near-miss column above should not be interpreted as a validated relaxed-matching rule. The paired examples below illustrate which semantic differences the official judge accepts or rejects. All verdicts shown are those of the official pairwise judge; a near miss is a close pair that it rejected.

Forecast-time question ✓

Will the Trump administration extend/renew its temporary waiver of Russian-oil sanctions beyond its ∼April 11 expiry (rather than letting it lapse or re-imposing sanctions) during April 2026?

Pre-cutoff signals.

- **2026-03-13:** The Trump administration temporarily eased Russian-oil sanctions to cool oil prices during the escalating Iran war.
- **2026-03-13:** Reporting stated the waiver was explicitly time-limited "until April 11," fixing a dated decision point inside the target window.
- **2026-03-14:** Britain publicly split from the US over the oil-sanctions pause, raising the stakes of the renewal decision.

Claude Opus 4.8

Gold question

Will the US Treasury Department extend or renew its temporary general license permitting purchases of sanctioned Russian oil and petroleum products beyond the April 11, 2026 expiration date?

Outcome summary.

In March 2026 the US Treasury issued a 30-day general license permitting purchases of Russian oil loaded before March 12, responding to the oil shock after the Strait of Hormuz closed during the Iran war. The administration allowed the waiver to expire on April 11 without announcing a renewal, and sanctions snapped back into force (a new license followed on April 17).

Figure 12: Claude Opus 4.8 candidate, accepted by the official judge: the candidate asks whether the waiver is extended past April 11; the gold question resolves on exactly that decision, and the negative resolution (expiry) counts because polarity is irrelevant.

Forecast-time question ✓

Will the EU formally adopt its 20th package of sanctions against Russia before May 1, 2026?

Pre-cutoff signals.

- **2026-03-18:** Merz called for adopting the sanctions package quickly while Hungary blocked progress pending Druzhba oil-transit resumption.
- **2026-03-28:** Fico threatened to block EU sanctions and Ukraine's accession, confirming the dual-veto dynamic.
- **2026-03-31:** Reporting alleged Hungary had lobbied against EU Russia sanctions at Moscow's request, raising pressure on the blockers.

Claude Fable 5

Gold question

Will the European Union approve its €90 billion loan package for Ukraine and a new package of sanctions against Russia before 1 May 2026?

Outcome summary.

At the April 1 cutoff, both the €90 billion Ukraine loan and the 20th Russia sanctions package had been blocked by Hungary and Slovakia over the damaged Druzhba pipeline. Orbán's April 12 election defeat and the resumption of Druzhba flows around April 22–23 broke the deadlock, and the EU formally approved both measures on April 23.

Figure 13: Claude Fable 5 candidate, accepted by the official judge: the candidate asks about formal adoption of the 20th sanctions package. The matching criteria in the gold record allow sanctions-only formulations because the same Hungarian/Slovak veto bottleneck controlled both the loan and the sanctions decisions.

Forecast-time question ✓

Will Fidesz fail to win an outright parliamentary majority in Hungary's April 12, 2026 parliamentary election?

Pre-cutoff signals.

- **2026-03-30:** AP detailed Tisza's youth-driven challenge, Orbán's polling weakness, and Magyar's promise to move Hungary away from Moscow.
- **2026-03-31:** Reuters reported Fidesz trailing Tisza in many polls, with implications for Hungary's pro-Moscow policy line.
- **2026-03-31:** Leaked Szijjártó–Lavrov sanctions discussions underscored how tightly the election was tied to EU Russia policy.

GPT-5.5

Gold question

Will Viktor Orbán's Fidesz party win the largest number of seats in Hungary's April 2026 parliamentary election?

Outcome summary.

Hungary's April 12 parliamentary election pitted Orbán's Fidesz against Péter Magyar's center-right Tisza party. In a landslide upset, Tisza won, Orbán conceded on election night, and his 16-year rule ended, a result that reshaped the EU's internal balance on Russia policy.

Figure 14: GPT-5.5 candidate, accepted by the official judge: both questions track the same uncertainty, whether Orbán's Fidesz keeps its hold on power in the April 12 election. The gold record's boundary note treats winning the election, retaining a parliamentary majority, and finishing as the largest party as one electoral-outcome issue, and its positive examples include "Will Fidesz retain its parliamentary majority in the 2026 Hungarian general election?"; the candidate asks the same question with the opposite polarity, which does not affect matching.

Forecast-time question ✗

Will Russian forces capture a significant Donetsk Oblast city (e.g., fully seize Pokrovsk, Kostiantynivka, or another major town) during April 2026?

Pre-cutoff signals.

- **2026-03-31:** Russia claimed it could take all of Donbas within two months; Zelensky countered that Russian forces could not achieve this.
- **2026-03-31:** Zelensky warned the Iran war was weakening Ukraine's position as Western support diverted.
- **2026-03-31:** The EU's Kallas said Russia's ambitions extend beyond Donbas, framing the offensive stakes.

Claude Opus 4.8

Gold question

Will Russian ground forces be engaged in combat inside the city limits of Kostiantynivka, Donetsk Oblast, by 1 May 2026?

Outcome summary.

Kostiantynivka anchors Ukraine's fortified "fortress belt" in Donetsk Oblast. Russian forces advanced toward the belt through early 2026, and on April 21 Gerasimov stated that Russian units were fighting inside parts of the city, confirming urban combat within the city limits.

Figure 15: Claude Opus 4.8 candidate, a near miss rejected by the official judge: the candidate resolves on full capture of a major city, whereas the gold question resolves on the lower threshold of urban combat inside Kostiantynivka. Same offensive, different resolution threshold.

Forecast-time question ✗

Will the US formally announce a reduction in its NATO commitment—whether troop levels, defense spending obligations, or Article 5 guarantees—during April 2026?

Pre-cutoff signals.

- **2026-03-27:** Trump publicly stated the US does not "have to be there for NATO," questioning the Article 5 commitment after European allies declined to support the Iran war.
- **2026-03-28:** Reporting said Trump threatened to abandon NATO over European refusal to back the Iran campaign.

GLM 5.2

Gold question

Will Donald Trump publicly threaten to withdraw the United States from NATO in April 2026?

Outcome summary.

European NATO members' refusal to help unblock the Strait of Hormuz during the Iran war escalated transatlantic tensions. On April 1, Trump told reporters he was "absolutely without question" considering withdrawal — crossing from criticism of allies to an explicit withdrawal threat.

Figure 16: GLM 5.2 candidate, a near miss rejected by the official judge: the candidate asks for a formal commitment reduction (troops, spending, guarantees), while the gold question resolves on a public withdrawal threat — a related but distinct action arising from the same tensions.

Forecast-time question ✗

Will at least one drone linked to the Russia–Ukraine war enter, crash in, or be engaged over Finland, Estonia, Latvia, Lithuania, or Poland during April 2026, prompting an official military alert, protest, or apology?

Pre-cutoff signals.

- **2026-03-31:** Multiple drones entered Estonian airspace, with debris found in Tartu County and earlier crashes in Latvia, Lithuania, and Finland.
- **2026-03-31:** Reuters noted one Ukrainian drone that crashed in Finland carried an unexploded warhead.
- **2026-03-31:** Ukraine claimed Russia deliberately redirected Ukrainian drones toward Finland and the Baltics to stir tensions.

GPT-5.5

Gold question

Will a Russian military drone or missile enter the airspace of a NATO member state before 1 May 2026?

Outcome summary.

Russia's war repeatedly spilled into NATO territory in early 2026. On April 17, a Russian drone breached Romanian airspace near Tulcea County during an overnight bombardment of Ukrainian civilian infrastructure; Romania's Defence Ministry condemned the violation.

Figure 17: GPT-5.5 candidate, a near miss rejected by the official judge: the candidate covers drones of either side over five named countries (Finland, the three Baltic states, and Poland) and additionally requires an official response. The gold question is specifically about a Russian military drone or missile entering the airspace of any NATO member, with no response requirement — and it resolved in Romania, outside the candidate's listed countries.

These examples illustrate the boundary applied by the offline matcher. Equivalent operational formulations are accepted even when wording differs, but questions are rejected when they switch from the realized event to a neighboring mechanism, precondition, response, or downstream consequence. The most common near misses had this form: they identified a real weak-signal cluster, but asked about a different uncertainty inside the same policy or security episode.

E JUDGE ROBUSTNESS

Two ablations test the stability of the offline matcher used for the main results: sensitivity to the embedding-prefilter threshold, and agreement with two second judge models. Neither changes the official metric; both are reported so the matcher's two discretionary choices (the prefilter cutoff and the judge model) can be audited.

E.1 CHOICE OF THE PREFILTER THRESHOLD

The official pipeline sends to the LLM judge only candidate–gold pairs whose question texts have embedding similarity at least 0.60. We fixed this threshold before the main run using a calibration check: the positive examples in the gold records, which are written as formulations that should count as matches, almost all clear it (2,247 of 2,251, or 99.8%, have similarity at least 0.60 to their record's normalized question; median 0.86).

The main-run judgments suggest that lowering the threshold would add little. The judge's acceptance rate falls steeply as similarity approaches 0.60: 47% of pairs at similarity at least 0.80, 11% in $[0.75, 0.80)$, 1.9% in $[0.70, 0.75)$, 0.3% in $[0.65, 0.70)$, and 0.03% (4 of 11,779 pairs) in $[0.60, 0.65)$. Pairs below 0.60 were not judged, so matches there cannot be ruled out, but this gradient suggests they are rare.

Raising the threshold would also change little. Of the 246 accepted matches in the main five-agent runs, 230 (93.5%) have similarity at least 0.70. Restricting accepted matches to that range leaves the union of covered gold questions unchanged (42 of 329) and shifts the Coverage@100 and AP of the four frontier agents by less than one standard error. GPT-5-mini is the exception: 9 of its 28 accepts fall in $[0.60, 0.70)$, so its Coverage@100 falls from 0.016 to 0.012 and its AP from 0.0016 to 0.0011,

still far below the frontier agents. We nevertheless keep 0.60: manual inspection shows that several near-threshold accepts are clear matches under the gold records' boundary notes. They are typically paraphrases of compound, multi-clause gold questions whose surface forms embed far apart, so a higher cutoff would drop matches systematically rather than at random.

E.2 Second-Judge Agreement

The official judge is DeepSeek v4 Pro. To test whether the results depend on this choice, we re-judged candidate–gold pairs with two other judge models, GPT-5.5 (xhigh) and GLM 5.2 (xhigh), using the identical pairwise prompt. The re-judged pairs form two overlapping sets.

- **All official accepts (373 pairs).** Every pair that the official judge accepted for the seven generation agents, that is, the five main agents and the two DeepSeek reference agents of Appendix F. Of these, 285 come from the original judge run, which covered the four frontier agents and DeepSeek v4 Pro, and 88 come from separate runs for GPT-5-mini and DeepSeek v4 Flash.

- **Audit set (570 pairs).** The 285 accepted pairs of the original run together with 285 pairs that the official judge rejected in the same run. The rejected pairs were chosen greedily to match the accepted pairs in embedding similarity, so they lie close to the decision boundary.

Of the 373 official accepts, GPT-5.5 confirms 351 (94.1%) and GLM 5.2 confirms 348 (93.3%). Table 12 breaks these confirmations down by the agent that generated the candidate, for the five main agents. Confirmation of the DeepSeek reference agents' pairs, which bears on possible judge self-favoritism, is reported in Appendix F. On the audit set, each second judge agrees with the official verdict on 530 of the 570 pairs (93.0%). The disagreements fall on both sides of the official decision. GPT-5.5 rejects 20 of the 285 official accepts and accepts 20 of the 285 official rejects, and GLM 5.2 rejects 19 and accepts 21. Disagreement with the official judge is therefore small and roughly symmetric.

Forecast-time agent	Official accepts	GPT-5.5 confirms	GLM 5.2 confirms
Claude Fable 5	63	59	59
GPT-5.5	58	55	58
Claude Opus 4.8	52	44	46
GLM 5.2	45	41	41
GPT-5-mini	28	26	24

Table 12: Confirmation of the official accepted matches by two second judges (GPT-5.5 xhigh and GLM 5.2 xhigh), by generation agent, for the five main agents. Each second judge re-judged every accepted candidate–gold pair with the identical pairwise prompt.

Table 13 recomputes Coverage@100 for the five main agents when an official match must also be accepted by at least one of the two second judges (2-of-3) or by both (all 3). Because the second judges re-judged all official accepts but only a sample of official rejects, these rules can remove official matches but cannot add new ones. Coverage therefore can only fall as the rule tightens, and it falls only slightly. Claude Fable 5 retains the highest point estimate under the strictest rule, but the small differences within the frontier band remain unresolved. Under all three tested judging rules, the frontier agents retain higher mean Coverage@100 than the lower-cost control.

Agent	Official	2-of-3 judges	All 3 judges
Claude Fable 5	0.035	0.033	0.033
GPT-5.5	0.033	0.033	0.032
Claude Opus 4.8	0.029	0.025	0.024
GLM 5.2	0.026	0.024	0.022
GPT-5-mini	0.016	0.016	0.013

Table 13: Coverage@100 under consensus judging rules over the three judges (official DeepSeek v4 Pro, GPT-5.5 xhigh, GLM 5.2 xhigh). The 2-of-3 rule keeps an official match if at least one second judge also accepts it, and the all-3 rule keeps it only if both do. Entries are means across five runs, and the denominator is the 329-question gold pool.

F DeepSeek Reference Results

DeepSeek v4 Pro is the model used for retrospective mining, gold record construction, and official judging. DeepSeek-family agents therefore raise an authorship concern that applies to no other evaluated agent: an agent whose underlying model shares retrieval and phrasing priors with the retrospective mining model can be advantaged in ways unrelated to real discovery ability. The overlap is measurable: the mean maximum cosine similarity between an agent's Perigon queries and the mining runs' query pool is 0.905 ± 0.001 for DeepSeek v4 Pro, versus 0.81–0.84 for the four main frontier agents. We therefore exclude DeepSeek agents from the main comparison and report their scores here for reference only.

F.1 Reference Scores

Agent	AP	Coverage@100
DeepSeek v4 Pro (xhigh)	0.0130 ± 0.0024	0.036 ± 0.004
DeepSeek v4 Flash (xhigh)	0.0121 ± 0.0016	0.032 ± 0.004

Table 14: Reference scores of the DeepSeek agents at the $K = 100$ budget, under the identical protocol and judge as Table 1. Entries are mean $\pm$ one standard error across five runs. Because DeepSeek v4 Pro is used both to construct the gold pool and to judge matches, these scores are not directly comparable to the main results.

DeepSeek v4 Flash's point estimates fall within the numerical range of the frontier band in Table 1, and DeepSeek v4 Pro's lie slightly above its upper end (AP 0.0130 vs. 0.0128; Coverage@100 0.036 vs. 0.035), within one standard error; we do not make a formal pairwise-ranking or equivalence claim. DeepSeek v4 Flash is inside this range at under a tenth of the frontier agents' generation cost ($0.51 per run). Two audits check how much of this is an artifact of the judge. First, the second-judge study of Appendix E finds no consistent favoritism toward DeepSeek-generated candidates: GPT-5.5 confirms 66 of v4 Pro's 67 accepted pairs and all 60 of v4 Flash's, GLM 5.2 confirms 62 and 58, and consensus judging preserves both scores (v4 Pro 0.036/0.036/0.033 and v4 Flash 0.032/0.032/0.032 under the official, 2-of-3, and all-3 rules). Second, the query-prior overlap above indicates that the residual concern is on the generation side, specifically shared search behavior with the miner, rather than the judging side. This is why we treat these scores as reference values rather than rankings.

F.2 Cross-Miner Authorship Check

The strongest test of the authorship concern is to score every agent only on gold questions that a non-DeepSeek miner independently recovered. We ran an additional retrospective mining agent (GPT-5.5, medium reasoning) to the configured 150-loop cap (856 raw questions), matched its output against the 329-question gold pool with the identical embedding-prefilter-plus-pairwise-judge pipeline, and flagged a gold question as *independently recovered* if at least one GPT-5.5-mined question matched it. This independent miner recovers 189 of 329 gold questions (57.4%), skewed toward higher-support questions: 38 of the 50 questions at evaluation ranks 1–50 (76.0%) versus 86 of the 179 questions at ranks 151–329 (48.0%).

Agent	AP	Coverage@100
Claude Fable 5	0.0184 ± 0.0017	0.050 ± 0.005
GPT-5.5	0.0175 ± 0.0023	0.050 ± 0.005
DeepSeek v4 Pro	0.0163 ± 0.0028	0.046 ± 0.004
Claude Opus 4.8	0.0163 ± 0.0025	0.044 ± 0.004
GLM 5.2	0.0149 ± 0.0024	0.036 ± 0.005
DeepSeek v4 Flash	0.0133 ± 0.0018	0.042 ± 0.005
GPT-5-mini	0.0023 ± 0.0004	0.022 ± 0.001

Table 15: AP and Coverage@100 restricted to the 189 gold questions independently recovered by the GPT-5.5-medium mining agent, at the $K = 100$ budget. Entries are mean ± one standard error across five runs.

AP rises in absolute terms for every frontier agent on this 189-question subset, including DeepSeek v4 Pro from 0.0130 to 0.0163, so the result is not score shrinkage. What changes is DeepSeek's relative position: its full-pool point-estimate advantage disappears, with Claude Fable 5 and GPT-5.5 obtaining higher AP point estimates and overlapping uncertainty. On the subset at evaluation ranks 151–329, DeepSeek v4 Pro still has the highest point estimate, but we make no pairwise-ranking claim. The cross-miner check is therefore consistent with a gold-authorship effect on DeepSeek's aggregate relative performance, while leaving open whether its tail result reflects shared priors or a real retrieval advantage. This is the central reason we treat DeepSeek scores as reference values throughout rather than including them in the ranked main comparison.

G News Backend and Cutoff Enforcement

AnteBench uses Perigon as its historical news backend. Perigon describes its News API as structured news and event data from over 80,000 web sources, enriched with fields such as topics, entities, sentiment, source metadata, and article clustering (Perigon, 2026a). Its article endpoint returns JSON over HTTP and supports search over title, description, and article content, including Boolean operators and exact-phrase matching (Perigon, 2025; 2026b).

Cutoff-bounded search. The wrapper treats Perigon as a dated evidence source rather than an unrestricted search engine. During forecast-time question generation, searches and article reads are restricted to the pre-cutoff period. During retrospective question generation, searches are allowed over the target window because the task is explicitly hindsight-informed. The offline matcher has no Perigon or live-web access.

Residual leakage risk. Cutoff enforcement reduces tool-side leakage, but it does not make Perigon a frozen archival snapshot. Residual leakage through article updates, indexing, ranking, clustering, or metadata cannot be fully ruled out. This is one reason the live forward-testing deployment is important: it freezes submissions before the target window resolves.

Parametric knowledge. Restricting the tools does not address what a model already knows from training. The models underlying four of the five main agents have provider-documented training-data cutoffs strictly before the target window: GPT-5.5 (December 2025), GPT-5-mini (May 2024), and Claude Opus 4.8 and Claude Fable 5 (both January 2026). GLM 5.2's cutoff is not documented, and the model was released in June 2026, after the window closed, so parametric knowledge of April 2026 events cannot be excluded for it. Such leakage cannot explain a leading result for the GLM agent because its point estimates are low, but it may still affect comparisons involving that agent's performance, tool use, diversity, and surprisingness. The DeepSeek v4 models were announced on April 24, 2026, inside the target window, which is an additional reason their agent scores are treated as reference values only (Appendix F).

H PROMPT AND TOOL DETAILS

This section describes the agent-facing tools and prompt roles used in the question-to-question pipeline. The central design constraint is that temporal access should be enforced by the harness and tool wrappers, not only by natural-language instructions inside the prompt.

Tools available to agents. The two news-access tools are `perigon_search` and `perigon_read_article`. The search tool returns compact article records with source identifiers, titles, summaries, source metadata, and publication dates. The read tool takes an article identifier returned by search and returns article text and metadata. Agents finish by calling a stage-specific submit tool; free-form prose is not accepted as a final answer. The retrospective mining stage additionally exposes a `finish_question_mining` tool: the mining agent iterates submit-and-continue loops and terminates by declaring its list complete, subject to a hard loop cap, instead of stopping at a fixed question count.

Component	Tool access	Role
Forecast-time question generation	Perigon search/read; `submit_candidates`	Generate a ranked pre-cutoff question list for the target window.
Retrospective question mining	Perigon search/read; `submit_candidates`; `finish_question_mining`	Mine materialized target-window issues, express them as retrospective questions, and declare completion.
Gold record construction	Perigon search/read; `submit_question_card`	Verify resolution, normalize the question, define matching boundaries, and mine precursor signals.
Semantic clustering and deduplication	Embedding API, plus an LLM duplicate judge and representative selection; no news tools	Consolidate repeated retrospective candidates into the fixed gold-question pool.
Offline matching	No news tools; strict-JSON verdict (no tool call)	Match forecast-time questions to gold records.

Table 16: Tool access by pipeline component. News tools are available only in the two generation stages and in gold record construction; the offline matcher has no live retrieval access.

Tool	Main inputs	Output
`perigon_search`	Query, date window, source and language filters	Ranked article summaries with run-local source IDs.
`perigon_read_article`	Source ID returned by prior search	Article text and metadata for follow-up reading.
Stage submit tools	Structured JSON payload	Forecast-time or retrospective questions, or gold records.

Table 17: Tools available to agents. Search results are cited by source ID; the harness resolves those IDs to underlying sources after submission.

Temporal access rules. Forecast-time question generation is cutoff-bounded: all news access is restricted to material published strictly before the cutoff t. Retrospective candidate generation is intentionally hindsight-informed and can search the target window T_t to identify consequential issues that materialized there. Gold record construction uses target-window coverage to verify resolution timing, but separately mines precursor signals, which the construction prompt requires to predate the cutoff. The offline matcher receives only the forecast-time questions and the fixed gold records; it cannot search Perigon or the live web.

Search syntax and source filters. Perigon search supports Boolean queries over article title, description, and content. The practical instruction given to agents is that Boolean operators must be uppercase, e.g. `A AND B`; lowercase operator words are treated as ordinary query terms. The

prompts also encourage local-language searches when an issue is likely to be covered first or most precisely outside English-language global media.

Prompt roles. The four core prompts differ mainly in their information condition and required output. Forecast-time generation asks what should be on the watchlist before the window resolves. Retrospective question generation asks, after the window closes, which questions should have been asked at the cutoff about the issues that materialized. Record construction then turns each selected retrospective question into a structured object for matching and audit. The judge prompt finally performs semantic matching without tools.

Prompt	Time perspective	Output	Key constraints
Forecast-time question generation	Forward from cutoff	Ranked forecast-time questions	Consequential, uncertain, concrete, evidence-backed, target-window aligned, in-area.
Retrospective question generation	Backward after target window	Ranked retrospective candidates	Consequential outcomes that materialized, uncertain at the cutoff, askable in question form.
Gold record construction	Deep dive on one selected retrospective question	Normalized question plus timing, boundary, examples, signals	Remove incidental hindsight detail; define same-issue inclusions and neighboring exclusions.
Question judge	Offline after gold construction	Per-pair match verdict	One candidate–gold pair per call; polarity-neutral same-uncertainty matching; reject actor-only, topic-only, precondition, consequence, or adjacent-procedural-step matches.

Table 18: Prompt roles in the question-to-question pipeline. The prompt text appears verbatim in Appendix K.

Structured outputs and audit traces. Each generation pass ends in one structured submit call, except that a retrospective mining run ends when the agent calls `finish_question_mining`; each record-construction run ends in one `submit_question_card` call. Forecast-time and retrospective generation submit ranked question lists with rationale and source IDs; record construction submits normalized fields, matching examples, precursor signals, and causal mechanisms. The official judge is not an agent: it is a single call per candidate–gold pair that returns one JSON verdict with a rationale. A near-miss audit channel additionally records close rejected pairs. These traces make failures auditable: a missed gold question can reflect retrieval, question formulation, semantic clustering, record normalization, or judge-boundary behavior rather than a single opaque score.

I Runtime and Cost Details

This section collects operational accounting for the reported runs. Table 19 reports forecast-time generation cost, reasoning-token usage, and runtime endpoints used to contextualize the main coverage results. Tables 20 and 21 report the corresponding gold-side construction and offline judge accounting. Wall-clock runtime is tracked in run logs but is not used as an agent-comparison metric because provider routing, queueing, and concurrency differ across runs. When a table reports parallelism, the wall time is the observed elapsed time at that parallelism, not a serial sum of per-call runtimes.

Agent	Reasoning setting	Mean gen. cost	Mean reasoning tokens	Mean wall time
DeepSeek v4 Pro	xhigh	$0.67	120,844	1h 46m
GPT-5.5	xhigh	$27.34	82,454	52m
Claude Opus 4.8	high (adaptive thinking)	$16.33	62,045	1h 46m
Claude Fable 5	high (adaptive thinking)	$23.53	45,338	1h 11m
GLM 5.2	xhigh	$5.48	284,503	2h 45m
DeepSeek v4 Flash	xhigh	$0.51	112,781	57m
GPT-5-mini	high	$1.33	314,781	1h 28m

Table 19: Forecast-time question-generation cost, reasoning-token usage, and wall-clock runtime at $K = 100$; the DeepSeek rows correspond to the reference agents of Appendix F. All rows summarize the same five-run evaluation slice and exclude gold construction and judge cost. Costs are estimated from the recorded token counts at each provider's list price around the time of the runs, including GPT-5.5's long-context surcharge (prompts above 272K tokens). Where isolated billing records exist (the Claude runs and one GPT-5.5 pilot run), these estimates fall 2–8% below the billed amounts; per-model billing is not isolated for every run.

Gold-side stage	Workload	Model or tool	Parallelism	Cost	Wall time
Retrospective question mining	5 mining runs, 2,217 raw questions	DeepSeek v4 Pro, xhigh	–	n/a	–
Embedding and clustering gates	2,217 raw questions plus 428 normalized records	text-embedding-3-small	local	<$0.01	not timed
Gold record construction	428 selected clusters	DeepSeek v4 Pro, xhigh	$c = 100$	$18.17	–
Record duplicate audit	1,238 directional judgments (619 pairs, both directions)	DeepSeek v4 Pro, xhigh	$c = 100$	$1.79	–
LLM representative selection	multi-sink duplicate components	DeepSeek v4 Pro, xhigh	–	$0.31	–
Total external API cost	Final gold-side construction path (excl. mining)	mixed	–	$20.27	–

Table 20: Gold-question construction accounting for the finalized Eastern Europe / Russia gold pool. The table includes the final construction path used for the 335-record pool and excludes discarded exploratory runs. Here c denotes configured concurrency for parallel API recipes or calls. The mining phase ran interleaved with other experiments, so no isolated billing record exists for it; the embedding cost is a token-count estimate, and all other costs are provider wallet deltas. Local clustering and final ranking are effectively zero-cost relative to the external API stages.

Judge stage	Workload	Model or setting	Parallelism	Cost	Wall time
Embedding prefilter	822,500 candidate–gold pairs scored; 19,856 retained at ≥ 0.60	text-embedding-3-small	local	<$0.01	not timed
Pairwise judge	19,856 judged pairs	DeepSeek v4 Pro, xhigh	$c = 100$	$15.25	1h 25m

Table 21: Offline judge accounting for the original five-agent pairwise run (the four main frontier agents plus DeepSeek v4 Pro); GPT-5-mini and DeepSeek v4 Flash were judged in separate runs with identical settings. The judge cost is the delayed provider wallet delta. Wall time is observed elapsed time at the listed maximum parallelism.

J Question Lists

For auditability, this appendix lists the fixed question texts used in the Eastern Europe / Russia April 2026 experiment. The gold excerpt shows the 50 highest-support records of the 329-question

evaluation pool; the complete pool is released with the dataset. The forecast-time tables list the first run only for each of the four frontier agents; the quantitative results in the main text use five runs per agent.

Eastern Europe / Russia gold-question pool (excerpt). Table 22 lists the 50 highest-support retrospective gold questions, ordered by evaluation rank after removing out-of-window records. The ordering inherits the merged raw-submission support ranking of the full construction pool, with tie breaking as described in Appendix A.1. We include the normalized question text, resolution date, and news-attention surprisingness score. Surprisingness is the macro mean of included per-query natural-log ratios, $\ln((\text{post}+1)/(\text{pre}+1))$, from the 20-query Perigon count procedure with the final `post >= 5` inclusion rule; a dash marks questions with no query passing that rule.

Table 22: Eastern Europe / Russia retrospective gold-question pool: 50 highest-support records of the 329-question evaluation pool.

Rank	Gold question	Resolution date	Surprisingness
1	Will the Russian-occupied Zaporizhzhia Nuclear Power Plant experience a worker fatality from military activity or a loss of external power supply during April 2026?	2026-04-14	2.01
2	Will the European Union approve its €90 billion loan package for Ukraine and a new package of sanctions against Russia before 1 May 2026?	2026-04-23	0.25
3	Will Russia's Gazprom publicly claim between 1 April and 1 May 2026 that Ukrainian drones have attacked or attempted to attack the Russkaya compressor station of the TurkStream gas pipeline?	2026-04-02	2.81
4	Will Russia halt or suspend the transit of Kazakh crude oil to Germany via the Druzhba pipeline before 1 May 2026?	2026-04-22	1.63
5	Will the Russian government or military officially identify specific sites in any NATO country as potential military targets before 1 May 2026?	2026-04-15	1.33
6	Will the Russian government impose a ban or significant restrictions on gasoline exports before 1 May 2026?	2026-04-02	-0.70
7	Will Russia's GDP contract in Q1 2026?	2026-04-29	0.54
8	Will Ukrainian drone strikes on a Russian oil facility cause a major environmental disaster involving oil contamination of the Black Sea or toxic precipitation over populated areas before 1 May 2026?	2026-04-24	2.27
9	Will the Russian government announce before 1 May 2026 that the 2026 Victory Day parade in Moscow will be held without tanks, armored vehicles, or missile systems?	2026-04-28	5.47
10	Will the Russian government officially declare full control of Ukraine's Luhansk Oblast before 1 May 2026?	2026-04-01	2.28
11	Will Russia veto a UN Security Council resolution on reopening the Strait of Hormuz before 1 May 2026?	2026-04-07	0.58
12	Will Rumen Radev's party or coalition win the largest share of seats in Bulgaria's April 2026 parliamentary election?	2026-04-19	1.77
13	Will a Ukrainian attack drone strike a target located at least 1,500 kilometers from Ukrainian-controlled territory before 1 May 2026?	2026-04-25	0.38
14	Will Viktor Orbán announce that he will not take up his seat in the Hungarian parliament following the April 2026 election?	2026-04-25	1.22
15	Will a senior member of the British royal family make a publicly reported visit to Ukraine before 1 May 2026?	2026-04-23	4.76
16	Will Ukraine successfully disable or destroy the last operational railway ferry connecting Russia to occupied Crimea across the Kerch Strait before 1 May 2026?	2026-04-06	3.08
17	Will a senior Russian government official publicly threaten to use military force to protect Russian citizens in Moldova's breakaway Transnistria region before 1 May 2026?	2026-04-21	1.14

Continued on next page

Table 22: Gold-question pool excerpt (continued).

Rank	Gold question	Resolution date	Surprisingness
18	Will the UK government publicly confirm before 1 May 2026 that it has tracked and deterred a Russian submarine operation targeting critical undersea infrastructure in or near UK waters?	2026-04-09	3.26
19	Will Russia's Africa Corps (or affiliated Russian paramilitary forces) withdraw from a town or military position in Mali due to rebel or insurgent military pressure in April 2026?	2026-04-26	1.96
20	Will the US Treasury Department extend or renew its temporary general license permitting purchases of sanctioned Russian oil and petroleum products beyond the April 11, 2026 expiration date?	2026-04-11	0.90
21	Will Germany and Ukraine announce a new bilateral defense cooperation package worth at least €1 billion before 1 May 2026?	2026-04-14	2.28
22	Will a senior Ukrainian government official publicly offer Ukraine's military or naval expertise, or drone technology, to assist international efforts to reopen or secure the Strait of Hormuz before 1 May 2026?	2026-04-03	-0.19
23	Will a Moldovan court convict Vladimir Plahotniuc and sentence him to prison in the bank fraud case before 1 May 2026?	2026-04-22	3.58
24	Will Ukraine and Norway publicly sign or announce a bilateral defense cooperation framework before 1 May 2026?	2026-04-14	1.71
25	Will a Kosovo court convict and sentence to prison any ethnic Serb defendants for the September 2023 Banjska attack before 1 May 2026?	2026-04-24	3.55
26	Will Ukraine publicly accuse Israel of importing grain stolen by Russia from occupied Ukrainian territories before 1 May 2026?	2026-04-28	2.26
27	Will Belarus release Andrzej Poczobut from prison by May 1, 2026?	2026-04-28	4.18
28	Will the Russian government issue a formal public travel advisory warning its citizens against traveling to countries that have extradition treaties with the United States before 1 May 2026?	2026-04-01	3.47
29	Will Ukraine's military leadership issue binding rules establishing a maximum deployment period for frontline combat troops before 1 May 2026?	2026-04-30	1.53
30	Will the acting US ambassador to Ukraine resign or announce their departure before 1 May 2026?	2026-04-28	2.24
31	Will OPEC+ agree at its April 2026 meeting to increase collective oil production quotas?	2026-04-05	1.83
32	Will Ukraine extend martial law and general mobilization beyond their scheduled May 2026 expiration?	2026-04-28	0.34
33	Will Donald Trump publicly announce or threaten a reduction of US troops stationed in Germany during April 2026?	2026-04-29	4.22
34	Will Ukraine publicly claim to have struck a Russian naval vessel in Sevastopol Bay, occupied Crimea, before 1 May 2026?	2026-04-19	0.58
35	Will the EU's 20th sanctions package against Russia include a full ban on maritime services for tankers transporting Russian crude oil and petroleum products, to take effect without waiting for G7 coordination?	2026-04-22	2.06
36	Will Russian federal authorities formally designate the Nobel Peace Prize-winning human rights organization Memorial as an 'extremist' organization before 1 May 2026?	2026-04-09	3.71
37	Will Russia set a new monthly record for long-range drone and missile strikes on Ukraine in April 2026?	2026-04-30	0.42
38	Will Russia agree to pay compensation to Azerbaijan for the December 2024 downing of Azerbaijan Airlines Flight 8243 by 1 May 2026?	2026-04-15	2.91

Continued on next page

Table 22: Gold-question pool excerpt (continued).

Rank	Gold question	Resolution date	Surprisingness
39	Will a senior US official publicly confirm before 1 May 2026 that Russia is providing direct military assistance — including intelligence, weapons, or targeting support — to Iran for use against US forces in the ongoing conflict?	2026-04-30	-0.40
40	Will a head of state or government of a NATO eastern flank member state publicly express doubt about the United States' commitment to NATO's Article 5 collective defense obligation before 1 May 2026?	2026-04-24	1.35
41	Will Ukrainian President Volodymyr Zelenskyy publicly state before 1 May 2026 that US President Donald Trump cannot be relied upon as a security guarantor for Ukraine?	2026-04-20	1.02
42	Will German authorities publicly announce before 1 May 2026 the arrest of a foreign national suspected of spying for Russian intelligence services in connection with German military support for Ukraine?	2026-04-29	-0.30
43	Will the Central Bank of Russia propose or announce that commercial banks must hold reserves in Chinese yuan by 1 May 2026?	2026-04-28	3.61
44	Will a Russian military drone or missile enter the airspace of a NATO member state before 1 May 2026?	2026-04-17	0.43
45	Will the Moldovan parliament give final approval to withdraw from the Commonwealth of Independent States (CIS) before 1 May 2026?	2026-04-02	1.87
46	Will Kosovo's parliament fail to elect a new president, triggering snap parliamentary elections?	2026-04-28	1.46
47	Will a senior Russian government official publicly warn before 1 May 2026 that European countries hosting French nuclear-capable forces would become targets for Russian military action?	2026-04-23	0.72
48	Will an official Ukrainian anti-corruption body publicly call for the removal of Rustem Umerov from his position as National Security and Defense Council Secretary before 1 May 2026?	2026-04-30	1.18
49	Will Vladimir Putin propose a temporary ceasefire in Ukraine to coincide with Russia's May 9 Victory Day celebrations before 1 May 2026?	2026-04-29	6.32
50	Will Ukrainian officials announce by 30 April 2026 that Ukraine's forces have recaptured at least 400 square kilometers of territory from Russia since the start of 2026?	2026-04-05	-0.22

Forecast-time question lists. The following tables list the exact top-100 forecast-time questions generated by each agent in run 1. A black check mark indicates that the forecast-time question matched at least one gold question under the official pairwise judge.

GPT-5.5.

Table 23: GPT-5.5 forecast-time questions, run 1 top-100.

Rank	Forecast-time question	Match
1	Will Russia and Ukraine announce or observe a U.S.-mediated Easter or energy-infrastructure ceasefire at any point during April 2026?	
2	Will Fidesz fail to win an outright parliamentary majority in Hungary's April 12, 2026 parliamentary election?	✓
3	Will Ukrainian long-range drones or missiles again damage a major Russian oil export terminal, refinery, or related loading operation during April 2026?	
4	Will Ukraine, Russia, and the United States hold an official trilateral peace or ceasefire meeting before 2026-05-01?	

Continued on next page

Table 23: GPT-5.5 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
5	Will Russian forces capture or establish confirmed control over Hryshyne, Donetsk Oblast, before 2026-05-01?	
6	Will at least one drone linked to the Russia–Ukraine war enter, crash in, or be engaged over Finland, Estonia, Latvia, Lithuania, or Poland during April 2026, prompting an official military alert, protest, or apology?	
7	Will the EU approve either the €90 billion Ukraine loan or its 20th Russia sanctions package before 2026-05-01 despite Hungary/Slovakia objections?	✓
8	Will EU experts inspect, or will Russian oil transit resume through, the Ukrainian Druzhba pipeline route to Hungary and Slovakia before 2026-05-01?	
9	Will Rumen Radev's Progressive Bulgaria finish first in Bulgaria's April 19, 2026 parliamentary election?	✓
10	Will the U.S. or NATO publicly acknowledge before 2026-05-01 that air-defense interceptors or other PURL weapons planned for Ukraine are being delayed or diverted to Middle East operations?	
11	Will the UK, France, Belgium, Sweden, or another European state board or detain a sanctioned Russian shadow-fleet oil tanker before 2026-05-01?	
12	Will Moldova lift or downgrade its 60-day energy state of emergency before 2026-05-01?	
13	Will Donald Trump and Alexander Lukashenko hold a face-to-face or formally announced 'Board of Peace' meeting before 2026-05-01?	
14	Will credible reporting or official statements confirm that Russian military personnel—combat troops, air-defense crews, or military advisers—have deployed to Iran in connection with the Iran war before 2026-05-01?	
15	Will Russia and Ukraine conduct another prisoner exchange returning at least 100 Ukrainian captives before 2026-05-01?	✓
16	Will Finland approve legal amendments lifting its blanket ban on importing, possessing, or hosting nuclear weapons for NATO defense purposes before 2026-05-01?	✓
17	Will an IAEA-brokered local cease-fire enable repairs restoring Zaporizhzhia Nuclear Power Plant's damaged main 750 kV off-site power line before 2026-05-01?	
18	Will the United States extend, renew, or materially broaden its temporary Russian-oil sanctions waiver after the current license expires on April 11, 2026?	✓
19	Will a commercial tanker or cargo ship in the Black Sea or approaches to the Bosphorus be damaged by a drone, missile, mine, or explosive attack linked to the Russia–Ukraine war before 2026-05-01?	
20	Will Russian forces establish confirmed control of any part of Lyman, Donetsk Oblast, or a verified foothold inside the city before 2026-05-01?	
21	Will NATO or a NATO member publicly announce an additional air-defense or counter-UAV deployment to Estonia, Latvia, or Lithuania before 2026-05-01 in response to recent drone or airspace incidents?	
22	Will Poland formally authorize the temporary transfer or deployment of a Patriot battery or PAC-3 interceptor missiles to Middle East operations before 2026-05-01?	
23	Will Russia announce a new mobilization or compulsory reservist call-up for the Ukraine war, beyond the routine spring conscription cycle, before 2026-05-01?	
24	Will Georgia's ruling Georgian Dream party formally expand or file its Constitutional Court bid to ban additional pro-Western opposition parties before 2026-05-01?	
25	Will the United States and Georgia formally announce restoration of the suspended U.S.–Georgia strategic partnership, or a major sanctions-relief step for Georgian Dream officials, before 2026-05-01?	
26	Will credible reporting or official statements confirm before 2026-05-01 that Russian-made or Russia-upgraded attack drones have arrived in Iran or been used by Iran in its war with the United States/Israel?	✓
27	Will the U.S. administration publicly propose, endorse, or be credibly reported to be advancing a Ukraine peace/security-guarantee framework requiring Ukraine to withdraw from currently held Donbas territory before 2026-05-01?	
28	Will Moldova officially report another Russian drone, missile, or war-related debris entering Moldovan airspace or territory during [2026-04-01, 2026-05-01), prompting airspace restrictions, a diplomatic protest, or a security response?	
29	Will Russian forces establish a verified persistent foothold or confirmed control over any part of Kostiantynivka, Donetsk Oblast, before 2026-05-01?	✓

Continued on next page

Table 23: GPT-5.5 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
30	Will EU or Ukrainian officials state before 2026-05-01 that April Verkhovna Rada votes have satisfied EU-requested reforms needed to unlock, preserve, or accelerate a pending Ukraine Facility disbursement of roughly €4 billion?	
31	Will Ukraine conduct another publicly confirmed FP-5 Flamingo missile strike on a Russian defense-industrial, munitions, missile, aircraft, or military-production facility before 2026-05-01?	
32	Will Russian forces capture or establish confirmed control of at least one additional settlement in Sumy Oblast, beyond Mala/Malaya Korchakivka, before 2026-05-01?	
33	Will Lithuania and Belarus hold a formal official bilateral meeting at deputy foreign minister level or higher before 2026-05-01?	
34	Will Ukraine publicly claim and provide credible confirmation of another hit on a Russian S-300, S-400, or Pantsir air-defense system in occupied Crimea or the southern occupied theater before 2026-05-01?	
35	Will EU member states formally approve the EU joining the Special Tribunal agreement, or will Ukraine/Council of Europe announce operationalization of the Special Tribunal for the crime of aggression against Ukraine, before 2026-05-01?	
36	Will Armenia and Russia announce before 2026-05-01 an agreement to transfer South Caucasus Railway management to a third country, or will Armenia formally move to denounce the 2008 Russian Railways concession?	
37	Will Ukrainian or Moldovan authorities announce new arrests, charges, or an interdicted operation before 2026-05-01 in a Russian-linked assassination, sabotage, or destabilization plot targeting Ukrainian officials, military personnel, journalists, or public figures?	
38	Will the Georgian Orthodox Church's Holy Synod elect or formally designate a new Catholicos-Patriarch before 2026-05-01?	
39	Will a publicly confirmed group of Ukrainian children abducted, deported, or held by Russia/Russian-controlled authorities be returned to Ukraine or their families before 2026-05-01 through U.S.-, Qatar-, or other third-party-mediated efforts?	
40	Will Moldova complete the final parliamentary/legal steps to withdraw from the Commonwealth of Independent States' statutory bodies before 2026-05-01?	✓
41	Will Russia, Gazprom, Turkey, or European transmission operators report a suspension or material reduction of gas flows through TurkStream or Blue Stream before 2026-05-01, due either to claimed Ukrainian attacks/sabotage or to a Russian policy decision?	
42	Will Moldova be formally accepted or publicly listed as a participant in the Ukraine "Coalition of the Willing" before 2026-05-01?	
43	Will Russian forces establish a verified foothold inside the built-up area of Huliaipole, Zaporizhzhia Oblast, or confirmed control over any part of the town, before 2026-05-01?	
44	Will Czech or allied authorities publicly attribute the Pardubice/LPP Holding drone-warehouse arson case to Russian intelligence/proxies, or announce Russia-linked charges or extraditions in that case, before 2026-05-01?	
45	Will the EU, United States, United Kingdom, or at least three EU member states impose new targeted sanctions or visa bans on Georgian Dream officials, judges, or security figures over political prisoners or repression before 2026-05-01?	
46	Will official statements or credible public reporting confirm before 2026-05-01 that at least one new Russian long-range attack-drone ground-control or launch station has become operational on Belarusian territory?	
47	Will Ukraine conduct another publicly confirmed Storm Shadow or SCALP cruise-missile strike on internationally recognized Russian territory before 2026-05-01?	
48	Will Ukraine publicly confirm damage to or destruction of another Russian Navy or FSB military vessel in the Black Sea, Crimea/Novorossiysk, or Baltic theater before 2026-05-01?	
49	Will Armenia and Azerbaijan sign or ratify their initialed peace treaty before 2026-05-01?	
50	Will the United States formally notify NATO/Congress or publicly announce another reduction or redeployment of at least 500 U.S. troops from Romania, Poland, Estonia, Latvia, or Lithuania before 2026-05-01?	
51	Will Russia conduct a publicly confirmed strike on a Ukrainian dam, hydropower plant, reservoir, or water-supply facility that triggers emergency water restrictions, a major flooding/pollution alert, or cross-border impacts before 2026-05-01?	

Continued on next page

Table 23: GPT-5.5 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
52	Will Russia's State Duma and/or President Putin formally adopt or sign a bill expanding the use of Russian armed forces abroad to "protect" Russian citizens arrested or prosecuted overseas before 2026-05-01?	✓
53	Will any European "coalition of the willing" member publicly announce a concrete plan to deploy peacekeeping, reassurance, or security-guarantee troops to Ukraine before 2026-05-01?	
54	Will Ukraine carry out or claim an attack that damages the Kerch/Crimean Bridge itself or forces it to close for at least 24 consecutive hours before 2026-05-01?	
55	Will Russia strike Kyiv's main railway station or another major Ukrainian passenger railway hub in a way that causes casualties, a fire, or suspension of passenger operations before 2026-05-01?	
56	Will Russian forces establish confirmed control over a substantial part of Kupiansk city, beyond isolated small groups, or force a Ukrainian withdrawal from the city before 2026-05-01?	
57	Will Russian forces be credibly reported to capture or establish confirmed control over at least one additional settlement in Dnipropetrovsk Oblast beyond Sheviakivka before 2026-05-01?	
58	Will Russian forces be credibly reported to capture all remaining Ukrainian-held parts of Chasiv Yar, Donetsk Oblast, before 2026-05-01?	
59	Will Romania's Social Democratic Party decide to leave the governing coalition, or will Prime Minister Ilie Bolojan resign, before 2026-05-01?	
60	Will Poland formally secure access to, or enact a workaround for, the roughly €44 billion EU SAFE defense-loan package despite President Karol Nawrocki's veto before 2026-05-01?	
61	Will Hungary declassify or release the national security report alleging Ukrainian funding or interference in the Tisza opposition campaign before 2026-05-01?	
62	Will Ukraine's Verkhovna Rada pass a law or binding national rule limiting the use of force or physical detention by recruitment centers during mobilization before 2026-05-01?	
63	Will Armenia and Azerbaijan sign or formally announce an implementation agreement on the TRIPP/Zangezur/Nakhchivan transport corridor before 2026-05-01?	
64	Will Belarus release another publicly confirmed group of at least 100 political prisoners or detainees in a U.S.-brokered or sanctions-linked deal before 2026-05-01?	
65	Will Lithuania, the EU, or a Lithuanian transport/port operator publicly agree to allow, facilitate, or formally study renewed Belarusian potash transit/export through Lithuania before 2026-05-01?	
66	Will credible reporting or official statements say that Georgia's Kulevi/Black Sea Petroleum terminal or refinery continued accepting or processing Russian crude or Russian-linked tanker cargo after its sanctions-compliance pledge before 2026-05-01?	
67	Will Moldova formally publish, submit, or endorse a reintegration plan for Transnistria involving an international or external civilian administration, transitional authority, or convergence fund before 2026-05-01?	
68	Will European authorities announce additional arrests, extraditions, or criminal charges linked to the Russian-GRU-backed parcel bomb or parcel-fire network targeting the UK, Germany, Poland, or Lithuania before 2026-05-01?	
69	Will Polish authorities publicly report discovering, sealing, or destroying another Belarus-linked migrant-smuggling tunnel, or a major Belarus-directed violent migrant assault on border guards, before 2026-05-01?	
70	Will Armenia formally request a revision, drawdown, or withdrawal of Russian military-base or border-guard presence, beyond already departed border guards, before 2026-05-01?	
71	Will Poland formally request to join nuclear-sharing arrangements, host allied nuclear-capable assets, or negotiate inclusion under France's nuclear umbrella before 2026-05-01?	
72	Will Zelenskyy, the Verkhovna Rada, or Ukraine's Central Election Commission formally announce a date or legal mechanism for national elections or a peace referendum before 2026-05-01?	

Continued on next page

Table 23: GPT-5.5 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
73	Before 2026-05-01, will Hungary materially reduce or suspend natural-gas exports to Ukraine, beyond normal commercial fluctuations, as leverage over the Druzhba oil-transit dispute?	
74	Before 2026-05-01, will credible mapping, official statements, or major battlefield reporting indicate that Ukrainian forces have lost confirmed control of Pokrovsk or Myrnohrad, either through Russian capture of most/all of either city or an announced Ukrainian withdrawal?	
75	During April 2026, will official Russian/Belarusian statements or independent public intelligence/satellite imagery confirm that Oreshnik launchers themselves—not just air defenses or construction—are on combat duty at a specific Belarusian site?	
76	Before 2026-05-01, will Ukraine publicly claim, or will credible reporting confirm, an April strike that damages a Russian military aircraft plant, aircraft repair facility, strategic-aviation airbase, or parked military aircraft, excluding purely oil-refining or oil-export targets?	
77	Before 2026-05-01, will a new Baltic Sea undersea cable, pipeline, or power interconnector suffer damage or an outage that authorities investigate as possible sabotage or link to a Russian/shadow-fleet vessel?	
78	During April 2026, will Estonia, Latvia, Lithuania, Poland, or Finland officially accuse a Russian manned military aircraft of violating national airspace, triggering a NATO scramble, diplomatic protest, or public military response?	
79	Before 2026-05-01, will Poland, Estonia, Latvia, or Lithuania announce a concrete order, contract, emplacement, or fielding plan for anti-personnel or "smart" mine systems along borders with Russia or Belarus?	
80	Before 2026-05-01, will Georgian authorities arrest, convict, imprison, or order pre-trial detention for another nationally prominent opposition leader, civil-society leader, or pro-EU protest organizer?	
81	Before 2026-05-01, will Ukrainian forces be credibly reported to liberate at least one additional occupied settlement in the Zaporizhzhia/western-Donetsk southern theater, such as on the Oleksandrivka, Orikhiv, or Huliaipole axes?	
82	Before 2026-05-01, will the United States and Belarus announce an additional normalization step beyond the March prisoner/sanctions deal—such as reopening an embassy, naming or returning ambassadors, a new general license or sanctions-relief package, or formal high-level follow-up talks?	
83	During April 2026, will Moldova impose mandatory electricity consumption restrictions, rolling blackouts, or emergency industrial curbs because Russian strikes or grid instability again make import routes insufficient?	
84	Before 2026-05-01, will Russian forces establish a confirmed sustained foothold on a Dnipro River island or a new right-bank bridgehead in the Kherson area, beyond short failed landing attempts?	
85	Before 2026-05-01, will Ukraine officially reopen any part of its airspace to commercial passenger or cargo flights, or permit a publicly confirmed commercial flight from Lviv, Boryspil, or another Ukrainian airport?	
86	During April 2026, will Russia launch at least 900 one-way attack drones and/or missiles at Ukraine within any 24-hour period?	
87	Before 2026-05-01, will Russian oil companies, Transneft, port operators, or major buyers publicly report force majeure or a prolonged mid-April-or-later halt/material reduction of loadings from Ust-Luga or Primorsk due to Ukrainian drone damage?	
88	During April 2026, will credible Ukrainian, ISW, or major-media reporting describe a Russian battalion-sized or larger mechanized/motorized assault on the Lyman–Sloviansk/Kramatorsk axis or the northern approaches to Ukraine's Donetsk "Fortress Belt"?	
89	Before 2026-05-01, will Russian drone, missile, or naval-drone attacks force a publicly reported temporary suspension of operations or vessel movements at a Ukrainian Black Sea or Danube export port such as Odesa, Chornomorsk, Pivdennyi, Izmail, Reni, or Vylkove?	
90	During April 2026, will Ukraine's forces or affiliated partisan networks publicly claim and provide credible evidence of a strike or sabotage that disrupts a Russian railway line, locomotive, or fuel/ammunition train used to supply a front sector?	

Continued on next page

Table 23: GPT-5.5 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
91	Before 2026-05-01, will Transnistria's de facto authorities extend their economic state of emergency again, or impose/renew major gas, heating, industrial, or fuel restrictions after the current 30-day extension?	
92	Before 2026-05-01, will Saudi Arabia, Qatar, the UAE, or Kuwait publicly announce a concrete purchase, order, or funding package for Ukrainian interceptor drones, electronic-warfare systems, or counter-drone training tied to new air-defense cooperation?	
93	Before 2026-05-01, will Russia publicly confirm, or will credible reporting show, a Russian Navy escort, mobile fire group, or armed security team deployed aboard or alongside a Russian-flagged or Russian-linked shadow-fleet tanker or cargo vessel?	
94	Before 2026-05-01, will Bulgarian authorities, the EU Rapid Alert System, or EU/NATO partners publicly identify, remove, or formally respond to a coordinated Russian-linked disinformation or hybrid campaign tied to Bulgaria's April 19 parliamentary election?	
95	Before 2026-05-01, will Polish investigators or government officials release expert findings on the September 2025 drone incursion that establish the drone type/number, whether any were armed or explosive, or the cause of property damage from interception attempts?	
96	Before 2026-05-01, will Armenia publicly release a draft new Constitution whose preamble or text omits the current reference to the Declaration of Independence or other language Azerbaijan treats as a territorial claim to Nagorno-Karabakh/Artsakh?	
97	Before 2026-05-01, will Moldova's Constitutional Court rule on the Justice Ministry petition challenging Gagauzia's unilateral appointment of its electoral body or other contested provisions of Gagauzia's special-status law?	
98	During April 2026, will Poland, Latvia, Lithuania, Estonia, Finland, Romania, or Bulgaria publicly report or attribute a new sabotage, arson, or cyber-physical incident against railway infrastructure or rolling stock to Russian services, Russian proxies, or actors working for Russia?	
99	Before 2026-05-01, will Belarus and North Korea announce a concrete implementation step under their new treaty involving defense, military-technical cooperation, arms-related industry, sanctions-evasion logistics, or labor deployments?	
100	During April 2026, will Russia or Belarus launch an Oreshnik/RS-26-type intermediate-range ballistic missile against Ukraine or another European target?	

Claude Opus 4.8.

Table 24: Claude Opus 4.8 forecast-time questions, run 1 top-100.

Rank	Forecast-time question	Match
1	Will Viktor Orban's Fidesz lose its parliamentary majority (or Orban fail to remain PM) in Hungary's April 12, 2026 election?	✓
2	Will Russia and Ukraine agree to and begin observing a mutual halt on strikes against each other's energy infrastructure (an 'Easter'/energy truce) during April 2026?	
3	Will a Russia-linked drone/airspace incident on NATO territory (Baltic states or Finland) trigger an Article 4 consultation, a shoot-down, or a formal escalatory response during April 2026?	
4	Will Russian forces capture a significant Donetsk Oblast city (e.g., fully seize Pokrovsk, Kostiantynivka, or another major town) during April 2026?	
5	Will the United States publicly withdraw from or suspend its mediation of the Russia-Ukraine peace negotiations during April 2026?	
6	Will Ukraine agree to significantly reduce or pause its long-range strikes on Russian oil/energy infrastructure during April 2026?	
7	Will the EU approve or operationalize its ˜EUR90bn frozen-Russian-asset-backed financing for Ukraine during April 2026?	✓
8	Will the United States formally lift or further ease sanctions on Russia's oil/energy sector during April 2026?	
9	Will mass unrest over Serbia's disputed elections force a major political concession (snap national election call, government resignation, or comparable crisis) during April 2026?	

Continued on next page

Table 24: Claude Opus 4.8 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
10	Will a trilateral US-Russia-Ukraine leaders' meeting OR a Zelensky-Putin meeting take place, OR a formal comprehensive peace/ceasefire framework be signed, during April 2026?	
11	Will Donald Trump meet Alexander Lukashenko in person, OR will the US formally lift major Belarus sanctions (potash/financial) or reopen its Minsk embassy, during April 2026?	
12	Will Russian forces fully capture Kupiansk during April 2026?	
13	Will Ukraine suffer a major government crisis during April 2026 - e.g., the collapse of the Servant of the People majority, PM Sviridenko's resignation or a cabinet reshuffle, mass deputy resignations, or a stalled/suspended IMF program?	
14	Will Russian crude deliveries via the Druzhba pipeline to Hungary (and Slovakia) remain halted for a sustained period through April 2026, or will Hungary/Slovakia declare an energy emergency over the disruption?	
15	Will Russia experience a visible domestic fuel crisis in April 2026 - e.g., extending its gasoline export ban to diesel, or widely reported gasoline/diesel shortages - driven by Ukrainian refinery strikes?	
16	Will Armenia and Azerbaijan formally sign a peace treaty (or take a decisive, treaty-level step such as initialing/ratification) during April 2026?	
17	In Bulgaria's April 19, 2026 early parliamentary election, will Rumen Radev's new 'Progressive Bulgaria' coalition finish first (i.e., GERB-SDS lose its plurality)?	✓
18	Will the Transnistria energy/economic crisis escalate decisively during April 2026 - e.g., a renewed state of emergency, appeal for Russian intervention/aid, or a major Moldovan reintegration step?	
19	Will the EU formally adopt a new sanctions package against Russia (the prepared '20th' package) during April 2026?	✓
20	Will Georgia's Constitutional Court rule to outlaw the main opposition (UNM and satellites) OR will the EU/US impose new escalatory measures (sanctions or visa-free suspension) on Georgian Dream during April 2026?	
21	Will a Russian aerial campaign cause an extended near-nationwide or multi-region emergency blackout (or a single mass-casualty strike) in Ukraine during April 2026?	
22	Will Russia overtly enter the US-Israel-Iran war during April 2026 - e.g., deploy combat forces to Iran or directly engage US/Israeli forces (beyond intelligence/technical aid)?	
23	Will Russia announce a new partial or general mobilization wave (beyond the routine spring conscription) during April 2026?	
24	Will Ukraine formally agree to cede or withdraw its troops from the unoccupied part of Donetsk Oblast (Donbas) as part of a peace settlement during April 2026?	
25	Will the United States visibly divert, reduce, or suspend critical weapons deliveries (especially air-defense interceptors) to Ukraine during April 2026 because of demands from the US-Israel-Iran war?	
26	Will Russia exhibit an acute, officially visible economic shock during April 2026 — e.g., a major state-firm bailout/default, an emergency rate move, or official acknowledgment of recession?	
27	Will the Special Tribunal for the Crime of Aggression against Ukraine be formally launched/inaugurated or take a decisive operational step (e.g., first session or first charges) during April 2026?	
28	Will a Russian-orchestrated targeted assassination or major sabotage attack against military/political figures or critical infrastructure be successfully carried out (or foiled at high profile) in Ukraine or an EU state during April 2026?	
29	Will Russian forces capture a notable settlement inside Dnipropetrovsk Oblast (or make a confirmed significant advance there) during April 2026?	
30	During April 2026, will a Russian oil major declare force majeure on Baltic-port crude exports, OR will Russian seaborne crude exports remain at a multi-year low because of sustained Ukrainian strikes on terminals such as Primorsk and Ust-Luga?	
31	Will Kosovo's Assembly fail to elect a new president by the Constitutional Court's April 28, 2026 deadline, automatically triggering fresh parliamentary elections?	✓
32	During April 2026, will Moldova implement sustained rolling blackouts or otherwise sharply escalate the 60-day national energy emergency triggered by Russian strikes on Ukraine's grid?	

Continued on next page

Table 24: Claude Opus 4.8 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
33	During April 2026, will Slovakia take a concrete step toward snap elections - e.g., President Pellegrini approving (or referring to the Constitutional Court) the opposition's referendum petition, or a no-confidence move against Fico's government?	
34	Will Russia and Ukraine carry out a major prisoner-of-war exchange (e.g., several hundred per side or a 'large for large' swap) in connection with Orthodox Easter (April 12) during April 2026?	✓
35	During April 2026, will Ukrainian-origin drones (errant or Russian-redirected) cause a serious incident on NATO/EU territory - e.g., casualties, significant infrastructure damage, or a formal diplomatic rupture between Ukraine and a Baltic state/Finland?	
36	During April 2026, will the Republika Srpska/Dodik standoff sharply escalate - e.g., the US re-imposing sanctions on Milorad Dodik, or RS institutions taking a concrete secessionist/non-cooperation step?	
37	Will Ukraine carry out another naval-drone or missile strike that damages or sinks a Russian shadow-fleet oil tanker in the Black Sea (or near the Bosphorus/Turkish coast) during April 2026?	
38	Will Babis's new Czech government adopt or have parliament pass its public-media defunding or 'foreign agent' legislation, OR face a major escalation (no-confidence motion or sustained nationwide protests) during April 2026?	
39	Will Belarus release another significant batch of political prisoners (e.g., 200+ or prominent named detainees) under the US-brokered process during April 2026?	
40	Will a NATO/EU state (e.g., France, UK, or a Baltic state) seize, board, or detain a Russian shadow-fleet tanker during April 2026?	
41	During April 2026, will a local ceasefire be agreed to enable Zaporizhzhia NPP repairs, OR will the plant suffer a serious nuclear-safety incident (e.g., extended total loss of external power)?	
42	Will the EU formally open the first negotiating cluster (e.g., 'Fundamentals') in Ukraine's accession talks during April 2026, overcoming Hungary's veto?	✓
43	Will Ukraine enact a major change to its mobilization rules during April 2026 (e.g., lowering the conscription age, a new AWOL amnesty, or a recruitment-system overhaul)?	
44	Will Pashinyan's government take a major escalatory step against the opposition during April 2026 (e.g., banning/jailing a prominent opposition figure or party, nationalizing Karapetyan assets, or a sharp church-state confrontation)?	
45	During April 2026, will Russia/Belarus declare the nuclear-capable Oreshnik missile system on combat duty in Belarus, or conduct a demonstrative/test launch from Belarusian territory?	
46	During April 2026, will the Belarus/Russia-orchestrated hybrid-migration pressure on Poland/Lithuania escalate materially - e.g., a mass crossing surge, a border-guard clash/casualty, discovery of further smuggling tunnels, or new emergency border closures?	
47	During April 2026, will the 'Coalition of the Willing' make a formal decision or announcement to deploy a reassurance/peacekeeping military force to Ukraine (beyond planning)?	
48	During April 2026, will the EU or a member state take a formal step (investigation, exclusion from sensitive meetings, sanctions referral, or comparable measure) in response to allegations that Hungary shared classified EU Council information with Russia?	
49	During April 2026, will Ukraine carry out a successful deep strike that destroys or seriously damages a high-value Russian strategic military asset (e.g., a strategic bomber, A-50 AWACS, major warship, or strategic air base)?	
50	Will Russian forces capture a significant settlement in Kharkiv or Sumy oblast (e.g., Vovchansk, or a notable town inside the renewed northern offensive zone) during April 2026?	
51	Will Russian gas supply via TurkStream to Hungary and/or Serbia be halted or significantly disrupted (by Ukrainian strikes on the Russkaya/feeder compressor stations or a related incident) during April 2026?	
52	Will Ukraine's anti-corruption bodies formally charge/name Andriy Yermak (or another top official) as a suspect in the Midas/Mindich case, OR will the government's NABU audit trigger a major institutional clash, during April 2026?	

Continued on next page

Table 24: Claude Opus 4.8 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
53	Will the Azerbaijan-Russia diplomatic rift escalate to concrete measures (e.g., expulsions, arrests of nationals, suspension of agreements, or a sharp public rupture) during April 2026?	
54	Will Poland secure access to EU SAFE defense-loan financing despite President Nawrocki's veto (via override, court ruling, or government workaround), OR will the Nawrocki-Tusk conflict escalate into a major constitutional confrontation, during April 2026?	
55	Will Russia's hybrid 'Narva People's Republic' destabilization campaign in Estonia trigger a concrete escalation (e.g., Estonian emergency/security measures, diplomatic expulsions, an arrest of organizers, or a separatist incident) during April 2026?	
56	Will Lukoil complete a binding/OFAC-approved sale of its international (foreign) assets before the US Treasury's May 1, 2026 deadline, rather than the deadline being extended again or the assets forced into wind-down?	
57	Will Romania introduce fuel rationing, declare an energy/fuel emergency, or see its Lukoil-owned Petrotel refinery halt operations because of the Lukoil sanctions and oil-price shock during April 2026?	
58	Will Russia detain, charge, dismiss, or otherwise move directly against Sergei Shoigu, or arrest/charge another minister- or deputy-minister-level Russian defense/security official, during April 2026?	
59	Will Russian forces capture a significant settlement on the Zaporizhzhia-oblast front (e.g., Huliaipole or Orikhiv) during April 2026?	
60	Will it become evident during April 2026 that Russia is reaping a major oil windfall from the Iran-driven price shock — e.g., Urals/ESPO crude sustainably trading well above the $60 price cap, record/elevated monthly export earnings, or higher crude output/Asian shipments — despite Ukrainian strikes having cut a large share of its export capacity?	
61	Will the United States and Russia announce a concrete bilateral deliverable unrelated to a Ukraine ceasefire during April 2026 — e.g., a Trump-Putin bilateral summit/announced date, a prisoner/detainee swap, restored direct flights, or an economic/Arctic cooperation arrangement?	
62	Will the US or EU impose new sanctions/designations on Russia — or will Trump publicly reverse his conciliatory stance toward Moscow — specifically in response to Russian military or intelligence support for Iran during April 2026?	
63	Will the US announce a significant reduction of forces or security commitment affecting NATO's eastern flank, or will a major transatlantic security rupture occur (e.g., suspended intelligence/operational cooperation, or a key European ally publicly breaking with Washington), during April 2026?	
64	Will Russian military or technical personnel be killed by US/Israeli strikes in Iran, or will a direct US/Israel-Russia military incident occur, during April 2026?	
65	During April 2026, will Russia impose a nationwide ban/permanent block on Telegram and/or WhatsApp, OR will the internet-censorship crackdown trigger significant multi-city protests with mass arrests?	
66	During April 2026, will Ukraine (Zelensky or the Rada) announce, schedule, or pass enabling legislation for a wartime presidential election or peace referendum?	
67	During April 2026, will North Korea deploy a significant new contingent of troops (or confirm a major new deployment) to Russia for the war against Ukraine?	
68	During April 2026, will Donald Trump publicly confirm (or formally decline) attending Russia's May 9 Victory Day parade in Moscow?	
69	During April 2026, will Armenia take a decisive step to strip Russia of control over a strategic asset (e.g., the South Caucasus Railway) or otherwise trigger a sharp Yerevan-Moscow rupture?	
70	During April 2026, will Ukraine recapture a significant settlement or make a confirmed notable counter-advance on a front axis (e.g., the Oleksandrivka axis)?	
71	Will a Ukrainian-attributed operation kill or critically target a senior Russian military, security, political, or propaganda figure inside Russia (or deep in occupied territory) during April 2026?	
72	Will Russia return a significant group of forcibly deported/transferred Ukrainian children to Ukraine during April 2026?	

Continued on next page

Table 24: Claude Opus 4.8 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
73	Will the US (Trump administration) formally recognize, or commit in a published/signed framework to recognizing, Russian sovereignty over Crimea or other occupied Ukrainian territory during April 2026?	
74	Will Georgia's escalating opposition 'national mobilization' campaign culminate in a major government crackdown during April 2026 — e.g., a violent dispersal/state of emergency or the arrest of additional prominent opposition leaders?	
75	Will there be a direct Russian naval/military confrontation with a NATO state in the Baltic Sea (e.g., Russia deploying warships to escort or contest interdiction of shadow-fleet tankers) during April 2026?	
76	Will a Ukrainian long-range strike cause civilian casualties or significant structural damage within the city limits of Moscow or St. Petersburg during April 2026 (beyond routine airport delays)?	
77	At its April 2026 monetary-policy meeting, will the Bank of Russia cut its key interest rate again (rather than hold or hike)?	✓
78	Will Lithuania (or the EU) reopen one or more closed road border crossings with Belarus, or otherwise reach a normalization deal on the border/truck dispute, during April 2026?	
79	During April 2026, will Ukrainian strikes on Novorossiysk-area oil infrastructure cause a confirmed disruption to Black Sea crude exports, including the Caspian Pipeline Consortium (CPC) handling Kazakh oil?	
80	Will Russia and China sign a binding Power of Siberia 2 pipeline gas contract (or announce a firm investment/construction decision) during April 2026?	
81	During April 2026, will Russia conduct or formally announce a nuclear weapons test or a demonstrative strategic-weapon launch (e.g., Burevestnik/Sarmat) in response to the post-New START environment and US testing moves?	
82	During April 2026, will a Russia-linked drone/missile airspace violation or drifting Black Sea sea mine cause a serious incident on Romanian (NATO) territory or waters — e.g., a shoot-down, casualty, infrastructure damage, or formal Article 4/escalatory response?	
83	During April 2026, will Serbia's NIS (Gazprom Neft-owned) oil company lose its US sanctions waiver and face a crude-supply halt/refinery shutdown or fuel emergency, OR will Gazprom complete a forced divestment of NIS?	
84	During April 2026, will there be a direct military exchange between Azerbaijan and Iran — e.g., a further Iranian drone/missile strike on Azerbaijani territory or an Azerbaijani retaliatory strike — as the US-Israel-Iran war spills into the Caspian?	
85	During April 2026, will the EU formally open a new negotiating cluster with Moldova (an intergovernmental conference opening one of the remaining clusters), advancing its accession independently of Ukraine?	
86	During April 2026, will Bulgaria take a decisive step on the Lukoil Neftohim Burgas refinery — e.g., a state takeover/special-management escalation, a forced sale, or an operational shutdown — driven by US sanctions and the May 1 deadline?	
87	During April 2026, will Russian strikes on civilian/grain shipping or Odesa-region ports cause a major disruption to Ukraine's Black Sea export corridor (e.g., a vessel sunk, mass casualties on a merchant ship, an insurer/operator suspension, or a formal corridor halt)?	
88	Will Romania's Bolojan coalition government fall during April 2026 (PM resignation, successful no-confidence vote, or coalition breakup)?	✓
89	During April 2026, will the standoff between Chisinau and Gagauzia escalate decisively (e.g., Gagauzia declaring an emergency/non-cooperation step, Chisinau dissolving Gagauz bodies, or new arrests/charges of Gagauz officials)?	
90	During April 2026, will France/Europe deliver — or Ukraine combat-deploy — the new SAMP/T NG air-defense system against Russian ballistic missiles?	
91	During April 2026, will Armenia and Azerbaijan (with the US/Turkey) take a concrete operational step on the TRIPP/Zangezur corridor — e.g., construction start, an opened transit route, or a signed implementation/operator agreement?	
92	During April 2026, will there be a concrete step toward a US takeover or restart of the Nord Stream pipelines (e.g., an announced US-Russia/Germany arrangement, a consortium/ownership deal, or formal US bid)?	
93	During April 2026, will a Russian (or Russian-redirected) drone or missile penetrate Polish airspace and trigger a NATO/Polish shoot-down, Article 4 consultation, or comparable formal escalatory response?	

Continued on next page

Table 24: Claude Opus 4.8 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
94	During April 2026, will a Ukrainian strike cause confirmed significant damage to, or a multi-day closure of, the Kerch Strait (Crimean) Bridge?	
95	During April 2026, will Russia introduce a new one-off windfall/excess-profit tax or a formal forced 'voluntary' contribution scheme on oligarchs/large business to support the budget or war effort?	
96	During April 2026, will the UK and Russia carry out a round of reciprocal diplomatic expulsions (beyond the already-reported expulsion of a single British diplomat)?	
97	During April 2026, will a Russian strike kill foreign nationals or directly hit a Western diplomatic mission, consulate, or international aid facility in Ukraine, prompting a formal diplomatic protest?	
98	Will a major European Ukraine backer (France, Germany, Italy, UK, Poland, or the EU collectively) publicly announce a reduction, suspension, delay, or new conditionality on its military or financial support to Ukraine during April 2026?	
99	Will Ukraine sign or bring into operation a major new non-Western (e.g., Gulf-state) air-defense or anti-drone supply/co-production partnership during April 2026?	
100	Will North Korea and Russia announce a major new escalation in their cooperation during April 2026 - e.g., a new bilateral agreement/treaty, a leader-level meeting, or an officially confirmed large new deployment of troops or munitions?	

Claude Fable 5.

Table 25: Claude Fable 5 forecast-time questions, run 1 top-100.

Rank	Forecast-time question	Match
1	Will Peter Magyar's Tisza party win more parliamentary seats than Viktor Orban's Fidesz in Hungary's April 12, 2026 parliamentary election?	✓
2	Will Russia and Ukraine agree to any ceasefire (an Easter truce and/or a mutual halt to strikes on energy infrastructure) that takes effect before May 1, 2026?	
3	Will Russian forces capture (be credibly assessed as controlling) Pokrovsk, Kostiantynivka, or Huliaipole before May 1, 2026?	
4	Will the EU formally adopt its 20th package of sanctions against Russia before May 1, 2026?	✓
5	Will the EU make the first disbursement of the agreed €90 billion loan to Ukraine before May 1, 2026?	✓
6	Will crude oil flows through the Druzhba pipeline to Hungary and/or Slovakia resume before May 1, 2026?	✓
7	Will a new round of direct Russia-Ukraine negotiations, or a trilateral US-Russia-Ukraine meeting (at leader or senior-envoy level), take place before May 1, 2026?	
8	Will a Ukrainian strike force the suspension of operations at a major Russian oil export terminal or refinery complex (e.g., Ust-Luga, Primorsk, Novorossiysk, Kirishi) during April 2026?	
9	Will a NATO eastern-flank member invoke Article 4 over an airspace/hybrid incident, or will NATO/allied forces shoot down a drone or aircraft inside allied airspace, before May 1, 2026?	
10	Will a local ceasefire around the Zaporizhzhia nuclear plant be agreed, or its lost main 750kV external power line be repaired/restored, before May 1, 2026?	
11	If Fidesz loses the April 12 vote, will Orban publicly concede (or the result be certified without formal legal challenge alleging fraud) before May 1, 2026 — or conversely, will the losing side formally contest the result?	✓
12	Will the US and Belarus take a further concrete normalization step before May 1, 2026 (e.g., release of additional political prisoners beyond the 250, additional sanctions relief, or an announced restoration of embassy operations)?	
13	Will US Vice President JD Vance travel to Hungary before the April 12, 2026 election in a show of support for Viktor Orban?	✓

Continued on next page

Table 25: Claude Fable 5 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
14	Will Transnistria experience a full halt of gas supplies / new formal state of emergency, or will Chisinau and Tiraspol announce a concrete new energy-or-reintegration arrangement, before May 1, 2026?	
15	Will Russia and Ukraine complete the agreed 500-for-500 prisoner exchange (or conduct another exchange of at least ˜200 POWs each) before May 1, 2026?	✓
16	Will Poland enact a legal pathway to access its ˜€44 billion in EU SAFE defense loans (via a new/amended law, veto override, or alternative government mechanism) before May 1, 2026?	
17	Will the US announce a further easing/extension of Russian oil sanctions relief (or formally reverse the March waiver) before May 1, 2026?	✓
18	Will Russia conduct at least one single-day (or overnight) attack on Ukraine involving 1,000 or more drones and missiles combined (per Ukrainian Air Force figures) during April 2026?	
19	Will British forces board, detain, or seize at least one Russian shadow-fleet vessel in UK waters (or will the UK publicly confirm such an operation) before May 1, 2026?	
20	Will Donald Trump and Vladimir Putin hold an in-person bilateral meeting (summit) before May 1, 2026?	
21	Will President Zelensky (or the Verkhovna Rada) formally announce or schedule a Ukrainian presidential election and/or a national referendum on peace terms before May 1, 2026?	
22	Will Ukraine's government or Verkhovna Rada formally adopt 'key changes' to mobilization and/or AWOL (SZCh) policy — via a passed law or government decree — before May 1, 2026?	
23	Will the EU formally (not just 'informally') open at least one negotiating cluster of Ukraine's accession talks before May 1, 2026?	✓
24	Will Russia publicly announce — or the US publicly confirm — an agreement under which Russia halts intelligence-sharing or military support to Iran before May 1, 2026?	
25	Will Armenia and Azerbaijan sign their initialed final peace treaty before May 1, 2026?	
26	Will the EU formally ease, suspend, or lift any of its sanctions on Belarus before May 1, 2026?	
27	Will the EU propose/adopt an extension of the visa-free suspension to ordinary Georgian citizens, or will the EU or US impose new sanctions on Georgian Dream officials, before May 1, 2026?	
28	Will Slovakia implement additional concrete measures against Ukraine over the Druzhba pipeline dispute before May 1, 2026 — e.g., extending/expanding its diesel export-sale restrictions, halting electricity or fuel exports to Ukraine, or formally vetoing an EU decision in support of Ukraine?	
29	Will a concrete formal step toward US (or US-backed) acquisition, control, or restart of the Nord Stream pipelines be announced before May 1, 2026?	
30	Will Rumen Radev's 'Progressive Bulgaria' coalition win the most votes (finish first) in Bulgaria's snap parliamentary election on April 19, 2026?	✓
31	Will a Moldovan court convict Vladimir Plahotniuc and sentence him to a prison term (prosecutors seek 25 years) in the 'billion-dollar theft' case before May 1, 2026?	✓
32	Will the US Treasury (OFAC) license/approve completion of the sale of Lukoil's international assets (e.g., to Carlyle) by the extended May 1, 2026 deadline?	
33	Will the US extend the OFAC operating waiver for Serbia's Gazprom-controlled NIS beyond its April 17, 2026 expiry, or will a sale of Gazprom's NIS stake (e.g., to MOL) be announced before May 1, 2026?	
34	Will the US officially divert, suspend, or visibly cut deliveries of Patriot interceptors or other PURL-funded weapons earmarked for Ukraine in order to support the Middle East/Iran war, before May 1, 2026?	
35	Will Timur Mindich (and/or Oleksandr Tsukerman) be extradited from Israel — or will Israel formally approve their extradition — to Ukraine before May 1, 2026?	
36	Will Ukraine publicly accept or sign a US-mediated peace framework document that includes Ukrainian withdrawal from the remainder of the Donbas in exchange for security guarantees, before May 1, 2026?	

Continued on next page

Table 25: Claude Fable 5 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
37	Will Moldova's Constitutional Court rule on the government's petition against Gagauzia's power to appoint its own electoral body — or will new Gagauz People's Assembly elections be formally scheduled — before May 1, 2026?	
38	Will a new incident of suspected sabotage damaging an undersea cable or pipeline in the Baltic Sea (with a vessel detained or sabotage formally suspected by authorities) occur during April 2026?	
39	Will the US administration officially announce a reduction or withdrawal of US troops stationed in Europe that affects NATO's eastern flank (Poland, Romania, the Baltics, or Germany-based forces supporting them) before May 1, 2026?	
40	Will the US and Russia publicly announce the start of formal negotiations (or hold a first round of talks) on a successor strategic nuclear arms-control arrangement before May 1, 2026?	
41	Will Vilnius (or Kaunas) airport suspend flight operations at least once during April 2026 due to contraband balloons or similar incursions from Belarus?	
42	Will the Petrotel Lukoil refinery in Romania formally resume operations (restart processing) before May 1, 2026?	
43	Will Russia formally announce a new wave of military mobilization (a presidential decree or official government announcement of partial/general mobilization beyond routine conscription) before May 1, 2026?	
44	Will at least one tanker carrying Russian oil (or a Russian shadow-fleet vessel) be struck by a drone or explosion in the Black Sea/Bosphorus area or the Mediterranean during April 2026?	
45	Will Ukrainian anti-corruption bodies (NABU/SAPO) serve a notice of suspicion to — or bring formal charges against — Andriy Yermak in connection with the Midas/Mindich case before May 1, 2026?	
46	Will the EU formally open accession negotiations on at least one cluster with Moldova (e.g., via an intergovernmental conference), decoupled from Ukraine's blocked track, before May 1, 2026?	
47	Will the EU (or G7 acting with the EU) formally ease any energy-related sanctions on Russia — e.g., raising or suspending the oil price cap, or granting new exemptions for Russian oil/gas purchases — before May 1, 2026?	
48	Will Georgian authorities formally initiate a ban on major opposition parties (e.g., a Constitutional Court submission) or arrest/sentence another nationally prominent opposition leader before May 1, 2026?	
49	Will the Georgian Orthodox Church's Holy Synod elect a new Catholicos-Patriarch (successor to Ilia II) before May 1, 2026?	
50	Will authorities in an EU eastern-flank state (Poland, Czechia, Slovakia, the Baltics, Romania or Bulgaria) publicly attribute a new land-based sabotage, arson, or rail/infrastructure attack (excluding undersea cables) to Russia or Russian-recruited agents during April 2026?	
51	Will Serbia formally call early parliamentary elections (presidential decree dissolving parliament and/or officially scheduling a snap vote, e.g., for June 28) before May 1, 2026?	
52	Will Kosovo hold its snap parliamentary election — or at minimum have the election date formally announced — before May 1, 2026?	✓
53	Will the Crimean (Kerch) Bridge be hit by a Ukrainian attack causing visible damage and/or a traffic closure of 24 hours or more during April 2026?	
54	Will Russia be officially confirmed (by Russia, the US, NATO, or Ukrainian intelligence with US corroboration) to have deployed military personnel or transferred major weapons systems (e.g., air-defense systems, combat aircraft) to Iran before May 1, 2026?	
55	Will Russia (or Belarus) launch an Oreshnik intermediate-range ballistic missile — either a combat strike on Ukraine or a declared test/demonstration launch from Belarusian territory — during April 2026?	
56	Will the US government formally reimpose (reinstate) sanctions on Milorad Dodik and/or other Republika Srpska officials or RS-linked entities before May 1, 2026?	
57	Will Russian forces be credibly assessed (e.g., by ISW or Ukraine's General Staff) as controlling Lyman or Kupiansk before May 1, 2026?	
58	Will Ukraine sign a firm purchase contract and/or make a first (pre)payment for Saab Gripen E and/or Dassault Rafale fighter jets before May 1, 2026?	

Continued on next page

Table 25: Claude Fable 5 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
59	Will the US administration officially announce a reduction or withdrawal of US troops serving in Kosovo (KFOR) before May 1, 2026?	
60	Before May 1, 2026, will Russia and Azerbaijan either (a) announce a formal settlement of the AZAL plane crash dispute (official Russian acceptance of responsibility, compensation, and/or prosecutions), or (b) suffer a new formal diplomatic rupture (ambassador recall/expulsion or suspension of high-level ties)?	
61	Will at least one European 'coalition of the willing' state formally approve (via government decision or parliamentary vote) the deployment of its troops to Ukraine as part of a reassurance/peacekeeping force before May 1, 2026?	
62	Will Moldova (government or Moldelectrica) implement scheduled rolling power cuts / load-shedding for consumers at any point during April 2026, amid its declared 60-day energy emergency?	
63	Will Russian forces be credibly assessed (e.g., by ISW, DeepState, or Ukraine's General Staff) as having captured Siversk or entered the urban outskirts of Sloviansk or Kramatorsk before May 1, 2026?	
64	Will Ukrainian forces be credibly assessed (e.g., by ISW or DeepState) as having recaptured Huliaipole — or forced a Russian withdrawal from / encirclement of the town — before May 1, 2026?	
65	Will Estonia or Latvia close part of its airspace or suspend operations at a major airport (e.g., Tallinn or Riga) at least once during April 2026 due to drone incursions or related airspace threats?	
66	Will Armenian authorities arrest or criminally charge another nationally prominent opposition leader (e.g., a Kocharyan-bloc figure or 'Strong Armenia'/Karapetyan-linked leader), or formally bar an opposition party or its candidate list from the June 7 parliamentary election, before May 1, 2026?	
67	Will Russia formally extend its fuel export restrictions before May 1, 2026 — e.g., extending the export ban to oil-producing companies and/or imposing a ban on diesel exports — beyond the gasoline export ban that took effect April 1?	✓
68	Will Vladimir Putin publicly announce, before May 1, 2026, a unilateral ceasefire timed to Victory Day in May (analogous to his April 2025 announcement of a May 8–10 truce)?	✓
69	Will the Bucharest Court of Appeal rule, before May 1, 2026, to return the coup-attempt indictment against Călin Georgescu and Horațiu Potra to prosecutors (or otherwise terminate/suspend the case) — rather than sending it to trial on the merits?	
70	Will US, South Korean (NIS), or Ukrainian intelligence publicly confirm a new deployment of North Korean troops (on the order of 1,000 or more) to Russia or Russian-occupied Ukraine during April 2026?	
71	Before May 1, 2026, will Russia and Ukraine (directly or via US/Qatari mediation) announce a formal agreement or mechanism on returning deported Ukrainian children, or will an officially confirmed return of at least 50 children occur in a single announced tranche?	
72	Will Russia announce, before May 1, 2026, the seizure, nationalization, or transfer to 'external management' of assets of at least one major Western company (or Euroclear/EU-linked assets in Russia), explicitly framed as retaliation for the EU's use of frozen Russian assets to fund Ukraine?	
73	Before May 1, 2026, will Valerii Zaluzhnyi be dismissed/recalled as Ukraine's ambassador to the UK, or publicly announce a formal political step (creating/joining a party, declaring a presidential candidacy, or resigning to enter politics)?	
74	Will Russian authorities (Roskomnadzor) impose a formal nationwide block of Telegram in Russia before May 1, 2026?	
75	Will gas deliveries through the TurkStream pipeline to Europe be suspended (zero or near-zero flow for at least 24 hours) as a result of Ukrainian strikes during April 2026?	
76	Will Hungary release the seven detained Ukrainian Oschadbank employees (and/or return the seized ˜$82 million in cash and gold), or alternatively formally indict them, before May 1, 2026?	
77	Will the United States and Russia carry out a bilateral prisoner exchange (release of at least one American citizen held in Russia and/or a Russian national held by the US) before May 1, 2026?	

Continued on next page

Table 25: Claude Fable 5 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
78	Will the US and Russia announce a concrete normalization step — e.g., resumption of direct flights, return/reopening of diplomatic properties or consulates, a formalized parliamentary-ties framework, or a signed economic/energy cooperation agreement — before May 1, 2026?	
79	Will Donald Trump publicly accept (or the White House confirm) the invitation for him to attend Moscow's May 9 Victory Day celebrations, before May 1, 2026?	
80	Will Ukraine's Verkhovna Rada pass at least one of the government-backed, IMF/EU-conditioned reform bills (e.g., the new taxes on digital services, self-employed entrepreneurs, or foreign parcels) before May 1, 2026?	✓
81	Will Ukraine's military leadership (Syrskyi/Zelensky) or ISW explicitly assess during April 2026 that Russia's anticipated spring-summer offensive has begun (large-scale offensive operations beyond current infiltration tactics)?	
82	Will a Russian general officer or senior security/defense official be killed or wounded in an assassination attempt or bombing inside Russia (attributed by Moscow or Kyiv to Ukraine or its partners) during April 2026?	
83	Will Lithuania, Poland, or Latvia reopen at least one currently closed border crossing with Belarus (or announce a formal agreement to do so) before May 1, 2026?	
84	Before May 1, 2026, will the Special Tribunal for the Crime of Aggression against Ukraine reach a formal milestone — the EU's approval as a founding member/co-founder of the Enlarged Partial Agreement, the tribunal's official inauguration in The Hague, or the announcement of its first indictments?	
85	Before May 1, 2026, will Russia officially announce that it will conduct — or actually conduct — a nuclear explosive test (e.g., at Novaya Zemlya), beyond missile/delivery-system tests?	
86	Before May 1, 2026, will Germany reverse course and approve Taurus cruise-missile deliveries to Ukraine, or will a UK–Germany 'ring exchange' (Storm Shadow/Taurus swap arrangement benefiting Ukraine) be formally announced?	
87	Will the US Senate hold a floor vote on (or pass) the Graham–Blumenthal Russia sanctions/secondary-tariffs bill before May 1, 2026?	
88	Will Ukraine's General Staff (or credible visual evidence) confirm that a Russian Navy warship (frigate, submarine, corvette or larger) was damaged, disabled or sunk by a Ukrainian strike during April 2026?	
89	Before May 1, 2026, will Iran conduct another military strike on Azerbaijani territory, or will Baku and Tehran undergo a formal diplomatic rupture (ambassador expulsion/recall, severing of ties, or an Azerbaijani military response)?	
90	Will Armenia and the United States sign their intergovernmental agreement implementing the TRIPP ('Trump Route') corridor — or will construction on the Armenian section be formally launched — before May 1, 2026?	
91	Before May 1, 2026, will Ukraine undergo a formal governance reset — dismissal/resignation of Prime Minister Svyrydenko or the Verkhovna Rada speaker, a formally announced new coalition or government of national unity, or a failed/called no-confidence vote in the Cabinet?	
92	Will Russia (Gazprom) and China (CNPC) sign a binding contract, final investment decision, or formal construction-launch agreement for the Power of Siberia 2 pipeline before May 1, 2026?	
93	Will Russian forces be credibly assessed (by ISW, DeepState, or Ukraine's General Staff) as having captured Orikhiv or entered the outskirts of Zaporizhzhia city before May 1, 2026?	
94	Will Ukraine's Supreme Court issue a ruling on Petro Poroshenko's lawsuit against the presidential sanctions imposed on him (in either direction) before May 1, 2026?	
95	Will Russia conduct (and announce) a test launch of the Sarmat heavy ICBM, or another strategic delivery-system test/demonstration (Burevestnik or Poseidon), during April 2026?	
96	Before May 1, 2026, will Russian forces be credibly assessed (by ISW, DeepState, or Ukraine's General Staff) as fully controlling Vovchansk, or as having captured any settlement within 15 km of Sumy city?	
97	Will a single Russian missile/drone strike (or one overnight attack wave) kill 20 or more civilians in Ukraine during April 2026 (per Ukrainian official casualty figures)?	✓

Continued on next page

Table 25: Claude Fable 5 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
98	Before May 1, 2026, will Hungary or Ukraine expel (declare persona non grata) at least one of the other country's diplomats, recall its ambassador, or summon the ambassador and announce formal retaliatory measures, in connection with the Zakarpattia/Tisza espionage scandal?	
99	During April 2026, will a Ukrainian strike on Russian electricity infrastructure (power plant or substation) cause a mass power outage inside Russia — e.g., a regional capital losing power for 100,000+ customers or a Russian region declaring an emergency over the outage?	
100	During April 2026, will a civilian commercial vessel (foreign- or Ukrainian-flagged cargo ship or tanker serving Ukraine's ports) be struck by a Russian drone or missile in or near a Ukrainian Black Sea/Danube port?	

GLM 5.2.

Table 26: GLM 5.2 forecast-time questions, run 1 top-100.

Rank	Forecast-time question	Match
1	Will Viktor Orbán's Fidesz party lose its parliamentary majority in the April 12, 2026 Hungarian elections, resulting in a government led by Péter Magyar's Tisza party?	✓
2	Will the US extend, expand, or formally reverse its March 2026 easing of sanctions on Russian oil exports and shipping during April 2026?	✓
3	Will the EU unblock the €90 billion loan for Ukraine and/or approve the 20th sanctions package against Russia during April 2026?	✓
4	Will Ukraine agree to pause or significantly reduce long-range strikes on Russian oil and energy infrastructure during April 2026, whether through a bilateral arrangement with partners or as part of an Easter energy truce with Russia?	
5	Will trilateral negotiations between Russia, the United States, and Ukraine on ending the war formally resume during April 2026?	
6	Will the US Pentagon formally divert or redirect air defense missiles or other military systems previously allocated to Ukraine toward Middle East operations during April 2026?	
7	Will a NATO member state invoke Article 4 (consultations) or formally escalate the response to Russian-directed drone incursions into allied airspace during April 2026?	
8	Will the ongoing anti-government protests in Serbia escalate to force a major concession, early election, or change of government during April 2026?	
9	Will Russia launch a major new ground offensive in the Donbas or elsewhere in Ukraine during April 2026, attempting to fulfill its stated ultimatum for Ukraine to leave the region?	
10	Will Armenia and Russia reach an agreement during April 2026 to transfer control of the South Caucasus Railway from Russian Railways to a third country (e.g., Kazakhstan, UAE, or Qatar), or will Armenia formally move to denounce the 2008 concession agreement?	
11	Will Russia formally announce a new wave of forced mobilization or compulsory conscription for the war in Ukraine during April 2026?	
12	Will the US formally announce a reduction in its NATO commitment—whether troop levels, defense spending obligations, or Article 5 guarantees—during April 2026?	
13	Will Russia restrict, reduce, or shut down oil flows through the Caspian Pipeline Consortium (CPC) from Kazakhstan during April 2026?	
14	Will the UK military board, detain, or seize a sanctioned Russian shadow fleet vessel in UK waters during April 2026?	
15	Will Russia's domestic economic pressures—including the gasoline export ban, budget sequestration, and delayed military payments—trigger visible public protests, fuel rationing, or a major policy reversal during April 2026?	
16	Will the US formally lift additional Belarus sanctions, or will Belarus and Lithuania announce a formal normalization or border agreement during April 2026?	

Continued on next page

Table 26: GLM 5.2 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
17	Will Russia escalate military support to Iran beyond drones—providing missiles, air defense systems, or direct military advisors—during April 2026?	
18	Will Russia take kinetic or military action (missile strikes, cyberattacks, or border incidents) against a NATO country in retaliation for Ukraine's use of that country's airspace for drone routes to Baltic ports during April 2026?	
19	Will Russian oil companies formally declare force majeure on crude oil exports from Baltic ports (Ust-Luga, Primorsk) due to Ukrainian drone strikes, or will a major Russian oil export terminal remain shut down for more than two consecutive weeks during April 2026?	
20	Will the US successfully pressure any NATO eastern flank member state (Poland, Romania, or a Baltic state) to redeploy its own air defense systems to the Middle East, or will the dispute over air defense redeployment publicly fracture NATO cohesion during April 2026?	
21	Will the Druzhba oil pipeline dispute be resolved with the resumption of Russian oil transit through Ukraine, or will Slovakia formally veto the next EU sanctions package over the pipeline issue during April 2026?	✓
22	Will Georgia's opposition mobilization campaign or the broader political crisis force a major concession from the Georgian Dream government, snap elections, or a change of government during April 2026?	
23	Will Moldova's energy emergency and security situation escalate during April 2026 to include a government collapse, a broader state of emergency beyond energy, a security incident in or around Transnistria, or Moldova formally requesting international security assistance?	
24	Will the EU formally announce the confiscation of the principal (not just profits) of frozen Russian sovereign assets to fund Ukraine during April 2026?	
25	Will Belarus and North Korea announce a formal military cooperation or arms transfer agreement—including missiles, drones, or artillery systems—during April 2026?	
26	Will Russia and Belarus escalate their migration weaponization campaign to create a new border crisis at the Polish or Baltic borders during April 2026?	
27	Will Ukraine and Russia conduct a major prisoner-of-war exchange (involving 200+ prisoners on either side) timed for Easter weekend (April 5–12, 2026)?	✓
28	Will Ukraine formally agree to or begin a military withdrawal from the Donbas region as part of a US-brokered peace deal during April 2026?	
29	Will Ukraine's parliament fail to pass IMF-mandated tax and anti-corruption reforms, causing the IMF to suspend or delay disbursement under its four-year credit program during April 2026?	
30	Will Russia formally announce asset nationalization or mandatory financial contributions from major businesses/oligarchs to fund the war effort during April 2026?	
31	Will the US and Russia reach a formal or de facto agreement during April 2026 in which Russia reduces intelligence or military support to Iran in exchange for the US reducing military support to Ukraine?	
32	Will Russia's central bank implement emergency measures—such as capital controls, sharp interest rate hikes above current levels, or ruble peg changes—due to currency or fiscal instability during April 2026?	
33	Will Russian electronic warfare (GPS jamming/spoofing) targeting the Baltic states cause a major civilian aviation safety incident—such as a near-miss, emergency diversion, or crash—during April 2026?	
34	Will Bulgaria's April 19, 2026 snap parliamentary election produce a governing coalition that includes pro-Russian or Eurosceptic parties (Revival, the Bulgarian Socialist Party, or Rumen Radev's Progressive Bulgaria), or result in another fragmented parliament unable to form a stable government during April 2026?	
35	Will Russia formally demand unrestricted transit through Lithuania to the Kaliningrad exclave, invoke its updated nuclear doctrine over threats to Kaliningrad's isolation, or take military action against Lithuania in response to Ukrainian drone strikes targeting Kaliningrad during April 2026?	

Continued on next page

Table 26: GLM 5.2 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
36	Will Russia's ongoing defense-ministry corruption purge escalate during April 2026 to a formal criminal case against Sergei Shoigu, his removal from his Security Council or Military-Industrial Commission roles, or a major reshuffle of senior Ground Forces command?	
37	Will a Russian-directed assassination or attempted assassination of a Ukrainian-linked figure, European official, or defector be carried out in a European country, or will a major Russian-directed sabotage attack on critical infrastructure (energy, undersea cables, or transport) occur in Europe, during April 2026?	
38	Will additional European NATO members (beyond Italy, Spain, France, and Poland) formally restrict US access to their airspace, bases, or territory for military operations against Iran, or will the US retaliate against these restrictions with formal diplomatic/military consequences, during April 2026?	
39	Will Hungary formally implement its threat to halt or significantly reduce gas deliveries to Ukraine during April 2026, in retaliation for the Druzhba oil pipeline dispute?	
40	Will Russian forces capture a significant Ukrainian-held city or strategic stronghold (e.g., Pokrovsk, Kostiantynivka, or another previously defended urban center) during April 2026?	
41	Will the EU energy crisis caused by the Iran war force any EU member state to implement formal energy rationing, price controls, or announce resumption/increase of Russian pipeline gas imports during April 2026?	
42	Will any Gulf state (Saudi Arabia, UAE, Qatar, Kuwait, or Jordan) formally transfer Patriot PAC-3 missiles or other air defense interceptors to Ukraine during April 2026, following Ukraine's deployment of counter-drone teams and defense cooperation agreements?	
43	Will the EU formally initiate or advance disciplinary proceedings against Hungary—including Article 7 voting rights suspension, Article 4(3) financial sanctions, or withholding of EU funds—during April 2026?	
44	Will Iran act on its declaration of Ukraine as a 'legitimate war target' by directly targeting Ukrainian personnel, assets, or interests—either in the Gulf region or through proxy attacks—during April 2026?	
45	Will Romania's ruling coalition collapse or will the government fall following the PSD's internal referendum on April 20, 2026, on whether to withdraw from the governing coalition?	✓
46	Will Russia stage a sabotage incident on the TurkStream or Blue Stream gas pipelines (as Ukraine's HUR warns), or will Hungary formally invoke NATO Article 5 consultations over attacks on TurkStream infrastructure, during April 2026?	
47	Will France formally announce the deployment of nuclear-armed aircraft to a European ally (e.g., Poland, Germany, or Greece) or establish a formal nuclear sharing/cooperation arrangement with another European state during April 2026?	
48	Will Ukraine's government under PM Yulia Svyrydenko fall, be reshuffled, or will Zelenskyy replace the Prime Minister, as a result of the ongoing parliamentary crisis and breakdown of executive-legislative relations during April 2026?	
49	Will Turkey formally restrict, suspend, or impose new conditions on the passage of Russian-linked or sanctioned oil tankers through the Bosporus Strait or Turkish exclusive economic zone in the Black Sea during April 2026?	
50	Will Russia formally deploy Oreshnik intermediate-range ballistic missile systems to Belarus, or confirm their operational deployment there, during April 2026?	
51	Will the IAEA successfully broker a local cease-fire at the Zaporizhzhia Nuclear Power Plant during April 2026 to allow emergency repairs to damaged power infrastructure, or will the plant experience a critical safety incident due to power loss?	
52	Will Donald Trump hold a face-to-face meeting with Belarusian leader Alexander Lukashenko (at a Board of Peace session or elsewhere), or will the US formally announce the reopening of its embassy in Minsk, during April 2026?	
53	Will the EU and partner states formally establish or inaugurate the Special Tribunal for the Crime of Aggression against Russia during April 2026?	
54	Will Poland's presidential veto of the EU SAFE defense loan implementing law be resolved (via new legislation, administrative workaround, or presidential reversal) during April 2026, or will it continue to block Poland's access to the €44 billion European defense loan facility?	

Continued on next page

Table 26: GLM 5.2 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
55	Will the United States re-impose sanctions on Milorad Dodik and/or his associates, or will the Republika Srpska take a formal step toward secession from Bosnia-Herzegovina, during April 2026?	
56	Will the United States impose new sanctions, military measures, or significant diplomatic consequences on Russia for its intelligence and military support to Iran that contributed to attacks wounding American service members, during April 2026?	
57	Will a major Russian oil company (Lukoil, Rosneft, or Gazprom Neft) face a debt crisis, state bailout, forced restructuring, or will the Lukoil foreign-asset divestiture deadline trigger sanctions enforcement or another extension, during April 2026?	
58	Will Russia deploy combat troops or military personnel to Iran in roles that directly engage US or Israeli forces, during April 2026?	
59	Will Russia formally implement a comprehensive ban on Telegram and/or significantly expand internet restrictions—including VPN blocks, regional mobile internet shutdowns, or restrictions on WhatsApp—during April 2026?	
60	Will Russia take direct military action, deploy naval assets, or formally threaten retaliation against Israel for striking Russian-Iranian arms transfer routes in the Caspian Sea during April 2026?	
61	Will Slovakia formally veto or block the advancement of Ukraine's EU accession negotiation clusters during April 2026?	
62	Will Russia formally operationalize long-range drone ground control stations or bases in Belarus during April 2026?	✓
63	Will Russia begin deploying naval convoys, armed escort teams, or protective equipment on board its shadow fleet or merchant ships during April 2026?	
64	Will Russia's parliament formally pass legislation authorizing the president to use military force to protect or free Russian citizens facing prosecution abroad during April 2026?	✓
65	Will Russia formally extend its ammonium nitrate and nitrogen fertilizer export restrictions beyond April 21, 2026, or explicitly condition fertilizer exports on Western sanctions relief during April 2026?	
66	Will any additional EU member state government leader publicly endorse normalizing relations with Russia or resuming Russian pipeline gas imports during April 2026, following the Belgian PM's public call?	
67	Will the ICC, UN Security Council, or another major international body formally initiate proceedings or issue a resolution specifically targeting Russia's forced conscription of Ukrainians in occupied territories during April 2026?	
68	Will Azerbaijan and Armenia formally sign a peace treaty, or will the peace process definitively collapse (with either side formally withdrawing from negotiations or a new military escalation along the border) during April 2026?	
69	Will the EU formally enact sanctions targeting Uzbek cotton cellulose suppliers that provide critical propellant materials for Russian missile production during April 2026?	
70	Will Armenia formally announce its withdrawal from the CSTO, apply for EU membership or candidate status, or sign a formal EU defense/security partnership agreement during April 2026?	
71	Will Ukrainian forces conduct a significant counteroffensive that recaptures a strategically important town, city, or territory from Russian control during April 2026?	
72	Will Russia formally announce a withdrawal or significant downsizing of its Africa Corps presence from Mali or the broader Sahel region during April 2026?	
73	Will Georgia formally complete the phase-out of Russian crude oil imports through its Kulevi refinery, or will the EU formally sanction the Kulevi refinery for processing Russian oil, during April 2026?	
74	Will India and Russia announce a major new defense procurement agreement, energy deal, or formal bilateral trade commitment during April 2026?	
75	Will Russia's regular spring conscription cycle (beginning April 1, 2026) trigger mass draft evasion at a scale prompting a major government crackdown, or visible anti-conscription protests in multiple Russian cities during April 2026?	
76	Will Kosovo's parliament fail to elect a president by the Constitutional Court's April 28, 2026 deadline, automatically triggering the country's third parliamentary election in 15 months?	✓

Continued on next page

Table 26: GLM 5.2 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
77	Will Ukrainian military drones crashing in Baltic NATO member states (Estonia, Latvia, Lithuania) during April 2026 trigger a formal NATO policy response, a bilateral diplomatic incident between Ukraine and a Baltic state, or a demand to alter Ukrainian drone flight corridors?	
78	Will South Korea formally join the NATO-backed Priority Request List (PURL) program or announce the supply of lethal weapons (e.g., 155mm artillery ammunition) to Ukraine during April 2026?	
79	Will Russia's wave of industrial strikes, wage arrears, and factory closures — including in monotowns like Boksitogorsk — escalate during April 2026 to force direct Kremlin intervention, spread to multiple sectors or regions, or trigger a major policy response such as wage subsidies or price controls?	
80	Will the United States and Ukraine formally sign a drone technology cooperation or joint production agreement — including the proposed $35-50 billion drone deal — during April 2026?	
81	Will Japan formally announce a policy change allowing the sale or transfer of lethal weapons and military equipment to Ukraine during April 2026?	
82	Will the revelation of Russia's 'Gamechanger' plot — a proposed staged assassination attempt on Viktor Orbán to influence Hungary's April 12 election — lead to a formal investigation by EU authorities, Hungarian prosecutors, or allied intelligence services during April 2026, or will evidence emerge that the plot was attempted or implemented?	
83	Will the leaked recordings of Hungarian FM Szijjártó sharing confidential EU information with Russia's Lavrov and offering to block EU sanctions trigger formal criminal proceedings, an EU-level investigation into sanctions sabotage, or enforcement action against Hungarian officials during April 2026?	
84	Will German Chancellor Merz's admission that Germany is 'not defense-capable' trigger a formal emergency defense spending package, Bundestag vote on emergency defense measures, or reinstatement of conscription during April 2026?	
85	Will the Russian drone incident targeting the French aircraft carrier Charles de Gaulle during NATO drills in Sweden escalate to a formal NATO response, French military retaliation, or a significant diplomatic crisis between Russia and NATO during April 2026?	
86	Will the UN, EU, or US formally initiate sanctions enforcement proceedings, condemnations, or new sanctions targeting Russia's deployment of 100,000+ North Korean forced laborers during April 2026?	
87	Will Russia's railway freight crisis — described by Kremlin-linked analysts as structural, with the civilian economy at its weakest since 2009 — trigger emergency government intervention, visible disruption to military supply lines, or major service rationing during April 2026?	
88	Will Ukraine formally withdraw from or significantly reduce its participation in NATO joint military exercises or overseas training programs during April 2026?	
89	Will Ukraine formally launch production or announce a concrete, funded timeline for domestically-developed air defense interceptors, including anti-ballistic missile components, during April 2026?	
90	Will Ukraine or a European country formally deport Russian volunteer fighters (e.g., Freedom of Russia Legion members) who fought for Ukraine back to Russia, or will Ukraine establish a formal legal framework protecting anti-Kremlin Russian volunteers from deportation during April 2026?	
91	Will the US formally notify any Eastern European NATO ally (Poland, Romania, Baltic states) of significant multi-year delays in delivering contracted weapons systems under Foreign Military Sales agreements, or will any ally formally suspend, cancel, or renegotiate its FMS procurement contracts, during April 2026?	
92	Will Ukraine's government launch a formal criminal investigation or prosecution into unauthorized drone exports by private defense manufacturers, or formally tighten export controls and oversight of defense technology production facilities, during April 2026?	
93	Will Ukraine's government announce emergency fiscal measures — including budget sequestration, new emergency taxation, suspension of social programs, or a formal request for accelerated emergency international financial assistance — during April 2026 due to depleting financial reserves?	

Continued on next page

Table 26: GLM 5.2 forecast-time questions, run 1 top-100. (continued).

Rank	Forecast-time question	Match
94	Will the US or EU announce new sanctions, export controls, or enforcement actions targeting Central Asian countries (Kyrgyzstan, Kazakhstan, Uzbekistan, or others) for serving as transshipment or re-export hubs for arms, ammunition, or dual-use goods to Russia during April 2026?	
95	Will Russia's ongoing defense ministry corruption purge cause visible disruptions to military procurement, weapons production, or front-line supply during April 2026, as Belousov's technocratic team imposes stricter financial controls?	
96	Will Ukraine's parliament or Defense Ministry formally adopt major mobilization reforms—including changes to service duration limits, AWOL/desertion penalties, rotation policies, or mobilization age thresholds—during April 2026?	
97	Will Russia's accelerating depletion of central bank gold reserves and the National Wealth Fund trigger a formal fiscal emergency declaration, credit rating downgrade, emergency budget revision, or acceleration of gold sales beyond current rates during April 2026?	
98	Will NATO formally attribute drone incursions into Baltic or Finnish airspace to Russia's deliberate electronic warfare tactics and announce specific countermeasures—such as counter-EW deployments, sanctions, or a formal démarche against Russia—during April 2026?	
99	Will Russian attacks on Ukrainian Black Sea and Danube port infrastructure cause a major disruption to grain exports—such as a multi-day port closure, a significant spike in global grain prices, or a new international maritime security initiative—during April 2026?	
100	Will France seize an additional Russian-linked shadow fleet vessel, or will Russia formally retaliate against France—through diplomatic expulsions, economic measures, cyberattacks, or military posturing—for its tanker seizures during April 2026?	

K PROMPT TEXT

For reproducibility, we include the core prompts used in the pipeline. In these prompts, "card" means structured gold record.

Retrospective Question Generation Prompt

```
I am retrospectively constructing gold forecasting questions for
    Metaculus, and I want them to represent the questions that
    should have been asked at a past cutoff.

The questions will be evaluated as follows:

1. At {target_start}T00:00:00Z, a prediction-side agent researches
    the world and posts a set of questions to Metaculus. These
    questions target uncertainties expected to materialize during [{
    target_start}, {target_end}).
2. After the target window, a hindsight-side agent reviews the
    consequential issues that actually materialized in [{
    target_start}, {target_end}): open uncertainties that should
    have been on a forecaster's watchlist at {target_start}T00:00:00
    Z, regardless of which direction they ultimately resolved.
3. Measure what fraction of the hindsight-side gold questions were
    covered by the prediction-side questions: how well the
    prediction-side question pool anticipated what actually mattered.


Your task is step 2.

You are working in the **geopolitical / military** scope (inter-
    state conflict, diplomatic shifts, alliances, military posture,
    sanctions, territorial disputes, major elections that materially
     affect foreign-policy posture, government changes, energy
```

```
    crises, hybrid attacks, migration shocks, and other
    consequential developments) within the area `{area_name}`.

The hindsight review time is: {target_end}T00:00:00Z.

You are looking BACK at the target window: [{target_start}T00:00:00Z
    , {target_end}T00:00:00Z). Both endpoints are UTC start-of-day;
    the start is inclusive and the end is exclusive (the half-open
    `[start, end)` convention applies uniformly).

You have access to tools (perigon_search, perigon_read_article) for
    searching and reading news in multiple languages. Use the tools
    to understand the consequential issues that materialized in the
    target window, then formulate the questions that should have
    been asked at {target_start}T00:00:00Z.

English-only searches may miss events primarily covered in local-
    language press. Consider mixing local-language ISO-639-1 codes
    into `perigon_search` (e.g., `language=["en", "fr"]`, `language
    =["zh", "ja"]`) when relevant to the area.

## Your task

Identify gold questions that meet ALL of the following:

1. **Consequential**: a serious analyst working in this area would
    have wanted to put this on the table BEFORE its outcome was
    determined. Once the resolution materialized (whether
    affirmative or negative -- happened, didn't happen, was
    cancelled, fell through, etc.), it mattered for decisions,
    budgets, strategy, risk, coordination, markets, policy, or
    public attention.
2. **Genuinely uncertain at the cutoff**: the outcome was not pre-
    determined at {target_start}T00:00:00Z. A reasonable forecaster
    at the cutoff would not have given it ~100% probability.
    Surprises that contradicted prior near-certain expectations are
    valid gold questions.
3. **Concrete and verifiable**: the question has identifiable actors
     and a clear binary criterion ("did the issue resolve in the
    affirmative or negative" -- either polarity counts as a valid
    resolution).
4. **Resolves within the target window**: the issue resolved or
    materially became determinate within [{target_start}, {
    target_end}).
5. **In-area, in-scope**: the question must be primarily about the
    area `{area_name}` and within scope. Cross-area questions count
    if and only if their primary actor or resolving locus is inside
    this area.

## What to avoid

Do not submit questions that:

- Were essentially pre-determined at the cutoff: scheduled summits/
    elections/anniversaries with foregone conclusions, announcements
     that just played out as already revealed, etc. (violates
    condition 2)
- Are too vague to be verified. (violates condition 3)
- Resolve clearly outside [{target_start}, {target_end}) (e.g. a
    question whose resolution can only occur after target_end, or a
    question whose resolution must have occurred before target_start
    ) (violates condition 4)
```

```
## Evidence

Every gold question must be tied to specific evidence from the tools
    . Use the evidence field to cite the concrete tool results that
    show why the issue mattered and how it resolved. Prefer evidence
     that points to the actual mechanism behind the question, not
    just broad topical similarity.

For each evidence item, cite the source by tool and id only:

- `source_tool`: `"perigon"`.
- `source_id`: the short id returned by `perigon_search`, e.g. `"
    a0007"`.
- `relevance`: one sentence explaining why that source supports the
    gold question.

Do not copy URLs into the evidence field. The harness resolves
    source ids to URLs after submission.

## Output

When you are done, call `submit_candidates` exactly once with **a
    list of distinct gold questions, ordered from rank 1 (most
    important to have asked at {target_start}T00:00:00Z, given what
    later materialized) down to the lowest rank still meeting the
    conditions above**.

Each gold question must include:

- `rank`: a 1-based integer matching the question's position in the
    submitted list. Number them explicitly: the first question gets
    `rank=1`, the second `rank=2`, the third `rank=3`, and so on.
    The list of `rank` values must form `[1, 2, 3, ..., N]` exactly
    (no gaps, no duplicates).
- `title`: one-line label for the question.
- `question`: a forecasting-style yes/no question or short
    hypothesis that would have been natural to ask at {target_start}
    T00:00:00Z.
- `rationale`: 4-6 sentences explaining why this gold question
    should have been on the watchlist at the cutoff, using hindsight
     about what later materialized.
- `evidence`: list of concrete tool results that most directly
    support the gold question. For each item, include only `
    source_tool`, `source_id`, and `relevance`.

You are not required to produce fully specified resolution criteria.
     What matters is that the question captures a consequential
    uncertainty that, in hindsight, should have been asked at the
    cutoff.

**There is no fixed N. Submit as many as you can defend, but each
    one must clear all conditions above. Do not pad the list with
    weak questions; the evaluation rewards covering consequential
    hindsight questions, not list length.**
If this run is extending prior submissions and no additional
    defensible gold question remains, call `finish_question_mining`
    instead of `submit_candidates`, with a brief reason. Do not call
     it in the same run as `submit_candidates`.

## How to work
```

```
- Search broadly, then follow consequential threads that
    materialized in the target window. Look across the major sub-
    themes inside the area + scope.
- When a realized issue looks important, investigate it further
    using additional searches and article reads.
- Convert realized issues into forecast questions that would have
    been askable at {target_start}T00:00:00Z.
- Look for both:
  - direct continuations of the situation visible at the cutoff, and
  - second-order or cross-domain consequences that became important
    during the target window.
- Actively try to identify questions that would have been easy to
    miss at the cutoff, not just the most obvious headline
    continuations.
```

Gold Record Construction Prompt

```
You are constructing a gold benchmark card for HorizonBench.

HorizonBench evaluates early question discovery: whether an ex ante
    question pool identified the core consequential uncertainties
    that, in hindsight, should have been on a forecaster's watchlist
     at a past cutoff.

Your task is to take one hindsight-selected candidate question and
    turn it into a canonical benchmark card.

Area:
{area_name}

Forecaster cutoff:
{target_start}T00:00:00Z

Target window:
[{target_start}T00:00:00Z, {target_end}T00:00:00Z)

Effective today:
{target_end}T00:00:00Z

Candidate hindsight question:
title: {question_title}
question: {question_text}
rationale: {question_rationale}
evidence: {question_evidence}

You may use news search and reading tools.

# Benchmark philosophy

The card should capture the core consequential uncertainty that a
    competent forecaster should have tracked at the cutoff.

The benchmark does not require an ex ante forecaster to predict the
    exact administrative milestone, implementation step, or wording
    by which the uncertainty later resolved. A candidate question
    can match if it identifies the same watchlist issue.

For example, approval, unblocking, authorization, first
    implementation step, initial tranche release, or ordinary
    realization of the same policy package may all belong to the
```

```
    same core issue if they concern the same actors, same object,
    same target, and same main bottleneck.

However, do not make the boundary so broad that it covers merely
    related topics. Different policy packages, different funding
    sources, different recipients, different domains, or separate
    causal bottlenecks should remain distinct.

Do not count a formulation as matching merely because it is a
    precursor, warning, evidence signal, bargaining condition,
    causal driver, operational signal, consequence, or contextual
    signal for the gold question. A match must concern the same
    policy object, decision or outcome axis, and watchlist issue;
    shared actors, rationale, causal mechanism, strategic theme,
    causal context, or broader story are not sufficient.

# What to produce

Produce one canonical gold question card with the fields below:

- normalized_question
- question_summary
- question_timing
- normalization_note
- matching.boundary_note
- matching.positive_examples
- matching.negative_examples
- precursor_signals
- causal_mechanisms

# Procedure

1. Verify materialization and timing

Use authoritative coverage to verify what materially happened, when
    it happened, and whether it resolved within the target window.

Identify:
- resolved_date_utc
- primary actors
- target or object
- concrete development that made the uncertainty determinate
- sources supporting the timing and resolution

2. Write the normalized_question

Write one minimal yes/no forecasting question.

It should:
- capture the primary surprise or impact axis
- be askable at the cutoff
- avoid future-created referents that did not exist at the cutoff
- avoid incidental hindsight details
- avoid unnecessary thresholds, mechanisms, actor details, or dates
- preserve all load-bearing actors, targets, and constraints

Default to the broader stable issue identity when exact amounts,
    dates, administrative steps, or implementation mechanisms could
    vary modestly without changing the core issue.

3. Define the matching boundary
```

```
The matching boundary should define the same core consequential
    uncertainty, not just exact paraphrases of normalized_question.

INCLUSIONS:
Include exact paraphrases and closely adjacent formulations that
    would put a forecaster on the same watchlist item. This may
    include approval, unblocking, authorization, implementation,
    first operational step, or ordinary realization of the same
    issue.

EXCLUSIONS:
Exclude only genuinely different issue axes. Do not exclude a
    formulation merely because it asks about a different
    administrative or implementation milestone of the same issue.
Do exclude formulations that are only precursors, warnings, evidence
     signals, bargaining conditions, causal drivers, operational
    signals, consequences, or contextual signals for the same
    broader situation.
A positive example must concern the same decision or outcome axis,
    not merely the same causal mechanism, strategic theme, rationale
    , or broader policy context.
Do not treat warnings, evidence signals, bargaining conditions, or
    downstream consequences as matches unless the candidate question
     asks about the same concrete decision or outcome axis.

Use this self-check:
If a competent forecaster had generated this candidate question at
    the cutoff, would they deserve credit for tracking the same
    important watchlist issue? If yes, it should usually be included.


4. Write positive and negative examples

matching.positive_examples:
Write 5 to 7 yes/no questions that should count as matches. Include
    close variants that identify the same core issue, even if the
    exact resolution milestone differs.

matching.negative_examples:
Write 3 to 5 clear non-matches. Prefer clear negatives over subtle
    near-misses. Good negatives involve a different policy package,
    different actor, different recipient, different funding source,
    different domain, or separate causal bottleneck.

Do not include opposite-polarity versions of the same issue as
    negatives.

Do not include implementation or realization steps of the same issue
     as negatives.

5. Mine precursor signals

Find at least 8 distinct pre-cutoff signals that, in hindsight,
    supported tracking this issue.

Each signal should:
- be available before the cutoff
- be tied to a source and date
- explain why it pointed toward the core issue
- not rely on information revealed after the cutoff

These signals are supporting evidence, not proof that the issue was
    obvious at the time.
```

```
6. Group causal mechanisms

Cluster the precursor signals into a small number of named causal
    mechanisms explaining why the issue was worth tracking.

# Output format

Call `submit_question_card` exactly once. Do not return free-form
    prose instead of the tool call.

Use this schema:

- `normalized_question`: the canonical yes/no question.
- `question_summary`: 2-4 sentences of context plus how the issue
    resolved.
- `question_timing`: `{{resolved_date_utc, first_reported_date_utc,
    first_reported_source, confirmed_date_utc, confirmed_sources,
    timing_confidence}}`.
- `normalization_note`: 2-4 sentences explaining what was kept,
    broadened, or removed from the candidate hindsight question and
    why.
- `matching`: `{{boundary_note, positive_examples, negative_examples
    }}`. Both example lists contain `{{formulation, reason}}` items.
- `precursor_signals`: list of `{{signal_date, signal, significance,
     source}}` items.
- `causal_mechanisms`: list of short mechanism names.

Use UTC dates in `YYYY-MM-DD` format. Use only real source URLs
    found through the tools. Do not fabricate sources.
```

Forecast-Time Question Generation Prompt

```
I am posting forecasting questions to Metaculus, and I want them to
    anticipate consequential events well.

The questions I post will be evaluated as follows:

1. At {target_start}T00:00:00Z, research the world and post a set of
     questions to Metaculus. These questions target uncertainties
    expected to materialize during [{target_start}, {target_end}).
2. Enumerate the consequential issues that resolved in [{
    target_start}, {target_end}): open uncertainties that should
    have been on a forecaster's watchlist in advance, regardless of
    which direction they ultimately resolved.
3. Measure what fraction of the step-2 events were covered by the
    step-1 questions: how well the question pool anticipated what
    actually mattered.

Your task is step 1.

You are working in the **geopolitical / military** scope (inter-
    state conflict, diplomatic shifts, alliances, military posture,
    sanctions, territorial disputes, major elections that materially
     affect foreign-policy posture, government changes, energy
    crises, hybrid attacks, migration shocks, and other
    consequential developments) within the area `{area_name}`.

The current time is: {target_start}T00:00:00Z.
```

```
You are looking FORWARD at the target window: [{target_start}T00
    :00:00Z, {target_end}T00:00:00Z); imagine what could materialize
     there. Both endpoints are UTC start-of-day; the start is
    inclusive and the end is exclusive (the half-open `[start, end)`
     convention applies uniformly).

You have access to tools (perigon_search, perigon_read_article) for
    searching and reading news in multiple languages. Search tools
    let you choose the publication-date window that is useful for
    your research.

English-only searches may miss events primarily covered in local-
    language press. Consider mixing local-language ISO-639-1 codes
    into `perigon_search` (e.g., `language=["en", "fr"]`, `language
    =["zh", "ja"]`) when relevant to the area.

## Your task

Identify candidates that meet ALL of the following:

1. **Consequential**: a serious analyst working in this area would
    want to put this on the table BEFORE its outcome is determined.
    Once the resolution materializes (whether affirmative or
    negative -- happens, doesn't happen, is cancelled, falls through
    , etc.), it would matter for decisions, budgets, strategy, risk,
     coordination, markets, policy, or public attention.
2. **Genuinely uncertain in advance**: the outcome is not pre-
    determined. A reasonable forecaster would not give it ~100%
    probability. Low-probability surprises that contradict prior
    near-certain expectations are valid candidates.
3. **Concrete and verifiable**: the candidate has identifiable
    actors and a clear binary criterion ("did the issue resolve in
    the affirmative or negative" -- either polarity counts as a
    valid resolution).
4. **Meaningfully diverse across the set**: each candidate should
    differ substantively from the others. Avoid near-duplicates and
    slight rephrasings of the same underlying scenario.
5. **Resolves within the target window**: expected to resolve
    approximately within [{target_start}, {target_end}).
6. **In-area, in-scope**: the candidate must be primarily about the
    area `{area_name}` and within scope. Cross-area candidates count
     if and only if their primary actor or resolving locus is inside
     this area.

## What to avoid

Do not submit candidates that:

- Are essentially pre-determined: scheduled summits/elections/
    anniversaries with foregone conclusions, announcements that just
     played out as already revealed, etc. (violates condition 2)
- Are too vague to be verified. (violates condition 3)
- Are duplicates or near-duplicates of another candidate in your set
    . (violates condition 4)
- Are expected to resolve clearly outside [{target_start}, {
    target_end}) (e.g. a candidate whose resolution can only occur
    after target_end, or a candidate whose resolution must occur
    before target_start) (violates condition 5)

## Evidence

Every candidate must be tied to specific evidence from the tools.
    Use the evidence field to cite the concrete tool results that
```

```
    support the candidate. Prefer evidence that points to the actual
     mechanism behind the candidate, not just broad topical
    similarity.

For each evidence item, cite the source by tool and id only:

- `source_tool`: `"perigon"`.
- `source_id`: the short id returned by `perigon_search`, e.g. `"
    a0007"`.
- `relevance`: one sentence explaining why that source supports the
    candidate.

Do not copy URLs into the evidence field. The harness resolves
    source ids to URLs after submission.

A modest leap from evidence to candidate is fine; stretch your
    imagination beyond what articles literally say.

## Output

When you are done, call `submit_candidates` exactly once with **a
    list of distinct candidates, ordered from rank 1 (most likely to
     materialize within [{target_start}, {target_end}) AND most
    consequential if it does) down to the lowest rank still meeting
    the conditions above**.

Each candidate must include:

- `rank`: a 1-based integer matching the candidate's position in the
     submitted list. Number them explicitly: the first candidate
    gets `rank=1`, the second `rank=2`, the third `rank=3`, and so
    on. The list of `rank` values must form `[1, 2, 3, ..., N]`
    exactly (no gaps, no duplicates).
- `title`: one-line label for the candidate.
- `question`: a forecasting-style yes/no question or short
    hypothesis. Prefer wording that is specific enough to be
    meaningful, but do not force artificial precision if the
    underlying uncertainty is still early and fuzzy.
- `rationale`: 4-6 sentences explaining why the candidate is worth
    asking about.
- `evidence`: list of concrete tool results that most directly
    support the candidate. For each item, include only `source_tool
    `, `source_id`, and `relevance`.

You are not required to produce fully specified resolution criteria.
     What matters is what evidence visible at the current time
    supports taking the candidate seriously.

**There is no fixed N. Submit as many as you can defend, but each
    one must clear all conditions above. Do not pad the list with
    weak candidates; the evaluation rewards covering real events,
    not list length.**


## How to work

- Search broadly, then follow promising threads. Look across the
    major sub-themes inside the area + scope.
- When a thread looks potentially important for the target interval,
     investigate it further using additional searches and article
    reads.
- Look for both:
  - direct continuations of the current situation, and
```

```
  - second-order or cross-domain consequences that may become
    important by the target interval.
- Actively try to find candidates that are easy to miss at the
    current time, not just the most obvious headline continuation.
```

Pairwise Question Judge Prompt (official)

```
I am an admin at Metaculus, and I want to retrospectively measure
    the quality of one submitted forecasting question.

You are working in the geopolitical / military scope (inter-state
    conflict, diplomatic shifts, alliances, military posture,
    sanctions, territorial disputes, major elections that materially
     affect foreign-policy posture, government changes, energy
    crises, hybrid attacks, migration shocks, and other
    consequential developments) within the area {area_name}.

Your effective today is {target_end}. You are looking BACK at the
    period: [{target_start}, {target_end}) - half-open UTC start-of-
    day boundaries.

You will be given:
- One forecast-time candidate question submitted before the target
    window.
- One retrospective gold-question card: an actual issue that
    resolved during the target window.

Decide whether the candidate and gold card match. They match iff ALL
     of the following hold:

1. Semantic alignment to canonical: the candidate question's actor +
     action + scope/threshold corresponds to the card's
    normalized_gold_question or to any of its matching.
    positive_examples. Wording need not be literal - paraphrases,
    alternative time-framings within the period, and slightly
    different specific instances all count, as long as the
    underlying scenario is the same. Polarity is irrelevant. Minor
    differences in path, mechanism, negotiating stance, or
    procedural route do not block a match when the candidate still
    asks about the same concrete resolution target or primary
    decision; do not extend this to preconditions, downstream
    consequences, background pressures, or merely adjacent
    procedural steps.
2. No rejected match: the candidate does NOT semantically match any
    matching.negative_examples. If a candidate fits both a positive
    and a negative example, the negative one wins.
3. Not a generic-prior lucky label: the candidate is not merely
    naming the right actor or event class from broad priors without
    engaging the specific scenario named by the card.

Reject when the candidate shares only actor/topic/event class but
    proposes a different specific outcome; when it asks only a
    precondition, downstream consequence, background pressure, or
    adjacent procedural step; when a required joint occurrence is
    only partly covered; or when the match is only a broad label.

Polarity is NOT a reject reason. The benchmark measures whether the
    right underlying issue was on the watchlist, not whether the
    direction was predicted correctly.

When in doubt, return no_match.
```

```
Output only strict JSON with this schema:
{"match": true|false, "rationale": "2-4 sentences explaining the
    decision"}
```

Near-Miss Audit Judge Prompt (qualitative audit only)

```
I am an admin at Metaculus, and I want to retrospectively measure
    the quality of the questions submitted to the platform.

The overall process is as follows:

1. At some past date, a large set of questions was created (
    researched from news and other resources) and posted to
    Metaculus. Those questions target uncertainties expected to
    materialize during [{target_start}, {target_end}) -- half-open:
    the start is inclusive, the end is exclusive (UTC start-of-day
    boundaries).
2. Enumerate the consequential issues that resolved in [{
    target_start}, {target_end}): open uncertainties that should
    have been on a forecaster's watchlist in advance.
3. Measure what fraction of the step-2 issues were covered by the
    step-1 questions: how well the question pool anticipated what
    actually mattered.

Your task is step 3.

You are working in the **geopolitical / military** scope (inter-
    state conflict, diplomatic shifts, alliances, military posture,
    sanctions, territorial disputes, major elections that materially
     affect foreign-policy posture, government changes, energy
    crises, hybrid attacks, migration shocks, and other
    consequential developments) within the area `{area_name}`.

Your effective "today" is {target_end}.
You are looking BACK at the period: [{target_start}, {target_end})
    -- half-open UTC start-of-day boundaries.

## Your task

You will be given:

- A **list of N candidate forecasting questions** submitted at step
    1. Each candidate has:
  - `question` (a yes/no question or short hypothesis)
  - `rationale` (4-6 sentences explaining why the candidate was
    worth asking about, typically with dated signals from the news)
- A **pool of "truth" cards** -- the actual issues that resolved
    during [{target_start}, {target_end}) from step 2. Each card has
    :
  - `card_id` (string identifier -- use this in your output)
  - `event_occurred_date_utc` (UTC date the event occurred)
  - `event_statement` (one sentence naming actor + action +
    threshold)
  - `event_summary` (2-4 sentences of context + the event)
  - `matching.boundary_note` (INCLUSIONS / EXCLUSIONS text --
    variants that all count as canonical, and nearby events that do
    NOT count)
  - `matching.positive_examples` (5-9 phrasings spanning either
    polarity, all identifying the same underlying issue this card
    resolves on, with reasons)
```

```
  - `matching.negative_examples` (5-9 genuinely-distinct issue
    classes with reasons; opposite-polarity wording of the same
    issue is intentionally NOT here -- those live in `matching.
    positive_examples`)

For each (candidate, card) pair, decide whether they match. A pair
    **matches** iff ALL of the following hold:

1. **Semantic alignment to canonical**: the candidate question's
    actor + action + scope/threshold corresponds to the card's `
    event_statement` or to any of its `matching.positive_examples`.
    Wording need not be literal -- paraphrases, alternative time-
    framings within the period, and slightly different specific
    instances all count, as long as the underlying scenario is the
    same. Polarity (affirmative vs negative wording) is irrelevant
    to alignment -- the issue is what matters, not which direction
    the candidate predicted. Minor differences in path, mechanism,
    negotiating stance, or procedural route do not block a match
    when the candidate still asks about the same concrete resolution
     target or primary decision; do not extend this to preconditions
    , downstream consequences, background pressures, or merely
    adjacent procedural steps.
2. **No rejected match**: the candidate does NOT semantically match
    any of the card's `matching.negative_examples`. The negative
    list is exhaustive of explicit exclusions; if a candidate fits
    BOTH a positive example AND a negative example, the negative one
     wins -> not a match.
3. **Not a generic-prior "lucky label"**: the candidate is not a
    phrase that names the right actor / event class but rests only
    on broad priors without engaging the specific scenario named by
    the card.

A single candidate may match multiple cards (rare -- happens when
    two cards capture related-but-distinct events). A single card
    may match multiple candidates (the agent submitted variants).

Walk through every candidate and check it against every card; then
    sweep cards that received zero matches to catch wording-variant
    misses.

Note: this judgment does NOT evaluate whether the candidate's
    evidence is signal-connected to the card's pre-cutoff signal
    cluster. That axis is intentionally out of scope here.

## What to reject

**Polarity is NOT a reject reason.** The benchmark measures whether
    the right underlying ISSUE was on the watchlist, not whether the
     direction was predicted correctly. Match against `
    event_statement` / `matching.positive_examples` -- the issue --
    not against the specific direction the event took. For example:
    both "Will summit happen?" and "Will summit be cancelled?" match
     a card whose underlying issue is summit happen/cancel,
    regardless of which polarity ultimately materialized.

Do NOT record a match when:

- The candidate shares only the actor or only the event-class with
    the card, but the proposed specific outcome differs (right
    neighborhood, wrong scenario).
- The candidate matches any `matching.negative_examples` formulation
     of the card -- that veto stands even if a `matching.
    positive_examples` formulation also matches.
```

```
- The card's `event_statement` requires a JOINT occurrence ("X AND Y
    ") and the candidate asks only about ONE of the conjuncts. The
    candidate must cover the joint event to match.
- The candidate is a generic-prior label hit ("Iran tensions
    escalate", "ME peace deal stalls") without naming the specific
    scenario the card resolves on.
- The candidate asks only about a precondition, downstream
    consequence, background pressure, or adjacent procedural step
    rather than the same concrete resolution target / primary
    decision.

The benchmark's design penalizes near-miss false positives more than
     misses, so when in doubt, do not record a match.

## Output

Submit your verdict via the `submit_window_batch_near_miss_verdict`
    tool call, exactly once. The tool expects:

  {{
    "matches": [
      {{
        "candidate_rank": <1-based global rank exactly as shown in
    the Candidate heading>,
        "card_id": "<card_id from the cards section>",
        "rationale": "<2-4 sentences. Quote or paraphrase the
    relevant part of the candidate question, name the card
    event_statement or accepted_formulation it aligns with, explain
    why the core uncertainty is the same, and state why no
    rejected_formulation applies.>"
      }},
      ... one entry per (candidate, card) match ...
    ],
    "near_misses": [
      {{
        "candidate_rank": <1-based global rank exactly as shown in
    the Candidate heading>,
        "card_id": "<card_id from the cards section>",
        "rationale": "<briefly explain why this was close but not
    counted as a match>"
      }},
      ... one entry per close rejected pair worth human audit ...
    ],
    "summary_rationale": "<2-4 sentences summarizing the verdict
    across all candidates and cards in scope>"
  }}

Use the global candidate ranks printed in the headings. If the batch
     starts at
`### Candidate 11`, the first candidate in that batch has `
    candidate_rank: 11`,
not `candidate_rank: 1`.

List true matches in `matches`. Also list close rejected pairs in `
    near_misses` when they are worth human audit. Do NOT list
    ordinary non-matches. Do NOT output any free text outside the
    tool call. Do NOT call any other tool.
```