BRIEF: BELIEF REVISION FOR ITERATIVE EVIDENCE-BASED FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Most forecasting agents that retrieve evidence use it in the same way. Once the search stops, everything collected is placed in a single context, and the model reads a forecast off it. Holding the evidence corpus fixed, we show that this practice makes poor use of it. Presenting the corpus all at once yields a Brier loss of 19.34, processing it in rounds lowers this to 18.13, and revising an explicit forecast belief as evidence arrives lowers it further to 16.52. Motivated by this finding, we introduce BRIEF (Belief Revision for Iterative Evidence-based Forecasting), a framework in which a belief state is initialized from multi-perspective research, revised under a Bayesian-inspired principle as new evidence arrives, and used to direct further research through the gaps and conflicts it leaves unresolved. A final refinement stage integrates candidate judgments and applies calibration and target-specific structural adjustments. Compared with twelve baselines spanning reasoning models, prompting and search strategies, and deep research systems, BRIEF achieves the best aggregate performance on FutureX, ForecastBench, and Prophet Arena, with the lowest variance across three runs. Relative to the strongest baseline, it improves the FutureX score by 7.86 points and reduces ForecastBench Brier loss by 5.3%, and it achieves the lowest Brier loss on Prophet Arena. These results indicate that how evidence is used is itself a design decision for forecasting agents, and that a maintained belief state offers an effective way to connect research decisions, evidence interpretation, and final forecast construction.

1 INTRODUCTION

Forecasting requires translating incomplete and heterogeneous evidence into judgments about unresolved events. Historical frequencies, current developments, and external expectations may each inform a prediction, but their implications must be assessed together. Large language models offer a promising foundation for this process through their ability to draw on retrieved information and reason across sources (Lewis et al., 2020; Zou et al., 2022). Building on this, retrieval-augmented forecasting systems combine search, structured reasoning, and aggregation to improve predictive performance (Halawi et al., 2024; Hsieh et al., 2024; Alur et al., 2025). Most of them share the same final step. Once the search is finished, the collected evidence is placed in a single context, and the model reads a forecast off it. This raises a question that is rarely examined on its own. How effectively do these systems turn the evidence they collect into a forecast?

Providing evidence does not ensure that a model uses it appropriately (Joren et al., 2025). Long-context studies show that models can underuse relevant information depending on where it appears in the input (Liu et al., 2024). In forecasting, Yuan et al. (2025) use EvolveCast to show that models often respond inconsistently or conservatively to new information, exposing a gap between receiving evidence and appropriately revising a prediction. Studies of probabilistic reasoning further show that the same evidence can lead a model to different updates depending on whether it is presented jointly or sequentially (Chen et al., 2026). How evidence should be presented to a forecaster is therefore an empirical question rather than an implementation detail.

The same evidence, processed differently, yields different forecasts. We compare three evidence-processing strategies on Prophet Arena (Yang et al., 2026), using the same frozen evidence snapshot for each question across all three settings (Section 5.1). One-shot integration places the entire corpus in a single context and asks for one forecast, as in the retrieve-then-answer pipelines above. Sequen-

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.

tial accumulation splits the same corpus into three rounds and re-forecasts after each round, keeping the text of earlier rounds in context. Belief revision, the strategy used by our method, maintains an explicit forecast belief and revises it as evidence arrives, so that the previous judgment, and not only the previous text, is carried forward. Brier loss and negative log-likelihood (NLL) both penalize inaccurate probability estimates (Brier, 1950; Gneiting & Raftery, 2007), so lower values indicate better forecasts. Brier loss falls from 19.34 under one-shot integration to 18.13 under sequential accumulation and 16.52 under belief revision, and NLL gives the same ordering. Since the three settings read the same evidence, the differences reflect how that evidence is processed. Presenting everything at once and reading off a single forecast, the most common design, is the weakest of the three in this comparison, and updating a forecast as evidence arrives is the strongest.

A belief state must also fit the structure of what is being predicted. Sequential belief updating has proven effective on binary questions (Murphy, 2026), the setting in which other forecasting agents have also been developed and evaluated (Halawi et al., 2024; Alur et al., 2025), where the belief reduces to a single probability. Many forecasting targets are richer. FutureX asks for concrete outcomes, from a choice among listed options to open-ended answers, and a Prophet Arena event can carry several markets that need not be mutually exclusive. For such targets, a single probability no longer describes the forecaster's state, so the belief must track a provisional answer in the required format, whether an option, a value, or a probability for each market. This requirement propagates through the pipeline. Research must cover what each part of the target depends on, sufficiency must be judged for the target as a whole, and the final answer must remain coherent, with probabilities that respect the ordering of thresholds or sum to one over mutually exclusive and exhaustive outcomes. The comparison above supports this view, as belief revision lowers Brier by 31.2% relative to sequential accumulation on these multi-outcome events. We therefore introduce BRIEF (Belief Revision for Iterative Evidence-based Forecasting), which makes such a structured belief the control signal of the entire pipeline through multi-perspective belief initialization, Bayesian-inspired revision, belief-guided sufficiency control, and target-aware forecast refinement.

BRIEF first frames the prediction target and decomposes research into complementary perspectives (Figure 1), including historical base rates (Tetlock & Gardner, 2016), drivers that could move the outcome, disconfirming evidence, and external expectations such as market prices (Wolfers & Zitzewitz, 2004). Executing the resulting queries initializes the belief state, which records a provisional forecast, the evidence supporting and opposing it, and the questions it leaves unresolved. As further evidence arrives, revision takes the current belief as the starting point and interprets new information in light of what is already known, following a Bayesian-inspired principle without requiring explicit likelihood estimation or exact posterior computation. The research loop is then driven by the belief itself. An evidence critic examines it for unsupported assumptions and conflicting information, and a sufficiency controller decides whether the remaining gaps warrant a targeted query or whether the evidence supports stopping under the remaining budget. Once research terminates, forecast refinement integrates five candidate judgments (Wang et al., 2023) generated from the shared evidence and terminal belief, after which calibration (Platt, 1999) and conditional structural adjustments adapt the result to the target.

Our contributions are threefold.

- We show that the standard practice of presenting all collected evidence at once and reading a single forecast off it makes poor use of that evidence. Holding the corpus fixed, processing evidence in rounds improves the forecast, and revising an explicit forecast belief improves it further, reducing Brier loss from 19.34 to 16.52 on identical evidence.
- We propose BRIEF, a unified forecasting framework that places an explicit belief state at the center of the pipeline. The belief is initialized from multi-perspective research, revised under a Bayesian-inspired principle as evidence arrives, used to decide what to investigate next and when to stop, and finally converted into the required answer through candidate aggregation, calibration, and structural adjustment.
- BRIEF achieves the best aggregate performance on FutureX, ForecastBench, and Prophet Arena against twelve evaluated baselines, improving the FutureX score by 7.86 points and reducing ForecastBench Brier loss by 5.3% relative to the strongest baseline. We isolate the contribution of each component through ablations and show that the gains carry over to three weaker open-weight backbones.

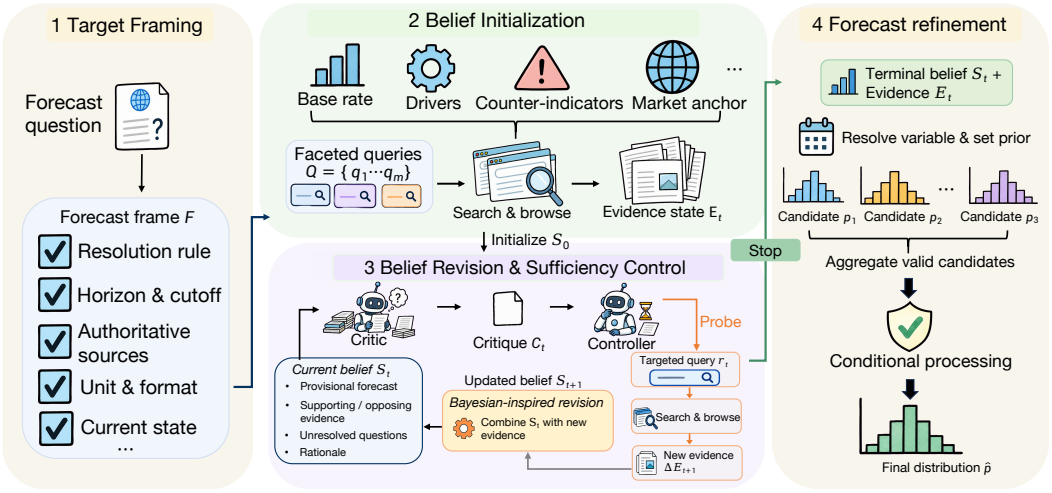


Figure 1: Overview of BRIEF. Multi-perspective research initializes a belief state, targeted research supports iterative revision, and forecast refinement produces the final forecast.

2 Related Work

LLM-based Forecasting. LLM-based forecasting has progressed from neural models trained to forecast world events (Zou et al., 2022) to retrieval, filtering, and aggregation pipelines (Halawi et al., 2024), and more recently to agents that combine search with simulation (Hsieh et al., 2024). AIA Forecaster integrates agentic search, forecast reconciliation, and calibration (Alur et al., 2025), while BLF sequentially updates linguistic beliefs during tool use (Murphy, 2026). Halawi et al. (2024), AIA Forecaster, and BLF were all developed and evaluated on binary questions, whereas recent benchmarks also require concrete outcome predictions (Zeng et al., 2026) and probabilities for multiple related markets (Yang et al., 2026). BRIEF is designed for these heterogeneous targets, connecting multi-perspective belief initialization, separate modules for evidence criticism and sufficiency control, and target-aware refinement. Table 1 summarizes these differences.

Table 1: Comparison with LLM forecasting systems, as described in their original papers.

	Halawi	AIA	BLF	BRIEF
Adaptive research?	✗	✓	✓	✓
Role-guided research?	✗	✗	✗	✓
Explicit belief state?	✗	✗	✓	✓
Sufficiency-based stopping?	✗	✗	✗	✓
Forecast aggregation?	✓	✓	✓	✓
Non-binary targets?	✗	✗	✗	✓

Evidence Utilization and Adaptive Retrieval. Evidence utilization depends on information position (Liu et al., 2024) and must be distinguished from context sufficiency (Joren et al., 2025). EvolveCast evaluates how models update forecasts given new information (Yuan et al., 2025), while Chen et al. (2026) find that batch and sequential processing differ in Bayesian consistency and that greater consistency does not always improve performance. Self-RAG learns adaptive retrieval and self-critique (Asai et al., 2024), and CRITIC verifies and revises outputs through external tools (Gou et al., 2024). We compare evidence-processing strategies on a frozen corpus, and BRIEF directs research through the questions its belief state leaves unresolved, stopping once the evidence supports an informed forecast even when outcome uncertainty remains.

Forecast Aggregation and Calibration. Pooling predictive distributions affects calibration and dispersion (Gneiting & Ranjan, 2013), while Platt scaling (Platt, 1999) and temperature scaling (Guo et al., 2017) map model scores to calibrated probabilities. Recent forecasting agents combine aggregation and calibration (Alur et al., 2025; Murphy, 2026). BRIEF applies these techniques to candidates sharing a terminal belief and evidence record, with conditional monotonic projection for ordered thresholds and normalization for mutually exclusive and exhaustive outcomes.

3 METHOD

3.1 PROBLEM FORMULATION

Given a forecasting question q and an evidence cutoff τ, the agent uses only evidence made public before τ to produce $\hat{a} \in \mathcal{A}_q$. The output space $\mathcal{A}_q$ accommodates outcome predictions, binary probabilities, and vectors of market probabilities.

Retrieve-then-answer pipelines (Lewis et al., 2020; Halawi et al., 2024) map collected evidence directly to a forecast, $q \to E \to \hat{a}$. BRIEF instead maintains an evidence archive E_t and an explicit belief state S_t that comprises a provisional forecast, supporting and opposing evidence, and unresolved questions. The belief is revised as evidence arrives and guides further research. Once research stops, BRIEF integrates candidate forecasts based on the terminal state and evidence. Figure 1 illustrates the pipeline, and Algorithm 1 in Appendix A gives the full procedure.

3.2 FORECAST FRAMING

Before retrieval, BRIEF constructs a structured forecast frame $F = f_{\text{frame}}(q)$ through model-based analysis. The frame specifies the resolution criterion, forecasting horizon, units, and answer format, and identifies the relevant deciding organization, authoritative sources, and status quo. These aspects clarify what is being predicted, how the outcome is determined, and which information to prioritize.

Frame elements may be inferred from the question when they are not stated explicitly, and they guide subsequent retrieval and verification. They serve as heuristic dimensions for decomposing the question and keep the prediction target consistent across research rounds.

3.3 MULTI-PERSPECTIVE RESEARCH

Given q and F, BRIEF generates complementary research queries $Q = f_{\text{plan}}(q, F) = \{q_1, \ldots, q_m\}$. Query planning is guided by forecasting roles rather than generic topics. Base-rate queries examine comparable situations to establish an initial expectation. Driver queries seek factors that could shift the outcome away from this baseline. Disconfirmer queries seek evidence against plausible or currently favored hypotheses. Market-anchor queries identify external expectations, such as market probabilities, when available. These roles are extensible heuristics, and additional perspectives, such as the latest published value and the resolution basis, are added when relevant to the question.

Executing the queries yields the initial evidence and belief:

$$E_0 = \text{Research}_{V_\tau}(Q), \qquad S_0 = f_{\text{init}}(q, F, E_0), \tag{1}$$

where V_τ restricts retrieval to evidence admissible before the cutoff (Appendix B.2 describes the leakage controls). Initialization combines historical expectations with current conditions and question-specific evidence, recording both the provisional forecast and the gaps that may require further investigation.

3.4 BELIEF REVISION AND SUFFICIENCY-CONTROLLED RESEARCH

At round t, a critic assesses the current belief against the evidence, identifying unsupported assumptions, conflicting claims, and repeated information. A separate controller decides whether these issues warrant further research:

$$\begin{aligned} C_t &= f_{\text{critic}}(q, F, S_t, E_t), \\ z_t &= f_{\text{suff}}(q, F, S_t, C_t, B_t) \in \{\text{stop}, \text{probe}\}, \end{aligned} \tag{2}$$

where B_t is the remaining research budget. The critic interprets evidence but cannot initiate retrieval, while the controller determines whether to continue. Sufficiency means having adequate support for an informed forecast, not certainty about the outcome. Prediction stability alone does not justify stopping if material conflicts remain. Appendix E illustrates this distinction with two cases.

When $z_t = \text{probe}$, the controller generates a targeted query $r_t = f_{\text{probe}}(q, F, S_t, C_t)$. Research returns additional evidence $\Delta E_{t+1} = \text{Research}_{V_\tau}(Q, r_t, E_t)$, which is merged into E_t to form E_{t+1}, retaining source provenance and flagging overlapping reports. BRIEF then updates its belief:

$$S_{t+1} = f_{\text{update}}(q, F, S_t, C_t, E_t, \Delta E_{t+1}). \tag{3}$$

Each update records the revised assessment and the evidence-based reasons for changing or retaining it. The update follows a Bayesian-inspired principle, interpreting new evidence relative to the current belief and what is already known. For a probabilistic target Y, this corresponds to

$$P(Y \mid E_t, \Delta E_{t+1}) \propto P(\Delta E_{t+1} \mid Y, E_t)\, P(Y \mid E_t). \tag{4}$$

This relation guides the model's reasoning. BRIEF does not estimate explicit likelihoods or compute exact posteriors, since enforcing consistency with stated likelihoods does not reliably improve task performance (Chen et al., 2026). In particular, repeated or dependent reports are not counted as independent support.

The revised belief and evidence are reassessed in the next round. The loop terminates when the controller signals stop, the budget is exhausted, or the round limit is reached, yielding (S_T, E_T).

3.5 Forecast refinement

After research, BRIEF refines the final prediction through candidate aggregation, probability calibration, and conditional structural adjustment, without additional retrieval. Deterministic preparation resolves quantities such as dates and forecasting horizons to form $\tilde{q} = h(q, F)$. We generate $k = 5$ candidate forecasts from the terminal belief and evidence:

$$a_1, \ldots, a_k \sim \pi(\cdot \mid \tilde{q}, F, S_T, E_T), \qquad \bar{a} = \mathrm{Agg}_{\mathcal{A}_q}(a_1, \ldots, a_k). \tag{5}$$

The candidates share the same research record and capture variation in its interpretation. Aggregation consolidates these judgments into a base forecast in the required output space, combining probabilities for probabilistic tasks and candidate answers for outcome-prediction tasks.

For probabilistic outputs, BRIEF applies Platt scaling (Platt, 1999) to correct systematic bias in the probability estimates:

$$p_{\mathrm{cal}} = \sigma(\alpha\ \mathrm{logit}(p) + \beta)\,, \tag{6}$$

where p is the input probability, σ is the logistic function, and α, β are fitted calibration parameters. Aggregation combines candidate judgments, whereas calibration adjusts their probability scale.

BRIEF also refines probability vectors when the outcome structure imposes explicit constraints. For ordered thresholds $u_1 < \cdots < u_J$ with $p_j = P(Y \geq u_j)$, the probabilities must satisfy $p_1 \geq \cdots \geq p_J$. When the option labels can be parsed into such a threshold sequence, BRIEF applies threshold-specific adjustments followed by monotonic projection using the pool-adjacent-violators algorithm (PAVA) (Ayer et al., 1955). These nested event probabilities need not sum to one, and normalization applies only to mutually exclusive and exhaustive outcomes. Structural adjustments are skipped when their prerequisites are absent, and non-probabilistic predictions retain the answer format specified by $\mathcal{A}_q$.

4 Experiments

We evaluate BRIEF across three forecasting benchmarks and conduct ablations to examine the contribution of each component.

4.1 Experimental Settings

Benchmarks. We use three complementary benchmarks. FutureX (Zeng et al., 2026) is a live benchmark for forecasting agents covering diverse real-world domains. It requires concrete outcome predictions and organizes questions into four tiers by answer structure and event volatility, namely single-choice (L1), multiple-answer (L2), and open-ended questions with low (L3) or high (L4) volatility. ForecastBench (Karger et al., 2025) is a dynamic probabilistic forecasting benchmark drawing questions from prediction platforms and real-world time series. We evaluate its binary-event questions, which require estimating the probability of an outcome by a specified resolution date. Prophet Arena (Yang et al., 2026) separates information collection from forecast generation to support controlled analysis of predictive capabilities. We use its frozen evidence snapshots, restrict retrieval to the supplied corpus, and elicit a probability for each market associated with an event. Appendix D.1 provides further details on the benchmarks and evaluation sets. Together, these benchmarks assess outcome prediction across difficulty levels, binary probability estimation, and multi-market forecasting under fixed evidence.

Table 2: Forecasting performance. Results are means over three runs with standard deviations in gray, except for the single-run reasoning models. Best in bold, second best underlined.

	FutureX ↑					ForecastBench ↓		Prophet Arena ↓	
Method	L1	L2	L3	L4	Score	Brier	NLL	Brier	NLL
General-purpose reasoning models									
Kimi K3	75.00	78.95	32.95	49.61	53.02	15.78	46.99	<u>16.67</u>	51.24
GPT-5.5	66.67	63.16	39.98	45.00	49.29	17.00	50.47	17.13	<u>49.88</u>
DeepSeek-V4-Pro	66.67	63.16	47.13	43.35	50.78	19.85	58.39	18.06	55.24
GLM-5.2	75.00	57.89	27.83	33.72	40.92	19.41	56.61	17.57	51.55
Prompting and search strategies									
No tools	41.67 ±0.00	42.11 ±5.26	22.95 ±5.13	19.44 ±7.08	27.25 ±5.39	17.45 ±1.56	57.06 ±7.86	22.19 ±1.21	70.55 ±9.57
Direct search	66.67 ±14.43	66.67 ±6.08	42.27 ±6.88	58.61 ±13.30	56.12 ±4.67	17.73 ±0.61	53.79 ±3.38	19.03 ±0.57	56.56 ±3.50
ReAct + search	72.22 ±9.62	64.91 ±15.19	53.56 ±7.78	53.40 ±2.48	57.63 ±0.87	17.05 ±1.51	54.76 ±6.87	18.76 ±0.71	55.46 ±1.89
Self-consistency	69.44 ±19.25	70.18 ±3.04	46.23 ±4.05	64.07 ±9.67	60.48 ±2.55	16.81 ±0.79	49.85 ±2.67	17.68 ±0.72	51.43 ±1.92
Halawi et al.	66.67 ±8.33	66.67 ±3.04	42.10 ±7.94	48.98 ±11.16	52.22 ±6.73	18.27 ±1.08	53.97 ±3.32	18.63 ±1.52	54.38 ±3.61
Deep research systems									
Open Deep Research	80.56 ±4.81	80.70 ±8.04	<u>59.44</u> ±8.25	52.50 ±7.95	63.03 ±3.93	<u>13.76</u> ±2.88	44.17 ±8.65	19.67 ±0.39	58.31 ±1.88
GPT Researcher	72.22 ±4.81	52.63 ±5.26	41.88 ±7.16	45.54 ±9.54	48.53 ±5.20	15.88 ±1.57	47.81 ±3.81	17.72 ±0.94	51.40 ±2.80
DeerFlow	**91.67** ±0.00	<u>82.46</u> ±3.04	52.91 ±2.87	<u>70.14</u> ±2.01	<u>69.59</u> ±1.32	14.44 ±0.54	<u>44.02</u> ±1.26	18.94 ±1.82	54.88 ±4.71
BRIEF (ours)	<u>83.33</u> ±0.00	**85.96** ±3.04	**67.32** ±0.89	**79.32** ±1.10	**77.45** ±0.67	**13.03** ±0.11	**40.30** ±0.49	**16.60** ±0.20	**48.62** ±0.50

Baselines. We include four general-purpose reasoning models as model-capability references, namely Kimi K3 (Kimi Team, 2026), GPT-5.5 (OpenAI, 2026), DeepSeek-V4-Pro (DeepSeek-AI, 2026), and GLM-5.2 (GLM-5 Team, 2026). We further compare BRIEF with eight method baselines. Five are prompting and search strategies, namely no-tools forecasting, direct search, ReAct with search (Yao et al., 2023), self-consistency (Wang et al., 2023), and the forecasting pipeline of Halawi et al. (2024). The other three are deep research systems, namely Open Deep Research (LangChain, 2025), GPT Researcher (Elovic, 2023), and DeerFlow (ByteDance, 2025). All methods are evaluated on the same question set within each benchmark. The deep research systems retain their respective planning and evidence-processing procedures, so these comparisons assess the combined effects of evidence acquisition and forecast construction.

Implementation. BRIEF and all method baselines use DeepSeek-V4-Flash (DeepSeek-AI, 2026) as the backbone, while the general-purpose reasoning models serve as their own backbones. All methods search the web with Serper (Serper, 2026) on FutureX and ForecastBench when retrieval is enabled, and on Prophet Arena they read only the supplied evidence corpus. BRIEF generates $k = 5$ candidate forecasts per question, and each method is run three times, except for the reasoning models, which are run once. Appendix C gives the full model and inference settings and the procedure for each baseline.

Evaluation metrics. For FutureX, we average answer-type-specific scores within each level and combine them with the benchmark's tier weights (Zeng et al., 2026):

$$S_{\text{FutureX}} = 0.1S_1 + 0.2S_2 + 0.3S_3 + 0.4S_4, \tag{7}$$

where S_ℓ is the mean score at level ℓ. On the two probabilistic benchmarks, we report the Brier score (Brier, 1950) and negative log-likelihood (NLL) (Gneiting & Raftery, 2007). For ForecastBench,

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2, \tag{8}$$

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N}\big[y_i \log p_i + (1 - y_i)\log(1 - p_i)\big], \tag{9}$$

where N is the number of questions, p_i is the predicted probability, and $y_i \in \{0, 1\}$ is the outcome. For Prophet Arena, we treat each market as a binary outcome, since markets need not be mutually exclusive, apply the same losses, and average them first over markets within each event and then over events. We use the natural logarithm and clip probabilities to $[\epsilon, 1 - \epsilon]$. All scores are multiplied by 100; higher is better for FutureX, and lower is better for Brier and NLL.

Table 3: Effect of removing each component on all three benchmarks. Arrows show the change relative to full BRIEF, and orange marks degradation.

	ForecastBench ↓		Prophet Arena ↓		FutureX ↑
Setting	Brier	NLL	Brier	NLL	Score
BRIEF (full)	13.01	40.07	16.73	48.94	76.90
w/o evidence critic	14.39 ↑1.38	43.02 ↑2.95	18.64 ↑1.91	53.75 ↑4.81	62.60 ↓14.30
w/o role guidance	13.42 ↑0.41	40.26 ↑0.19	17.30 ↑0.57	50.39 ↑1.45	55.90 ↓21.00
w/o sufficiency control	14.28 ↑1.27	42.84 ↑2.77	17.51 ↑0.78	50.90 ↑1.96	47.61 ↓29.29
w/o forecast refinement	14.13 ↑1.12	44.12 ↑4.05	17.67 ↑0.94	52.94 ↑4.00	74.08 ↓2.82

4.2 Main Results

Table 2 compares BRIEF with four general-purpose reasoning models, five prompting and search strategies, and three deep research systems.

BRIEF ranks first on every aggregate metric and is the most stable method across runs. Its mean FutureX score of 77.45 exceeds that of DeerFlow, the strongest baseline, by 7.86 points. On ForecastBench, it lowers Brier by 5.3% relative to Open Deep Research and NLL by 8.5% relative to DeerFlow. On Prophet Arena, it attains the lowest mean Brier and NLL. On every aggregate metric, its standard deviation is the smallest among the methods run three times. Appendix D.2 reports the individual runs.

Systems that research well do not necessarily forecast well. DeerFlow and Open Deep Research are the two strongest baselines on FutureX and ForecastBench, where each system gathers its own evidence, yet on Prophet Arena their mean Brier falls behind that of every general-purpose reasoning model. Because Prophet Arena fixes the evidence corpus and BRIEF shares its backbone with all method baselines, BRIEF's lower Brier and NLL relative to these baselines mainly reflect how evidence is turned into probabilities, suggesting that much of its advantage on this benchmark comes from how it processes evidence.

Aggregating several judgments is a consistent source of improvement. Self-consistency, which samples multiple judgments rather than conducting additional research, is the strongest prompting strategy on every aggregate metric and has the lowest Prophet Arena Brier among all method baselines. BRIEF builds on the same principle, conditioning its candidates on a shared belief before calibration and structural adjustment.

BRIEF's advantage grows with question difficulty. It leads on L2, L3, and L4, and its margin over the best baseline on each tier widens from 3.50 points on L2 to 7.88 on L3 and 9.18 on L4, the tiers that carry 90% of the weighted score.

4.3 Component Ablations

We remove one component at a time and evaluate each variant on all three benchmarks (Table 3). The variants remove the evidence critic, replace forecasting-role guidance with generic topical query planning, replace sufficiency control with a fixed research plan, or produce a single forecast without aggregation, calibration, or structural adjustment.

Each research component contributes substantially to outcome prediction. On FutureX, removing the evidence critic, role guidance, or sufficiency control lowers the score by 14.30, 21.00, and 29.29 points, respectively, and each of these variants falls below DeerFlow (Table 2). Sufficiency control has the largest effect, suggesting that deciding how much to research matters most on this benchmark. Consistent with this, BRIEF searches more and stops early less often on harder FutureX tiers (Appendix D.3).

The evidence critic also improves probabilistic forecasts. Removing it raises Brier and NLL by 1.38 and 2.95 on ForecastBench and by 1.91 and 4.81 on Prophet Arena, the largest increases among the research components.

Forecast refinement mainly improves probability estimates. Removing it lowers the FutureX score by 3.7% in relative terms, whereas Brier and NLL rise by 5.6% to 10.1% on ForecastBench and Prophet Arena. The increase is larger under NLL, suggesting that refinement reduces overconfidence.

All four refinement operations reduce error, with candidate aggregation contributing most. We remove one operation at a time on Prophet Arena (Table 4). Using a single candidate instead of aggregation raises Brier by 3.20, more than removing projection for ordered thresholds (2.04), the median for multi-outcome questions (1.47), or reference-class anchoring (1.04). This is consistent with the strong performance of self-consistency among the baselines (Table 2).

Table 4: Refinement operations on Prophet Arena. Δ is relative to a separate full BRIEF run.

Setting	Brier ↓	Δ
BRIEF (full)	**16.52**	–
w/o aggregation	19.72	+3.20
w/o anchoring	17.56	+1.04
w/o median	17.99	+1.47
w/o projection	18.56	+2.04

5 Analysis

This section first shows that presenting all evidence at once leads to worse forecasts than processing it progressively, then tests whether BRIEF's gains hold with weaker backbones, and finally separates the contributions of evidence acquisition and forecast construction.

5.1 Same Evidence, Different Forecasts

Table 5: The same evidence processed in three ways on Prophet Arena. All settings read an identical frozen corpus, so differences reflect how the evidence is used.

	Overall		Binary		Multi-outcome	
Setting	Brier↓	NLL↓	Brier↓	NLL↓	Brier↓	NLL↓
One-shot integration	19.34	54.97	21.44	60.93	12.57	35.66
Sequential accumulation	18.13	53.73	20.27	58.03	11.20	39.82
BRIEF	**16.52**	**48.49**	**18.95**	**54.69**	**7.71**	**26.90**

We first test whether presenting all evidence at once makes good use of it. Prophet Arena's frozen corpus fixes the evidence available for each question, and we compare three ways of processing it (Table 5). One-shot integration presents all evidence in a single context, sequential accumulation introduces the same evidence over three rounds while retaining earlier text, and BRIEF runs its full pipeline on the corpus, revising an explicit forecast belief as evidence arrives. If presentation did not matter, one-shot integration and sequential accumulation would perform alike, since they see the same corpus.

Overall, presenting all evidence at once is the weakest way to use it: on identical evidence, sequential accumulation lowers Brier from 19.34 to 18.13 and NLL from 54.97 to 53.73.

BRIEF makes better use of the same evidence. It lowers Brier to 16.52 and NLL to 48.49, 8.9% and 9.8% below sequential accumulation, and still outperforms it without forecast refinement (17.67 and 52.94, Table 3).

BRIEF's gains are largest on multi-outcome questions. There, sequential accumulation even raises NLL above one-shot integration (39.82 versus 35.66), whereas BRIEF lowers Brier by 31.2% and NLL by 32.4% relative to sequential accumulation.

5.2 Generalization to Weaker Backbones

To test whether BRIEF depends on a strong backbone, we run it with three open-weight models of 3B to 9B parameters from the Llama 3 (Grattafiori et al., 2024) and GLM-4 (Team GLM et al., 2024) families, keeping the configuration of the main experiments, and compare it with direct prediction by the same model on ForecastBench (Table 6).

Table 6: Direct prediction versus BRIEF with weaker backbones on ForecastBench. Market and dataset columns give Brier on each question type.

	Brier↓		NLL↓		Market Brier↓		Dataset Brier↓	
Backbone	Direct	BRIEF	Direct	BRIEF	Direct	BRIEF	Direct	BRIEF
Llama-3.2-3B	26.47	23.99 ↓2.48	76.03	67.15 ↓8.88	27.23	23.77 ↓3.46	25.70	24.21 ↓1.49
Llama-3.1-8B	26.58	23.34 ↓3.24	74.88	65.59 ↓9.29	27.89	22.24 ↓5.65	25.27	24.44 ↓0.83
GLM-4-9B	25.15	22.46 ↓2.69	70.38	64.17 ↓6.21	25.50	23.20 ↓2.30	24.80	21.72 ↓3.08

BRIEF consistently improves all three backbones. Relative to direct prediction, it lowers Brier by 2.48 to 3.24 points and NLL by 6.21 to 9.29 points, about 9% to 12% on both metrics. BRIEF also lowers Brier on both market and dataset questions, by up to 5.65 and 3.08 points, respectively.

BRIEF can compensate for a smaller backbone. With BRIEF, Llama-3.2-3B outperforms direct prediction with the larger Llama-3.1-8B and GLM-4-9B on both metrics, suggesting that much of BRIEF's advantage comes from how it organizes research and revises beliefs rather than from backbone capacity.

5.3 Separating Evidence Acquisition from Forecast Construction

We freeze the evidence that BRIEF and DeerFlow collect on FutureX and combine each evidence source with two predictors (Table 7). The benchmark-prompt predictor produces a single answer with the prompt provided by FutureX, and the BRIEF predictor applies our question-specific frame, candidate aggregation, and structural processing. The diagonal cells correspond to the complete DeerFlow and BRIEF systems. Differences within a row reflect the evidence, and differences within a column reflect the predictor.

Table 7: Frozen evidence exchange on FutureX. Higher weighted scores are better.

	Evidence source	
Predictor	DeerFlow	BRIEF
Benchmark prompt	68.93	70.86
BRIEF predictor	71.42	**76.90**

BRIEF improves both the evidence it collects and the way it uses evidence. Under the benchmark prompt, replacing DeerFlow evidence with BRIEF evidence raises the score by 1.93 points. On DeerFlow evidence, replacing the benchmark prompt with the BRIEF predictor raises it by 2.49 points.

The two improvements are complementary. Using both raises the score from 68.93 to 76.90, a gain of 7.97 points that exceeds the sum of the two single gains (4.42). The BRIEF predictor gains 6.04 points on BRIEF evidence, more than twice its gain on DeerFlow evidence, suggesting that evidence acquisition and forecast construction in BRIEF reinforce each other.

6 Conclusion

We introduced BRIEF, a forecasting framework that connects multi-perspective research, belief revision, sufficiency control, and forecast refinement. Its central design is to maintain an evolving assessment that both incorporates new evidence and decides what to investigate next. Across three forecasting benchmarks, BRIEF achieves the best aggregate performance among the evaluated methods and is the most stable across runs, and ablations show that each component contributes. Our analyses show that how evidence is processed affects forecast quality even when the evidence itself is held fixed, with one-shot integration the weakest of the strategies compared and BRIEF the strongest. They further show that evidence acquisition and forecast construction in BRIEF reinforce each other, and that its gains carry over to weaker open-weight backbones. Together these findings support designing forecasting agents around the connection between evidence acquisition, belief revision, and final forecast construction, rather than treating retrieval and prediction as separate stages.

AI USE STATEMENT

We used generative AI tools to assist with language editing and code development. The research design and experimental analysis were conducted by the authors. All AI-assisted text and code were reviewed and revised by the authors, and the code was checked through execution. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

This study uses publicly available benchmarks and web sources and does not involve newly recruited human participants or non-public personal data. The proposed method is evaluated for research purposes. Its forecasts may reflect errors or biases in the underlying models and evidence, and their use in consequential decisions would require further validation and human judgment.

REPRODUCIBILITY STATEMENT

The main text describes the proposed method, experimental setup, and evaluation metrics. The appendix provides further details on the algorithm, benchmark construction, and evidence filtering. We plan to make the implementation and evaluation code publicly available to support reproduction of the results.

REFERENCES

Rohan Alur, Bradly C. Stadie, Daniel Kang, Ryan Chen, Matt McManus, Michael Rickert, Tyler Lee, Michael Federici, Richard Zhu, Dennis Fogerty, Hayley Williamson, Nina Lozinski, Aaron Linsky, and Jasjeet S. Sekhon. AIA Forecaster: Technical report. *arXiv preprint arXiv:2511.07678*, 2025.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647, 1955.

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

ByteDance. DeerFlow. `https://github.com/bytedance/deer-flow`, 2025.

Chacha Chen, Matthew Jörke, Adam Goliński, Masha Fedzechkina, Guillermo Sapiro, Sinead Williamson, and Nicholas Foti. LLMs are not (consistently) Bayesian: Quantifying internal (in)consistencies of LLMs' probabilistic beliefs. *arXiv preprint arXiv:2605.06915*, 2026.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Assaf Elovic. GPT Researcher. `https://github.com/assafelovic/gpt-researcher`, 2023.

GLM-5 Team. GLM-5: From vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

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

Tilmann Gneiting and Roopesh Ranjan. Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782, 2013.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

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.

Elvis Hsieh, Preston Fu, and Jonathan Chen. Reasoning and tools for human-level forecasting. In *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pp. 48–57, 2024.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025.

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*, 2025.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

LangChain. Open Deep Research. `https://github.com/langchain-ai/open_deep_research`, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Kevin Murphy. Agentic forecasting using sequential Bayesian updating of linguistic beliefs. *arXiv preprint arXiv:2604.18576*, 2026.

OpenAI. GPT-5.5 System Card. `https://openai.com/index/gpt-5-5-system-card/`, 2026.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Serper. Serper: Google search API. `https://serper.dev`, 2026.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Random House, 2016.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

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

Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. LLM-as-a-Prophet: Understanding predictive intelligence with Prophet Arena. In *International Conference on Learning Representations*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Zhangdie Yuan, Zifeng Ding, and Andreas Vlachos. Do language models update their forecasts with new information? *arXiv preprint arXiv:2509.23936*, 2025.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, et al. FutureX: An advanced live benchmark for LLM agents in future prediction. In *International Conference on Learning Representations*, 2026.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27293–27305, 2022.

A Detailed BRIEF Algorithm

A.1 Overall Procedure and State Representation

Algorithm 1 expands the procedure in Section 3. Initial queries and later probes share a research budget B, with at most R controller rounds. The task-specific output specification $\mathcal{M}_q$ defines the output space $\mathcal{A}_q$ and its validity constraints. For budget accounting, $\mathrm{Research}_{V_\tau}$ also returns the remaining budget.

Algorithm 1 BRIEF inference procedure

Require: Question q, cutoff τ, research budget B, maximum controller rounds R, output specification $\mathcal{M}_q$
Ensure: Forecast $\hat{a}$

$F \leftarrow \textsc{Frame}(q)$
$\rho \leftarrow \textsc{Route}(q, F)$ ▷ series or event
$K \leftarrow \textsc{Roles}(\rho)$
$Q \leftarrow \textsc{Plan}(q, F, K)$
$(E_0, B_0) \leftarrow \mathrm{Research}_{V_\tau}(Q; B)$
$E_0 \leftarrow \textsc{MergeDeduplicate}(E_0)$
$S_0 \leftarrow \textsc{InitializeState}(q, F, E_0)$
$t \leftarrow 0$
while $t < R$ **and** $B_t > 0$ **do**
  $C_t \leftarrow \textsc{Critic}(q, F, S_t, E_t)$
  $(c_t, r_t) \leftarrow \textsc{Controller}(q, F, S_t, C_t, E_t, B_t)$
  $z_t \leftarrow \textsc{OperationalDecision}(c_t, r_t, B_t)$
  if $z_t = \mathrm{stop}$ **then**
    break
  end if
  $(\Delta E_{t+1}, B_{t+1}) \leftarrow \mathrm{Research}_{V_\tau}(Q, r_t, E_t; B_t)$
  $E_{t+1} \leftarrow \textsc{MergeDeduplicate}(E_t, \Delta E_{t+1})$
  $S_{t+1} \leftarrow \textsc{ReviseState}(q, F, S_t, C_t, E_t, \Delta E_{t+1})$
  $t \leftarrow t + 1$
end while
$T \leftarrow t$
$\tilde{q} \leftarrow h(q, F)$ ▷ resolve dates and horizons
for $j = 1, \ldots, 5$ **do**
  $a_j \leftarrow \textsc{ForecastCandidate}(\tilde{q}, F, S_T, E_T, \mathcal{M}_q)$
end for
$\bar{a} \leftarrow \textsc{Aggregate}(a_1, \ldots, a_5, \mathcal{M}_q)$
$\tilde{a} \leftarrow \textsc{CalibrateIfApplicable}(\bar{a}, \mathcal{M}_q)$ ▷ Platt scaling
$\hat{a} \leftarrow \textsc{ApplyConstraints}(\tilde{a}, F, \mathcal{M}_q)$
return $\hat{a}$

The terminal index T counts completed retrieval–revision steps, not controller assessments. A controller round includes the critic assessment and the decision to stop or probe. If the first decision is to stop, one controller round has occurred but $T = 0$, and refinement uses (S_0, E_0). All five candidates share the terminal record, with no further retrieval.

The forecast frame F remains fixed while the evidence and belief evolve. The archive E_t retains dated evidence with source provenance and records overlaps across retrieved material. The belief S_t contains the provisional forecast, supporting and opposing evidence, and unresolved questions. Separating these objects allows research to adapt without changing the prediction target or its resolution conditions.

A.2 Question Routing and Role-Guided Research

Research roles depend on the resolution target rather than the surface answer format. A binary question about whether a published quantity exceeds a threshold is therefore treated as a SERIES

question. Table 8 summarizes the research directions for each route. It describes planning heuristics rather than mandatory coverage.

Table 8: Research roles for different resolution targets.

Scope	Resolution target	Research roles
Shared	Any target.	Base rate, drivers, disconfirmers, and market anchors.
Series	A published value, band, or threshold.	Resolution source, latest value, release timing, components, and advance estimates.
Event	A distinct outcome, entity, or action.	Status quo, evidence distinguishing possible outcomes, and announcement source.

A.3 Critic–Controller Research Control

The critic assesses the current belief against the evidence without retrieving information or revising the belief. It identifies unsupported assumptions, conflicts, and repeated information. The controller uses this assessment and the remaining budget to decide whether further research is justified. Section C.3 gives the corresponding prompt templates.

Let c_t indicate whether the record supports an informed forecast. A probe r_t is a single query addressing a searchable issue that could materially affect the forecast, or $\varnothing$ if no such query is proposed. The operational decision is

$$z_t = \begin{cases} \text{probe}, & \neg c_t \wedge r_t \neq \varnothing \wedge B_t > 0, \\ \text{stop}, & \text{otherwise}. \end{cases} \tag{10}$$

Stopping research is distinct from evidence sufficiency. Uncertainty that retrieval cannot reduce is not a reason to continue searching, while a stable forecast alone does not justify stopping when material conflicts remain. The budget or round limit may also force a stop even when evidence is insufficient. The terminal state is then passed to the output procedure in Section C.2.

B Benchmark Construction and Temporal Integrity

B.1 Benchmark Sources and Time Boundaries

Evidence availability is determined by the forecast time, not the resolution time. Let t_f denote the forecast time, τ the evidence cutoff, and t_r the resolution time. The forecast horizon is $t_r - t_f$, but evidence must have been public before τ. Table 9 gives these boundaries, with $\tau = t_f$. The cutoff is taken from the evaluation setup and benchmark metadata, not inferred by the forecasting model.

Table 9: Forecast and evidence time boundaries for each benchmark.

Benchmark	Forecast time and evidence cutoff $t_f = \tau$	Resolution time t_r
FutureX	Evaluation-window as-of date.	Question `end_time`.
ForecastBench	Freeze or base date of the scored question.	Question resolution date or market close.
Prophet Arena	`snapshot_time`.	`close_time`.

Evaluation questions are reconstructed from the released benchmark artifacts. FutureX uses resolved weekly records, preserving question text, answer options, difficulty levels, and ground truth. ForecastBench metadata are joined to scored records by question identifier to reconstruct the official zero-shot market prompt, including background and resolution criteria; the supplied freeze-time crowd value is not included. Prophet Arena preserves each event's market–source mapping and uses only the supplied frozen corpus, without live web retrieval.

All methods receive the same question collection within each benchmark. Dataset exclusions are determined before inference and applied uniformly; output failures do not become dataset exclusions. FutureX retains invalid answers for scoring under its official evaluator. Probabilistic comparisons require valid probability outputs.

B.2 Evidence Admissibility and Leakage Control

Temporal filtering combines search-date restrictions with content checks. On FutureX and ForecastBench, Serper restricts search results using τ, and DeepSeek-V4-Flash checks retrieved content for information unavailable before that cutoff. Content checks are needed because an older page may have been updated with later information. Likewise, a statistic's reference period does not establish when it became public.

BRIEF checks evidence before it enters the forecasting context. Search-result dates are screened against τ, fetched pages are checked for later information and reports of the target outcome, and extracted facts are checked again before entering the evidence archive. Reports of the realized outcome for the target event, or the target quantity and period, are excluded. Historical observations and forecasts remain admissible when available before τ; mentioning the target period alone does not make a statement an outcome report.

The following illustrative prompt summarizes the content checks in abridged form, with representative output fields.

Evidence filtering

You receive the forecasting question, its resolution criterion, the evidence cutoff τ, and a retrieved passage or extracted fact with its source and available date information. Assess whether the content can be used to forecast the target outcome.

Exclude information that became available after the cutoff and reports of the realized outcome for the target event or target quantity and period. Check what the content reports, not only the page's publication date, since an older page may contain later updates. Distinguish when information became public from the period it describes.

Retain historical observations and forecasts available before the cutoff. Do not treat a prediction as a realized outcome or exclude content solely because it mentions the target period. Do not answer the forecasting question.

Return a decision to retain or exclude the content and a brief reason, using the fields `decision` and `reason`.

B.3 Scope of the Comparison Controls

The same evidence cutoff and temporal filtering requirements apply to BRIEF and all retrieval-enabled baselines on FutureX and ForecastBench. Their web evidence is subject to Serper date restrictions and DeepSeek-V4-Flash content filtering, while each baseline retains its own planning and evidence-processing procedure. These shared controls do not imply identical internal filtering stages or execution order. Baselines without retrieval do not require web filtering. On Prophet Arena, the supplied frozen corpus determines evidence availability, so the online filtering procedure is not applied.

C Implementation and Baseline Details

C.1 Model and Inference Settings

BRIEF and the eight method baselines use DeepSeek-V4-Flash, while the four standalone model baselines retain the named models in Section 4.1. Retrieval uses Serper on FutureX and ForecastBench and the supplied frozen corpus on Prophet Arena, subject to Section B.

Research is bounded by the settings in Table 10. Each controller round includes one stop-or-probe decision, including a decision that ends research. The controller may stop before reaching the round limit when further research is not warranted.

Table 10: BRIEF inference settings.

Setting	Value
Initial subquestions	At most 6
Controller rounds	At most 5, with early stopping
Initial research budget B	20 budget units per question
Final forecast candidates k	5

The research budget is shared across initial research, direct source access, and subsequent probes. Each initial subquestion or probe consumes one unit when dispatched for research. Direct source access is charged by the number of searches and page fetches it performs, with a minimum charge of one unit. Thus, a budget unit does not always correspond to a single search request. No further research is initiated once the budget is exhausted.

In the weaker-backbone experiments, the selected backbone replaces DeepSeek-V4-Flash in all BRIEF reasoning stages, rather than only in final forecast generation. The evidence-filtering model remains as specified in Section B.2.

C.2 Output Formats and Forecast Refinement

The output specification $\mathcal{M}_q$ preserves each benchmark's required answer type. Five candidates share the terminal belief and evidence without further retrieval, as defined in Section 3.5. During candidate generation, reference-class anchoring asks the model to state a base-rate estimate for the forecasting horizon and justify any adjustment using question-specific evidence. This is a generation instruction, distinct from the probability calibration applied after aggregation.

Table 11: Forecast outputs and candidate aggregation.

Benchmark	Output	Candidate aggregation
FutureX	Option selections or numerical answers in the task's original format.	Most frequent parsed answer, with ties resolved by first occurrence.
ForecastBench	One binary-event probability in $[0, 1]$.	Mean by default; median for questions from `yfinance` and `dbnomics`.
Prophet Arena	One probability in $[0, 1]$ per listed market, aligned with its label.	Per-market median for non-threshold questions with at least three outcomes; per-market mean otherwise.

Aggregation follows the answer type and outcome structure in Table 11. For probability vectors, candidates are combined separately for each market rather than across markets. The aggregated probabilities then undergo Platt scaling as defined in Equation 6, before any structural adjustment.

Structural constraints apply only when the outcome labels support them. For parsed ordered thresholds, threshold-specific adjustment is followed by PAVA to enforce monotonicity. These nested probabilities are not normalized to sum to one. Normalization applies only to mutually exclusive and exhaustive outcomes; structural adjustments are otherwise bypassed.

Final response adapters map outputs to the required format without retrieving new evidence or revising the forecast. Unmappable outputs remain invalid rather than being replaced by a neutral prediction, and are handled under the evaluation policy in Section B.1.

C.3 Core Prompt Templates

Initialization forms S_0; the critic and controller assess the current state, and revision incorporates additional evidence to produce S_{t+1}. The following illustrative templates summarize the core instructions in Section 3, with representative output fields.

Evidence inputs retain source references and follow Section B.2. Retrieval is executed separately by Algorithm 1, not within these prompts.

Belief initialization

You receive the question q, forecast frame F, and initial evidence E_0. Construct a provisional forecast in the answer format specified by F.

Use relevant historical expectations and current conditions to interpret the question-specific evidence. Identify a base rate when the evidence supports one; do not invent a numerical base rate when it does not. Separate evidence supporting the forecast from evidence against it, linking each item to its source. Record unresolved questions that could materially affect the forecast.

Keep the resolution criterion, horizon, and units fixed. Do not retrieve new information or present unsupported assumptions as observations.

Return one JSON object with `forecast`, `supporting_evidence`, `opposing_evidence`, `unknowns`, and `rationale`. Evidence fields contain source-linked lists; `unknowns` lists open questions, and `rationale` briefly explains the provisional forecast. This object is the initial belief state S_0.

Evidence critic

You receive q, F, the current belief S_t, and the evidence archive E_t. Assess whether the evidence supports the current forecast.

Identify useful evidence for and against the forecast, unsupported assumptions, and conflicts among sources. Distinguish independent information from repeated or dependent reports so that repetition does not count as new support. For each issue, identify the relevant claim and source and explain why it matters to the forecast. Summarize what the evidence establishes and what remains unresolved.

Do not revise the belief, retrieve information, propose queries, or decide whether research should stop. Your assessment is passed to the controller and, if research continues, to belief revision.

Return one JSON object with `useful_evidence`, `unsupported_assumptions`, `conflicts`, `repeated_evidence`, and `synthesis`. Use lists for the first four fields, leaving them empty when no issue or evidence applies, and a brief summary for `synthesis`. This object is C_t.

Sufficiency controller

You receive q, F, S_t, the critic assessment C_t, the evidence E_t, and the remaining budget B_t. Decide whether the record supports an informed forecast, not whether the outcome is certain.

Continue only if a remaining issue is both searchable and consequential to the forecast. A stable prediction is not sufficient reason to stop while material conflicts remain. Uncertainty that further retrieval cannot reduce is not a reason to keep searching.

If further research is warranted and budget remains, propose exactly one targeted query that addresses the unresolved issue and is not already answered by E_t. Otherwise, return no probe. Do not revise the belief or execute retrieval. Budget exhaustion forces a stop but does not establish that the evidence is sufficient.

Return one JSON object with `answerable` (a Boolean sufficiency assessment), `rationale` (a brief justification), `unknowns` (remaining consequential questions), and `probe` (one object or `null`). A probe contains a `query` and a `type` chosen from `fact`, `conflict`, or `track_record`. These address a missing fact, a source disagreement, or source reliability, respectively.

Belief revision

You receive q, F, the current belief S_t, the critic assessment C_t, the previous evidence E_t, and the additional evidence ΔE_{t+1}. Revise the belief using the new evidence in the context of what was already known.

Identify which new observations address the critic's concerns, contradict earlier evidence, or add information relevant to the forecast. Do not count overlapping or dependent reports as independent support. Reassess the forecast and record the evidence-based reason for changing or retaining it. Carry forward still-relevant supporting and opposing evidence, remove resolved questions from the unknowns, and record consequential gaps that remain.

Keep the forecast frame fixed and do not retrieve further information. Return the complete updated state, not only the changes, using the same fields as initialization: `forecast`, `supporting_evidence`, `opposing_evidence`, `unknowns`, and `rationale`. The rationale explains how the additional evidence affects the previous assessment. This object is S_{t+1}.

The controller's `answerable` field corresponds to c_t, and its probe query to r_t in Equation 10. The operational rule determines whether retrieval occurs. After revision, the new belief and updated evidence return to the critic and controller in the next round.

C.4 Baseline Implementations and Comparison Controls

The method baselines differ in how they acquire evidence and construct a forecast. Table 12 summarizes their procedures.

Table 12: Forecasting procedures of the method baselines.

Baseline	Forecasting procedure
No tools	One forecast from the question without retrieved evidence.
Direct search	One retrieval pass followed by one forecast, without belief revision.
ReAct + search	A single agent alternates assessment and search.
Self-consistency	Multiple forecasts from a shared evidence set, followed by aggregation.
Halawi et al.	The published evidence-filtering and forecast-aggregation pipeline.
Open Deep Research	The native supervisor and researcher loop.
GPT Researcher	Native planning, evidence compression, and report generation.
DeerFlow	The native lead-agent graph and tool loop.

Method-specific instructions are retained rather than replaced with a common prompt. The comparisons therefore assess evidence acquisition together with forecast construction, as stated in Section 4.1. Shared evidence cutoffs and filtering requirements are described in Section B.3.

D Additional Results

D.1 Evaluation Sets

We construct the evaluation sets from publicly available Hugging Face releases. FutureX comprises 59 questions from one weekly prediction round. For ForecastBench, we randomly sample 100 data-based and 100 market-based questions from the resolved binary-event questions. For Prophet Arena, we use the released Prophet-Arena-Subset-100 and its supplied evidence snapshots. These tasks require models to interpret resolution criteria, integrate evidence available before the specified cutoff, and produce concrete outcome predictions or probability estimates.

D.2 Run-level Results

Tables 13–15 give the individual results underlying the three-run summaries in Table 2. They report tier and overall scores for FutureX and Brier and NLL for ForecastBench and Prophet Arena. Figures 2–4 summarize the means and standard deviations for FutureX Score and Brier on the two probabilistic benchmarks. On these three metrics, BRIEF combines the best mean performance with the smallest reported standard deviation.

Table 13: FutureX tier scores and overall scores for the three runs ($\times 100$, $\uparrow$). The overall score uses the tier weights in Equation 7.

	L1			L2			L3			L4			Score		
Method	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
No tools	41.67	41.67	41.67	36.84	47.37	42.11	17.14	26.89	24.82	11.83	25.82	20.66	21.41	32.03	28.30
Direct search	83.33	58.33	58.33	73.68	63.16	63.16	34.56	44.44	47.80	43.26	66.84	65.71	50.74	58.54	59.09
ReAct + search	83.33	66.67	66.67	47.37	73.68	73.68	62.08	51.77	46.82	50.74	53.79	55.66	56.73	58.45	57.71
Self-consistency	91.67	58.33	58.33	73.68	68.42	68.42	41.56	48.57	48.57	52.91	69.35	69.95	57.53	61.83	62.07
Halawi et al.	75.00	66.67	58.33	63.16	68.42	68.42	32.94	46.28	47.07	36.10	55.24	55.60	44.45	56.33	55.88
Open Deep Research	83.33	75.00	83.33	73.68	78.95	89.47	63.77	64.62	49.92	54.39	59.33	43.78	63.96	66.41	58.72
GPT Researcher	75.00	75.00	66.67	47.37	57.89	52.63	50.15	37.75	37.74	56.27	38.03	42.30	54.53	45.62	45.44
DeerFlow	91.67	91.67	91.67	78.95	84.21	84.21	55.28	49.71	53.73	71.68	67.87	70.87	70.21	68.07	70.48
BRIEF (ours)	83.33	83.33	83.33	84.21	89.47	84.21	68.34	66.81	66.81	78.06	79.81	80.10	76.90	78.20	77.26

Table 14: ForecastBench Brier scores and NLL for the three runs ($\times 100$).

	Brier $\downarrow$			NLL $\downarrow$		
Method	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
No tools	19.24	16.58	16.51	66.13	52.23	52.83
Direct search	17.93	17.05	18.21	55.32	49.91	56.14
ReAct + search	16.45	18.76	15.93	51.51	62.65	50.11
Self-consistency	17.62	16.75	16.05	52.31	50.24	47.01
Halawi et al.	19.50	17.85	17.46	57.51	53.50	50.91
Open Deep Research	17.09	12.18	12.02	54.11	38.33	40.07
GPT Researcher	17.55	14.45	15.63	51.92	44.40	47.10
DeerFlow	14.85	14.65	13.84	44.82	44.68	42.57
BRIEF (ours)	13.14	13.01	12.93	40.87	40.07	39.96

Table 15: Prophet Arena Brier scores and NLL for the three runs ($\times 100$).

	Brier $\downarrow$			NLL $\downarrow$		
Method	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
No tools	22.55	20.84	23.19	64.72	65.33	81.59
Direct search	19.67	18.84	18.59	60.60	54.81	54.27
ReAct + search	19.08	19.24	17.95	55.18	57.47	53.72
Self-consistency	18.51	17.35	17.19	53.63	50.50	50.15
Halawi et al.	20.34	18.10	17.45	58.43	53.20	51.50
Open Deep Research	19.22	19.90	19.90	56.15	59.20	59.59
GPT Researcher	18.43	18.07	16.65	53.80	52.07	48.32
DeerFlow	20.93	17.34	18.56	59.98	50.70	53.96
BRIEF (ours)	16.71	16.73	16.37	48.89	48.94	48.04

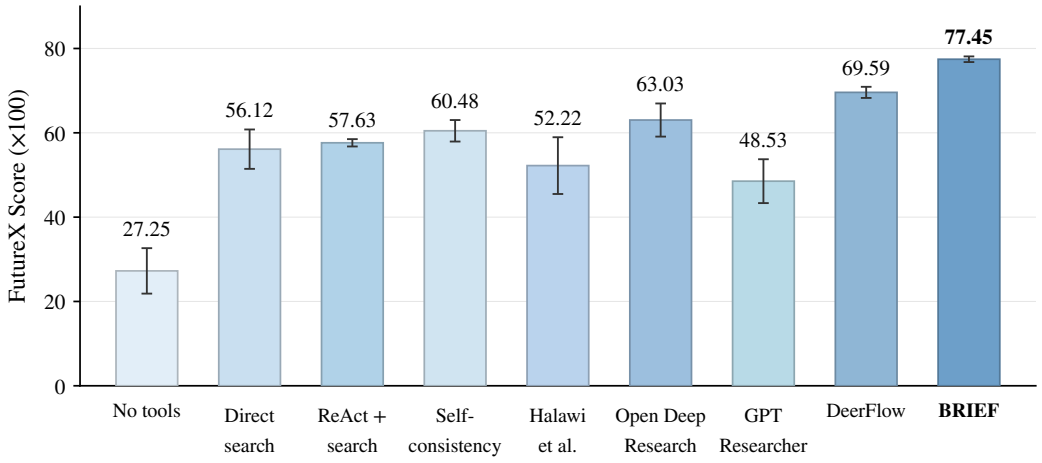


Figure 2: FutureX Score across three runs for BRIEF and the eight method baselines. Bars show means and error bars indicate one standard deviation. Scores are multiplied by 100; higher is better.

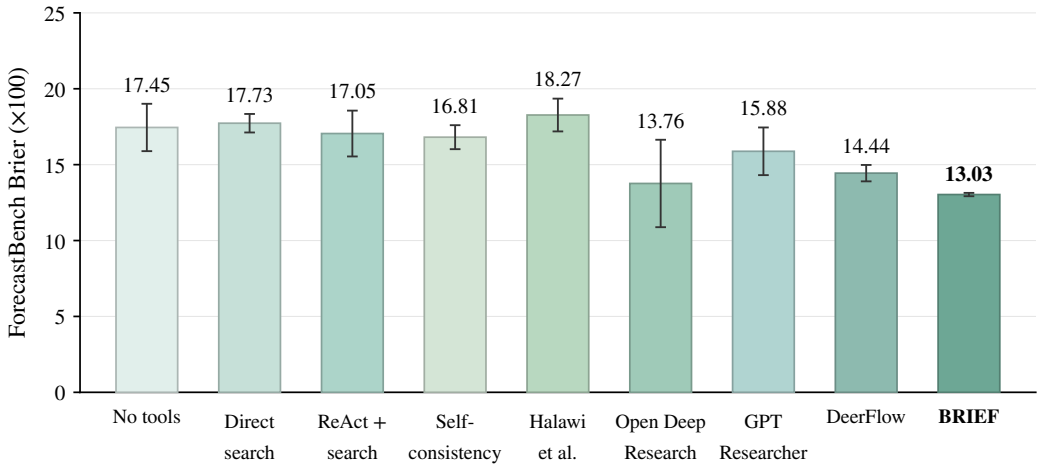


Figure 3: ForecastBench Brier scores across three runs for the same methods. Bars show means and error bars indicate one standard deviation. Scores are multiplied by 100; lower is better.

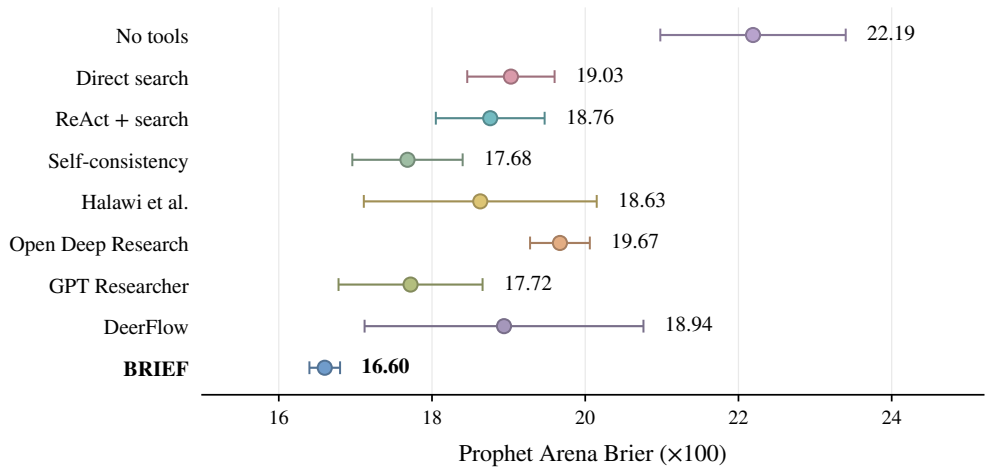


Figure 4: Prophet Arena Brier scores across three runs for the same methods. Circles show means and horizontal error bars indicate one standard deviation. Scores are multiplied by 100; lower is better.

D.3 Retrieval Use and Stopping

BRIEF uses more retrieval on the higher FutureX tiers (Table 16 and Figure 5). Median search calls increase from 18 per question at L1 to 45 at L4, while median page fetches increase from 16 to 32. The controller stops after judging the evidence sufficient on 83% of L1 questions, compared with 33% of L4 questions. These results describe resource use and stopping behavior across tiers, rather than isolate the effect of sufficiency control.

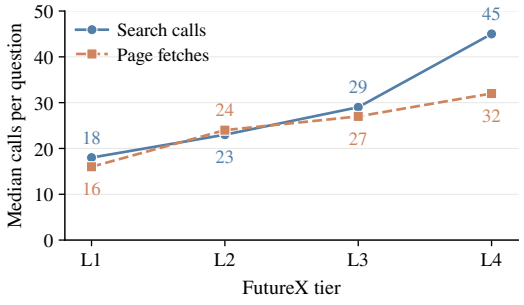


Figure 5: Median search calls and page fetches per question on FutureX. Lines connect tier summaries.

Table 16: BRIEF research cost and stopping behavior by FutureX tier, from the research run underlying the main results.

Level	Search calls	Page fetches	Controller rounds	Early stop (%)	Budget used	Research time (s)
L1	18	16	1.58	83	9.5	319
L2	23	24	2.10	65	11.6	480
L3	29	27	2.77	69	13.9	595
L4	45	32	3.93	33	17.3	782

Search calls, page fetches, and research time are medians across questions; controller rounds and budget use are means. Controller rounds count stop-or-probe decisions, including the stopping decision. Early stop is the percentage of questions for which the controller judged the evidence sufficient and stopped research. Budget use is measured in budget units, with a maximum of 20 per question.

E QUALITATIVE EXAMPLES

The following cases illustrate the distinction between evidence sufficiency and operational stopping in Algorithm 1. In the first case, BRIEF stops after obtaining relevant historical evidence for an unpublished statistic. In the second, additional research resolves measurement conflicts but leaves a recent-price gap when the round limit is reached.

We select these two FutureX cases from archived runs. Each case summarizes the retained research trajectory and reports the forecast from its first answer-stage replay, rather than reproducing a complete execution log. The resolved outcomes are shown only for retrospective evaluation. These examples illustrate individual decisions, not the frequency or causal effect of the behaviors discussed.

E.1 STOPPING BEFORE A STATISTICAL RELEASE

The first case concerns a statistic whose reference period had ended but whose value was not yet public. The relevant decision was whether the available evidence supported a forecast, rather than whether retrieval could recover the unreleased answer.

Case Study 1. U.S. Employment Cost Index

Question

What will be the quarter-over-quarter percentage increase in the U.S. Employment Cost Index (ECI) for civilian workers in Q2 2026? The evidence cutoff was July 29, 2026, before the scheduled July 31 release.

Forecast frame and relevant evidence

The forecast frame identified the target as the seasonally adjusted three-month change in total compensation from March to June 2026. This excluded year-over-year growth, annualized rates, and wages alone. Role-guided research sought recent quarterly releases and the release schedule. The evidence critic retained the relevant quarterly sequence of 0.9%, 0.8%, 0.7%, and 0.9% for the quarters ending in June, September, and December 2025 and March 2026, respectively.

The critic also distinguished the published seasonally adjusted headline from changes computed using unadjusted index levels. It preferred the headline because it matched the resolution criterion. The resulting evidence supported a forecast in the requested statistical measure, rather than an estimate derived from a different growth rate.

Stopping decision

At its first assessment, the sufficiency controller judged the available evidence adequate and issued no further probe. This decision combined relevant historical observations with confirmation that the target release was still unavailable. The absence of the unpublished result was not, by itself, a reason to continue retrieval.

Forecast and resolved outcome

BRIEF forecast **0.9%**. The subsequently released benchmark outcome was also **0.9%**.

The main lesson is that sufficiency depends on evidence relevant to the forecast target, not on observing the outcome itself. The frame specified which quarterly measure mattered, role-guided research retrieved matching historical observations, and the critic excluded incompatible measures. This gave the controller a basis for stopping while uncertainty about the forthcoming release remained. It illustrates the distinction in Section 3.4 between an informed forecast and a certain outcome; the correct answer alone does not isolate any component's effect.

E.2 STOPPING WITH AN UNRESOLVED EVIDENCE GAP

The second case separates uncertainty about how to compute the target from uncertainty about a required input. Unlike the unpublished ECI value, a recent pre-cutoff stock price was a potentially retrievable observation. Its absence therefore remained a reason for further research.

Case Study 2. CGN Power Market Capitalization

Question

On July 30, 2026, what will be the total market capitalization of CGN Power (003816) on the Shenzhen Stock Exchange, in hundreds of millions of yuan, to two decimal places? The evidence cutoff was July 29, 2026.

Forecast frame and evidence gaps

The forecast frame distinguished total market capitalization from free-float capitalization and identified the required output unit. Initial evidence left three issues unresolved: the applicable share count, whether the reported measure used all issued shares or only A-shares, and the most recent price available before the cutoff. Role-guided research addressed these components separately.

What additional research resolved

Further evidence supported a total of 50,498,805,700 issued shares. Historical quote records were consistent with multiplying the Shenzhen A-share closing price by total A+H shares. The evidence critic also rejected inconsistent price snippets and identified source labels that confused billions of yuan with hundreds of millions of yuan. These checks resolved the share-count and measurement conflicts.

The recent-price gap remained. A later assessment identified the freshest retrieved quote as 4.290 yuan on March 26, 2026, several months before the forecast date. Further research did not recover the July quote sought by the controller. BRIEF therefore distinguished a resolved calculation rule from an unresolved forecasting input.

Stopping decision

The controller continued to judge the evidence insufficient. Research ended at the five-round limit, not because the missing price had been found or the evidence had become sufficient.

Forecast and resolved outcome

The final forecast was **2146.20**, compared with the benchmark outcome of **2100.75**, both in units of 10^8 yuan. The forecast used an assumed target-day closing price near 4.25 yuan, not a verified July quote.

The resolved conflicts made the calculation consistent with the target, but did not make the evidence complete. The remaining price gap explains both the controller's continued requests for research and the assumption required in the final forecast. This is the operational distinction in Algorithm 1: the system must still produce an answer when a limit is reached, without treating that limit as evidence of sufficiency. Together, the cases illustrate why an unavailable outcome and a missing historical input should not lead to the same stopping judgment.