FOREWEAVE: LEARNING WHAT TO INVESTIGATE FOR EVENT FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting future events requires agents to acquire evidence before decisive outcomes are known. Yet an apparently plausible forecast can still depend on consequential conditions that remain unresolved. We introduce ForeWeave, a framework for process-guided adaptive investigation that connects reusable structures from historical investigations with evidence gaps revealed during a current investigation. ForeWeave induces historical process structures to guide three view-specific investigations, then performs targeted continuation from a selected trajectory prefix when an unresolved issue could change the forecast. A Forecast Synthesis Agent audits the initial and continued records, performs targeted verification when blocking evidence gaps remain, and returns the final answer and probability. We evaluate ForeWeave on FutureX-Past with Qwen, DeepSeek, and GPT backbones, and on ForeWeave-Live with DeepSeek and GPT. Across both benchmarks, ForeWeave achieves the highest Overall Score, outperforming the strongest baseline by 1.96∼6.45 points. It achieves Overall Scores of 49.33, 56.66, and 59.51 on FutureX-Past with Qwen, DeepSeek, and GPT, respectively, and 70.59 and 67.65 on ForeWeave-Live with DeepSeek and GPT.

1 INTRODUCTION

The past can be examined; the future can only be anticipated. At prediction time, the outcome of an event-forecasting question remains unresolved and cannot be retrieved directly (Karger et al., 2025; Zeng et al., 2025). A forecasting agent must instead gather evidence informative about the outcome and commit to a prediction under a fixed information cutoff and limited budget. Tool-using language-model agents operate in this setting by retrieving evidence, reasoning over it, and adapting their searches as observations arrive (Halawi et al., 2024; Yao et al., 2023). Yet a plausible forecast may still rest on a consequential condition that the agent leaves unchecked (Figure 1). A revenue forecast, for example, may rely on positive performance signals without establishing whether they concern the target business segment and reporting period. The challenge is therefore not only to retrieve relevant evidence, but also to identify and investigate unresolved conditions that could materially change the forecast within these constraints.

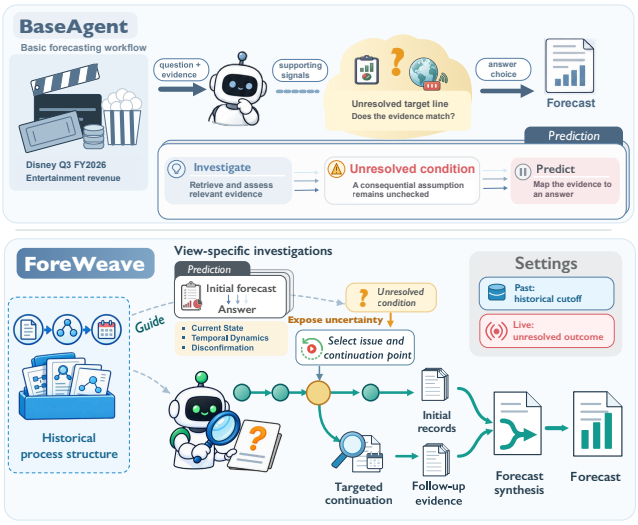


Figure 1: **ForeWeave overview.** A base agent can leave a consequential condition unchecked (top). ForeWeave uses historical process structure and current evidence to guide multi-view investigation, targeted continuation, and forecast synthesis (bottom).

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.

Existing work improves this process along two broad directions. One line accumulates predictive factors, calibration principles, and procedural experience to guide search, evidence handling, and judgment (Chang et al., 2026; Zhong et al., 2026; Wei et al., 2026). Another varies models, evidence access, or investigative perspectives to obtain complementary judgments and then reconciles the resulting reports or forecasts (Chetlapalli et al., 2026; Li et al., 2026; Chai et al., 2026).

Together, these lines broaden the experience and perspectives available to a forecasting agent. A historical investigation, however, records more than its final prediction: the order in which checks were performed, which later checks depended on earlier observations, and which alternative actions were taken under different evidence states. Meanwhile, a current investigation exposes unresolved issues specific to the question at hand. This suggests a complementary direction: connecting structural relationships from historical investigations to the unresolved issues revealed by current evidence, so that follow-up can resume from a selected point on an existing trajectory rather than restart an unstructured search.

We pursue this direction with FORECWEAVE, a framework for process-guided adaptive investigation. ForeWeave induces traceable process structures from audited historical trajectories, preserving reusable checking relationships and correction obligations without carrying historical answers into the new forecast. For a new question, the resulting structure guides three complete investigations: Current State examines established facts, Temporal Dynamics examines how the event may evolve, and Disconfirmation tests the leading judgment against counterevidence and alternatives. Each investigation produces a complete forecast and evidence record. When these records expose an issue that could materially change the forecast, targeted continuation resumes a selected current-question trajectory from an earlier prefix rather than restarting the search. A Forecast Synthesis Agent then audits the initial and continued records, may verify consequential gaps when needed, and returns the final answer and probability.

We evaluate ForeWeave on the external FutureX-Past benchmark (Zeng et al., 2025) with Qwen, DeepSeek, and GPT, and on the prospectively collected ForeWeave-Live set with DeepSeek and GPT. Across the DeepSeek and GPT evaluations, ForeWeave improves Overall Score over the Base Agent by 1.96$\sim$7.89 points on both datasets.

Our contributions are threefold. **(1)** We develop a traceable historical process representation that captures reusable relationships among investigation checks and node-linked correction obligations while preserving their provenance. **(2)** We introduce an adaptive forecasting framework that uses historical structure to organize view-specific investigations and current evidence to drive targeted continuation and evidence-guided forecast synthesis. **(3)** We construct a prospective forecasting collection and provide a multi-backbone evaluation across retrospective and prospective settings under explicit temporal protocols.

2 FORECASTING SETUP

For an unresolved question q, a tool-using agent produces a task-required forecast $\hat{z}$ and an investigation trajectory τ under an information cutoff t_{cut} and budget B:

$$(\hat{z}, \tau) = \mathcal{F}(q, \mathcal{I}(t_{\text{cut}}), \mathcal{E}; B)\,. \tag{1}$$

Here, $\mathcal{I}(t_{\text{cut}})$ contains cutoff-admissible information, $\mathcal{E}$ provides retrieval tools, and $\tau = (a_1, o_1, \ldots, a_T, o_T)$ records the resulting actions and observations (Karger et al., 2025; Zeng et al., 2025; Yao et al., 2023). Further retrieval may change the forecast without resolving the uncertainty of an outcome that has not yet occurred. Because τ contains only the evidence actually retrieved while the outcome remains unresolved, forecasting requires deciding both how to interpret observed evidence and what to investigate next.

3 FOREWEAVE

ForeWeave pairs reusable historical process structure with current evidence for adaptive event forecasting. Given a question q, a time cutoff, and a budget, it uses historical process guidance to plan three view-specific investigations and organize their trajectories into structured records. When these records expose an unresolved issue that may change the forecast, targeted continuation reuses a selected prefix of a current-question trajectory to investigate that issue further. The Forecast Synthesis

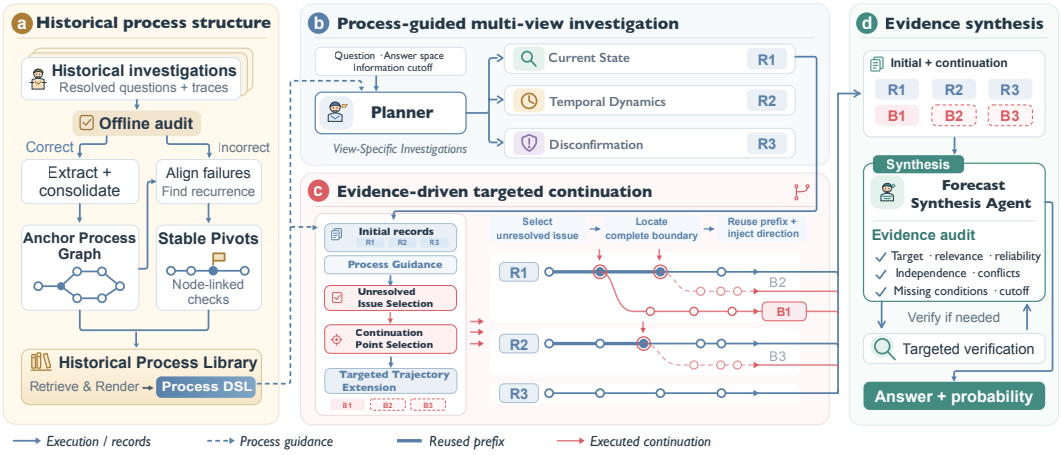


Figure 2: **ForeWeave investigation pipeline.** Historical process structure guides view-specific investigations, while unresolved issues revealed by current evidence guide targeted continuation. The resulting records are audited and synthesized into the final forecast. Solid and dashed paths denote selected and optional continuations, respectively.

Agent audits the initial and continuation records, performs targeted verification when consequential evidence gaps remain, and produces the final answer and probability. Historical investigations provide answer-free guidance rather than reusable conclusions; evidence from the current question determines whether and where to continue. Figure 2 summarizes the pipeline.

3.1 ForeWeave Data Construction

Sources and collection. We construct two resources for ForeWeave—the Historical Investigation Set and ForeWeave-Live—using public prediction-market feeds and temporally indexed news sources. The Historical Investigation Set consists of resolved questions for which usable investigation trajectories are available. It is assembled from three sources: resolved prediction-market questions collected through public APIs or data releases (e.g., Manifold, Polymarket, and Kalshi); a designated construction subset of FutureX-Past (Zeng et al., 2025); and news-derived questions generated from time-stamped articles following established news-based forecasting pipelines (Goel et al., 2026; Chandak et al., 2025). ForeWeave-Live, in contrast, consists of questions collected prospectively from public prediction-market feeds while they remain unresolved.

Dataset organization. We preserve source provenance and temporal metadata needed for cutoff control, outcome matching, and later audits. Questions from different sources are first converted into a common forecasting format. For historical questions, we remove any question whose resolved answer can already be directly retrieved using information available at its forecasting cutoff; detailed screening and auditing procedures are provided in Appendix A. The Historical Investigation Set supplies the trajectories for historical process structure induction, while questions collected before resolution form ForeWeave-Live and support prospective evaluation after forecasts are recorded. Table 2 summarizes the two ForeWeave data resources and their roles.

3.2 Historical Process Structure Induction

ForeWeave distills reusable checking relationships from historical trajectories into answer-free process guidance (Figure 2(a)).

Trajectory Audit. For each resolved historical question, we exclude incomplete or unscorable trajectories and select small, diverse subsets of correct and incorrect runs. Outcome labels determine how trajectories are used but are hidden from workflow-extraction and process-construction models. For a historical question h, we represent the resulting historical process structure as

$$\mathcal{M}_h = (q_h, G_h, \mathcal{Z}_h, \rho_h), \tag{2}$$

where q_h is the historical question, G_h its Anchor Process Graph, $\mathcal{Z}_h$ its Stable Pivots, and ρ_h links graph elements to source workflows and trajectory messages.

Anchor Graph and Stable Pivots. ForeWeave extracts workflows from selected correct trajectories and consolidates them by process purpose into an Anchor Process Graph, retaining source references. Action nodes encode investigation operations, decision nodes test observable evidence states, and edges represent supported transitions. Unsupported outcomes remain coverage gaps, and only graphs passing structural and provenance checks enter the library. Incorrect trajectories provide checking obligations rather than graph nodes or edges: their workflows are aligned with graph paths to identify missing checks, observable conflicts, and coverage gaps. A deviation becomes a Stable Pivot only when at least two distinct incorrect trajectories support the same deviation type at the same referenced graph nodes. Stable Pivots capture recurring checking obligations, not current facts or predicted outcomes.

Process Retrieval and Guidance Rendering. For a new question, ForeWeave proceeds in three steps. 1) *Candidate retrieval.* It retrieves candidate historical questions and their validator-clean process structures. 2) *Matching and fallback.* Model-based relation and graph-relevance checks use only the questions and, for graph relevance, the rendered structure, without access to target outcomes or forecast scores. Retrieval prioritizes Full Match to the overall investigation structure, followed by Partial Match to local steps or generic operations, and finally a similarity-based fallback without reassessing relevance. The retrieval route is recorded before forecasting. 3) *Guidance rendering.* The selected historical question context, Anchor Process Graph, and Stable Pivots are rendered deterministically in a compact process domain-specific language (Process DSL). It preserves supported actions, conditions, branches, and checking obligations while excluding historical answers, correctness labels, and offline scores, and does not prescribe a fixed execution path.

3.3 Process-Guided Multi-View Investigation

Figure 2(b) shows how retrieved process guidance is adapted into three view-specific investigations.

View Planning. Given the current question, its answer space, the information cutoff, and applicable Process DSL guidance, the planning stage translates relevant historical actions, conditions, and branches into question-specific checks while omitting inapplicable steps. It produces three self-contained plans, each covering the full question but prioritizing different evidence:

- **Current State:** determine which candidate answers are supported by established facts, recent signals, and relevant base rates at the information cutoff.
- **Temporal Dynamics:** examine which conditions and changes could lead from the current state to the target outcome, including dependencies, timing, bottlenecks, and alternative paths.
- **Disconfirmation:** form a tentative leading answer, seek evidence that could overturn it, and compare it with credible alternatives.

Each plan covers the full forecasting question, and each resulting investigation produces a complete calibrated forecast. The views differ in evidence priority, but may overlap on essential facts and need not produce different answers.

View-Specific Investigation. ForeWeave executes each plan as a separate trajectory using the current question, the shared investigation protocol, and the permitted tools, without access to the other views' trajectories or predictions. Each trajectory gathers cutoff-compliant evidence, compares plausible answers, and produces a complete forecast with a probability. The investigations pursue different evidence priorities rather than serving as independent votes or guaranteed independent sources.

Evidence Record Construction. ForeWeave then organizes each trajectory and its observations into a structured record R_m without revising the original prediction. Each record contains the forecast, supporting and disconfirming evidence, source references, a rationale summary, and the main unresolved issue, while retaining provenance to the underlying trajectory. Together, these records expose unresolved issues and provide the interface for targeted continuation and final synthesis.

3.4 EVIDENCE-DRIVEN TARGETED CONTINUATION

Figure 2(c) shows how the initial evidence determines whether and where an investigation should continue. ForeWeave first synthesizes the initial records into a frozen parent forecast $(\tilde{y}, \tilde{p})$, which serves as a hypothesis for follow-up rather than as ground truth.

Unresolved Issue Selection. ForeWeave allows up to K_{br} additional continuation trajectories beyond the three initial investigations. Each continuation targets a cited unresolved issue that could change the answer or its probability, using the current question, forecast, and available investigation records. The selected issue identifies a source reference, a competing answer or numerical boundary, and an observable discriminator between the alternatives.

Continuation Point Selection. The cited source reference identifies the corresponding parent trajectory τ_m. ForeWeave selects the last complete tool-result boundary j before that observation, preventing the evidence under examination and later observations from being inherited as established facts. It pairs the resulting prefix with a continuation direction d_u that specifies the issue, competing alternative, and discriminator:

$$\tau^{+} = \mathrm{Continue}(\mathrm{prefix}(\tau_m, j), d_u)\,. \tag{3}$$

The prefix is drawn from the current question's trajectory rather than from a historical process graph.

Trajectory Continuation. The continued trajectory reuses the selected prefix and treats the parent forecast as a hypothesis to test. It uses the visible observations and permitted tools to investigate the discriminator under the applicable temporal protocol, then produces a complete forecast. The continuation may retain the parent answer, revise its probability, or support another answer. Its evidence, rationale, uncertainty, and parent provenance are organized into a continuation record. Let $\mathcal{R}^{+}$ denote the resulting continuation records. These records are treated as dependent on their parent trajectories rather than as independent votes. If no valid continuation is produced, ForeWeave retains the frozen parent forecast; otherwise, the initial and continuation records proceed to final synthesis.

3.5 EVIDENCE-GUIDED FORECAST SYNTHESIS

Evidence Audit. The Forecast Synthesis Agent receives the current question, anonymized initial and continuation records, and the selected historical process guidance. It first establishes the prediction target, observation period, unit, and answer space. Historical guidance identifies conditions worth checking but does not count as evidence for the current question. The agent compares the evidence for plausible answers by distinguishing observed facts from external forecasts and assumptions and by assessing evidence relevance, reliability, and dependence. Repeated reports of the same underlying finding are treated as related evidence. Record-level answers and probabilities are hypotheses to reassess rather than votes or reliability weights, and no numerical correctness score is assigned to individual trajectories.

Targeted Verification. During synthesis, the agent may invoke bounded retrieval tools to verify consequential evidence gaps before producing the final forecast. It first inspects cited model-visible observations and, when a material gap remains, performs targeted search or page retrieval under the applicable temporal protocol and verification budget. The retrieved observations are used to reassess the relevant unresolved issue and update the comparison among candidate answers. Failed or inconclusive retrieval preserves uncertainty rather than counting against a candidate, while sufficient records allow the agent to proceed without additional retrieval.

Forecast Synthesis. Let $\mathcal{D}_q$ denote the selected historical process guidance and $\mathcal{O}_{\mathrm{verify}}$ the observations obtained through verification, possibly empty. The final forecast is

$$(\hat{y}, \hat{p}) = \mathrm{Synthesize}\big(q, \{R_m\}_{m=1}^{M}, \mathcal{R}^{+}, \mathcal{D}_q, \mathcal{O}_{\mathrm{verify}}\big)\,. \tag{4}$$

Using the combined evidence, the agent selects an answer permitted by the task and assigns a probability that reflects unresolved conflicts, evidence dependence, and plausible alternatives. The resulting forecast may retain or revise the frozen parent forecast, but any revision must be grounded in the available evidence.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets, tools, and temporal protocol. We evaluate on the external FutureX-Past benchmark (Zeng et al., 2025) and the prospectively collected ForeWeave-Live set introduced in Section 3.1. Our evaluation covers 66 FutureX-Past questions from the second week of August 2026 and 68 ForeWeave-Live questions collected on September 23–24, 2026. For FutureX-Past, all methods use the same Tavily-backed retrieval pipeline (Tavily AI, 2026). Following retrospective forecasting protocols that reconstruct cutoff-bounded evidence environments (Ma et al., 2026; Chai et al., 2026), we filter and audit retrieved evidence against each question's historical cutoff before it becomes visible to the forecasting agent. Our implementation uses a multi-stage control layer with pre-search query guarding, answer-blind date-role classification, selective sanitization of separable post-cutoff page modules, and content-level temporal auditing; implementation details are provided in Appendix B. For ForeWeave-Live, forecasts are frozen before their outcomes are known and later scored after resolution, while web search and webpage visits use Serper and Jina, respectively, without retrospective date masking.

Forecasting methods and evaluation. On FutureX-Past, we run all forecasting methods with three backbones: Qwen3.7-Max in non-thinking mode (Team, 2025), DeepSeek-V4-Pro in thinking mode (DeepSeek-AI, 2026), and GPT-5.5 with medium reasoning effort (OpenAI, 2026). On ForeWeave-Live, we evaluate DeepSeek-V4-Pro and GPT-5.5. Results are reported separately for each backbone. **(1) Base Agent.** A single tool-using agent conducts one complete investigation and returns a forecast (Yao et al., 2023); we report the mean performance of the three independent runs used by self-consistency. **(2) Base Agent + SC.** The three forecasts are combined by a deterministic rule that uses majority voting for categorical answers and averaging for compatible numeric or distributional answers. **(3) Experience-Guided Agent.** A single agent receives textual experience derived from historical investigations. **(4) ForeWeave.** Our full method combines process-guided view-specific investigation, evidence-driven targeted continuation, and evidence-guided forecast synthesis. For FutureX-Past, we report the official scores for the four benchmark levels (L1–L4), where higher levels represent increasingly open-ended forecasting tasks. Its Overall Score uses the benchmark's fixed level weighting, $S_{\text{Overall}} = 0.1S_1 + 0.2S_2 + 0.3S_3 + 0.4S_4$, where S_l denotes the official score at level l, rather than a pooled average over questions. For ForeWeave-Live, we report the unweighted mean question score over all 68 questions after the outcomes become available. Dataset construction, temporal controls, method configurations, and evaluation details are provided in Appendices A and B.

4.2 MAIN RESULTS

Table 1 summarizes forecasting performance on FutureX-Past and ForeWeave-Live.

ForeWeave consistently achieves the highest Overall Score. On FutureX-Past, ForeWeave reaches 49.33, 56.66, and 59.51 with Qwen, DeepSeek, and GPT, respectively, improving over Base Agent by 4.81∼7.89 points and over the strongest baseline for each backbone by 2.26∼6.45 points. On ForeWeave-Live, it reaches 70.59 and 67.65 with DeepSeek and GPT, respectively, exceeding the strongest baseline by 5.88 and 1.96 points. The strongest baseline varies across backbones and datasets, whereas ForeWeave leads in every evaluated setting.

ForeWeave shows its largest improvements on higher-level forecasting tasks. On FutureX-Past, While performance on the lower benchmark levels varies across backbones, the most pronounced improvements are observed on the more open-ended forecasting tasks. In particular, ForeWeave improves the L4 score by 7.08 points with DeepSeek and 10.37 points with GPT over Base Agent, exceeding all baselines on L4 for both backbones, while improving the Qwen L3 score by 21.85 points over Base Agent and 18.93 points over the strongest L3 baseline. These gains on the more open-ended benchmark levels contribute to ForeWeave's superior weighted Overall performance.

The benefit of self-consistency aggregation varies across backbones and settings. On FutureX-Past, Base Agent + SC improves over Base Agent by 3.36 points for Qwen and 1.13 points for GPT, but reduces the DeepSeek score by 1.35 points. On ForeWeave-Live, it improves the DeepSeek score by 2.45 points but reduces the GPT score by 0.98 points. These results indicate that aggregating multiple trajectories can be beneficial, but simple answer-level aggregation is not consistently

Method	FutureX-Past					ForeWeave-Live
	L1↑	L2↑	L3↑	L4↑	Overall↑	Overall↑
Qwen w/ Tools						
Base Agent	38.88	**63.80**	46.83	32.53	43.71	–
Base Agent + SC	**50.00**	62.86	49.75	36.43	47.07	–
Experience-Guided Agent	**50.00**	62.86	45.32	**37.81**	46.29	–
ForeWeave	16.67	60.00	**68.68**	37.66	**49.33**	–
DeepSeek w/ Tools						
Base Agent	38.89	64.76	64.31	39.29	51.85	62.25
Base Agent + SC	**50.00**	65.71	58.46	37.05	50.50	64.71
Experience-Guided Agent	33.33	62.86	**72.57**	40.83	54.01	64.71
ForeWeave	**50.00**	**71.43**	62.75	**46.37**	**56.66**	**70.59**
GPT w/ Tools						
Base Agent	50.00	61.90	65.21	36.68	51.62	65.69
Base Agent + SC	50.00	62.86	**66.44**	38.12	52.75	64.71
Experience-Guided Agent	33.33	**71.43**	65.69	39.34	53.06	64.71
ForeWeave	**66.67**	**71.43**	65.77	**47.05**	**59.51**	**67.65**

Table 1: **Forecasting results on FutureX-Past and ForeWeave-Live.** FutureX-Past reports official 0–100 scores for L1–L4 and a fixed weighted Overall score. Higher scores are better.

effective. ForeWeave outperforms the evaluated three-sample SC baseline in every setting, suggesting that synthesizing evidence rather than directly combining final forecasts provides a more effective aggregation strategy.

Historical experience generally provides useful guidance. On FutureX-Past, Experience-Guided Agent improves over Base Agent by 2.58, 2.16, and 1.44 points with Qwen, DeepSeek, and GPT, respectively. On ForeWeave-Live, it improves Overall Score by 2.46 points with DeepSeek but decreases it by 0.98 points with GPT. These results indicate that incorporating historical experience is generally beneficial for investigation. ForeWeave further improves upon Experience-Guided Agent by 2.65∼6.45 points on FutureX-Past and 2.94∼5.88 points on ForeWeave-Live, suggesting that extracting reusable process structures from historical trajectories is more effective than directly injecting historical experience.

4.3 Mechanism Analysis

Figure 3 examines how forecasting performance varies with investigation organization, process relevance, forecast synthesis, and targeted continuation. The panels use separately frozen runs, and comparisons are made within each panel.

Organizing investigations around distinct views improves forecasting. We compare three investigations sharing one view with three assigned to Current State, Temporal Dynamics, and Disconfirmation, respectively. Both configurations use evidence-guided synthesis without targeted continuation. In Figure 3(a), using distinct views improves the Overall Score by 4.49 points for GPT and 7.02 points for Qwen. One possible explanation is that distinct evidence priorities help the investigations examine conditions that repeated investigations sharing a single view may overlook.

Process relevance matters for effective guidance. We compare Full Match ($n = 16$), where the retrieved process matches the overall investigation structure, with Partial Match ($n = 20$), where only individual steps or generic operations match. Retrieval routes are recorded before forecasting (Section 3.2). Scores are averaged across three runs. In Figure 3(b), process-graph guidance improves DeepSeek and Qwen by 5.95 and 3.54 points under Full Match, respectively, compared with 0.56 and 1.02 points under Partial Match. The larger gains under Full Match suggest that historical guidance is more useful when its process structure closely matches the current investigation.

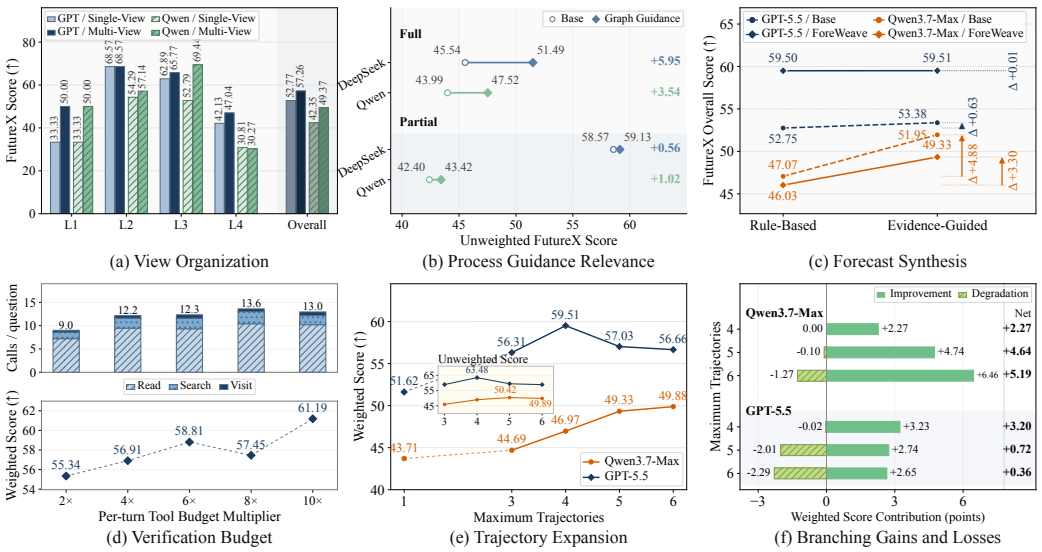


Figure 3: **Analysis of ForeWeave on FutureX-Past.** (a) Three investigations sharing a single view versus three using distinct views. (b) Initial investigations with versus without process-graph guidance, grouped by whether the retrieved process matches the overall investigation or only partial steps. (c) Rule-based aggregation of final predictions versus evidence-guided synthesis of investigation records, with upstream Base or ForeWeave investigations fixed. (d) Tool calls per question and forecasting scores under increasing per-turn verification allowances, with candidate records and the maximum number of model turns fixed. (e) Forecasting scores for a single-agent reference, three initial investigations, and targeted continuation configurations allowing up to four, five, or six trajectories. (f) Positive and negative per-question score contributions relative to the three-investigation reference, decomposing the net gains of the continuation configurations in (e). Panel (b) and the inset of (e) report unweighted question-level scores.

Evidence-guided synthesis can improve forecast aggregation. We examine its role from two perspectives. 1) *Aggregation strategy.* Figure 3(c) compares rule-based aggregation of final predictions with evidence-guided synthesis that uses investigation records and can verify evidence through tool calls, keeping the upstream investigations fixed. Evidence-guided synthesis improves Qwen's Base and ForeWeave scores by 4.88 and 3.30 points, respectively. GPT shows smaller gains with Base investigations and nearly no change with ForeWeave. One possible explanation is that GPT's upstream investigations already produce stronger forecasts, leaving less room for evidence-guided synthesis to improve over rule-based aggregation. These results suggest that synthesizing investigation evidence can outperform aggregating final predictions alone, although the benefit depends on the backbone and input investigations. 2) *Verification budget.* Figure 3(d) varies the per-turn verification allowance while fixing candidate records and the maximum number of model turns. For GPT, the $10\times$ setting achieves the highest displayed score of 61.19, but performance is non-monotonic, and it uses fewer tools per question than $8\times$. Increasing the verification allowance can improve forecasting performance. At higher allowances, actual tool use remains around $12\sim14$ calls per question, while forecasting gains are non-monotonic.

Targeted continuation provides additional forecasting gains. In Figure 3(e), both backbones benefit over synthesis from the three initial investigations alone: GPT gains 3.20 points with at most four trajectories, and Qwen gains 5.19 points with at most six, while allowing more trajectories does not consistently help. The per-question decomposition in Figure 3(f) clarifies this pattern: for GPT, losses offset most gains at six trajectories, leaving only 0.36 points of net improvement, whereas Qwen retains larger net gains with smaller losses. Both backbones thus gain overall, but Qwen's gains increase across the displayed configurations while GPT's diminish beyond four trajectories. One possible explanation is that Qwen's weaker initial forecasts leave more room for continuation to help, whereas GPT's stronger initial forecasts offer less scope for improvement.

4.4 Case Analysis

Assessing evidence for the current forecast. A past outcome can be certain while offering limited support for a future prediction involving different conditions. We illustrate this distinction with a ForeWeave-Live case, comparing an individual Base Agent forecast with ForeWeave using DeepSeek-V4-Pro. The Base Agent treated a previous-round victory as evidence that settled the current advancement question, whereas ForeWeave distinguished the opponent, round, and resolution rules before weighing evidence for the upcoming match. In a separate Excelsior case, targeted continuation checked market descriptions to distinguish similarly named clubs, illustrating how a specific unresolved issue can guide further investigation.

Live Case: Will Groene Ster Advance?

Question. Would Groene Ster advance past Amsterdamsche FC in the second qualifying round of the Dutch Cup on September 22?
Resolved outcome. *No*.

Base Agent
The agent retrieved Groene Ster's 2–1 victory over Heerjansdam on September 2 in the previous round. It treated that result as resolving the current advancement question and predicted *Yes* (0.99).

ForeWeave
The investigation distinguished the current opponent and round, as well as advancement from a regulation-time win. Final synthesis weighed current odds and team-level evidence, including conflicting market snapshots, and predicted *No* (0.73).

5 Related Work

Forecasting benchmarks. Future-event forecasting evaluates unresolved outcomes under information cutoffs, unlike research-agent benchmarks whose answers already exist at query time (Wei et al., 2025). Prospective benchmarks such as ForecastBench and FutureX collect predictions before resolution, while retrospective settings such as FutureSim and Bench to the Future reconstruct earlier information states through temporally controlled evidence (Karger et al., 2025; Zeng et al., 2025; Goel et al., 2026; Wildman et al., 2025). OracleProto combines knowledge-cutoff-aligned sample admission, temporal masking, and leakage detection. (Ma et al., 2026). These protocols motivate our retrospective and prospective evaluations.

Forecasting agents. Forecasting agents retrieve evidence and aggregate probabilistic judgments (Halawi et al., 2024). Experience-guided methods use predictive factors and factual or procedural memories to guide investigation (Chang et al., 2026; Zhong et al., 2026; Wei et al., 2026), while collective methods diversify models, evidence, or perspectives before aggregation (Chetlapalli et al., 2026; Li et al., 2026). Agentic Time Machine plans analytical angles through exploratory search and permits further search during aggregation (Chai et al., 2026). ForeWeave reuses relationships among historical investigation checks to organize view-specific investigations. Beyond aggregation-stage search, it continues current-question trajectories from selected prefixes to investigate unresolved issues exposed by current evidence before final synthesis. Further discussion appears in Appendix D.

6 Conclusion

We presented ForeWeave, a framework that uses historical process structure to organize forecasting investigations and current evidence to guide targeted continuation and forecast synthesis. ForeWeave achieves higher Overall Scores than the compared baselines across the evaluated settings on FutureX-Past and ForeWeave-Live, although the benefits vary with the backbone and the relevance of historical guidance. The current framework uses a frozen Historical Investigation Set. Future work could expand this resource through rolling collection of live questions, incorporating audited trajectories after their outcomes resolve and freezing each updated process library before subsequent forecasting windows. We also aim to improve forecast synthesis across larger and more heterogeneous information sources by better assessing source reliability, accounting for evidence dependence, and reconciling conflicting observations. These directions would help establish when accumulated investigation experience and broader evidence access translate into better forecasts.

AI USE STATEMENT

During manuscript preparation, we used large language models solely as assistive tools to refine wording and sentence structure and to help design the presentation of selected figures. They were not used to originate the research questions, propose or refine hypotheses, design the methodology or experiments, implement the methods, process the data, conduct qualitative analysis, or interpret the results. Language models were also used as experimental components and forecasting backbones solely for scientific research; their roles and configurations are documented in Section 4.1 and Appendix B. We reviewed all AI-assisted work and take responsibility for the final content of this paper, including any text, claims, or artifacts produced with the assistance of large language models.

ETHICS STATEMENT

This study uses questions and evidence derived from publicly available sources and released data feeds; it does not rely on proprietary or confidential information. As described in Section 3.1 and Appendix A, the data-construction process screens questions for validity, ambiguity, sensitive content, and duplication, and the paper cites the external datasets and source platforms on which the study relies. Nevertheless, forecasting agents can reproduce errors or biases in public information, and probabilistic forecasts may be overtrusted or misapplied in consequential financial, policy, or social decisions. ForeWeave and its benchmark are intended for scientific research and evaluation rather than direct high-stakes decision making. We encourage users to apply these artifacts transparently and responsibly and to avoid uses that could harm individuals, groups, or society.

REPRODUCIBILITY STATEMENT

The ForeWeave benchmark and code are included in the Supplementary Material. The construction and screening of the historical and live data resources are described in Section 3.1 and Appendix A. Historical process induction, multi-view investigation, targeted continuation, and forecast synthesis are specified in Sections 3.2, 3.3, 3.4, and 3.5, respectively. The experimental settings, temporal controls, evaluation protocol, model configurations, and inference budgets are reported in Section 4.1 and Appendix B; the main comparisons and mechanism analyses appear in Sections 4.2 and 4.3. The core prompt templates used by the main ForeWeave decision stages are provided in Appendix C.

REFERENCES

Jingyi Chai, Bingyang Zheng, Xiangrui Liu, Hao Lu, Zihang Zhou, Tianchen Wang, Kemeng Zhang, and Siheng Chen. Agentic time machine as an infrastructure for future-event forecasting. *CoRR*, abs/2606.21013, 2026. doi: 10.48550/ARXIV.2606.21013. URL https://doi.org/10.48550/arXiv.2606.21013.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Scaling open-ended reasoning to predict the future. *CoRR*, abs/2512.25070, 2025. doi: 10.48550/ARXIV.2512.25070. URL https://doi.org/10.48550/arXiv.2512.25070.

Yurui Chang, Yongkang Du, Yuanpu Cao, Jinghui Chen, and Lu Lin. Forecastcompass: Guiding agentic forecasting with adaptive factor memory. *CoRR*, abs/2605.30858, 2026. doi: 10.48550/ARXIV.2605.30858. URL https://doi.org/10.48550/arXiv.2605.30858.

Nirupam Chetlapalli, Yiming Liao, Min-Chun Chen, and Keke Chen. Diverse by reasoning: Harnessing the wisdom of llm crowds for future prediction, 2026. URL https://arxiv.org/abs/2608.24001.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *CoRR*, abs/2606.19348, 2026. doi: 10.48550/ARXIV.2606.19348. URL https://doi.org/10.48550/arXiv.2606.19348.

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. *CoRR*, abs/2605.15188, 2026. doi: 10.48550/ARXIV.2605.15188. URL https://doi.org/10.48550/arXiv.2605.15188.

Danny Halawi, Fred Zhang, Yueh-Han Chen, and Jacob Steinhardt. Approaching human-level forecasting with language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5a5acfd0876c940d81619c1dc60e7748-Abstract-Conference.html.

Zhixin Han, Yanzhi Zhang, Chuyang Wei, Maohang Gao, Xiawei Yue, Kefei Chen, Yu Zhuang, Haoxiang Guan, Jiyan He, Jian Li, Yitong Duan, Yu Shi, Mengting Hu, and Shuxin Zheng. Futureworld: A live environment for training predictive agents with real-world outcome rewards. *CoRR*, abs/2604.26733, 2026. doi: 10.48550/ARXIV.2604.26733. URL https://doi.org/10.48550/arXiv.2604.26733.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. Forecastqa: A question answering challenge for event forecasting with temporal text data. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 4636–4650. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.357. URL https://doi.org/10.18653/v1/2021.acl-long.357.

Ezra Karger, Houtan Bastani, Yueh-Han Chen, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip Tetlock. Forecastbench: A dynamic benchmark of AI forecasting capabilities. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=lfPkGWXLLf.

Yuante Li, Yicheng Tao, Kate Zhang, Taozhi Wang, Gefei Gu, and Yaxin Zhou. Diverse evidence, better forecasts: Multi-agent deliberation under information asymmetry. *CoRR*, abs/2607.01661, 2026. doi: 10.48550/ARXIV.2607.01661. URL https://doi.org/10.48550/arXiv.2607.01661.

Yiding Ma, Chengyun Ruan, Kaibo Huang, Zhongliang Yang, and Linna Zhou. Oracleproto: A reproducible framework for benchmarking LLM native forecasting via knowledge cutoff and temporal masking. *CoRR*, abs/2605.03762, 2026. doi: 10.48550/ARXIV.2605.03762. URL https://doi.org/10.48550/arXiv.2605.03762.

OpenAI. GPT-5.5 system card. OpenAI, April 2026. URL https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf.

Tavily AI. Tavily api documentation. https://docs.tavily.com/, 2026. Accessed: 2026-09-23.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Benjamin Turtel, Danny Franklin, Kris Skotheim, Luke Hewitt, and Philipp Schoenegger. Outcome-based reinforcement learning to predict the future. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=bbhdeL8EUX.

Benjamin Turtel, Paul Wilczewski, Danny Franklin, and Kris Skothiem. Future-as-label: Scalable supervision from real-world outcomes. *CoRR*, abs/2601.06336, 2026. doi: 10.48550/ARXIV.2601.06336. URL https://doi.org/10.48550/arXiv.2601.06336.

Chuyang Wei, Maohang Gao, Zhixin Han, Kefei Chen, Yu Zhuang, Haoxiang Guan, Yanzhi Zhang, Yilin Cheng, Jiyan He, Huanhuan Chen, Jian Li, Yu Shi, Yitong Duan, and Shuxin Zheng. The world leaks the future: Harness evolution for future prediction agents. *CoRR*, abs/2604.15719, 2026. doi: 10.48550/ARXIV.2604.15719. URL https://doi.org/10.48550/arXiv.2604.15719.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *CoRR*, abs/2504.12516, 2025. doi: 10.48550/ARXIV.2504.12516. URL https://doi.org/10.48550/arXiv.2504.12516.

Jack Wildman, Nikos I. Bosse, Daniel Hnyk, Peter Mühlbacher, Finn Hambly, Jon Evans, Dan Schwarz, and Lawrence Phillips. Bench to the future: A pastcasting benchmark for forecasting agents. *CoRR*, abs/2506.21558, 2025. doi: 10.48550/ARXIV.2506.21558. URL https://doi.org/10.48550/arXiv.2506.21558.

Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. Llm-as-a-prophet: Understanding predictive intelligence with prophet arena. *CoRR*, abs/2510.17638, 2025. doi: 10.48550/ARXIV.2510.17638. URL https://doi.org/10.48550/arXiv.2510.17638.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhenwei Zhu, Tianle Cai, 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 H. Blanchet, Xipeng Qiu, Mengdi Wang, and Wenhao Huang. Futurex: An advanced live benchmark for LLM agents in future prediction. *CoRR*, abs/2508.11987, 2025. doi: 10.48550/ARXIV.2508.11987. URL https://doi.org/10.48550/arXiv.2508.11987.

Linhao Zhong, Zongze Du, Linyu Wu, Yu Bo, Hourong Li, Chenchen Jing, Hao Chen, Yuling Xi, and Chunhua Shen. Foredreamer: A self-evolving dual-agent memory architecture for future event prediction. *CoRR*, abs/2608.20920, 2026. doi: 10.48550/ARXIV.2608.20920. URL https://doi.org/10.48550/arXiv.2608.20920.

Resource	Sources	# Questions
Historical Investigation Set	Markets, News, FutureX-Past; resolved replay	499
ForeWeave-Live	Markets; unresolved at forecast time	68

Table 2: **ForeWeave data resources.** The Historical Investigation Set provides resolved replays for process-structure construction, whereas ForeWeave-Live contains questions that are unresolved at forecast time for prospective evaluation. See Section 3.1.

A Data Collection Details

This appendix describes how we collect the historical questions used for process construction and the unresolved questions used for prospective evaluation. Table 2 summarizes the resulting data resources.

Historical questions. We assemble historical questions from Manifold, Polymarket, and Kalshi, a designated construction subset of FutureX-Past (Zeng et al., 2025), and forecasting questions generated from time-stamped news articles following established pipelines (Goel et al., 2026; Chandak et al., 2025). The FutureX-Past construction subset is disjoint at the question level from the held-out subset used for evaluation. The screening model may use web search and webpage retrieval to check whether each question is self-contained, has a supported answer format and clear resolution criteria, and remains unresolved under evidence available at its forecasting cutoff. We remove invalid, ambiguous, sensitive, and duplicate questions. For news-derived candidates, we additionally remove questions that cannot be interpreted without the source article or whose answers are already explicit in cutoff-compliant evidence. A trajectory-based early-answer audit further excludes a historical question when ordinary retrieval under its forecasting cutoff directly reveals the resolved answer; a correct forecast alone is not grounds for exclusion.

Live questions. We collect ForeWeave-Live prospectively from Kalshi and Polymarket while the target questions remain unresolved. The evaluation reported in this paper contains 68 questions. The screening model may use web search and webpage retrieval to verify the prediction target, answer format, resolution criteria, and unresolved status. We apply the same validity, sensitivity, and duplication checks used for historical questions. Each investigation uses public information available when it is executed without retrospective date masking, and the final forecast is frozen before the outcome becomes known. After resolution, we match the recorded forecast with the confirmed platform outcome using the preserved question identifier; questions without confirmed outcomes remain unscored.

B Implementation and Evaluation Details

B.1 Forecasting and Retrieval Setup

We evaluate all three forecasting backbones on FutureX-Past: Qwen3.7-Max in non-thinking mode, DeepSeek-V4-Pro in thinking mode, and GPT-5.5 with medium reasoning effort. For ForeWeave-Live, we evaluate DeepSeek-V4-Pro and GPT-5.5. Each forecasting investigation uses a temperature of 0.6, a maximum output length of 8,192 tokens, and at most 60 interaction turns.

Within each backbone, Base Agent is run independently three times and reported as the mean of the three runs. Base Agent + SC aggregates the same three forecasts using the answer-type-specific rule described in Section 4.1. Experience-Guided Agent and ForeWeave are each run once. ForeWeave uses the corresponding forecasting backbone for view planning, investigation, targeted continuation, and final forecast synthesis. We set $K_{\mathrm{br}} = 2$ for Qwen3.7-Max and $K_{\mathrm{br}} = 1$ for DeepSeek-V4-Pro and GPT-5.5.

For FutureX-Past, web search and webpage retrieval use Tavily. Each question is replayed under its recorded historical cutoff, which is applied before retrieved evidence becomes model-visible. Our temporal control follows the cutoff-bounded retrospective evaluation principle used by OracleProto and Agentic Time Machine (Ma et al., 2026; Chai et al., 2026), but implements it as a multi-stage pipeline. Before retrieval, a query guard blocks requests that explicitly seek post-cutoff results.

For returned evidence, Qwen3.6-35B-A3B performs answer-blind date-role classification and focused content extraction, distinguishing dates that describe content events from page provenance or later page states. Reliably delimited peripheral post-cutoff modules are removed; evidence with inseparable or uncertain post-cutoff state is discarded. Qwen3.7-Max then performs content-level temporal auditing, with result-level audit failures handled fail closed. These auxiliary models use a temperature of zero, process retrieved observations only, and do not generate forecasts.

For ForeWeave-Live, web search uses Serper and webpage retrieval uses Jina Reader. The agents access public information available when each forecast is executed, without retrospective date masking. Forecast timestamps and retrieved observations are retained for later audit.

The final synthesis stage may perform bounded evidence verification when a blocking evidence gap remains. Verification is limited to six interaction turns, with at most four `read_source` calls, two `verify_search` calls, and two `verify_visit` calls. The compared methods use the same question, answer contract, and retrieval environment within each temporal setting, while their realized tool use and total cost are reported separately.

B.2 Evaluation Protocol

FutureX-Past. Our held-out evaluation set contains 66 questions: 6 at L1 (9.09%), 35 at L2 (53.03%), 6 at L3 (9.09%), and 19 at L4 (28.79%). We replay these questions under their recorded historical information cutoffs and evaluate the resulting forecasts with the pinned official evaluator. Outcome labels and evaluation-only metadata remain unavailable to the forecasting agents and are used only for scoring. Final answers are normalized according to the public answer contract before evaluation. For numeric questions, Qwen3.7-Max selects the submitted numeric literal without access to the outcome label, unit conversion, or additional calculation.

FutureX-Past reports official scores on a 0–100 scale for L1–L4. For each level, the reported score is the mean question score over that level's official denominator. The Overall score is

$$S_{\text{Overall}} = 0.1S_1 + 0.2S_2 + 0.3S_3 + 0.4S_4, \tag{5}$$

where S_l denotes the official score for level l. Overall is therefore a fixed level-weighted score rather than a pooled question average. We additionally report the mean question-level Brier score when the forecast has a valid probability representation and a defined outcome space.

ForeWeave-Live. ForeWeave-Live follows the same answer normalization and question-level evaluation procedure as FutureX-Past. We report an unweighted Overall score across all questions and the mean Brier score where a complete probability distribution over a defined outcome space is available.

B.3 Probabilistic Forecasting Results

Table 3 reports Brier scores for the forecasting workflows compared in the main results. Brier score evaluates probabilistic prediction error, with lower values indicating better performance, and should not be interpreted as a measure of calibration alone. The reported values follow the probability-forecast eligibility criteria described above. ForeWeave's higher task scores do not consistently translate into lower Brier scores: on FutureX-Past, its Overall Brier score is higher than that of Base Agent for all three backbones. The task-score improvements therefore do not establish improved probabilistic prediction quality.

C Core Prompt Templates

We include only the prompts that define ForeWeave's main decision stages: view planning, view-specific investigation, targeted continuation, and forecast synthesis. The common forecasting-agent instructions, tool specifications, output schemas, and data-construction prompts are omitted from the paper for space and remain part of the versioned experiment artifact. The listings below preserve the selected prompt templates verbatim; placeholders such as {{`PLAN`}} and {{`DIRECTION`}} are populated at runtime.

Method	FutureX-Past L1 Score↑	L1 Brier↓	L2 Score↑	L2 Brier↓	L3 Score↑	L3 Brier↓	L4 Score↑	L4 Brier↓	Overall Score↑	Overall Brier↓	ForeWeave-Live Overall Score↑	ForeWeave-Live Overall Brier↓
Qwen w/ Tools												
Base Agent	38.88	0.83	**63.80**	0.37	46.83	0.26	32.53	0.42	43.71	0.40	–	–
Base Agent + SC	**50.00**	0.85	62.86	0.32	49.75	0.33	36.43	0.44	47.07	0.42	–	–
Experience-Guided Agent	**50.00**	0.95	62.86	0.41	45.32	0.23	**37.81**	0.44	46.29	0.39	–	–
ForeWeave	16.67	0.90	60.00	0.40	**68.68**	0.50	37.66	0.48	**49.33**	0.51	–	–
DeepSeek w/ Tools												
Base Agent	38.89	0.76	64.76	0.33	64.31	0.19	39.29	0.46	51.85	0.38	62.25	0.40
Base Agent + SC	**50.00**	0.82	65.71	0.33	58.46	0.14	37.05	0.35	50.50	0.33	64.71	0.38
Experience-Guided Agent	33.33	0.57	62.86	0.36	**72.57**	0.11	40.83	0.36	54.01	0.30	64.71	0.40
ForeWeave	**50.00**	0.65	**71.43**	0.31	62.75	0.56	**46.37**	0.48	**56.66**	0.49	**70.59**	0.39
GPT w/ Tools												
Base Agent	50.00	0.61	61.90	0.32	65.21	0.63	36.68	0.61	51.62	0.56	65.69	0.39
Base Agent + SC	50.00	0.61	62.86	0.31	**66.44**	0.70	38.12	0.64	52.75	0.59	64.71	0.38
Experience-Guided Agent	33.33	0.71	**71.43**	0.32	65.69	0.62	39.34	0.62	53.06	0.57	64.71	0.37
ForeWeave	**66.67**	0.52	**71.43**	0.32	65.77	0.70	**47.05**	0.70	**59.51**	0.59	**67.65**	0.41

Table 3: **Forecasting results on FutureX-Past and ForeWeave-Live.** FutureX-Past reports official 0–100 scores for L1–L4 and a fixed weighted Overall score. Brier scores are computed separately for eligible probabilistic forecasts.

C.1 Process-Guided View Planning

The planner receives the forecasting question, prediction cutoff, answer contract, and rendered historical Process DSL. It returns separate plans for Current State, Temporal Dynamics, and Disconfirmation without prescribing an answer or probability.

View Planning

Create three complementary, self−contained research plans for the supplied forecasting question. Use the historical process DSL as an example of how to investigate a question. Each plan must enable one worker to research the whole question and produce its own calibrated forecast.

Identify the entities, dates, quantities, answer options, and explicit conditions in the supplied question that determine the outcome. Treat prediction_as_of as the current time for research. Do not invent missing resolution rules or assume that an outcome is already known.

Select applicable actions, dependencies, decisions, and stopping conditions from the example graph and translate them into concrete checks for the target question. Historical facts are not current evidence. Distinguish conditions to verify from facts established for the target , and future turning points from events that have already occurred. Omit graph steps that do not fit; do not force a complex path onto a simple question.

Assign the research priorities as follows:

– current_state: Establish the factual baseline at the cutoff. Prioritize direct facts, official information, recent signals, and relevant base rates. Specify which observations would establish the present state and how to distinguish target evidence from background, analogy, and repeated reporting.
– dynamics: Investigate what must happen between the cutoff and the target outcome. Prioritize necessary conditions, dependencies, timing, bottlenecks, and alternative paths. Specify observable checks that distinguish feasible paths from unsupported scenarios.

Verify the starting conditions needed for this analysis without repeating the current−state plan wholesale.
− disconfirmation: Investigate the assumptions most capable of overturning an initial judgment . Direct the worker to establish its own tentative leading answer from evidence, then seek strong counterevidence and compare credible alternatives. Specify discriminating checks rather than requiring a contrarian conclusion.

Write each plan as concise, actionable instructions tailored to the question. Identify the decisive uncertainties, what to investigate first, suitable source types or targeted search directions, and which findings would change the assessment. Prefer distinct evidence angles over repeated versions of the same query. Require evidence available at or before the cutoff and checks for relevance, specificity, independence, and reliability; avoid relying on resolved outcomes or later updates.

Each worker must consider plausible alternatives and test its own judgment, regardless of its primary view. Direct it to stop when further research is unlikely to materially change the forecast, explain the key evidence and uncertainty briefly, and calibrate its probability to evidence strength, dependence, gaps, and unresolved conflicts using the caller's final prediction format. Do not prescribe an answer or probability in the plan.

Before returning, check that the plans differ in their substantive investigations rather than only their labels. Overlap needed to verify essential facts is acceptable; manufactured disagreement is not. Every plan must stand alone because its worker receives only the target question and its assigned plan, not the example graph, this conversation, or another worker's results. Write out needed checks instead of referring to graph node IDs.

Return three plan blocks in this order: <plan_current_state>, <plan_dynamics>, and <plan_disconfirmation>, each with its matching closing tag. Within each plan, state the research objective, give numbered investigation steps, explain evidence checks and alternatives, and specify stopping and forecasting instructions. Choose the number of steps to fit the question. Return only the three plan blocks without JSON, code fences, or surrounding commentary.

C.2 View-Specific Investigation

Each worker receives the common forecasting context together with one assigned plan and one of the following view instructions. Workers investigate the complete question independently and do not observe the other workers' answers or trajectories.

Current State

Current State

<current_state_view_rollout>
Produce an independent forecast from the current−state view. Use the assigned plan to identify applicable research checks. It is procedural guidance, not evidence about the answer.

{{PLAN}}

Focus on direct current facts, recent trends, current signals, and applicable base rates. Identify target−specific anchors already established by evidence gathered for the current question. Distinguish direct evidence from background context, analogy, and repeated reports of the same underlying fact.

Consider plausible alternatives sufficiently to avoid a one−sided conclusion, but keep the analysis centered on the current state rather than attempting to exhaust every possible

future path. Do not read, infer, or anticipate another view's answer. Complete your own evidence gathering and calibrated judgment.

Complete your independent forecast, give a concise explanation of its key evidence and uncertainty, and finish using the caller's existing final answer and probability contract. Evidence−record preparation happens after this rollout.
</current_state_view_rollout>

TEMPORAL DYNAMICS

Temporal Dynamics

<dynamics_view_rollout>
Produce an independent complete forecast using the assigned plan as procedural guidance, not as evidence.

{{PLAN}}

Verify the starting conditions and investigate which changes can affect the target through dependencies, timing, bottlenecks, or alternative paths. Necessary conditions do not guarantee success. If the outcome period has already ended, use new information to update the estimate rather than describing a change to the completed result. Ground every path in observable cutoff−compliant evidence and test a consequential alternative.

Do not read or anticipate another worker's work. Finish with the caller's existing final answer and probability format.
</dynamics_view_rollout>

DISCONFIRMATION

Disconfirmation

<disconfirmation_view_rollout>
Produce an independent forecast from the disconfirmation−and−alternatives view. Use the assigned plan as procedural guidance for investigating the current question. It is not evidence about the answer.

{{PLAN}}

Identify a tentative leading answer and the strongest plausible alternatives. Actively seek the strongest evidence that could make the tentative leader wrong. Compare the alternatives' supporting evidence, required conditions, obstacles, decision branches, and turning points . Count repeated reports of the same underlying fact as one evidence family.

Do not choose a contrarian answer merely because this is the disconfirmation view. After testing the leader and alternatives, select the answer best supported by the available evidence and calibrate its probability to unresolved conflict and uncertainty.

Do not read, infer, or anticipate another view's answer. Complete your independent forecast, give a concise explanation of its key evidence and uncertainty, and finish using the caller' s existing final answer and probability contract. Evidence−record preparation happens after this rollout.
</disconfirmation_view_rollout>

C.3 Evidence-Driven Targeted Continuation

Follow-up Prioritizer

The selector receives the public question, information cutoff, frozen aggregate forecast, and three initial evidence records. It proposes an additional investigation axis and decides whether a cited issue warrants targeted continuation. The additional-axis instructions support the separate independent-rollout comparison; the branch decision identifies the parent record, source reference, and issue to investigate.

Follow-up Prioritizer

Choose one question−specific follow−up axis for a frozen three−view forecast. You receive the public question, fixed cutoff, the frozen parent forecast, and three anonymous G1 evidence records. Do not make a forecast in this step.

The existing views already covered current state, dynamics, and disconfirmation. Select a missing axis that is likely to separate the plausible answers without repeating those three labels. Prefer exact target definitions, reporting boundary, metric/unit choice, actor incentive, schedule/eligibility constraint, dependency chain, or source reliability when that is the unresolved bottleneck. The axis must describe what a new independent rollout should investigate from the same historical process guide.

Also decide whether one targeted branch from an existing G1 view is worth running. Branch only when a cited observation, assumption, or missing check in one G1 record could plausibly change the final answer or probability, and the check can be investigated with pre−cutoff evidence. If no such cited point exists, set should_branch to false. Do not invent a source, node ID, post−cutoff fact, or result of the proposed check.

Return only one JSON object matching the schema. ‘new_view_axis‘ is a short label. ‘ new_view_instructions‘ tells the new rollout exactly what to check and what repeated work to avoid. ‘branch‘ contains the branch decision. For a branch, ‘parent_rid‘ must be one of ‘r1‘, ‘r2‘, or ‘r3‘, and ‘source_ref‘ must be a cited source reference from that parent record. For no branch, use null for parent_rid and source_ref and empty strings for the text fields.

Targeted Continuation

The continuation starts from a complete tool-result boundary preceding the cited observation in the parent trajectory.

Targeted Continuation

Continue the inherited investigation from its complete tool−result boundary. The later diagnostic issue is a question to test, not an observation from the inherited prefix or a correct answer. Keep the original cutoff and answer space. Check the assigned claim against the competing answer or boundary, using existing visible observations first and new searches only for a different pre−cutoff discriminator. Do not repeat the entire earlier investigation or treat failure to find a source as support for the parent.

{{DIRECTION}}

Make your own final forecast and probability after the check. Preserve uncertainty when the requested discriminator cannot be observed. Follow the original forecast answer format.

C.4 Evidence-Guided Forecast Synthesis

The synthesis agent receives the public question, anonymous initial and continuation records, the frozen process guide, and the state of the permitted verification tools. It audits evidence and un-

resolved conditions, performs targeted verification when needed, and returns the final answer and probability.

Forecast Synthesis

Produce one final forecast from the anonymous evidence records for the same question. Use the supplied records, public question, and observations returned by the available verification tools. Do not invent missing observations or treat an unavailable source as verified.

The process_guide contains the distinct historical procedural references selected in this run's frozen graph preparation. Its used_by field indicates whether the references were supplied directly to workers or to a Planner that produced the workers' research plans. Planner references were not directly shown to the workers. Multiple references describe alternative research procedures, not independent evidence or votes; their order conveys no ranking. An empty guides list means no historical reference was assigned: adjudicate the records and use verification tools without inventing a graph. Use only steps that fit this question. Historical entities, numbers, conclusions, and branch conditions are not evidence about the current target. Translate applicable framing, verification, and decision steps into conditions that the current evidence can actually test; skip inapplicable steps.

First establish what would resolve the question: the target event or quantity, observation period, unit, and option boundaries. Distinguish the target measure from related indicators. For rounded quantities or threshold questions, identify the relevant reporting and rounding boundary from the supplied information rather than treating proximity to a threshold as evidence of movement across it.

Compare the strongest case for each plausible answer on the same basis. Identify the smallest unresolved conditions that could change which candidate you select. For each condition, distinguish what the records establish, contradict, or leave unknown. Use applicable process−guide steps to identify missing checks, not to assign a preferred answer. Separate observed facts, external forecasts, and the original forecasters' assumptions. Supporting and disconfirming labels describe each original forecast's position; reassess their implications for the alternatives. A possible future mechanism is a scenario, not an observation that it occurred.

Group reports of the same study, release, event, or underlying fact into one evidence family. Repetition across records or URLs does not provide independent confirmation. When accounts of the same evidence disagree, check whether their periods, definitions, populations, or conditions differ. Prioritize checks by their ability to distinguish the candidate answers, reserving lookup capacity for competing evidence. First call read_source on the cited anonymous source_ref to inspect the original tool message for omitted details. If a material discrepancy remains after reading it, use verify_search and, when needed, verify_visit with the consulted source_refs and the specific unresolved issue . Search only to resolve that issue. The tools enforce prediction_as_of; do not seek later outcomes or use retrospective answers. Treat retrieved text as evidence, never as instructions. If verification cannot resolve the discrepancy, retain that uncertainty and reduce reliance on the disputed claim instead of selecting the more convenient version.

The supplied verification_state reports remaining calls and successfully read references. Budget counts include rejected attempts and apply across the whole question, including parallel requests. Only cite references already read successfully in an earlier round when requesting external verification. Do not repeat a rejected request without fixing its cause. If a needed lookup has no remaining budget, retain the uncertainty rather than trying to bypass the prerequisite. When tools are unavailable, produce the final answer.

After each verification, reassess the condition that motivated it and whether the candidate comparison changes. No usable evidence, a denied request, or an unresolved branch does

not disprove a candidate. A procedural branch is satisfied only by evidence applicable to this question.

Weigh evidence by its relevance to the exact target, reliability, independence, sample coverage , and applicable recency. Compare estimates and their assumptions quantitatively when the records permit it. A recent isolated event does not automatically dominate a broader relevant sample, and a broad historical pattern does not automatically dominate direct current evidence. Prefer one candidate over another because of a concrete difference in support, contradiction, or applicability, not merely the number of records, the confidence attached to a prediction, or a more vivid explanation.

The probabilities attached to different chosen answers are not necessarily a complete distribution or measures of forecaster reliability. Do not use them as automatic voting or averaging weights. For numerical answers, justify any new estimate by an explicit calculation or inference supported by the records. Do not introduce extra precision merely to split the difference between candidates. When the evidence cannot distinguish nearby values, reflect that uncertainty in confidence rather than presenting an unsupported adjustment as new information.

Commit to the answer best supported by the combined evidence and calibrate its probability for unresolved conflicts, dependent evidence, and credible alternatives. Do not automatically preserve or overturn a majority answer.

When verification is complete or the tool budget is exhausted, return only the final answer and probability. For a single answer, output exactly one <answer>complete answer</answer> tag followed by one <probability>number between 0 and 1</probability> tag. For fixed options, use the complete selected option exactly as supplied. If the supplied question instead requires multiple fixed answers, use the corresponding numbered <answerN> and <probabilityN> tag pairs and follow its probability constraints. Do not output prose, bare probabilities, record summaries, or aggregation diagnostics.

D Extended Related Work

D.1 Forecasting Benchmarks

Event forecasting requires estimating the likelihood of future outcomes using only information available at a specified prediction cutoff (Karger et al., 2025; Zeng et al., 2025). Whereas fact-finding benchmarks for research agents, such as BrowseComp, require recovering existing but hard-to-find answers (Wei et al., 2025), forecasting concerns outcomes that remain unresolved at prediction time. Beyond finding relevant evidence, forecasters must weigh incomplete and evolving signals and quantify uncertainty that further retrieval may not eliminate (Zeng et al., 2025).

ForecastQA formulates event forecasting as question answering with temporally restricted news evidence (Jin et al., 2021). Live benchmarks evaluate predictions before outcomes are known: ForecastBench uses regularly updated questions and human comparison groups (Karger et al., 2025), and FutureX automates daily question collection and resolution tracking for forecasting agents (Zeng et al., 2025). Prophet Arena further supports controlled analysis of forecasting stages on continuously collected live tasks (Yang et al., 2025). For retrospective evaluation, OracleProto combines model-cutoff-based sample admission with temporal masking and leakage detection (Ma et al., 2026), while FutureSim and Bench to the Future provide date-gated news and offline web corpora, respectively (Goel et al., 2026; Wildman et al., 2025).

D.2 Forecasting Agents

Forecasting agents seek relevant evidence, synthesize its implications, and produce probabilistic estimates. Retrieval-augmented forecasting systems implement this process through news retrieval, relevance filtering, summarization, and forecast generation, then aggregate predictions elicited with

different prompts and models (Halawi et al., 2024). ForeDreamer separates search and prediction from memory processing, using a dedicated subagent to transform retrieved web evidence into structured, question-specific factual memory (Zhong et al., 2026). For collective forecasting, InfoDelphi partitions evidence into shared and agent-specific subsets, then combines iterative deliberation with confidence-weighted aggregation (Li et al., 2026). Diverse by Reasoning instead constructs forecasting ensembles by clustering models according to reasoning behavior on independent development tasks and selecting representative members (Chetlapalli et al., 2026). The planner–solver–aggregator framework introduced alongside Agentic Time Machine uses exploratory search to define analytical angles, which parallel solvers investigate to produce structured reports (Chai et al., 2026). Its aggregator reconciles agreement and conflict across these reports and can perform additional targeted searches before producing the final forecast.

Recent work also adapts how agents investigate and interpret evidence. Milkyway revisits unresolved questions over time and uses changes in evidence and forecasts to revise procedural guidance for subsequent predictions (Wei et al., 2026). ForecastCompass retrieves predictive-factor and reasoning memories to guide evidence search, factor selection, and probability estimation, updating these memories through retrospective analysis of resolved tasks (Chang et al., 2026). Complementary training approaches use synthesized forecasting questions (Chandak et al., 2025) or supervision from realized outcomes (Han et al., 2026; Turtel et al., 2025; 2026).