From Hindsight to Foresight: Adapting Reusable Reasoning Structures from Recurring Financial Events

Sangjin Jin*
sj.jin@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Junhyeong Lee*
jun.lee@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Daehee Han*
day.han@lab.kakaobank.com
KakaoBank
Seongnam, Republic of Korea

Kangmin Kim
kmin1015@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Suhwan Park
suhwan@unist.ac.kr
UNIST
Ulsan, Republic of Korea

Seonghun Jeong
bentley.j@lab.kakaobank.com
KakaoBank
Seongnam, Republic of Korea

Daekyung Park
daekyungpark@linqalpha.com
LinqAlpha
New York, United States

Lukasz Szpruch
L.Szpruch@ed.ac.uk
University of Edinburgh
Edinburgh, United Kingdom

Zhangyang Wang
atlaswang@utexas.edu
The University of Texas at Austin
Austin, United States

Chanyeol Choi
jacobchoi@linqalpha.com
LinqAlpha
New York, United States

Youngjun Kwak†
vivaan.yjkwak@lab.kakaobank.com
KakaoBank
Seongnam, Republic of Korea

Yongjae Lee†
yongjaelee@unist.ac.kr
UNIST
Ulsan, Republic of Korea
LinqAlpha
New York, United States

Abstract

Forecasting real-world events requires reasoning under uncertainty from information available at a fixed cutoff. Recent work has improved how large language models (LLMs) retrieve evidence and produce probabilistic forecasts, but each question is treated independently even when similar events recur. Resolved occurrences carry experience for later forecasts, yet past outcomes and case summaries do not preserve the relationships that made an earlier occurrence informative. We develop ReFoRM (Recurrent-event Forecasting with Reasoning Memory), which organizes reasoning from resolved events into predictive paths, retrieves relevant paths for a later occurrence, and reconstructs their relations using evidence at the new forecast cutoff. We evaluate ReFoRM on the Recurring Financial Events Set, which contains 300 questions across 40 recurring financial event families and uses earlier occurrences as hindsight for 100 forecast targets. Across five LLMs and six baselines, ReFoRM improves categorical and probabilistic forecasting performance. Further analyses show that validating historical relations with current evidence and reconstructing them enhances forecasting. Reasoning evaluation also demonstrates stronger evidence grounding, logical validity, and mechanism specificity.

CCS Concepts

• **Computing methodologies** → **Natural language processing**;
• **Information systems** → *Information retrieval.*

Keywords

event forecasting, financial forecasting, large language model

1 Introduction

Financial forecasting is central to decision-making in investment, risk management, and economic policy [10]. Large language models (LLMs) increasingly support financial analysis by interpreting text, integrating evidence, and reasoning over quantitative information [29, 33]. Recent benchmarks evaluate these capabilities on knowledge-intensive tasks [19, 36], while forecasting benchmarks require predictions from temporally ordered evidence [17, 38]. Models must therefore separate predictive signals from background information and form probabilistic judgments before resolution.

Many financial questions also recur. Firms report quarterly results, statistical agencies publish the same indicators, and central banks revisit policy decisions on regular schedules. Although evidence and conditions change, each occurrence adds to a history of comparable forecasting problems. Research on financial analysts suggests that task-specific experience and industry knowledge improve subsequent forecasts [2, 6]. Forecasts for later occurrences can therefore draw on knowledge accumulated earlier.

These histories contain more than outcomes. Repeated occurrences reveal which factors matter, how they interact, and through which economic channels they affect the target. Information propagates through links between firms, while macroeconomic and monetary-policy announcements are interpreted through recurring surprise and policy channels [1, 7, 14]. These regularities can structure later forecasts as relationships and conditions change.

*These authors contributed equally to this work.
†Co-corresponding authors.

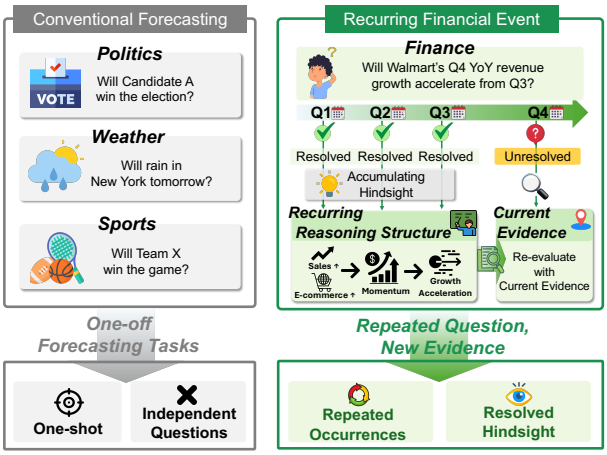


Figure 1: Motivation from resolved events. The same outcome can arise from different reasoning paths. Relationships from resolved recurring events help distinguish supported from unsupported reasoning in new instances.

Figure 1 illustrates this intuition. Earlier instances of a recurring earnings question reveal relationships among economic conditions, firm-specific drivers, and the outcome. A later instance presents new evidence, but these relationships guide relevant comparisons and the interpretation of competing signals. Prior work likewise shows that structured comparisons with historical cases and similar past economic states can inform later forecasts [9, 16]. Resolved cases thus provide structured hindsight for later forecasts.

LLM-based forecasting has focused on retrieving current evidence, decomposing questions, and aggregating predictions [15, 31]. These advances help analyze unresolved events, but rarely carry structured reasoning from resolved occurrences into later forecasts. Pretraining and retrieval can make historical information accessible, but do not by themselves identify which factors and relationships should guide the next forecast. The open challenge is to reuse reasoning from resolved occurrences while re-evaluating the underlying relationships using evidence available for the current event.

We introduce ReFoRM (Recurrent-event Forecasting with Reasoning Memory), which turns hindsight from resolved occurrences into forecasting structures. It constructs hindsight graphs from resolved occurrences [5], retrieves predictive paths relevant to a new occurrence, and reconstructs their relations under current evidence. Resolved events can therefore guide a new forecast without treating their outcomes or premises as facts about the target occurrence.

We evaluate ReFoRM on the *Recurring Financial Events Set*, which pairs resolved occurrences with later forecast targets from the same financial event families. Across five LLMs and six baselines, ReFoRM improves categorical and probabilistic forecasting performance. Additional analyses evaluate its reasoning quality and examine when relation reconstruction leads to forecasting gains.

Our contributions are summarized as follows.

- We formulate how resolved hindsight informs recurring financial forecasts and introduce ReFoRM to retrieve paths and reconstruct their relations under current evidence.
- We introduce the Recurring Financial Events Set, which pairs resolved and target occurrences for controlled evaluation across recurring financial event families.
- Across five LLMs, ReFoRM outperforms six forecasting baselines. Further analyses show stronger reasoning and explain when its forecasting gains arise.

2 Related Work

Language models for finance. Financial language models have been developed for domain-specific text understanding, sentiment analysis, and decision support [33]. FinBen evaluates models across a broad collection of financial tasks, while BizBench and XFinBench emphasize quantitative and knowledge-intensive reasoning [19, 29, 36]. Their results show substantial progress on document understanding but persistent weaknesses on complex reasoning and forecasting. Macroeconomic forecasting has also examined whether LLMs can complement conventional time-series models [3]. We focus on probabilistic forecasting for recurring financial events, where resolved occurrences provide structured hindsight for subsequent forecasts within the same event family.

Language models for forecasting. ForecastQA introduced forecasting questions paired with temporally ordered news [17]. AutoCast paired tournament questions with dated news, and AutoCast++ improved evidence retrieval [31, 38]. Search, decomposition, and forecast aggregation can bring LLMs closer to human forecasting performance [15]. FOReCAst broadens this evaluation with Boolean, timeframe, and quantity questions that assess both prediction accuracy and confidence calibration [35]. ForecastBench, Prophet Arena, and FutureX extend evaluation to live or continually refreshed questions [18, 21, 34]. These works improve how an unresolved event is researched and scored. ReFoRM instead uses resolved occurrences as structured hindsight for later forecasts and reconstructs their relations using evidence available for the current event.

Reasoning from resolved events. Prior cases can inform subsequent decisions. Studies of financial analysts link task-specific experience and industry knowledge to improved forecasts [2, 6]. Case-based decision support and similarity-based macroeconomic forecasting also support structured historical comparisons [9, 16]. Agent-memory methods distill reusable knowledge from prior trajectories [24, 37]. ForecastCompass updates memories of predictive factors and probability calibration after resolution [4]. A complementary line of work evaluates reasoning beyond endpoint accuracy. ReCEval assesses the correctness and informativeness of reasoning steps [26]. Related evaluation frameworks assess whether generated claims are supported by cited sources [12] and whether stated reasoning is faithful to model decisions [25]. WorldReasoner adapts these criteria to event forecasting through separate outcome, source, and event-graph measures [5]. Taken together, these strands leave the relational structure of resolved cases underexplored as guidance for later forecasts. ReFoRM addresses this gap by organizing resolved reasoning into predictive paths and reconstructing their relations using evidence available for a new occurrence.

Table 1: Coverage of the 40 recurring financial event families.

Category	Event families and prediction target
Per category: 8 families and 60 questions 4 families recur 7 times and 4 families recur 8 times	
Corporate earnings	Apple, Amazon, Alphabet, Meta, Microsoft, NVIDIA, Tesla, and Walmart. *Target:* Quarterly YoY revenue-growth acceleration (**SEC**).
Energy & commodities	Brent and WTI crude, Henry Hub gas, U.S. gasoline and diesel, aluminum, copper, and wheat. *Target:* Monthly price return or change (**FRED**).
Macroeconomics	CPI, core CPI, PCE prices, payrolls, unemployment, retail sales, industrial production, and housing starts. *Target:* Monthly change or growth (**FRED**).
Markets & credit	S&P 500, NASDAQ, VIX, U.S. dollar, Treasury yield, IG and HY spreads, and financial stress. *Target:* Monthly return or change (**FRED**).
Monetary policy	Federal funds rate, Treasury and real yields, breakeven inflation, Federal Reserve assets, and M2. *Target:* Monthly rate change or growth (**FRED**).

3 Recurring Financial Events

Financial forecasting often concerns events that recur on public schedules, including earnings releases and macroeconomic announcements. We group comparable occurrences by fixed targets, units, and resolution rules, then order each family chronologically. Earlier resolved occurrences can therefore inform later targets.

3.1 Event Families and Coverage

We organize 40 recurring event families across corporate earnings, energy and commodities, macroeconomic indicators, markets and credit, and monetary policy. Each category contains eight issuers or series. An event family fixes the entity or series, target quantity, unit, comparison rule, and resolution source across dated occurrences. The forecast cutoff, target period, available evidence, numerical boundaries, and outcome vary over time. Each family recurs seven or eight times, yielding 300 questions and 60 per category. Table 1 lists the covered issuers and series with their target quantities and resolution sources for each category.

Table 2: Resolved examples of a three-class and a binary question. Target questions use the same mappings with the outcome withheld during forecasting.

Corporate earnings: Walmart, three classes	
Question	As of 2023-01-31, which range will Walmart's quarterly year-over-year revenue-growth acceleration a_τ for the fiscal quarter ending 2023-01-31 fall into?
Classes	*below* if $a_\tau < -0.496$; *within* if $-0.496 \le a_\tau < 2.930$; *above* if $a_\tau \ge 2.930$, in percentage points
Outcome	$a_\tau = -1.405 \rightarrow$ **below recent range**
Energy and commodities: U.S. diesel, binary	
Question	As of 2023-01-01, will the monthly retail diesel-price change for 2023-01 be at least -3.864%?
Classes	*yes* if the monthly change is at least -3.864%; *no* otherwise
Outcome	observed change $1.873\% \rightarrow$ **yes**

3.2 Forecasting Question Formulation

Questions are generated from immutable snapshots of 32 Federal Reserve Economic Data (FRED) series and SEC Company Facts for eight issuers, each snapshot recorded with its URL, retrieval time, and content hash [11, 22, 28]. Once the event families are fixed, question generation involves no manual selection or difficulty rating. The resulting questions therefore reflect the underlying series, and the recorded snapshots make their construction reproducible.

Class boundaries use only observations available before the cutoff. For each FRED series, the target is a percentage or absolute change between consecutive observations $u_{\tau-1}$ and u_τ. Three-class questions compare it with the first and third quartiles of the preceding eight changes, while binary questions use the median of the preceding four. For corporate families, let g_τ denote quarterly year-over-year revenue growth and $a_\tau = g_\tau - g_{\tau-1}$ its acceleration. Three-class questions apply the same quartile rule to the preceding eight acceleration values. Binary questions ask whether $a_\tau \ge 0$, equivalently $g_\tau \ge g_{\tau-1}$. The two formats distinguish whether growth accelerates from how unusual the acceleration is. This yields 190 three-class and 110 binary questions. Questions state the forecast cutoff, target period, unit, answer set, and resolution rule, while withholding the outcome value, label, and resolution explanation. Table 2 shows a resolved example of each type.

3.3 Chronological Split

The split is chronological within each family. The first five occurrences provide reference questions and the remaining two or three become forecast targets, yielding 200 reference questions and 100 targets. Reference cutoffs fall between January 2023 and August 2025, while target cutoffs fall between March 2025 and June 2026. Each target therefore has five eligible resolved cases from its own family, and questions from the same release remain on the same side of the split. Outcomes resolve at the first corresponding SEC filing or scheduled release for the target period. After each target resolves, we construct its reference DAG from post-resolution documents. It records supported factor-target relationships and is used only for analysis, never forecasting.

4 Forecasting with Recurring Structures

ReFoRM (Recurrent-event Forecasting with Reasoning Memory) adapts structures from resolved recurring events in four stages. It constructs hindsight graphs, uses their factors to retrieve current evidence, selects and reconstructs relevant paths under that evidence, and produces a probabilistic forecast. Figure 2 summarizes the full process from resolved occurrence to forecast.

4.1 Hindsight Graph Construction

Consider a target question $q_i = (x_i, \kappa_i)$, where x_i describes the event and κ_i specifies the metric, period, unit, options, forecast cutoff t_i, and resolution rule. Given evidence available before t_i, the task is to return a distribution p_i over $\mathcal{Y}_i$ without observing the realized label y_i at forecast time.

For each earlier occurrence q_j in the same family that resolved before t_i, we construct a hindsight graph from documents available by its resolution

$$G_j = (V_j, A_j, S_j). \tag{1}$$

The graphs in Equation 1 form $\mathcal{H}_i$. Nodes V_j contain factors and events, edges A_j record predictive relationships, and S_j links claims

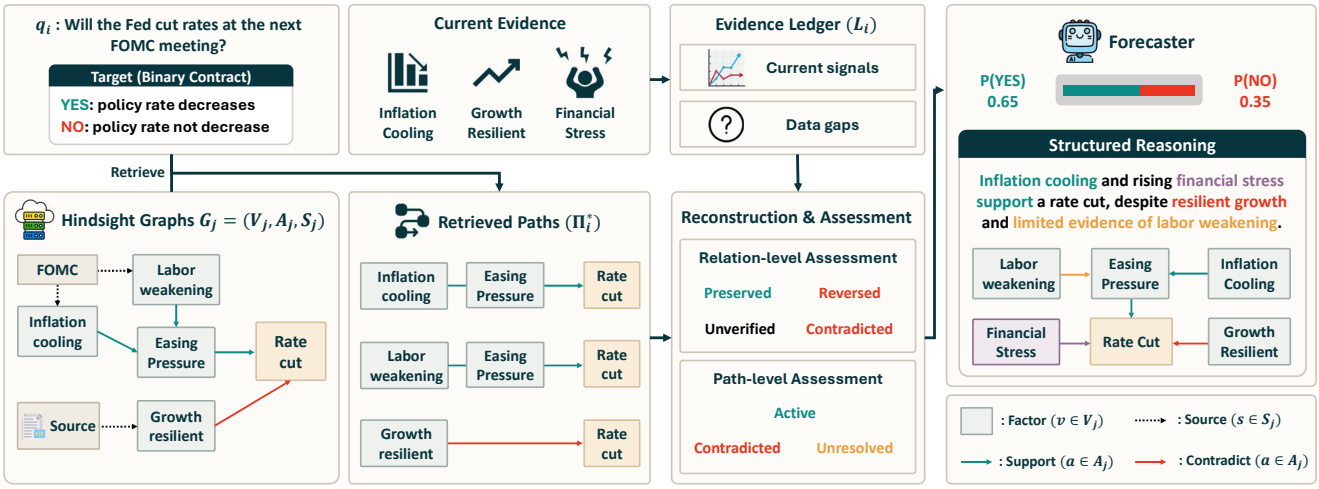


Figure 2: Overview of ReFoRM. A resolved occurrence is distilled into reusable factors, conditional paths, and alternatives after removing outcome-specific content. For a later occurrence in the same family, this structure guides relation testing, pre-cutoff evidence determines which relations remain supported, and the resulting assessment is mapped to the new answer boundaries.

to sources. Following [5], construction targets 8 events and a minimum depth of 3 while enforcing sourced claims, temporal order, acyclicity, and connectivity to the resolved target.

Each graph is reduced to $B_j = (\Pi_j, C_j)$. The set Π_j contains complete root-to-target paths, while C_j records their evidence requirements and the conditions under which they may apply or fail. The paths keep their factors, topology, relation semantics, and historical support. Resolved answers, realized values, and historical probabilities are removed. An answer-free review sequence W_j records checks for baselines, intermediate relations, competing explanations, and uncertainty. It is consulted after reconstruction only to check that the current reasoning is complete.

4.2 Retrieving Current Evidence

Factors in the hindsight graphs guide the search for evidence available before the forecast cutoff. Historical paths, relations, answers, and probabilities remain hidden during this search, allowing the factors to guide retrieval without exposing a historical claim or target implication. The retrieved documents form the evidence set $\mathcal{E}_i^{<t_i}$ and are organized into a current evidence ledger. The ledger gives an initial account of the target event using only the question and retrieved evidence and is defined as

$$L_i = \ell_\theta(q_i, \mathcal{E}_i^{<t_i}). \tag{2}$$

In Equation 2, ℓ_θ denotes ledger construction. The ledger L_i records the target operation, usable baseline, observed signals and their timing, supporting sources, and missing information. It remains separate from the subsequent path retrieval.

4.3 Path Retrieval and Reconstruction

Path retrieval compares the factors and target semantics in each eligible structure with the current ledger. Complete paths are selected intact so their relations remain connected to the surrounding factors and target. The selected paths and their corresponding review sequences are expressed as

$$(\Pi_i^*, W_i^*) = R\big(L_i, \{(B_j, W_j) \mid G_j \in \mathcal{H}_i\}\big). \tag{3}$$

In Equation 3, R denotes retrieval, Π_i^* is the selected set of complete paths, and W_i^* contains the associated review sequences. Every selected path keeps its topology and conditions, so retrieval proposes structures to examine rather than premises to accept.

The selected paths are then assessed with current evidence. Their nodes are instantiated and each relation is evaluated without assuming that the complete historical path remains valid. The resulting current structure is written as

$$\widehat{\Pi}_i = I_\theta(\Pi_i^*, L_i, \mathcal{E}_i^{<t_i}). \tag{4}$$

In Equation 4, I_θ jointly assesses nodes, relations, and paths, while $\widehat{\Pi}_i$ denotes the selected paths with their resulting states. Nodes record their current values and evidence. A relation is preserved when the same effect is supported, reversed when the opposite effect is supported, contradicted when the original relation fails without establishing its opposite, and unverified when evidence is insufficient. Paths are active, contradicted, or unresolved according to these joint assessments rather than a fixed aggregation rule.

Reconstruction may remain partial when evidence does not establish every relation. Evidence-supported segments may still organize the argument, while current evidence always takes precedence over historical structure. The review sequences W_i^* are consulted only after reconstruction to identify missing baselines, intermediate relations, competing explanations, and unresolved uncertainty.

4.4 Forecasting with the Current Structure

After reconstruction, the target operation, evidence ledger, reconstructed structure, and review sequences produce a reasoning trace before an option is selected or an estimate or probability is assigned. Evidence-backed current factors absent from retrieved paths may

Table 3: Forecasting results on 100 recurring financial event questions, as mean ± standard deviation over three runs. Best and second-best per model and metric are bolded and underlined.

	Gemini 2.5 Flash Lite			GPT-5 mini			DeepSeek V3.2		
	Acc ↑	Brier ↓	NLL ↓	Acc ↑	Brier ↓	NLL ↓	Acc ↑	Brier ↓	NLL ↓
No-memory baselines									
Direct Forecast	0.523 ±0.021	0.235 ±0.008	1.040 ±0.034	0.490 ±0.010	0.229 ±0.008	0.976 ±0.040	0.517 ±0.025	0.229 ±0.009	0.981 ±0.041
Structured Reasoning	0.387 ±0.031	0.285 ±0.015	1.210 ±0.055	0.477 ±0.051	0.237 ±0.013	1.006 ±0.043	0.463 ±0.035	0.238 ±0.007	0.999 ±0.022
Resolved-event memory									
Factor Memory	0.523 ±0.012	0.235 ±0.005	1.001 ±0.018	**0.520** ±0.036	0.227 ±0.010	0.969 ±0.033	0.520 ±0.026	0.226 ±0.002	0.987 ±0.011
Principle Memory	0.460 ±0.035	0.260 ±0.015	1.116 ±0.067	0.513 ±0.021	0.226 ±0.006	0.956 ±0.016	0.520 ±0.026	0.225 ±0.008	0.962 ±0.022
Case Memory	0.453 ±0.042	0.262 ±0.008	1.160 ±0.049	**0.520** ±0.020	0.229 ±0.008	0.981 ±0.027	0.520 ±0.020	0.233 ±0.011	0.993 ±0.026
Structure Memory	0.483 ±0.012	0.250 ±0.005	1.094 ±0.003	0.493 ±0.023	0.228 ±0.006	0.966 ±0.015	0.500 ±0.000	0.227 ±0.003	0.969 ±0.024
ReFoRM (ours)	**0.530** ±0.000	**0.212** ±0.000	**0.914** ±0.000	0.513 ±0.015	**0.217** ±0.008	**0.928** ±0.030	**0.583** ±0.025	**0.213** ±0.006	**0.886** ±0.019

	Llama 4 Maverick			Qwen Plus (2025-07-28)			Five-model average		
	Acc ↑	Brier ↓	NLL ↓	Acc ↑	Brier ↓	NLL ↓	Acc ↑	Brier ↓	NLL ↓
No-memory baselines									
Direct Forecast	0.433 ±0.025	0.238 ±0.006	0.976 ±0.028	0.523 ±0.021	0.228 ±0.005	0.984 ±0.035	0.497 ±0.012	0.231 ±0.004	0.991 ±0.021
Structured Reasoning	0.433 ±0.012	0.246 ±0.005	0.996 ±0.010	0.440 ±0.000	0.244 ±0.003	1.019 ±0.013	0.440 ±0.006	0.250 ±0.001	1.046 ±0.005
Resolved-event memory									
Factor Memory	0.453 ±0.023	0.239 ±0.013	1.000 ±0.041	0.530 ±0.017	0.233 ±0.008	1.017 ±0.054	0.509 ±0.008	0.232 ±0.003	0.995 ±0.011
Principle Memory	0.427 ±0.040	0.241 ±0.006	0.994 ±0.013	0.507 ±0.015	0.234 ±0.006	1.024 ±0.059	0.486 ±0.018	0.237 ±0.005	1.011 ±0.020
Case Memory	0.440 ±0.026	0.242 ±0.008	1.002 ±0.023	0.517 ±0.032	0.237 ±0.005	1.050 ±0.029	0.490 ±0.012	0.241 ±0.002	1.037 ±0.012
Structure Memory	0.450 ±0.026	0.245 ±0.008	1.004 ±0.021	0.533 ±0.040	0.232 ±0.004	1.025 ±0.027	0.492 ±0.008	0.236 ±0.001	1.012 ±0.007
ReFoRM (ours)	**0.527** ±0.025	**0.227** ±0.003	**0.946** ±0.014	**0.597** ±0.015	**0.217** ±0.006	**0.936** ±0.036	**0.550** ±0.003	**0.217** ±0.002	**0.922** ±0.009

Table 4: Category-level Brier scores pooled across five models using one run per model. Best results are bolded per column.

Method	Earnings	Energy	Macro	Markets	Policy	Avg.
Direct Forecast	0.2457	**0.2183**	0.2466	0.2116	0.2126	0.2270
Structured Reasoning	0.2443	0.2693	0.2606	0.2376	0.2187	0.2461
Factor Memory	0.2452	0.2261	0.2396	0.2272	**0.1897**	0.2256
Principle Memory	0.2588	0.2226	0.2588	0.2227	0.2167	0.2359
Case Memory	0.2667	0.2300	0.2655	0.2257	0.2236	0.2423
Structure Memory	0.2717	0.2250	0.2576	0.2110	0.2281	0.2387
ReFoRM (ours)	**0.2391**	0.2186	**0.2250**	**0.2013**	0.1967	**0.2161**

also enter this stage. The trace is

$$z_i = f_\theta(q_i, \mathcal{E}_i^{<t_i}, L_i, \widehat{\Pi}_i, W_i^*). \tag{5}$$

In Equation 5, f_θ denotes reasoning and z_i is the resulting trace. It connects evidence-backed drivers to the target, considers counterevidence, and states what remains uncertain.

Once the reasoning trace is complete, low, central, and high estimates are formed in the target unit and mapped to the public option boundaries. This mapping produces the final forecast distribution

$$p_i = g_\theta(z_i, \mathcal{E}_i^{<t_i}, \kappa_i). \tag{6}$$

In Equation 6, g_θ denotes numerical estimation and boundary mapping, while p_i is the resulting distribution. Weak magnitude support produces a less concentrated forecast. A final consistency check ensures that the probabilities sum to one and that the central estimate lies within the modal option. Historical probabilities are not reused, and no post-hoc pooling or calibration is applied.

5 Experiments

Our experiments test whether structures from resolved occurrences improve later forecasts. We compare ReFoRM with six baselines across five LLMs and financial categories, then evaluate the validity of their relations and the quality of their reasoning. Further analyses and case studies examine how structural assessments affect predicted options and probability distributions. Additional implementation details and experimental configurations are available in the linked code repository.

5.1 Experimental Setup

Evaluation Data. The evaluation uses the 100 targets in Section 3. All methods receive the same target specification, and all evidence is restricted to documents available before the forecast cutoff. Methods using resolved events draw from the same five eligible occurrences, and target outcomes are never supplied.

Forecasting Models. To assess performance across diverse LLMs, we evaluate Gemini 2.5 Flash Lite [8], GPT-5 mini [27], DeepSeek V3.2 [20], Llama 4 Maverick, and Qwen Plus (2025-07-28) [32] on questions with forecast cutoffs between March 2025 and June 2026. The set spans five model families and includes both proprietary and open-weight models, with knowledge cutoffs ranging from May 2024 to March 2025. All cutoffs precede the earliest forecast cutoff, preventing overlap with target outcomes. Reasoning effort is medium where supported, and all other inference settings are fixed across methods for each model throughout evaluation.

Baselines. We compare ReFoRM with six baselines. Direct Forecast and Structured Reasoning use no resolved event. The remaining

baselines each expose one form of historical information, which separates the value of retrieving past content from the value of reconstructing its structure for the current event

- **Direct Forecast** follows the standard retrieval-augmented forecasting used in prior language-model forecasters [31].
- **Structured Reasoning** adapts the search-enabled graph reasoning setup for current evidence [5].
- **Factor Memory** adapts the factor-level representation of forecasting experience from resolved events [4].
- **Principle Memory** extracts reusable forecasting principles from resolved events [24, 37].
- **Case Memory** adapts episode-level experience retrieval from experiential and agent-memory systems [30, 37].
- **Structure Memory** directly retrieves an outcome-redacted historical event DAG from a resolved occurrence [5].

All methods retrieve current evidence from frozen pre-cutoff document collections. Factor Memory and ReFoRM share the same factor-guided results, while the remaining methods retrieve with the target question. Structure Memory separately tests direct use of a resolved graph without reconstruction.

5.2 Financial Forecasting

Forecasting Metrics. We evaluate each forecast distribution p_i with metrics that capture both categorical correctness and probabilistic quality. For question q_i, let p_{ik} denote the probability assigned to outcome $k \in \mathcal{Y}_i$, and let $\hat{y}_i = \text{argmax}_{k \in \mathcal{Y}_i} p_{ik}$ denote the predicted outcome. Across N questions, accuracy and NLL are $\text{Acc} = N^{-1} \sum_{i=1}^{N} \mathbb{I}[\hat{y}_i = y_i]$ and $\text{NLL} = -N^{-1} \sum_{i=1}^{N} \log p_{iy_i}$. We use the Brier score [13] to evaluate the full distribution.

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathcal{Y}_i|} \sum_{k \in \mathcal{Y}_i} (p_{ik} - \mathbb{I}[k = y_i])^2 . \tag{7}$$

Accuracy records whether $\hat{y}_i$ matches the realized outcome y_i, while NLL penalizes low probability assigned to y_i. Equation 7 measures squared error across the outcomes in p_i, capturing probability quality beyond the predicted outcome alone. Together, they distinguish identical outcomes with different probability assignments.

We evaluate every method on the same 100 questions across three random seeds. Table 3 reports the mean and sample standard deviation across runs, with all methods paired by question within each model and run throughout the analysis.

Table 3 compares all methods on recurring financial events. ReFoRM improves the average across the five models from 0.497 to 0.550 in accuracy, from 0.231 to 0.217 in Brier score, and from 0.991 to 0.922 in NLL relative to Direct Forecast. It also obtains the lowest Brier score and NLL for every model. Together, these results show improvements in both categorical decisions and forecast distributions, with probabilistic gains consistent across every model. Since neither structured reasoning over current evidence nor direct use of resolved event information matches this performance, the results support reconstructing historical relations under current evidence.

Table 4 shows that ReFoRM achieves the lowest average Brier score and the best result in three of the five categories, while outperforming Direct Forecast in four categories. The gains span earnings, macroeconomics, financial markets, and monetary policy, while the energy result is nearly tied. This breadth shows that the overall improvement is not confined to one type of financial event. The remaining variation motivates examining when relation reconstruction improves a forecast.

5.3 Reasoning Evaluation

Forecasting scores do not show whether a prediction follows from valid reasoning. We therefore compare relations from each reasoning trace with the post-resolution DAGs defined in Section 3.3.

Table 5: Reasoning scores for a single Gemini 2.5 Flash Lite single run. The last column is the share of a method's correct forecasts whose trace scored at or below 2.5.

Method	Gnd.	Val.	Algn.	Prob.	Mech.	Mean	≤ 2.5
No-memory baselines							
Direct Forecast	3.49	2.70	2.93	2.86	2.69	2.93	24%
Structured Reasoning	3.01	2.71	3.07	2.54	2.96	2.86	29%
Resolved-event memory							
Factor Memory	3.58	2.89	3.25	2.81	3.31	3.17	19%
Principle Memory	3.32	2.86	3.07	2.82	2.77	2.97	30%
Case Memory	3.22	2.86	3.13	2.75	2.83	2.96	24%
Structure Memory	3.13	2.72	3.09	2.53	2.85	2.86	37%
ReFoRM (ours)	**4.08**	**3.55**	**3.52**	**3.78**	**3.74**	**3.73**	**0%**

Relation Validity. Reasoning can still rely on relations that do not hold after resolution. Following [23], each trace is decomposed into atomic relations supporting the forecast and frozen before evaluation against the reference DAG by a blinded GPT-5.6 Terra judge. For question q and method m, $\mathcal{A}_{qm}$ denotes the frozen relations and G_q the reference DAG. We set $v(a, G_q) = 1$ only when G_q entails the same factors, direction, target implication, and horizon. Verified Relation Precision measures the supported share.

$$\text{VRP}(m) = \frac{\sum_q \sum_{a \in \mathcal{A}_{qm}} v(a, G_q)}{\sum_q |\mathcal{A}_{qm}|}. \tag{8}$$

Method identity, forecast choice, probabilities, and correctness are withheld during verification. Under Equation 8, ReFoRM achieves the highest VRP at 56.4%, exceeding Factor Memory at 50.0% and Structure Memory at 33.6%, the two strongest baselines under this metric. ReFoRM therefore uses a larger share of relations that remain valid after resolution. Its advantage supports selective reconstruction rather than direct transfer of resolved-event structure.

LLM Evaluation of Reasoning Quality. Accuracy and Brier evaluate forecast outputs, not whether a prediction follows from the stated reasoning. On a three-way question, an unsupported claim about a threshold crossing can select the correct option by chance and receive the same forecasting score as a well-supported rationale. We therefore evaluate the reasoning traces directly and relate their scores to correctness. Claude Opus 5 grades all 700 traces from the Gemini 2.5 Flash Lite run, with a separate judge instance for each question. The judge receives each trace's evidence, reasoning, prediction, and probabilities, but not the resolved outcome or forecasting scores. Each trace is rated from one to five on evidence

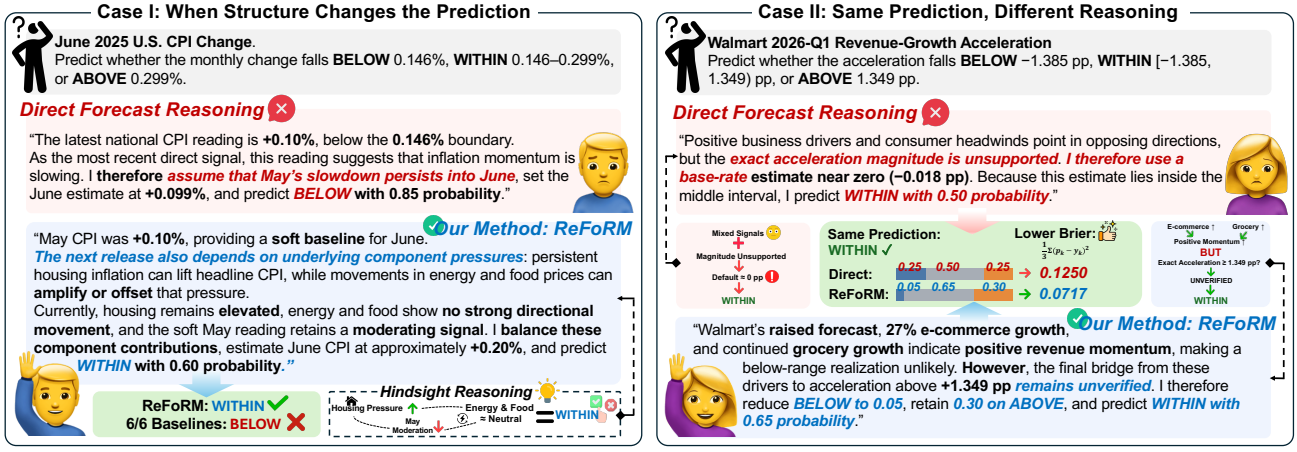


Figure 3: Two target occurrences under Direct Forecast and ReFoRM. In the first, evaluating the transferred relations against current evidence changes the predicted interval. In the second, both methods select the same interval, but only ReFoRM states which driver is established and where the bridge to the target remains unverified.

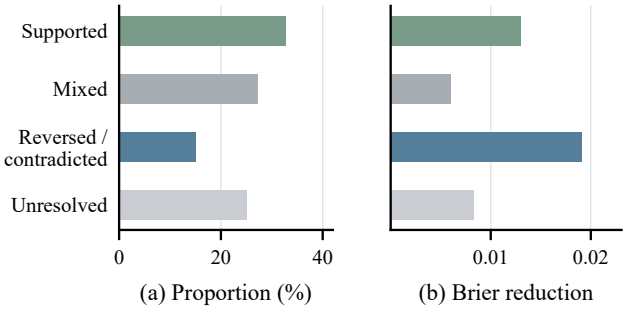


Figure 4: Relation assessments and Brier reduction across five models from a single run. Panel (a) shows proportions, and panel (b) shows reduction over Direct Forecast.

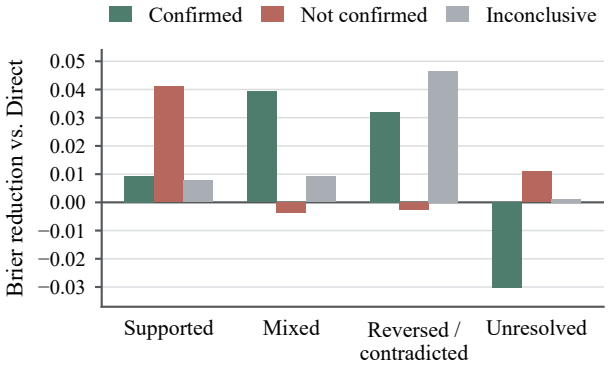


Figure 5: Forecasting gains by post-resolution confirmation across five models using one run per model. Bars show ReFoRM's Brier reduction relative to Direct Forecast.

grounding, logical validity, prediction alignment, probability justification, and mechanism specificity, where prediction alignment asks whether the reasoning entails the selected option. Six binary flags record categorical failures not captured by mean scores, and every rating requires a supporting span from the trace.

ReFoRM has the highest mean on all five dimensions, 3.73 against 2.86 to 3.17, with paired differences by question from 0.57 to 0.88 and standard errors below 0.09, and it records the lowest or tied-lowest rate of every flag. Its largest margins occur in probability justification and mechanism specificity. None of its correct forecasts has a trace scored at or below 2.5, compared with 19% to 37% for the baselines. Its forecasting errors among otherwise strong reasoning are more consistent with limitations in the available pre-cutoff evidence than with invalid reasoning. Pooled over methods, traces scoring at least 4.0 reach 0.630 accuracy and 0.192 Brier, compared with 0.475 and 0.248 below 2.5. These judge scores are not ground truth and cover only one forecasting model and one random seed, so they provide supporting evidence rather than a benchmark-wide estimate of reasoning quality.

5.4 Effectiveness of Structural Reconstruction

We examine structural reconstruction at two levels. Quantitative results relate ReFoRM's assessments of historical relations to probabilistic performance and post-resolution confirmation. Case studies show how reconstruction changes individual forecasts.

Quantitative Analysis. For each model and question, we assign the saved current structure to one of four aggregate profiles without using its prediction, outcome, or score. Here, changed means reversed or contradicted. *Supported* has preserved but no changed relations, *mixed* has both, *reversed or contradicted* has changed but no preserved relations, and *unresolved* has neither. Repeated relations across retrieved paths are counted once. Figure 4 shows the largest Brier reduction relative to Direct Forecast for reversed or contradicted assessments at 0.019, compared with 0.013, 0.006, and 0.008 for supported, mixed, and unresolved assessments, respectively. This motivates the confirmation analysis that follows.

A blinded judge compares each profile with its reference DAG without observing the model, prediction, probabilities, or score. Confirmation requires persistence for supported, both states for

mixed, change for reversed or contradicted, and continued uncertainty for unresolved. Other states are not confirmed and insufficient DAG evidence is inconclusive. Over conclusive cases, the rates are 87.8%, 22.0%, 50.8%, and 3.7%. The low last rate reflects that post-resolution evidence adjudicates at least one unverified relation. Figure 5 shows that confirmed mixed and reversed or contradicted assessments reduce Brier score by 0.039 and 0.032, whereas the cases that are not confirmed show no improvement. The distinction is weaker for supported and unresolved assessments. Together, the two analyses suggest that the gains depend more on reassessing relations than on carrying them forward unchanged.

Case Studies. Figure 3 traces two representative forecasts under the same question and current evidence. In the CPI case, Direct Forecast extrapolates the latest aggregate reading and selects the lower interval. ReFoRM instead relates housing, energy, and food components to the next release and selects the middle interval. In the Walmart case, both methods select the middle interval, but ReFoRM assigns less probability to the lower interval after separating supported revenue drivers from an unresolved link to growth. Together, the cases show that structural reconstruction can change the selected option or refine the probability distribution when the selected option remains unchanged.

6 Conclusion

This paper introduces ReFoRM, which reconstructs forecasting structures from resolved financial events using evidence available for the current forecast. It represents predictive factors, relations, evidence requirements, competing explanations, counterevidence, and unresolved uncertainty without treating historical claims as current facts. Across five LLMs, ReFoRM improves categorical and probabilistic forecasting on recurring financial questions. Relation-level analysis further shows that its gains are strongest when historical relations are reassessed under current evidence. Our evaluation also distinguishes forecasting performance from reasoning validity. Future work can incorporate expert-curated hindsight structures as forecasting guidance or domain knowledge for evaluating financial reasoning beyond outcome accuracy.

References

[1] Torben G Anderson, Tim Bollerslev, Francis X Diebold, and Clara Vega. 2003. Micro effects of macro announcements: Real-time price discovery in foreign exchange. *American economic review* 93, 1 (2003), 38–62.

[2] Daniel Bradley, Sinan Gokkaya, and Xi Liu. 2017. Before an analyst becomes an analyst: Does industry experience matter? *The Journal of Finance* 72, 2 (2017), 751–792.

[3] Andrea Carriero, Davide Pettenuzzo, and Shubhranshu Shekhar. 2024. Macroeconomic forecasting with large language models. *arXiv preprint arXiv:2407.00890* (2024).

[4] Yurui Chang, Yongkang Du, Yuanpu Cao, Jinghui Chen, and Lu Lin. 2026. ForecastCompass: Guiding Agentic Forecasting with Adaptive Factor Memory. *arXiv preprint arXiv:2605.30858* (2026).

[5] Yizhou Chi, Eric Chamoun, Zifeng Ding, and Andreas Vlachos. 2026. WorldReasoner: Evaluating Whether Language Model Agents Forecast Events with Valid Reasoning. *arXiv preprint arXiv:2606.11816* (2026).

[6] Michael B Clement, Lisa Koonce, and Thomas J Lopez. 2007. The roles of task-specific forecasting experience and innate ability in understanding analyst forecasting performance. *Journal of Accounting and Economics* 44, 3 (2007), 378–398.

[7] Lauren Cohen and Andrea Frazzini. 2008. Economic links and predictable returns. *The Journal of Finance* 63, 4 (2008), 1977–2011.

[8] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[9] Yiannis Dendramis, George Kapetanios, and Massimiliano Marcellino. 2020. A similarity-based approach for macroeconomic forecasting. *Journal of the Royal Statistical Society Series A: Statistics in Society* 183, 3 (2020), 801–827.

[10] Graham Elliott and Allan Timmermann. 2008. Economic forecasting. *Journal of Economic Literature* 46, 1 (2008), 3–56.

[11] Federal Reserve Bank of St. Louis. 2026. Federal Reserve Economic Data (FRED). https://fred.stlouisfed.org/.

[12] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 6465–6488.

[13] W Brier Glenn. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.

[14] Refet S Gurkaynak, Brian Sack, and Eric T Swanson. 2005. Do actions speak louder than words? The response of asset prices to monetary policy actions and statements. (2005).

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

[16] Stephen J Hoch and David A Schkade. 1996. A psychological approach to decision support systems. *Management Science* 42, 1 (1996), 51–64.

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

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

[19] Michael Krumdick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. BizBench: A quantitative reasoning benchmark for business and finance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8309–8332.

[20] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556* (2025).

[21] Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Lingyue Yin, Mingren Yin, Zhu Zhenwei, et al. 2026. FutureX: An Advanced Live Benchmark for LLM Agents in Future Prediction. In *International Conference on Learning Representations*, Vol. 2026. 19522–19563.

[22] Michael W McCracken and Serena Ng. 2016. FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics* 34, 4 (2016), 574–589.

[23] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*. 12076–12100.

[24] Siru Ouyang, Jun Yan, I Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long Le, Samira Daruki, Xiangru Tang, et al. 2026. Reasoningbank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, Vol. 2026. 94327–94354.

[25] Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 15012–15032.

[26] Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 10066–10086.

[27] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 System card. *arXiv preprint arXiv:2601.03267* (2025).

[28] U.S. Securities and Exchange Commission. 2025. EDGAR Application Programming Interfaces (APIs). https://www.sec.gov/search-filings/edgar-application-programming-interfaces.

[29] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024. Finben: A holistic financial benchmark for large language models. *Advances in neural information processing systems* 37 (2024), 95716–95743.

[30] Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for LLM agents. *Advances in Neural Information Processing Systems* 38 (2025), 17577–17604.

[31] Qi Yan, Raihan Seraj, Jiawei He, Lili Meng, and Tristan Sylvain. 2024. Autocast++: Enhancing world event prediction with zero-shot ranking-based context retrieval. In *International Conference on Learning Representations*, Vol. 2024. 34523–34539.

[32] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[33] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031* (2023).

[34] Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. 2026. Llm-as-a-prophet: Understanding predictive intelligence with prophet arena. In *International Conference on Learning Representations*, Vol. 2026. 107020–107072.

[35] Zhangdie Yuan, Zifeng Ding, and Andreas Vlachos. 2026. Introducing FOReCAst: The future outcome reasoning and confidence assessment benchmark. *Advances in Neural Information Processing Systems* 38 (2026).

[36] Zhihan Zhang, Yixin Cao, and Lizi Liao. 2025. XFinBench: Benchmarking LLMs in complex financial problem solving and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*. 8715–8758.

[37] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19632–19642.

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

Appendix

A Benchmark and Experimental Details

A.1 Benchmark Construction

The Recurring Financial Events Set follows a deterministic construction pipeline. We first register 40 financial event families. Each registry entry fixes the entity or series, target quantity, transformation, unit, release schedule, and authoritative resolution source. We then instantiate dated occurrences from immutable snapshots of 32 FRED series and SEC Company Facts for eight issuers. Every snapshot records its source URL, retrieval time, and content hash.

Table 6: Construction of the Recurring Financial Events Set. Each operation is applied uniformly after the event-family registry is fixed.

Stage	Operation	Recorded fields
Family registry	Fix the recurring target before viewing individual outcomes	Entity or series, metric, unit, transformation, schedule, and resolution source
Source snapshot	Read the registered SEC or FRED source and preserve the input	Source identifier, URL, retrieval time, content hash, and dated observations
Target formation	Apply the registered transformation to each dated occurrence	Forecast cutoff, target period, horizon, target value, and resolution date
Answer formation	Compute boundaries from observations available before that cutoff	Binary threshold or three-class lower and upper boundaries
Question record	Render the target as a fixed probabilistic forecasting task	Question, answer set, unit, comparison rule, and resolution rule
Chronological split	Order occurrences within each family and separate their roles	Five resolved references followed by two or three forecast targets

For a registered FRED series, let u_τ denote the observation for target period τ. The target change is fixed as either $d_\tau = 100(u_\tau/u_{\tau-1} - 1)$ in percent or $d_\tau = u_\tau - u_{\tau-1}$ in the series unit. A three-class question compares d_τ with the lower and upper quartiles of the preceding eight changes. A binary question compares it with the median of the preceding four changes. The transformation mode and unit are registered per family and cannot change across its occurrences.

For corporate events, let R_τ denote quarterly revenue and $g_\tau = 100(R_\tau/R_{\tau-4} - 1)$ its year-over-year growth rate. The target acceleration is $a_\tau = g_\tau - g_{\tau-1}$. Three-class questions compare a_τ with the quartiles of the preceding four accelerations, while binary questions ask whether $a_\tau \geq 0$, equivalently whether the current growth rate reaches the preceding quarter's growth rate. Resolution uses the first SEC 10-Q or 10-K containing the target quarter.

After the family registry is fixed, every eligible occurrence is instantiated uniformly. Each public record contains the cutoff, target period, answer set, numerical rule, unit, and resolution source, with its realized value, answer label, and resolution explanation retained in a separate resolution record. Within each family, the first five occurrences provide resolved references and the remaining occurrences form the test set. This produces 200 reference questions and 100 forecast targets, with 190 three-class and 110 binary questions overall.

A.2 Benchmark Question Examples

Table 7 shows one target from each financial category as it is presented at forecast time. The examples reproduce the public wording and numerical resolution rules used for evaluation.

Table 7: Representative target questions from the five financial categories.

Category	Public question	Answer and resolution rule
Corporate earnings	As of 2025-03-31, which range will Tesla's quarterly year-over-year revenue-growth acceleration for the fiscal quarter ending 2025-03-31 fall into?	Below if $a_\tau < -7.321$, within if $-7.321 \le a_\tau < 6.908$, and above otherwise, in percentage points. Resolve from the first SEC filing containing the quarter.
Energy and commodities	As of 2025-06-01, will the monthly U.S. retail diesel-price change in 2025-06 be at least -1.47%?	Yes if the percentage change in FRED series GASDESW is at least -1.47%, and no otherwise.
Macroeconomics	As of 2025-06-30, which range will the monthly U.S. CPI change in 2025-06 fall into?	Below if the CPIAUCSL change is below 0.146%, within if it is from 0.146% to 0.299%, and above otherwise.
Markets and credit	As of 2025-06-01, will the monthly Baa-minus-10-year Treasury spread change in 2025-06 be at least 0.155 percentage points?	Yes if the absolute change in FRED series BAA10Y reaches 0.155, and no otherwise.
Monetary policy	As of 2025-06-01, which range will the monthly effective federal-funds-rate change in 2025-06 fall into?	Below if the FEDFUNDS change is below -0.167, within if it is from -0.167 to 0, and above otherwise, in percentage points.

A.3 Models and Inference Configuration

We evaluate five LLM families and six baselines. The public model identifier is fixed across methods within each comparison. Native reasoning controls are set to medium where the model interface provides them; otherwise, the model's standard inference interface is used.

Table 8: Forecasting models used in the experiments.

Model	Public identifier
Gemini 2.5 Flash Lite	`google/gemini-2.5-flash-lite`
GPT-5 mini	`openai/gpt-5-mini`
DeepSeek V3.2	`deepseek/deepseek-v3.2`
Llama 4 Maverick	`meta-llama/llama-4-maverick`
Qwen Plus (2025-07-28)	`qwen/qwen-plus-2025-07-28`

All methods share the target specification, question IDs, scoring procedure, and output validation. Direct Forecast, Structured Reasoning, Case Memory, Principle Memory, and Structure Memory retrieve current evidence from the target question. Factor Memory and ReFoRM use factors from resolved occurrences to guide retrieval. Both conditions are fixed before inference and restricted to pre-cutoff documents.

B ReFoRM Inference Procedure

B.1 Current-Evidence Ledger

Prompt B.1: Current-evidence ledger

[SYSTEM]

You extract a pre-cutoff financial evidence ledger. You never produce an answer or probability. Return only schema-conforming JSON and do not add fields outside the requested schema.

[INPUT]

```
CURRENT CASE:
{question, context, target_specification, forecast_cutoff}

TARGET OPERATION:
{target_metric, target_period, unit, comparison_rule,
 resolution_rule, semantic_guard, predicate_or_intervals}

CURRENT PRE-CUTOFF EVIDENCE:
[{id, title, source, published_date, excerpt}, ...]
```

Historical paths, resolved outcomes, option mappings, and historical probabilities are withheld from this call.

[INSTRUCTION]

Read the current pre-cutoff evidence before any historical graph is shown. Build a compact factual ledger for the exact financial target. Record the latest usable target baseline, current drivers, the temporal relation of each driver to the target period, and information that remains missing. Do not forecast, choose an option, assign probabilities, or infer what a past occurrence would imply. Distinguish a level from a change, a broad outlook from the target period, and a leading signal from a direct target observation. A target baseline may be marked observed only when a supplied article directly supports it. Every factual signal must cite one or more supplied evidence IDs.

[OUTPUT FORMAT]

```
target_baseline: {
  status: observed | partial | unavailable,
  observation: string,
  evidence_ids: [evidence_id],
  assessment: string
}
current_signals: [{
  signal_id: string,
  factor: string,
  state: strengthening | weakening | elevated | depressed | stable | mixed | unknown,
  temporal_role: target_period | leading | lagging | structural,
  observation: string,
  evidence_ids: [evidence_id]
}]
data_gaps: [string]
ledger_summary: string
```

B.2 Relation Reconstruction

Prompt B.2: Relation reconstruction

[SYSTEM]

You instantiate retrieved financial subgraphs with current evidence. You never forecast, choose an answer, or assign probabilities. Preserve the supplied topology and return only compact schema-conforming JSON.

[INPUT]

```
CURRENT CASE:
{question, context, target_specification, forecast_cutoff}

TARGET OPERATION:
```

```
{target_metric, target_period, unit, semantic_guard, boundaries}

CURRENT EVIDENCE LEDGER:
{target_baseline, current_signals, data_gaps, ledger_summary}

RETRIEVED HINDSIGHT PATHS:
[{path_id, checkpoints, edges, mechanism,
  applicability_conditions, failure_conditions}, ...]

CURRENT PRE-CUTOFF EVIDENCE:
[{id, title, source, published_date, excerpt}, ...]
```

The routed subgraphs contain no historical answer or probability.

[INSTRUCTION]

Fill the supplied subgraphs from the hindsight graphs with current pre-cutoff evidence. Keep every node and edge and return only their current state. For nodes, record whether the current occurrence aligns with the factor, reverses it, leaves it unobserved, or uses it only as a structural placeholder. Also record the current value, timing, evidence IDs, and confidence. For edges, determine whether the historical relation is preserved, reversed, contradicted, or unverified, and record its lag and current support. Historical topology is structural knowledge, not current evidence, and direct current evidence overrides it. An unobserved mediator does not by itself invalidate an otherwise applicable mechanism. Mark each complete path active only when its material relations can support the current forecast, contradicted when current evidence rejects the mechanism, and unresolved when the available evidence cannot decide.

[OUTPUT FORMAT]

```
node_states: [{
  node_id,
  relation: ALIGNED | REVERSED | UNOBSERVED | STRUCTURAL,
  value,
  time: target_period | leading | lagging | structural | unknown,
  evidence_ids,
  confidence: high | medium | low
}]
edge_states: [{
  edge_id,
  relation: PRESERVED | REVERSED | CONTRADICTED | UNVERIFIED,
  lag: immediate | short | medium | long | unknown,
  support: direct | indirect | structural | none,
  evidence_ids,
  confidence: high | medium | low
}]
path_states: [{
  path_id,
  status: ACTIVE | CONTRADICTED | UNRESOLVED,
  effect_on_target: up | down | neutral | mixed | uncertain
}]
graph_synthesis: {active_path_ids, direction, assessment}
```

B.3 Evidence-Grounded Forecast Reasoning

Prompt B.3: Forecast reasoning

[SYSTEM]

You are a financial forecaster using pre-cutoff evidence. Return an evidence-grounded reasoning trace in the requested schema, never an answer choice, target number, or probability distribution.

[INPUT]

```
CURRENT CASE:
{question, context, target_specification, forecast_cutoff}
```

```
TARGET OPERATION:
{target_metric, target_period, unit, semantic_guard, boundaries}

CURRENT EVIDENCE LEDGER:
{target_baseline, current_signals, data_gaps, ledger_summary}

RECONSTRUCTED CURRENT STRUCTURE:
[{path_id, checkpoints_with_current_state,
  edges_with_current_state, path_state}, ...]

RESOLVED-EVENT REVIEW SEQUENCES:
[{reasoning_sequence, counterevidence_check,
  uncertainty_check, structural_lesson}, ...]

CURRENT PRE-CUTOFF EVIDENCE:
[{id, title, source, published_date, excerpt}, ...]
```

[INSTRUCTION]

Build a complete reasoning trace using the retrieved hindsight paths as an incomplete structural scaffold. Start with the exact target operation and the current baseline. Use helpful nodes, edges, or partial paths to organize current drivers and mechanisms. Add evidence-backed current factors and intermediate reasoning when the historical graph does not cover them. Test counterevidence, resolve competing explanations where possible, and end with an explicit bridge to the target quantity. Distinguish target-period magnitude support from directional support.

Current evidence always overrides historical structure. A relation marked UNVERIFIED is uncertain rather than false and may guide a hypothesis, but it is not a current fact. Never use an explicitly CONTRADICTED relation as positive support. When discussing a contradicted relation, express it as counterevidence rather than a driver or mechanism. Cite only current evidence IDs for factual claims and list only the paths that materially contributed to the current synthesis.

The worked checks describe how an earlier occurrence organized its argument. Use them only after reading the current evidence to check whether the current trace omitted a baseline, causal link, competing explanation, or uncertainty. Do not copy their historical facts, direction, estimate, conclusion, answer, or probability.

[OUTPUT FORMAT]

```
target_semantics: string
selected_evidence_ids: [evidence_id]
evidence_fit: {
  metric_match: direct | partial | weak,
  horizon_match: direct | partial | weak,
  magnitude_support: supported | partial | unsupported,
  assessment: string
}
current_new_factors: [{
  factor,
  effect_on_target: up | down | neutral | mixed | uncertain,
  evidence_ids,
  assessment
}]
causal_balance: {
  favored_direction: up | down | balanced | uncertain,
  used_path_ids,
  assessment
}
magnitude_readiness: {
  support: direct | derived | direction_only | insufficient,
  evidence_ids,
  assessment
}
reasoning_steps: [{
```

```
  step_type: baseline | driver | mechanism | counterevidence | target_bridge,
  statement,
  evidence_ids,
  effect_on_target,
  source_id
}]
counterevidence: string
uncertainty: string
```

B.4 Forecasting with the Current Structure

Prompt B.4: Forecasting with the current structure

[SYSTEM]

You are a financial forecast boundary auditor using pre-cutoff evidence. Return only the requested schema-conforming output. Do not introduce factual claims that are absent from the supplied reasoning and evidence.

[INPUT]

```
CURRENT CASE:
{question, context, target_specification, forecast_cutoff}

CURRENT REASONING TRACE:
{target_semantics, evidence_fit, current_new_factors,
 causal_balance, magnitude_readiness, reasoning_steps,
 counterevidence, uncertainty}

CURRENT PRE-CUTOFF EVIDENCE:
[{id, title, source, published_date, excerpt}, ...]
```

Any previous prediction, probabilities, target estimate, and option mapping are hidden from this call.

[INSTRUCTION]

Complete the final probabilistic forecasting stage. Audit the supplied reasoning against the current pre-cutoff evidence and verify the exact target operation. For change, return, growth, or acceleration targets, identify both quantities required by the public operation and express them in a common unit. Then form a coarse low, central, and high estimate in the target unit.

Compare the numerical central estimate with every public boundary. A positive estimate does not automatically imply the above-range option, and a negative estimate does not automatically imply the below-range option. Pay particular attention to negative boundaries. The arithmetically mapped central option must be the modal option. Allocate the remaining probability using the overlap of the estimate range with each interval and the uncertainty stated in the reasoning. The prediction, mapped option, and probability mode must agree.

Use only current evidence IDs. Do not invent a latent number merely to complete the mapping. Label magnitude support *direct* when pre-cutoff evidence contains an explicit quantitative expectation for the target period. Label it *derived* when current quantities support an auditable proxy or arithmetic estimate in the target unit. Use *direction only* when only the sign is supported and *insufficient* when even the direction lacks an auditable basis. When magnitude evidence is weak, keep the estimate broad and avoid a concentrated probability distribution.

[OUTPUT FORMAT]

```
target_operation_check: string
directional_signal: up | down | flat | mixed | uncertain
magnitude_assessment: {support, evidence_ids, rationale}
latent_target_estimate: {low, central, high, unit, basis}
boundary_checks: [{option, interval, compatibility, rationale}]
mapped_option: option
prediction: option
option_probabilities: [{option, probability}]
uncertainty: string
```

C Baseline Prompts

The six baseline prompts are shown separately because they expose different forms of resolved-event information and impose different constraints on its use.

Prompt C.1: Direct Forecast

[SYSTEM]

You are a financial forecaster using pre-cutoff evidence. Return only the requested structured forecast and cite supplied current evidence IDs for factual claims.

[INPUT]

`CURRENT CASE:`
`{question, context, target_specification, forecast_cutoff}`
`CURRENT PRE-CUTOFF EVIDENCE:`
`[{id, title, source, published_date, excerpt}, ...]`

[INSTRUCTION]

Lock the exact metric, horizon, unit, and option boundaries. Establish a current baseline, identify supported drivers and counterevidence, connect them to the target, and preserve uncertainty when magnitude support is weak. Reason only from the current question and evidence; no resolved-event information is available. Do not confuse level with change, growth with acceleration, or direction with a boundary crossing.

[OUTPUT FORMAT]

Reasoning steps, target estimate, selected option, and one probability distribution over the public options.

Prompt C.2: Structured Reasoning

[SYSTEM]

You construct a prospective financial DAG from current evidence and then forecast only from the validated graph and its linked evidence.

[INPUT]

`{CURRENT CASE}`, `{TARGET SPECIFICATION}`, and `{CURRENT EVIDENCE}`. No resolved-event information is supplied.

[INSTRUCTION]

First build a current-case DAG containing a historical baseline, current factors, directed relationships, countervailing paths, and a target bridge. Do not select an answer during graph construction. Validate all evidence IDs and remove unsupported nodes or relations. Then trace the main and countervailing paths to a target estimate and map it once to the public options.

[OUTPUT FORMAT]

Prospective DAG followed by a graph-grounded reasoning trace, prediction, and probability distribution.

Prompt C.3: Factor Memory

[SYSTEM]

You are a financial forecaster using pre-cutoff evidence. Historical factor cards are search and coverage guidance, not evidence for the current target.

[INPUT]

`{CURRENT CASE}`, `{FACTOR CARDS}`, and `{FACTOR-GUIDED CURRENT EVIDENCE}`.

[INSTRUCTION]

Use the cards to identify which current drivers and counterevidence should be checked. Do not transfer historical entities, dates, values, outcomes, or causal relations. Establish every current factual claim from supplied evidence, then connect the supported factors to the exact target and its boundaries.

[OUTPUT FORMAT]

Factor-grounded reasoning trace, target estimate, prediction, and probability distribution.

Prompt C.4: Principle Memory

[SYSTEM]

You are a financial forecaster using pre-cutoff evidence. The supplied principles are outcome-redacted reasoning guidance rather than current evidence.

[INPUT]

`{CURRENT CASE}`, `{FORECASTING PRINCIPLES}`, and `{CURRENT EVIDENCE}`.

[INSTRUCTION]

Apply the principles as general search and reasoning lessons. Independently establish the current baseline, drivers, counterevidence, and target bridge. Do not assume that a historical mechanism, direction, or magnitude still holds. Use only current evidence IDs for factual claims.

[OUTPUT FORMAT]

Principle-guided reasoning trace, target estimate, prediction, and probability distribution.

Prompt C.5: Case Memory

[SYSTEM]

You are a financial forecaster using pre-cutoff evidence. The retrieved resolved episode is an analogy, not evidence for the current occurrence.

[INPUT]

{CURRENT CASE}, {RESOLVED EPISODE}, and {CURRENT EVIDENCE}. The episode may contain its historical question, evidence, reasoning, and answer.

[INSTRUCTION]

Use the episode to identify potentially relevant comparisons, but never treat its answer, entity, dates, values, or historical facts as current evidence. Re-establish every factual premise from the current evidence and preserve uncertainty when the analogy does not determine the target-period magnitude.

[OUTPUT FORMAT]

Case-guided reasoning trace, target estimate, prediction, and probability distribution.

Prompt C.6: Structure Memory

[SYSTEM]

You are a financial forecaster using pre-cutoff evidence. The supplied outcome-redacted DAG is historical structure, not current evidence.

[INPUT]

{CURRENT CASE}, {HISTORICAL DAG}, and {CURRENT EVIDENCE}.

[INSTRUCTION]

Inspect which historical roles and relations appear relevant, but do not copy the graph's topology, direction, or support blindly. Use current evidence to decide which relations can organize the forecast. Unlike ReFoRM, do not produce a separate node-, edge-, and path-level reconstruction before forecasting.

[OUTPUT FORMAT]

DAG-guided reasoning trace, target estimate, prediction, and probability distribution.

D Evaluation Prompts and Rubrics

D.1 Reasoning-Quality Evaluation

Prompt D.1: Reasoning-quality judge

[SYSTEM]

You evaluate forecast reasoning, not whether the forecast later turns out to be correct. The resolved answer is deliberately withheld. Do not search for it, infer it from outside knowledge, or use information beyond the supplied packet.

[INPUT]

```
QUESTION:
{question_text, options, forecast_cutoff, target_specification}

TRACE-SPECIFIC EVIDENCE:
[{trace_id, retrieval_condition, documents}, ...]

TRACES:
[{trace_id, method, retrieval_condition, resolved_event_information,
  reasoning, prediction, probabilities}, ...]
```

The outcome, correctness, and all forecasting metrics are withheld.

[INSTRUCTION]

Score each trace from 1 to 5 on evidence grounding, logical validity, prediction alignment, probability justification, and mechanism specificity. Judge a trace only against the evidence and resolved-event information supplied to it. An ID available only to another method is not available to the trace being judged. Do not reward or penalize a method for the amount of retrieved evidence, and do not reward verbosity, jargon, or structural formatting. Read the reasoning first, decide which option it implies, and only then compare it with the emitted prediction. Every score and every positive flag must cite an exact span from the trace. Rank all seven traces within the question after scoring them independently.

[RUBRIC]

Dimension	Score 1	Score 5
Evidence grounding	Material claims are invented or misquoted	Every factual claim traces to supplied evidence
Logical validity	Non-sequential or self-contradictory argument	Each step follows and competing considerations are resolved
Prediction alignment	Reasoning points to another option	Reasoning entails the selected option and excludes alternatives
Probability justification	Probability conflicts with stated uncertainty	Support, gaps, and probability mass agree
Mechanism specificity	Generic language fits unrelated questions	Named drivers connect to the target quantity

[BINARY FLAGS]

```
unsupported_magnitude_leap: direction is converted to a precise magnitude without support
hallucinated_number: a material number appears nowhere in the supplied evidence
internal_contradiction: the trace relies on mutually inconsistent claims
boilerplate_only: the reasoning contains no question-specific mechanism
post_hoc_option_fit: the option appears chosen before the rationale is assembled
admits_own_gap: the trace explicitly identifies an inference it cannot verify
```

[OUTPUT FORMAT]

```
scores: {evidence_grounding, logical_validity, prediction_alignment,
         probability_justification, mechanism_specificity}
flags: {unsupported_magnitude_leap, hallucinated_number,
        internal_contradiction, boilerplate_only,
        post_hoc_option_fit, admits_own_gap}
decisive_quote: string
verdict_one_line: string
within_question_ranking: [trace_id]
```

D.2 Relation-Validity Evaluation

Prompt D.2: Relation-validity judge

[SYSTEM]

You verify forecast-supporting relations against a post-resolution reference DAG. Evaluate the relation itself rather than the fluency of the surrounding reasoning. Method identity and forecast performance are irrelevant.

[INPUT]

```
ATOMIC FORECAST RELATION:
{relation_id, source_factor, relation, target_factor,
 target_implication, forecast_horizon, supporting_trace_span}

POST-RESOLUTION REFERENCE DAG:
{nodes, directed_relations, evidence_spans, resolution_period}
```

Method identity, prediction, probabilities, correctness, and forecasting score are hidden. Relations that the forecaster explicitly rejected, treated as hypothetical, or left unresolved are not evaluated as adopted forecast support.

[INSTRUCTION]

Determine whether the reference DAG entails the relation at the same level of specificity. First align the source and target factors by their economic meaning, not lexical overlap. Then compare the relation direction and the implication for the forecast quantity. Finally, verify that the relationship operates over a compatible forecast horizon. Broader thematic similarity is insufficient. A relation is verified only when factor correspondence, direction, target implication, and horizon all agree. Cite the reference-DAG span supporting the decision and explain any mismatch precisely.

[DECISION RUBRIC]

```
factor_correspondence: required
direction_match: required
target_implication_match: required
horizon_match: required
verified = all four checks are satisfied
```

[OUTPUT FORMAT]

```
relation_id: string
factor_correspondence: boolean
direction_match: boolean
target_implication_match: boolean
horizon_match: boolean
verified: boolean
reference_span: string
justification: string
```

E Additional Results

E.1 Forecasting Gains by Relation Assessment

Table 9 expands Figure 4 to all methods. The four columns use the same 164 supported, 136 mixed, 75 reversed or contradicted, and 125 unresolved forecasts classified from ReFoRM's saved relation states. Positive values indicate a lower Brier score than Direct Forecast on the identical subset.

Table 9: Brier reduction relative to Direct Forecast within each forecast-time relation assessment.

Method	Supported	Mixed	Rev. / contr.	Unresolved
Structured Reasoning	−0.0209	−0.0153	−0.0155	−0.0230
Factor Memory	+0.0072	−0.0159	+0.0092	+0.0080
Principle Memory	−0.0111	+0.0007	−0.0141	−0.0135
Case Memory	−0.0226	−0.0070	−0.0240	−0.0097
Structure Memory	−0.0043	−0.0220	−0.0056	−0.0139
ReFoRM	**+0.0130**	**+0.0060**	**+0.0191**	**+0.0083**

ReFoRM improves over Direct Forecast in all four assessments and obtains the largest reduction in each column. The strongest gain occurs when relations are reversed or contradicted, showing the value of reconstructing historical structure when the current event departs from earlier occurrences.

E.2 Forecasting Gains by Post-Resolution Confirmation

Table 10 expands Figure 5 with the same subsets for every baseline. After forecasts are frozen, a blinded judge determines whether each assessment is confirmed by the reference DAG. Positive values again indicate improvement over Direct Forecast.

Table 10: Brier reduction relative to Direct Forecast by forecast-time assessment and post-resolution confirmation. Bold marks the largest reduction within each confirmation block and column.

Method	Supported	Mixed	Rev. / contr.	Unresolved
Confirmed				
Forecasts	137	27	33	4
Structured Reasoning	−0.0241	−0.0168	−0.0172	−0.0321
Factor Memory	+0.0028	+0.0124	+0.0123	−0.0398
Principle Memory	−0.0159	−0.0187	−0.0195	**+0.0137**
Case Memory	−0.0230	+0.0005	−0.0291	+0.0019
Structure Memory	−0.0123	−0.0584	−0.0239	−0.0046
ReFoRM	**+0.0094**	**+0.0393**	**+0.0319**	−0.0305
Not confirmed				
Forecasts	19	96	32	104
Structured Reasoning	+0.0046	−0.0126	−0.0178	−0.0274
Factor Memory	+0.0427	−0.0164	−0.0122	**+0.0127**
Principle Memory	+0.0205	**+0.0056**	−0.0136	−0.0120
Case Memory	−0.0081	−0.0111	−0.0210	−0.0151
Structure Memory	**+0.0501**	−0.0152	−0.0039	−0.0148
ReFoRM	+0.0410	−0.0037	**−0.0028**	+0.0109
Inconclusive				
Forecasts	8	13	10	17
Structured Reasoning	−0.0264	−0.0324	−0.0027	+0.0063
Factor Memory	−0.0016	−0.0711	**+0.0676**	−0.0099
Principle Memory	−0.0026	+0.0045	+0.0025	−0.0294
Case Memory	−0.0488	+0.0076	−0.0169	**+0.0202**
Structure Memory	+0.0041	+0.0031	+0.0492	−0.0101
ReFoRM	**+0.0077**	**+0.0092**	+0.0466	+0.0012

The confirmation contrast is clearest for mixed and reversed or contradicted assessments. ReFoRM reduces Brier by 0.0393 and 0.0319 when these assessments are confirmed, but the reductions become negative when they are not confirmed. This contrast links ReFoRM's largest calibration gains to relation assessments that remain valid after the event resolves.