Do Language Models Update their Forecasts with New Information?

Anonymous Authors[1]

Abstract

Prior work largely treats forecasting as a task in which predictions are stationary, ignoring how forecasts should evolve with new evidence. To address this we introduce EVOLVECAST, a framework for evaluating whether large language models (LLMs) revise forecasts appropriately given new information released after their training cutoff. Comparing LLMs against human forecasters, we find while models often track the correct update direction, their quantitative shifts are under-responsive. Furthermore, both verbalized and logits-based confidence estimates remain poorly calibrated compared to the human reference standard. These findings suggest current LLMs are limited in their response to new information; models treat new information as mere retrieval context rather than evidence that dynamically shifts posterior probabilities. EVOLVECAST highlights the critical need for robust mechanisms to incorporate external knowledge into belief updating.

1. Introduction

Large language models (LLMs) excel at factual recall and reasoning over static data (Lin et al., 2022; Guo et al., 2023). However, real-world decision-making is dynamic: it requires reasoning about uncertain future events where relevant information evolves over time. Forecasting is central to this process, requiring agents to adjust beliefs as new signals arrive. Unlike static fact retrieval, a robust forecaster must incorporate emerging evidence to prevent predictions from becoming obsolete. Figure 1 illustrates this process.

Prior studies explore LLM forecasting capabilities (Jin et al., 2020; Zou et al., 2022; Yuan et al., 2024a; Wang et al., 2025), but often frame it as a stationary problem. While some incorporate news (Ye et al., 2024; Halawi et al., 2024) or reinforcement learning for temporal understanding (Liu et al., 2025), they typically focus on prediction accuracy rather than their response to new information. Even confidence-aware benchmarks (Karger et al., 2024; Yuan et al., 2025) do not assess how models revise their forecasts when presented with information released after its training cutoff.

We introduce EVOLVECAST, a framework to evaluate how LLMs revise forecasts and confidence in response to new evidence. Unlike benchmarks measuring performance against a final resolution, EVOLVECAST measures alignment with aggregate human forecasts. We treat the human aggregate as a proxy for rational belief updating under uncertainty. This approach allows us to focus on the reasoning process of confidence updates rather than merely static accuracy.

We construct a benchmark of scenarios where information becomes available after model training cutoffs. We also evaluate a setting where models leverage previous human forecasts as context (see Figure 1), mimicking environments where forecasters observe market consensus. Our evaluation focuses on the direction and magnitude of prediction shifts and the calibration of updated confidence scores. Findings show that while LLMs often capture the correct update direction, their quantitative shifts are under-responsive and their confidence estimates fall short of human standards. EVOLVECAST provides a necessary framework for analyzing how models update forecasts as the world evolves.

2. EvolveCast: Evaluation Formalization

2.1. Task Definition

We formalize EVOLVECAST as a dynamic forecasting task. Let q denote a forecasting yes/no question about a future event. At time $t = 0$, we establish an initial belief p_0. When new information x_t becomes available at time t, the model produces a revised belief $p_t = P(\text{Yes} \mid q, x_t) \in [0, 1]$. The confidence update is defined as $\Delta p = p_t - p_0$. We then evaluate how well this update aligns with human shifts.

2.2. Evaluation Criteria

We assess the alignment between model updates (Δp) and reference updates (Δh) across three dimensions:

Directional Agreement. We use Mean Directional Accuracy (MDA) to measure if the model and reference forecasts

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

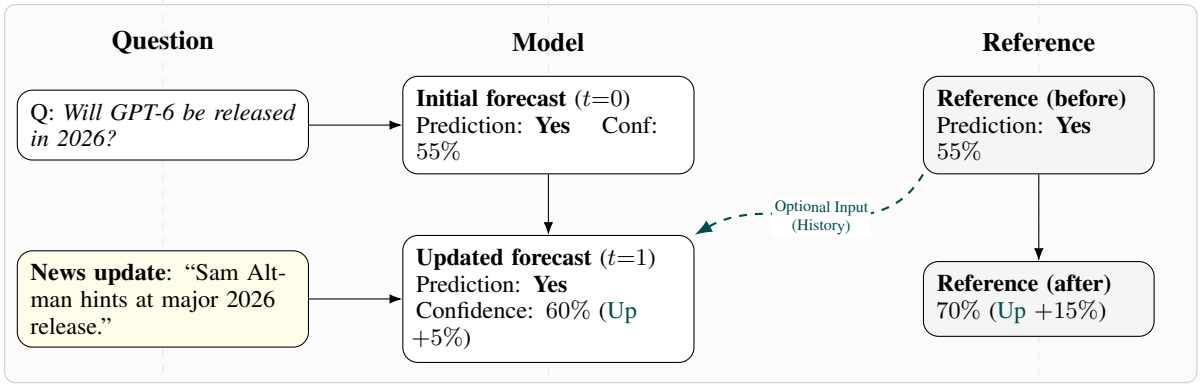


Figure 1. The EVOLVECAST framework evaluates the direction and magnitude of forecast updates ($P_0 \rightarrow P_t$) given new evidence (x_t). Here, the model successfully captures the human directional trend (`Up`) despite an under-responsive magnitude shift (+5% vs. +15%).

shift in the same direction:

$$\text{MDA} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\text{sign}(\tilde{\Delta}_p^{(i)}) = \text{sign}(\tilde{\Delta}_h^{(i)})\right]$$

where $\tilde{\Delta}$ represents a thresholded update to ignore negligible noise (see Appendix for details). We also report F1 scores for the directional labels `Up`, `Down`, and `Still`.

Magnitude Alignment. We measure the proximity of update magnitudes using Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\Delta p_i - \Delta h_i)^2$$

To evaluate relative shifts, we use Symmetric Mean Squared Percentage Error (SMSPE). We define percentage change δ as the difference divided by the mean of the two values:

$$\text{SMSPE} = \frac{1}{N}\sum_{i=1}^{N}\left(\delta_p^{(i)} - \delta_h^{(i)}\right)^2$$

Confidence Calibration. We evaluate calibration using the Brier Score: $\text{Brier}(p, h) = (p - h)^2$. To measure how calibration evolves, we calculate the average change in Brier scores before and after information is introduced:

$$\Delta\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}\left[(p_t^{(i)} - h_t^{(i)})^2 - (p_0^{(i)} - h_0^{(i)})^2\right]$$

A negative ΔBrier indicates that the model's confidence became better aligned with the reference standard after processing new information.

3. EvolveCast Dataset Construction

EVOLVECAST is constructed from Metaculus (`www.metaculus.com`), a platform where users submit probability estimates for future events. Metaculus aggregates these into a community prediction that is continuously updated until resolution. We filter for binary questions with at least 100 individual forecasts to ensure statistical reliability, spanning a wide range of domains..

3.1. Reference Forecasts and Confidence

Following prior work (Yuan et al., 2025), we use the aggregated human forecast distribution as a proxy for the latent probability of an event. For a community probability P, we compute a normalized confidence score:

$$h = \sigma\left(\frac{\ln P - \ln 0.5}{\ln 2}\right)$$

This scales the confidence relative to a 50% (chance-level) baseline. The reference update direction is then derived directly from the sign of the resulting confidence shift, $\Delta h = h_t - h_0$. While absolute aggregate human probabilities may occasionally carry idiosyncratic platform biases, the collective updates and directional shifts of the crowd reliably reflect rational belief revision as new signals emerge. Consequently, these aggregate human shifts serve as an ideal proxy for tracking evidence integration, prioritizing the logical trajectory of the belief update over static numerical perfection (Dezecache et al., 2022; Jeen et al., 2026).

3.2. News Retrieval and Alignment

We define each evaluation instance as a unique question-news pair, generated whenever a new community comment signals emerging evidence. For each comment timestamp, we retrieve relevant articles via the Google Search API and rank them by semantic similarity to the text using sentence embeddings. This pipeline yields 1,613 instances derived from 203 unique questions, allowing us to evaluate sequential belief updates over time for the same event. Manual validation confirms high relevance between the retrieved news and the forecasting topics.

Table 1. Main results. Top: **black-box** (verbalized) confidence. Bottom: **white-box** (logit-based) confidence. Columns are grouped as *Directional agreement* (MDA/Prec/Rec/F1), *Magnitude error* (MSE/SMSPE), and *Calibration change* (ΔBrier; negative is better).

Black-box (verbalized) confidence

	Directional agreement				Magnitude		Cal.
Model	MDA	Prec	Rec	F1	MSE	SMSPE	ΔBrier
Qwen backbone							
DS R1 Qwen-1.5B	0.2529	0.4413	0.2529	0.2559	0.0258	0.1033	0.0232
Qwen-2.5 1.5B	0.2372	0.4768	0.2372	0.2455	0.0262	0.1059	0.0238
DS R1 Qwen-7B	0.3534	0.4639	0.3534	0.3791	0.0232	0.0939	0.0210
Qwen-2.5 7B	0.2603	0.4361	0.2603	0.2654	0.0256	0.1015	0.0231
LLaMA backbone							
DS R1 LLaMA-8B	0.3360	0.4710	0.3360	0.3607	0.0237	0.0947	0.0214
LLaMA-3.1 8B	0.2199	0.3003	0.2199	0.2063	0.0267	0.1067	0.0244

White-box (logit-based) confidence

	Directional agreement				Magnitude		Cal.
Model	MDA	Prec	Rec	F1	MSE	SMSPE	ΔBrier
Qwen backbone							
DS R1 Qwen-1.5B	0.2461	0.4740	0.2461	0.2325	0.0260	0.1050	0.0236
Qwen-2.5 1.5B	0.2298	0.4505	0.2298	0.2360	0.0266	0.1065	0.0241
DS R1 Qwen-7B	0.3440	0.4422	0.3440	0.3650	0.0237	0.0951	0.0217
Qwen-2.5 7B	0.2545	0.4228	0.2545	0.2592	0.0259	0.1022	0.0234
LLaMA backbone							
DS R1 LLaMA-8B	0.3271	0.4595	0.3271	0.3510	0.0242	0.0960	0.0219
LLaMA-3.1 8B	0.2144	0.2925	0.2144	0.2018	0.0273	0.1082	0.0249

4. Experimental Setup

4.1. Models

We evaluate three reasoning models from the DeepSeek-R1 series (Guo et al., 2025): `DeepSeek-R1-Distill-Qwen` (1.5B and 7B) and `DeepSeek-R1-Distill-LLaMA-8B`, alongside their *base* counterparts (Qwen 2.5 (Bai et al., 2025) and Llama 3.1 (Dubey et al., 2024)). This selection allows us to isolate the effects of reasoning-style post-training across different architectures and parameter scales. All models use identical prompt structures and decoding settings.

4.2. Inference Procedure

Each question is evaluated under two conditions: (i) *question-only* and (ii) *question+news* (see Sec. 3). To avoid training cutoff leakage, we restrict our benchmark to questions posted after October 2023. We employ a dynamic anchoring approach: T_0 is the initial question date and T_1 is the news publication date. In both cases, the prompt explicitly instructs the model to treat the specified T as "today" and ignore information released after that date. This workflow mirrors real-world forecasting and ensures we measure belief revision ($\Delta p = p_t - p_0$) rather than simple fact recall.

We extract model confidence through two lenses:

Black-box (Verbalized) Confidence. The model provides a binary prediction and a probability on a 1–10 scale. This score is normalized to $[0, 1]$ to facilitate direct comparison with the human reference h. This mimics how human forecasters communicate uncertainty through explicit estimates (Xiong et al., 2023; Zhang et al., 2025).

White-box (logit-based) confidence. We also derive confidence estimates directly from the model's output probabilities. For a single generated answer $y = (y_1, \ldots, y_T)$, we compute the mean token probability $\hat{p} = \frac{1}{T}\sum_{t=1}^{T} P(y_t \mid y_{<t}, x)$, where $P(y_t \mid y_{<t}, x)$ is the model's predicted probability of token y_t given the context.

5. Results

Table 1 compares verbalized (black-box) and logit-based (white-box) confidence across settings. Neither approach emerges as a definitive winner: both show low directional agreement and poor calibration relative to human forecasters. While reasoning-tuned models (R1) consistently outperform base models, especially at larger scales (7B and 8B), the performance gap between black- and white-box methods within these models is marginal and inconsistent.

Several factors contribute to these distinct limitations. Verbalized probabilities show slightly greater stability across

Table 2. Ablation on news context. Models are evaluated with either only the most recent update (S = Single) or the full sequence of updates from T_0 to T_1 (A = Accumulated). Metrics are grouped as in Table 1.

	Directional agreement				Magnitude error		Calibration
Model	MDA	Prec	Rec	F1	MSE	SMSPE	ΔBrier
DS R1 Qwen-1.5B (S)	0.2529	0.4413	0.2529	0.2559	0.0258	0.1033	0.0232
DS R1 Qwen-1.5B (A)	0.2554	0.4564	0.2554	0.2512	0.0262	0.1052	0.0237
DS R1 Qwen-7B (S)	0.3534	0.4639	0.3534	0.3791	0.0232	0.0939	0.0210
DS R1 Qwen-7B (A)	0.3236	0.4565	0.3236	0.3473	0.0251	0.1024	0.0228
DS R1 LLaMA-8B (S)	0.3360	0.4710	0.3360	0.3607	0.0242	0.0960	0.0219
DS R1 LLaMA-8B (A)	0.3013	0.4687	0.3013	0.3196	0.0268	0.1083	0.0247

Table 3. Directional QA format results under the *Single* (S) and *Accumulated Updates* (A) settings. Metrics cover only directional agreement since this setting directly elicits Up/Down/Still labels.

Model	MDA	Prec	Rec	F1
DS R1 Qwen-1.5B (S)	0.3521	0.4401	0.3521	0.3762
DS R1 Qwen-1.5B (A)	0.3620	0.4366	0.3620	0.3838
DS R1 Qwen-7B (S)	0.4675	0.4538	0.4675	0.4581
DS R1 Qwen-7B (A)	0.4314	0.4387	0.4314	0.4320
DS R1 LLaMA-8B (S)	0.4923	0.4621	0.4923	0.4714
DS R1 LLaMA-8B (A)	0.4389	0.4383	0.4389	0.4351

prompt variations, but models frequently default to generic or conservative values (e.g., anchoring near 50%), which blunts the fidelity of their estimates. Logit-based confidences draw directly on internal activations but exhibit high sensitivity to prompt length, decoding hyperparameters, and normalization schemes, undermining their robustness. We observe a systematic pattern of conservatism across all models. This behavior suggests a fundamental tension between the model's pre-trained parametric memory and the external context. Instead of acting as Bayesian agents that update a prior based on the likelihood of new evidence, LLMs process news updates merely as additional tokens within the attention mechanism. The massive weight of pre-training data acts as an overwhelming anchor, preventing the short context window from shifting the posterior probability distribution as significantly as a rational update requires.

5.1. Ablation: Directional QA Prompting

To decouple core reasoning from the difficulties of numerical calibration, we prompt models to directly predict the update direction as a three-way classification task: `Up`, `Down`, or `Still`. This format bypasses the noise inherent in subtracting two separate probability estimates.

Table 3 shows that this direct method yields significantly higher MDA and F1 scores, particularly for larger scales. For instance, the DS R1 LLaMA-8B model achieves 0.49 MDA with single updates, a substantial jump from the 0.34 observed in the probabilistic setting. Consistent with the main results, accumulated news context again fails to improve performance for the 7B and 8B models, reinforcing that simply concatenating text does not translate to effective chronological belief integration. Crucially, the high performance in this directional setting serves as a vital validation of the benchmark itself. It confirms that the retrieved news snippets contain clear, accessible signals for updating beliefs. The underperformance in our primary probabilistic setting is therefore not a data retrieval failure, but rather stems from the models' inability to translate qualitative textual reasoning into calibrated probability shifts.

5.2. Ablation: Accumulated News Context

We investigate whether richer temporal context improves updating by providing the entire sequence of chronological news updates from T_0 to T_1 (Accumulated) rather than only the latest update (Single). As shown in Table 2, providing more history does not consistently improve performance. While the smallest model (Qwen-1.5B) shows minor gains in specific metrics, the larger Qwen-7B and LLaMA-8B models actually degrade in directional agreement when given the full sequence of updates.

This performance drop suggests that LLMs struggle to hierarchically weight evidence over time. Unlike human forecasters who dynamically prioritize recent signals and discount outdated info, models appear to treat all updates with similar salience, leading to diluted adjustments. Furthermore, longer contexts introduce volatility; models accumulate noise from the history, translating minor contextual variations into spurious updates even when reference forecasters remain stable. Extended qualitative analysis and confusion matrices are detailed in Appendix E.1.

6. Conclusion

We introduced EVOLVECAST to evaluate how LLMs revise forecasts with new evidence. While models often track update directions, their confidence shifts remain underresponsive compared to human standards, proving context retrieval alone cannot substitute for updating to news.

Impact Statement

This work evaluates how large language models update their forecasts when exposed to new information. Our experiments are conducted entirely on publicly available models and datasets. The forecasting questions and human reference data are sourced from Metaculus, an open platform with strict guidelines for question resolution and user conduct. The news snippets paired with questions are retrieved from publicly accessible web sources, and only short excerpts (title and headline) are included for research purposes. No private or sensitive data are used.

We recognize that forecasting research can have downstream societal implications. Forecasts produced by models may influence decision making in domains such as politics, economics, or health. Our goal is not to deploy automated forecasters, but to analyze their current limitations in belief updating and calibration. By highlighting where models fall short relative to human references, we aim to support the responsible use of AI in forecasting contexts and to discourage premature deployment of uncalibrated systems. All results should therefore be interpreted as a diagnostic study of model behavior, not as actionable forecasts.

References

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Dezecache, G., Dockendorff, M., Ferreiro, D. N., Deroy, O., and Bahrami, B. Democratic forecast: Small groups predict the future better than individuals and crowds. *Journal of Experimental Psychology: Applied*, 28(3):525, 2022.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Yu, L., Liu, Y., Li, J., Xiong, B., Xiong, D., et al. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*, 2023.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. *arXiv preprint arXiv:2402.18563*, 2024.

Jeen, S., Aitchison, M., and Mantic. Training llms to predict world events. *Thinking Machines Lab: News*, 2026. https://thinkingmachines.ai/news/training-llms-to-predict-world-events/.

Jin, W., Khanna, R., Kim, S., Lee, D.-H., Morstatter, F., Galstyan, A., and Ren, X. Forecastqa: A question answering challenge for event forecasting with temporal text data. *arXiv preprint arXiv:2005.00792*, 2020.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. Forecastbench: A dynamic benchmark of ai forecasting capabilities. *arXiv preprint arXiv:2409.19839*, 2024.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Liu, Z., Han, P., Yu, H., Li, H., and You, J. Time-r1: Towards comprehensive temporal reasoning in llms. *arXiv preprint arXiv:2505.13508*, 2025.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Roemmele, M., Bejan, C. A., and Gordon, A. S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI spring symposium series*, 2011.

Wang, Z., Zhou, X., Yang, Y., Ma, B., Wang, L., Dong, R., and Anwar, A. Openforecast: A large-scale open-ended event forecasting dataset. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5273–5294, 2025.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Ye, C., Hu, Z., Deng, Y., Huang, Z., Ma, M. D., Zhu, Y., and Wang, W. Mirai: Evaluating llm agents for event forecasting. *arXiv preprint arXiv:2407.01231*, 2024.

Yuan, C., Xie, Q., Huang, J., and Ananiadou, S. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM on Web Conference 2024*, pp. 1963–1974, 2024a.

Yuan, M., Chamoun, E., Aly, R., Whitehouse, C., and Vlachos, A. PRobELM: Plausibility ranking evaluation for language models. In *First Conference on Language Modeling*, 2024b. URL https://arxiv.org/pdf/2404.03818.

Yuan, Z., Ding, Z., and Vlachos, A. Forecast: The future outcome reasoning and confidence assessment benchmark. *arXiv preprint arXiv:2502.19676*, 2025.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a Machine Really Finish Your Sentence? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Zhang, C., Zhu, X., Li, C., Collier, N., and Vlachos, A. Reinforcement learning for better verbalized confidence in long-form generation. *arXiv preprint arXiv:2505.23912*, 2025.

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., and Hendrycks, D. Forecasting future world events with neural networks. *Advances in Neural Information Processing Systems*, 35: 27293–27305, 2022.

A. The Use of Large Language Models

Large language models were used in the preparation of this paper as writing assistants. Specifically, they were employed to refine phrasing, improve clarity, and suggest alternative structures for sections and subsections. Models were also used to draft prompt templates in an iterative process, which were then carefully reviewed, tested, and adjusted by the authors.

B. Reproducibility Statement

We have taken several steps to make our study as reproducible as possible. We provide a number of implementation details in the paper, including data construction steps, hyperparameter choices, and prompt templates. We also release the full set of processed question–news pairs used in our experiments, together with cleaned code, as supplementary material.

One limitation is that our news retrieval step relies on the Google Search API. Because search results may vary over time and are influenced by ranking algorithms, exact replication of the retrieval stage cannot be guaranteed. Despite this, we make available the aligned data used in our experiments so that downstream evaluation can be reproduced. We believe these materials, combined with the methodological details provided in the text, give future researchers the necessary resources to replicate and extend our work.

C. Example Metaculus Questions

To illustrate how human forecasts evolve over time, consider a question:

> *Q1: Will the US Senate weaken or eliminate the filibuster before January 3, 2029?*

For Q1, Figure 2 shows how community forecasts changed over time, while Figure 3 presents the histogram of the final forecast distribution.

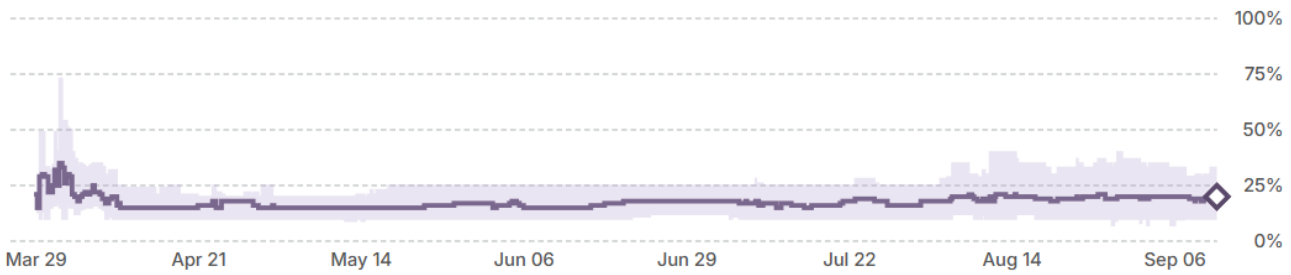


Figure 2. Community prediction trend for a Metaculus question on the US Senate filibuster issue.

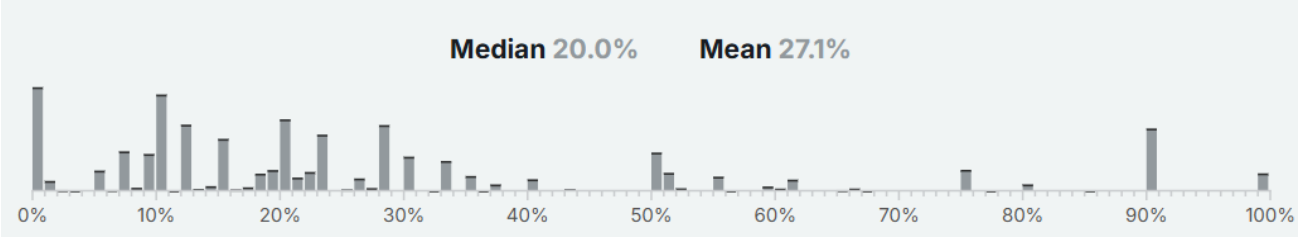


Figure 3. Histogram of final community forecasts.

D. Experimental Details

D.1. Implementation Details

Decoding parameters. All models are evaluated with identical sampling hyperparameters: `temperature` = 0.6 and `top_p` = 0.95. We set the maximum output length to `max_tokens` = 1024 for verbalized (black-box) confidence runs, and

max_tokens = 2048 for logit-based (white-box) runs.

Evaluation thresholds. Directional agreement metrics (Sec. **??**) use a minimal change threshold ϵ. For verbalized confidence we set ϵ=0, treating even small probability movements as valid shifts. For logit-based confidence we set ϵ=0.05 to reduce noise from minor fluctuations in token-level probabilities.

Implementation. All experiments are run with identical random seeds to ensure consistency across model conditions. In cases where models fail to produce valid JSON outputs, we assign the lowest possible score for that instance. A detailed list of hyperparameters and scripts used for inference are included in the supplementary code release.

News Retrieval Pipeline To align news with Metaculus comments, we filter out bot-generated or short entries (< 20 characters). We search for articles within a one-week window preceding a comment to account for publication delays. Using the Google Search API, we retrieve up to 100 articles per timestamp. We compute semantic similarity between the comment and article headlines using sentence embeddings (Reimers & Gurevych, 2019). We select the top-ranked article to serve as the associated news update.

Verbalized Confidence Anchors The 1–10 verbalized confidence scale uses the following anchors: 1: extremely unlikely, 2: very unlikely, 3: unlikely, 4: somewhat unlikely, 5: slightly unlikely, 6: slightly likely, 7: somewhat likely, 8: likely, 9: very likely, 10: extremely likely. Scores are mapped to $[0, 1]$ (e.g., $1 \rightarrow 0.1, 10 \rightarrow 1.0$) for comparison with reference probabilities.

D.2. Prompt Templates

We used standardized prompt templates across all models to ensure comparability. Each template contains three blocks: a <question> field with the forecasting question, a <metadata> field with the evaluation date and optional news update, and a <task> field describing the required output format. Prompts also include a <think> tag to encourage reasoning before producing the final answer when necessary.

Verbalized confidence. The model is instructed to provide a probability estimate on a 1–10 scale with descriptive anchors. We normalize the returned score to $[0, 1]$ for evaluation. An example template is:

```
<question>
${title}
</question>

<metadata>
- Today is ${start_time}.
- You do not have access to updates after ${start_time}.
- A recent news update:  "${best_news}"
</metadata>

<task>
You are an AI model predicting the likelihood of future events, now incorporating
recent news.
Your task is to estimate the probability of the following event occurring, given
the recent news update.

Provide your confidence score on a scale of 1 to 10, where:
1 - Extremely unlikely
2 - Very unlikely
3 - Unlikely
4 - Somewhat unlikely
5 - Neutral (50-50 chance)
6 - Somewhat likely
7 - Likely
8 - Very likely
9 - Extremely likely
```

```
10 - Almost certain

Return the confidence score in this format after thinking:
{ "confidence": X }
</task>
<think>
```

Logit-based confidence. The model is prompted to output a binary answer ("Yes"/"No"). Token-level probabilities are then extracted directly from the output distribution. An example template is:

```
<question>
${title}
</question>

<metadata>
- Today is ${start_time}.
- You do not have access to updates after ${start_time}.
- A recent news update: "${best_news}"
</metadata>

<task>
You are an AI model predicting the likelihood of future events, now incorporating
recent news.
Your task is to answer if the following event will occur, given the recent news
update.
You must also provide an answer with your best guess.

Return the answer in this format after thinking:
{ "answer": "Yes" / "No" }
</task>
<think>
```

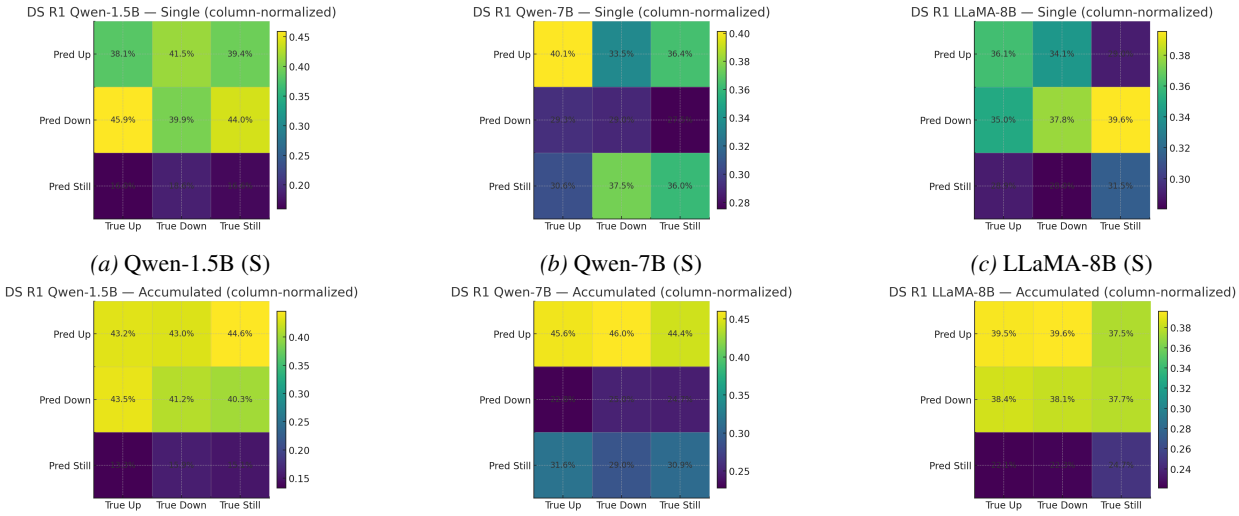
D.3. Additional Details

Rational Updating vs. Clairvoyance Evaluating models against a final binary resolution (e.g., whether an event occurred) can be misleading for belief dynamics. A model that guesses 100% probability early on might be "correct" about the result, but it would be hallucinating certainty given the limited evidence available at the time. By evaluating against the human aggregate, we distinguish rational updating (adjusting credence in proportion to evidence) from clairvoyance (predicting the outcome before sufficient evidence exists). This avoids penalizing models for maintaining justified uncertainty when evidence is thin, ensuring the metric measures reasoning quality rather than lucky guesses.

Comparison to Reasoning Benchmarks While our work involves reasoning under uncertainty, it differs from standard benchmarks like COPA (Roemmele et al., 2011), HellaSwag (Zellers et al., 2019), and PRobELM (Yuan et al., 2024b). These tasks evaluate causal inference or the ability to rank outcomes by plausibility based on static world knowledge. In contrast, EVOLVECAST requires dynamic belief revision where the "correct" answer is a shifting probability distribution rather than a static categorical choice.

Thresholding and Labeling Logic To ensure that negligible fluctuations in probability are not treated as meaningful updates, we apply a threshold $\epsilon = 0.05$ (5%) to all deltas. An update is only categorized as `Up` or `Down` if $|\Delta| \geq \epsilon$; otherwise, it is labeled `Still`. Precision, Recall, and F1 scores are calculated based on these three categorical labels to evaluate the model's qualitative alignment with human forecasters.

Metaculus Reliability and Filtering Metaculus community predictions on resolved binary questions show close alignment to observed outcomes, with roughly half of observed frequencies lying within the 90% credible interval around the ideal calibration line. To maintain high data quality, we discard questions marked as ambiguous or those where resolution criteria were modified post-submission. We only include questions with ≥ 100 individual forecasts to ensure the aggregated prediction is statistically robust.



(a) Qwen-1.5B (S) *(b)* Qwen-7B (S) *(c)* LLaMA-8B (S)

(d) Qwen-1.5B (A) *(e)* Qwen-7B (A) *(f)* LLaMA-8B (A)

Figure 4. Normalized confusion matrices for DeepSeek R1 models under Single (S) and Accumulated (A) news updates. Columns correspond to models; rows correspond to evidence settings. Values are column-normalized, showing $\mathrm{Pr}(\text{pred} \mid \text{true})$ in %.

Human Forecasts as Ground Truth Aggregated human forecasts provide informative signals of uncertainty reflecting the latent probability of an event at a specific time. Disagreement among forecasters reflects inherent ambiguity rather than noise, similar to how annotator disagreement is leveraged in tasks with multiple plausible interpretations. Aggregation smooths idiosyncratic biases while preserving collective uncertainty, providing a stable, dynamic signal suited for evaluating belief revision.

Data Validation We manually inspected a random subset ($N = 20$) of question-news pairs to validate the alignment process. In 90% of cases, the retrieved article was directly relevant to the forecasting topic and provided information post-dating the question start time. This suggests that model performance gaps are likely due to reasoning failures rather than retrieval noise.

Model Selection and Mapping We pair each DeepSeek-R1 distilled model with its corresponding base instruction model: R1-Distill-Qwen (1.5B and 7B) with Qwen 2.5, and R1-Distill-Llama-8B with Llama 3.1. This setup isolates the effects of reasoning-style post-training (distillation/RL) while controlling for architectural and scale-based variables. We exclude weaker, unrelated models to minimize variance caused by differing tokenizers or pretraining corpora.

Dynamic Anchoring and Leakage Mitigation By setting T_0 and T_1 relative to the question and news dates, we minimize the impact of "hindsight" where models might leverage future knowledge from pretraining. Our focus on the within-instance change (Δp) effectively cancels out static parametric knowledge and tests the model's direct responsiveness to evidence.

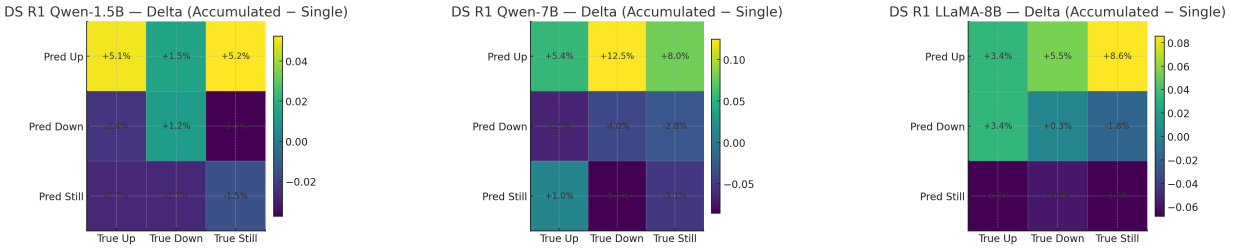
E. Additional Visualizations

E.1. Confusion Matrices and Error Analysis

To better understand the models' updating behavior, we visualize the confusion matrices for the Single vs. Accumulated conditions in Figure 4.

In the Single news condition, models show systematic confusion between `Up` and `Down` labels. When moving to the Accumulated condition, larger models (7B and 8B) develop a bias toward predicting movement even when humans remain stable. This suggests models interpret chronological history as a signal for change, leading to spurious updates rather than better integration of evidence. See Section E.2 for the corresponding delta heatmaps.

Figure 5. Delta confusion heatmaps (**A**−**S**) for **DS R1** models. Each plot shows the difference between column-normalized confusion matrices under Accumulated vs. Single updates, i.e., changes in $\Pr(\text{pred} \mid \text{true})$ (percentage points). Positive values indicate increased mass under Accumulated; negative values indicate decreased mass.



(a) DS R1 Qwen-1.5B (A−S) *(b)* DS R1 Qwen-7B (A−S) *(c)* DS R1 LLaMA-8B (A−S)

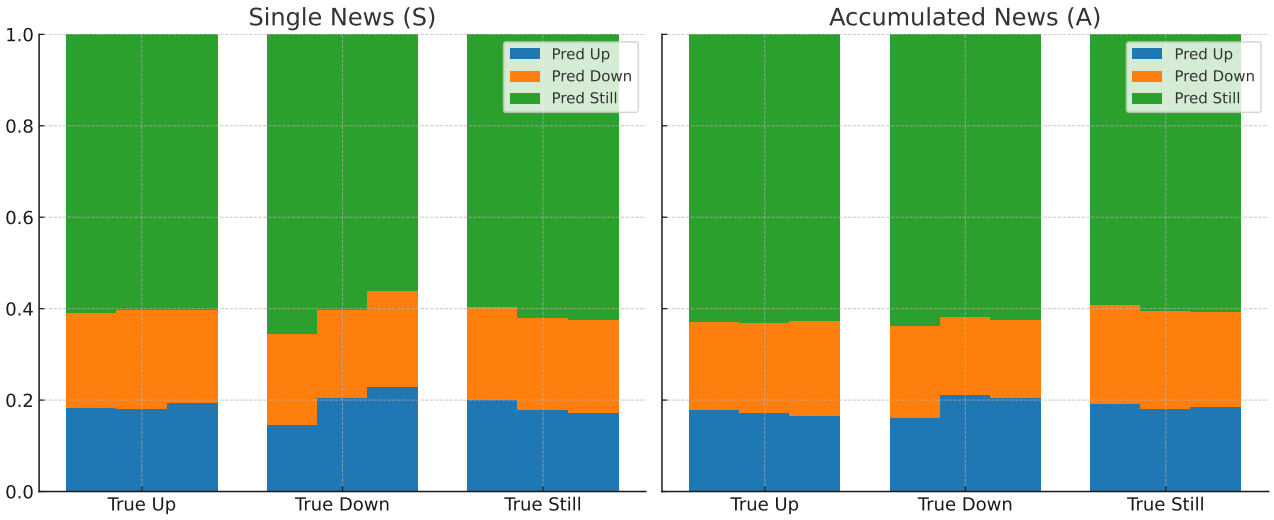


Figure 6. Normalized predicted distributions for the **Directional QA** setting. Bars show the distribution of predicted labels (Up = blue, Down = orange, Still = green) for each true label across DS R1 models under Single (S) and Accumulated (A) news conditions.

E.2. Delta Heatmaps: Accumulated minus Single

Figure 5 visualizes the difference between Accumulated and Single news contexts as *delta* heatmaps (A−S). Blue cells indicate reduced probability mass in the accumulated condition, while red cells indicate increased mass. Across models, the largest shifts occur in the "Still" column: accumulated updates tend to reduce correct "Still" predictions and redistribute probability into "Up" or "Down." This pattern complements the quantitative results in Table 2 and the confusion matrices in Fig. 4, confirming that accumulated context often introduces spurious directional movement rather than improving alignment with the reference.

E.3. Additional Visualizations for Direct Directional Prompting

The visualization complements Table 3 by highlighting systematic biases in how models allocate their predictions. The smaller DS R1 1.5B model produces noisy outputs and often overpredicts "Up," even when the true label is "Down" or "Still." By contrast, DS R1 7B and 8B show a strong conservative bias toward "Still" in the Single setting, but under Accumulated context they shift toward predicting "Up" more frequently. Together, these trends illustrate how accumulated evidence can introduce spurious movements and reduce directional accuracy, especially in larger models.

F. Direct Directional Prompting Template

For the direct directional prompting setting, models are asked to classify the impact of a news update as "Up," "Down," or "Still." This approach removes the need to compare before/after probabilities and directly probes whether the model can identify the directional effect of new information. The exact prompt is shown below:

```
<question>
${title}
</question>

<metadata>
- Today is ${start_time}.
- You do not have access to updates after ${start_time}.
- A recent news update:  "${best_news}"
</metadata>

<task>
You are an AI model analyzing how recent news impacts event predictions.
Your task is to determine whether the confidence in this event occurring should
increase, decrease, or remain the same after seeing the news.

Return the predicted_trend in this format after thinking:

{ "trend":  "Up" / "Down" / "Still" }
</task>
<think>
```

This minimal QA-style interface makes the task closer to classification benchmarks and avoids dependence on numeric probability estimation, which often suffers from poor calibration.

G. Additional Ablations

G.1. Alternative Confidence and Reference Signals

We briefly summarize two further ablations, with full results and visualizations later in Appendix G.2 and G.3: 1. *Similarity-Sensitive Confidence* implements a semantics-aware estimate (Kuhn et al., 2023) by clustering multiple generations into groups of similar answers and aggregating their confidence. In practice, since our task is binary, the clustering almost always reduces to two groups, limiting its usefulness. While such methods may prove more valuable in open-ended generation settings, here they do not offer additional insight over simpler probability estimates; 2. *Human Forecast Reference as Context* augments prompts with an aggregate human forecast at the corresponding time, providing the model with an explicit calibration anchor. However, we observe no measurable improvements, suggesting that models are unable to effectively exploit even strong external reference signals.

G.2. Similarity-Sensitive Confidence

In this ablation, we implement a semantics-aware confidence estimation method inspired by clustering-based approaches (Kuhn et al., 2023). The intuition is that if a model generates multiple plausible answers, the distribution of these generations can provide a more robust uncertainty estimate than any single output. Concretely, we sample N candidate outputs $\{y^{(1)}, \ldots, y^{(N)}\}$ and group them into K clusters based on cosine similarity of sentence embeddings (Reimers & Gurevych, 2019). Confidence for cluster C_k is then defined as

$$\hat{p}_k = \frac{\sum_{i:y^{(i)} \in C_k} P(y^{(i)} \mid x)}{\sum_{j=1}^{K} P(C_j)},$$

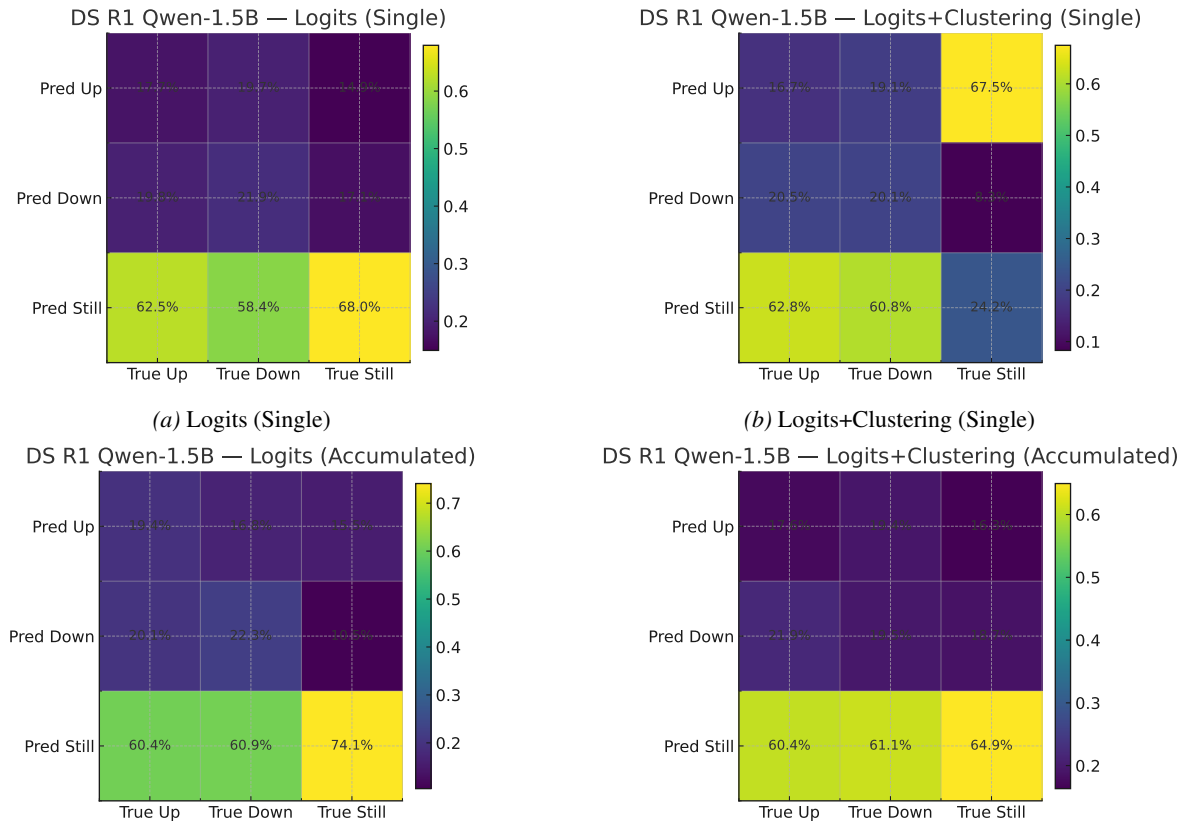
where $P(y^{(i)} \mid x)$ is the sequence-level probability of output $y^{(i)}$, computed as the average token probability across the sequence. For binary questions ($K = 2$), we report $\hat{p}_{\text{Yes}}$ as the final confidence estimate.

In practice, this approach does not yield benefits in EVOLVECAST, since binary Yes/No questions naturally collapse to two clusters. Thus, while clustering may offer richer signals in open-ended generation tasks (e.g., free-form QA or summarization), in binary forecasting it effectively reduces to re-labeling outputs without adding new information. Full quantitative results are shown in Table 4.

To further illustrate, Figs. 7 plots the confusion matrices for DS R1 Qwen-1.5B under both the baseline logits method and the clustering variant. The matrices show that clustering does not meaningfully shift the distribution of errors: the model still predicts "Still" excessively, and when it does move, the confusion between "Up" and "Down" persists.

Table 4. Comparison of logits-based confidence with and without semantic clustering for DS R1 Qwen-1.5B. Metrics are directional agreement (MDA/Prec/Rec/F1).

Method	MDA	Prec	Rec	F1
Logits (S)	0.252	0.494	0.252	0.236
Logits (A)	0.259	0.491	0.259	0.236
Logits+Clust. (S)	0.247	0.442	0.247	0.239
Logits+Clust. (A)	0.259	0.471	0.259	0.260



(a) Logits (Single) *(b)* Logits+Clustering (Single)

(c) Logits (Accumulated) *(d)* Logits+Clustering (Accumulated)

Figure 7. Normalized confusion matrices for DS R1 Qwen-1.5B under logits-based confidence and logits+clustering, for Single and Accumulated news. Values are column-normalized, showing $\Pr(\text{pred} \mid \text{true})$ (%). Clustering yields no qualitative change in error distribution.

G.3. Human Forecast Reference as Context

We also test an ablation where the model is provided with the contemporaneous aggregate human forecast as an additional context feature. Prompts are augmented with a line such as: *"Human forecast for this question on t is h_t%."* This setting

Table 5. Results for human forecast reference as context, comparing verbalized confidence and direct directional prompting under the Single Update setting. Direct prompting consistently outperforms verbalized confidence, but including human forecasts as anchors produces no measurable gains.

Model	Method	MDA	Precision	Recall	F1
DS R1 Qwen-1.5B	Verbalized	0.279	0.487	0.279	0.284
	Directional QA	0.350	0.441	0.350	0.376
DS R1 Qwen-7B	Verbalized	0.308	0.461	0.308	0.329
	Directional QA	0.457	0.462	0.457	0.457
DS R1 LLaMA-8B	Verbalized	0.307	0.479	0.307	0.327
	Directional QA	0.487	0.455	0.487	0.467

gives the model an explicit calibration anchor that, in principle, should simplify the task by showing where expert forecasters stood at the time.

The motivation is twofold. First, it simulates a collaborative human–AI forecasting scenario, where models can build on expert input rather than starting entirely from scratch. Second, it allows us to probe whether models meaningfully reason about how new evidence shifts beliefs relative to the anchor, or whether they simply mirror the human input.

Table 5 reports results under this setting for three DS R1 models using verbalized confidence (black-box) and direct directional prompting. While direct prompting again achieves higher directional agreement than verbalized probabilities, providing the human forecast reference itself does not yield any measurable improvement. This is somewhat surprising, as human forecasters rely heavily on such reference anchors; the lack of benefit here suggests that current models are unable to incorporate even strong external signals into their belief updating in a meaningful way.