Agentic Forecasting using Sequential Bayesian Updating of Linguistic Beliefs

Anonymous Authors[1]

Abstract

We present the Bayesian Linguistic Forecaster (BLF), an agentic system for binary forecasting that achieves state-of-the-art performance on the ForecastBench benchmark. The system is built on three ideas: (1) a *linguistic belief state* that combines a numeric probability with natural-language evidence summaries and is updated by the LLM at each step of an iterative tool-use loop; (2) *multi-trial aggregation* via shrunken logit-space averaging; and (3) *hierarchical Platt calibration* with per-source intercepts, which avoids over-shrinking extreme predictions on sources with skewed base rates. On 400 questions from the ForecastBench leaderboard, BLF outperforms all top public methods, including Cassi, GPT-5, Grok 4.20, and Foresight-32B, and is the only method to significantly beat a strong LLM-free crowd+empirical-prior baseline. Mixed-effects ablations across five base LLMs show that all three components contribute, with their relative importance depending on the base model and on whether a crowd anchor is available.

1. Introduction

We present BLF (Bayesian Linguistic Forecaster), an agentic system that achieves a new state of the art (SOTA) on ForecastBench (FB) (Karger et al., 2025). Our approach has three key ideas:

- **Linguistic belief states.** Most forecasting agents either search in parallel and reason once, or sequentially accumulate raw search results in context. BLF instead maintains a semi-structured belief state — a probability paired with natural-language evidence summaries — updated by the LLM at each step. We call this "Bayesian-style" updating: the slots mirror a sequential Bayesian update, but the actual update is an LLM forward pass and need not satisfy formal coherence (Qiu et al., 2025; Falck et al., 2024).

- **Multi-trial aggregation.** LLM forecasts have high variance across runs. We run $K{=}5$ independent trials and aggregate in logit space, with optional shrinkage toward an empirical or uniform prior, inspired by James–Stein / empirical Bayes.

- **Hierarchical calibration.** We apply Platt scaling (Platt, 1999) with per-source intercept offsets; this is critical when empirical priors induce source-specific biases, which global Platt scaling would over-shrink.

On 400 backtesting questions from FB, BLF achieves the highest Brier Index (BI) on every question type (Overall, Market, Dataset) compared to all top FB methods — including agentic systems with tools and search (Cassi (Cassi AI, 2025), Grok 4.20, Foresight-32B (Lightning Rod Labs, 2025)) and strong zero-shot baselines (GPT-5, Gemini-3.1-Pro). BLF is also the only method that significantly beats a strong LLM-free crowd+empirical-prior baseline (Tab. 1). On the difficulty-adjusted Brier Index, BLF (ABI=71.0) is close to the human superforecaster median (ABI=70.9) reported on the FB leaderboard.[1]

Beyond these three methodological ideas, we contribute (i) a four-layer date-leakage defense for backtesting with 1.5% residual leakage, and (ii) a paired mixed-effects analysis that quantifies each component's contribution while controlling for question difficulty (which accounts for 62% of the performance variance).

[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.

An extended version of this work, with additional analyses and a ∼50-page appendix, is concurrently under double-blind review at NeurIPS 2026.

[1]Human superforecaster value from https://www.forecastbench.org/leaderboards/; computed on a different question set, so not strictly comparable.

Algorithm 1 BLF agent loop

Require: Question q, cutoff date d, max steps T
$b_0 \leftarrow$ initial belief (p=0.5); $o_0 \leftarrow \emptyset$; $m_0 \leftarrow (q, b_0)$
for $t = 1, \ldots, T$ **do**
  $(a_t, b_t) \leftarrow \mathrm{LLM}(m_{t-1})$ *// update belief, choose action*
  if $a_t = \texttt{submit}(\hat{p})$ **then**
    return $\hat{p}$
  end if
  $o_t \leftarrow \mathrm{Env}(a_t; q, d)$ *// execute action*
  if a_t is web search **then**
    $o_t \leftarrow \mathrm{LeakFilter}(o_t, d)$ *// 2nd-pass LLM filter*
  end if
  $m_t \leftarrow m_{t-1} \oplus (a_t, o_t, b_t)$
end for
Force submit; **return** $b_T.p$

2. Method

BLF is an agentic system for binary question forecasting. Given a question q with cutoff date d (set equal to the forecast date f to avoid leakage), BLF runs an iterative tool-use loop (Algorithm 1) that maintains a structured *belief state*, aggregates K independent trials, and applies hierarchical calibration. We use Gemini-3.1-Pro as the default base LLM and Brave as the search engine, but evaluate alternatives in Sec. 3.

Belief state. At each step t, the LLM reads the message history m_{t-1} and produces both an action a_t and an updated belief b_t in a single generation: $(a_t, b_t) = \mathrm{LLM}(m_{t-1})$. The belief b_t is a semi-structured JSON object containing a probability $p \in [0, 1]$, a confidence level, key evidence for/against, and open questions. This contrasts with two common alternatives: (i) *NoBel*, which simply appends search/tool results to a growing context (no belief state); and (ii) *Batch*, which issues all search queries in parallel and reasons once at the end (no sequential search).

Agent loop and tools. At each step the agent selects from `web_search` (with automatic leak filtering), `summarize_results` (read and summarize retrieved pages), `lookup_url`, source-specific data tools (e.g. `fetch_ts_yfinance`, `fetch_wikipedia_section`), or `submit`. The loop runs for up to $T_{\max}$=10 steps; on `submit`, probabilities are clamped to $[0.05, 0.95]$ to bound worst-case Brier loss. A meta-controller selects the available tool set per question type. For market questions, the crowd price is injected as a prompt anchor; for dataset questions, a per-source empirical prior π_q provides a base-rate baseline.

Multi-trial aggregation. We run K=5 independent trials per question and consider two aggregators: (i) arithmetic mean, $\hat{p} = \frac{1}{K}\sum_k p_k$; and (ii) shrunken logit-space mean, $\hat{p} = \sigma\big(\alpha \cdot \frac{1}{K}\sum_k \mathrm{logit}(p_k)\big)$, where $\alpha \in [0, 1]$ depends on the between-trial variance. When α<1, predictions shrink toward the prior, reducing overconfidence when trials disagree. We find that logit-space averaging beats arithmetic averaging, and that adding shrinkage further helps when trials are noisy.

Hierarchical calibration. We map raw forecasts to calibrated probabilities via Platt scaling, selecting the L2 regularization strength by leave-one-out CV. On FB we use *hierarchical* Platt with per-source intercept offsets δ_s (also L2-regularized). The hierarchical variant matters most in the zero-shot regime when the empirical prior is on: global Platt over-shrinks extreme predictions on sources with skewed base rates, whereas hierarchical Platt preserves them.

3. Experiments and Results

Setup. Our two FB evaluation tranches (A: 2025-10-26, B: 2025-11-09) cover 400 unique questions and 791 binary resolution events (market questions resolve at a single date; dataset questions resolve at multiple horizons). Tranche dates were chosen to (1) lie after current frontier-LLM knowledge cutoffs, (2) maximize resolved questions, and (3) maximize overlap with publicly submitted forecasts to enable paired comparisons. Our primary metric is the Brier Index (Kucinskas et al., 2026), BI $= 100 \times (1 - \sqrt{\overline{\mathrm{BS}}})$, where higher is better and always-0.5 scores 50.

Comparison to SOTA. Fig. 1 compares BLF (Pro) against the top FB methods. BLF achieves the highest BI on every partition. Tab. 1 reports paired bootstrap comparisons that control for question difficulty (which absorbs 62% of the variance). BLF is significantly better than Cassi, GPT-5, and Foresight on Overall BI, and on Dataset questions where the sample is large. The exception is Grok: BLF leads in point estimate (73.3 vs. 70.5), but the paired bootstrap on Grok's single-tranche coverage is statistically indistinguishable (p=0.51) due to low sample size and high variance. Similarly, on Market questions, the paired bootstrap has lower power because only 200 events are available on A∪B, so most Market gaps do not reach significance despite BLF's lead in point estimates (83.8 vs. 79.5–82.0). The bottom block of Tab. 1 shows that BLF is the only method to beat the LLM-free baseline on Overall BI by a statistically significant margin (+3.4**).

Component ablation across base LLMs. We ablate each of the three BLF components (belief state, shrinkage aggregation, hierarchical calibration) and swap the base LLM among Pro-3.1, Flash-3.1, Sonnet-4.6, GPT-5, and Kimi-K2.5. Comparisons use a paired bootstrap mixed-effects analysis that controls for question difficulty; methods are

Table 1. Brier Index (higher = better) on FB A∪B for BLF (Pro), four external SOTA methods, and an LLM-free crowd+empirical-prior baseline. All methods use the crowd signal for market questions. Top block: absolute BI per (question type, tranche) plus paired Δ BI vs BLF (Δ<0 means BLF wins). Bottom block: paired Δ BI of each LLM-based method (incl. BLF) vs the crowd+empirical baseline (Δ>0 means it beats the baseline). Tranche A: 2025-10-26; B: 2025-11-09; All: A∪B. Grok submitted only A and Foresight only B; missing tranches are ×. Significance: $^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$, $^{ns}p\geq0.05$ (paired bootstrap, 5,000 resamples).

	Market				Data				All			
Method	A	B	All	Δ	A	B	All	Δ	A	B	All	Δ
BLF (Pro)	81.3	86.7	**83.8**		62.4	63.1	**62.7**		71.9	74.9	**73.3**	
Cassi	80.1	84.2	82.0	-1.7^{ns}	61.8	57.4	59.6	-3.1^{***}	71.0	70.8	70.8	-2.4^{*}
GPT-5 (ZS)	79.8	80.9	80.4	-3.4^{ns}	61.1	59.0	60.1	-2.7^{**}	70.5	70.0	70.2	-3.0^{*}
Grok	79.5	×	79.5	-1.9^{ns}	61.4	×	61.4	-1.0^{ns}	70.5	×	70.5	-1.4^{ns}
Foresight	×	81.0	81.0	-5.7^{***}	×	57.6	57.6	-5.5^{***}	×	69.3	69.3	-5.6^{***}
Baseline (no LLM, crowd+emp)	81.4	81.6	81.5		56.7	59.9	58.3		69.1	70.8	69.9	
Method vs. Baseline (no LLM): Δ>0 means the method beats the baseline.												
BLF (Pro)				$+2.3^{ns}$				$+4.5^{***}$				$+3.4^{**}$
Cassi				$+0.6^{ns}$				$+1.3^{ns}$				$+1.0^{ns}$
GPT-5				-1.1^{ns}				$+1.8^{*}$				$+0.3^{ns}$
Grok				-1.9^{ns}				$+4.7^{***}$				$+1.4^{ns}$
Foresight				-0.6^{ns}				-2.3^{ns}				-1.5^{ns}

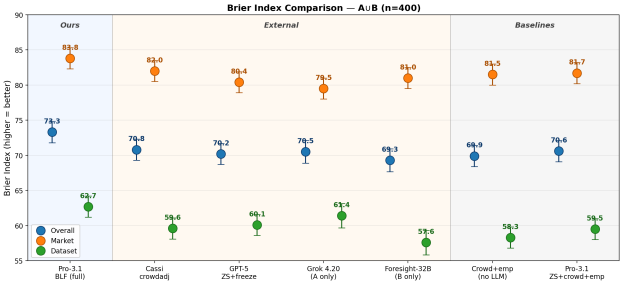

Figure 1. Comparison to external methods and baselines on FB A∪B in terms of Brier Index, split by Overall, Market, and Dataset. Error bars: 95% bootstrap CI. See Tab. 1 for pairwise significance.

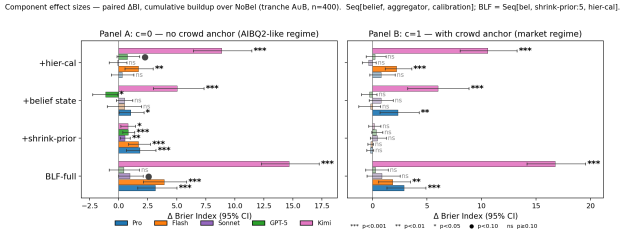

Figure 2. Cumulative buildup from a NoBel reference (search-enabled agent loop with sequential text accumulation, no belief state, no shrink-prior aggregation, no calibration) to BLF-full. The first three bars each swap a single axis on top of the previous treatment, in the notation Seq[*belief*, *aggregator*, *calibration*], so they sum to the cumulative *BLF-full* (vs. NoBel) bar at the bottom.

compared to a strong sequential search baseline similar to AIA (Alur et al., 2025). We test two regimes: **c=0** (no crowd anchor) and **c=1** (with crowd anchor); see Fig. 2.

The full BLF stack helps every model, but most dramatically Kimi K2.5, which improves from 64.4 (calibrated NoBel baseline) to 70.2 in the c=0 regime (+5.8 BI), and from 65.8 to 72.0 in the c=1 regime (+6.2 BI). The final BLF+Kimi performance is close to the overall best (BLF+Pro at 73.3), despite Kimi being ∼5× cheaper and an open-weights model. Kimi's calibrated-NoBel BI is ∼5 below Pro's, but this is itself the residual of a much larger gap on the *un*calibrated NoBel+Kimi baseline (∼12): hierarchical calibration alone closes a lot of that gap, but the further BLF additions add another ∼6 BI on top. BLF also helps Pro and Flash significantly, but the improvement on Sonnet and GPT-5 is not statistically significant despite comparable NoBel baselines.

Belief evolution. Fig. 3 shows how the agent's probability evolves across steps for 5 independent trials on the question "Will WorldAtlas.com display the Gulf of America before July 1, 2025?" (outcome: No). All trials start at p_0=0.5 but quickly diverge as different search queries surface different evidence. The high inter-trial variance (σ=0.20) motivates multi-trial aggregation: the trial mean (0.57) is closer to the true outcome than most individual trials. The lone "skeptic" trial discovers at step 9 that WorldAtlas uses static map images (not dynamic Google Maps), and revises sharply downward to p=0.28.

Tool usage. Fig. 4 shows the mean number of tool calls per question on FB A∪B, broken down by source. Web search dominates across all sources; source-specific data tools (market info, Wikipedia sections, time-series fetchers) are used selectively for their respective sources. DBnomics questions (which require predicting if the temperature at a future date f at a certain location x will be higher or lower

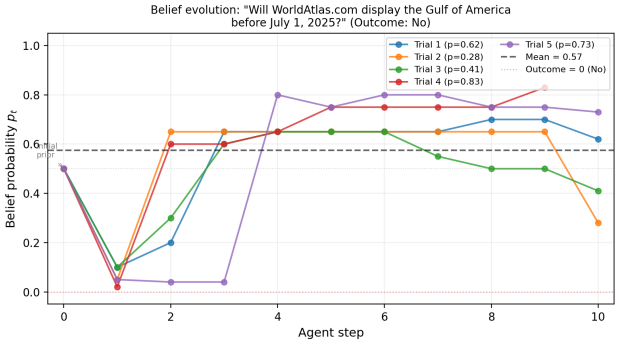

Figure 3. Belief evolution across 5 trials on a single AIBQ2 question (outcome: No). All trials start at p_0=0.5 and dip low at steps 1–2 (initial search finds skepticism), then diverge. High inter-trial variance (σ=0.20) motivates aggregation.

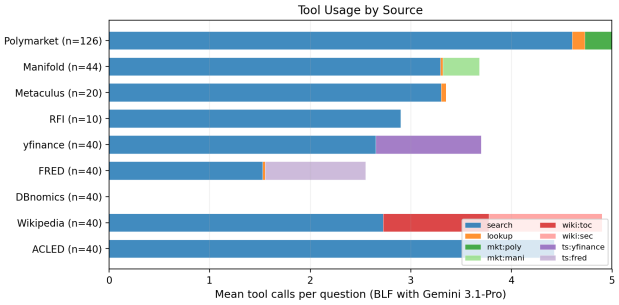

Figure 4. Mean tool calls per question by source for `BLF` (trial 1, A∪B). Source-specific tools are used selectively. DBnomics bypasses the LLM via a KNN model.

than its current value) are handled entirely by a simple KNN model applied to the retrieved historical temperature timeseries, since we found the LLM did poorly at this task. Polymarket questions use the most tools (∼5.5 per question), reflecting the additional market-info fetch on top of web search.

4. Related Work

Benchmarks. Zou et al. (2022) introduced Autocast; Halawi et al. (2024) showed retrieval-augmented GPT-4 approaches the human crowd's Brier on 5K+ questions. ForecastBench (Karger et al., 2025) provides a rolling benchmark with difficulty-adjusted scoring (Kucinskas et al., 2025). TFRBench (Ahamed et al., 2026), FutureX (Zeng et al., 2025), Metaculus FutureEval (Metaculus, 2026), and OpenEP (Guan et al., 2024) target reasoning, contamination resistance, and open-ended outcomes respectively. Paleka et al. (2025) and Li et al. (2026) analyse temporal-leakage and ignorance-simulation pitfalls; our four-layer leakage defense addresses several of these.

Agentic and tool-augmented systems. Hsieh et al. (2024) (RTF), AIA (Alur et al., 2025), CogForecast (Wang et al., 2025), and BayesAgent (Huang et al., 2026) all target agentic forecasting. AIA in particular uses adaptive iterative search with Platt calibration, but aggregates raw text rather than maintaining a structured belief state, and uses an LLM-based calibration that we found inferior to mean aggregation. BayesAgent constructs a verbalized PGM per query; `BLF` instead maintains a non-Markovian natural-language belief state updated from tool-call evidence, avoiding an explicit graph.

Zero-shot, prompt-based, and RL methods. Pratt et al. (2024) and Schoenegger et al. (2025) report negative results for sophisticated prompting on top of base models, motivating our agentic + calibration approach. Recent RL methods — Time-R1 (Liu et al., 2025), RLVR (Turtel et al., 2025), OpenForecaster (Chandak et al., 2026), Foresight Learning (Turtel et al., 2026) — improve the *base model*, while we improve the *agent harness*; the two directions are complementary.

Ensembles. Schoenegger et al. (2024) show that aggregating 12 diverse LLMs matches human-crowd accuracy. Consistent with their finding that ensemble gains require genuine diversity, our negative result that within-architecture ensembling does not help (see Sec. A).

5. Conclusion

We presented `BLF`, an agentic forecasting system that combines agentic tool use, structured belief updates, multi-trial aggregation with logit-space shrinkage, and hierarchical Bayesian calibration. On ForecastBench it outperforms all top public methods and is the only one to significantly beat a strong LLM-free baseline. Limitations include the restriction to binary outcomes and backtesting-only evaluation. Useful extensions include categorical and numerical outcomes, better tools (e.g. TimesFM (Das et al., 2024)), offline fine-tuning of the base model using RL, or online learning of the tool-use policy using bandits, and live evaluation on ProphetArena (Yang et al., 2025).

Impact Statement

This paper presents improved methods for probabilistic forecasting. Such systems can be applied to sensitive topics (e.g., geopolitical or military events) and should be audited by humans before deployment in such contexts.

References

Ahamed, M. A., Parmar, M., Goyal, P., Song, Y., Le, L. T., Cheng, Q., Li, C.-L., Palangi, H., Yoon, J., and Pfister, T. TFRBench: A reasoning benchmark for evaluating forecasting systems. *arxiv*, 2026. URL `https:`

//arxiv.org/abs/2604.05364.

Alur, R., Stadie, B. C., Kang, D., Chen, R., McManus, M., Rickert, M., Lee, T., Federici, M., Zhu, R., Fogerty, D., Williamson, H., Lozinski, N., Linsky, A., and Sekhon, J. S. AIA forecaster: Technical report. *arxiv*, 2025. URL https://arxiv.org/abs/2511.07678.

Cassi AI. Cassi: AI-powered forecasting. https://cassi-ai.com/, 2025.

Chandak, N., Goel, S., Prabhu, A., Hardt, M., and Geiping, J. Scaling open-ended reasoning to predict the future. *arxiv*, 2026. URL https://arxiv.org/abs/2512.25070.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting. In *ICML*, 2024. URL https://arxiv.org/abs/2310.10688.

Falck, F., Wang, Z., and Holmes, C. Is in-context learning in large language models bayesian? a martingale perspective. In *ICML*, June 2024. URL https://arxiv.org/abs/2406.00793.

Guan, Y., Peng, H., Wang, X., Hou, L., and Li, J. OpenEP: Open-ended future event prediction. *arxiv*, 2024. URL https://arxiv.org/abs/2408.06578.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. *arxiv*, 2024. URL https://arxiv.org/abs/2402.18563.

Hsieh, E., Fu, P., and Chen, J. Reasoning and tools for human-level forecasting. *arxiv*, 2024. URL https://arxiv.org/abs/2408.12036.

Huang, H., Shen, X., Wang, S., Meng, L., Liu, D., Duchene, D. A., Wang, H., and Bhatt, S. BayesAgent: Bayesian agentic reasoning under uncertainty via verbalized probabilistic graphical modeling. In *AAAI*, 2026. URL https://arxiv.org/abs/2406.05516.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *ICLR*, 2025.

Kucinskas, S., Bastani, H., and Karger, E. ForecastBench: An updated ranking methodology, 2025. URL https://forecastbench.org/assets/pdfs/forecastbench_updated_methodology.pdf.

Kucinskas, S., Bastani, H., and Reynolds, M. Making forecasting scores easier to interpret: Introducing the brier index, 2026. URL https://forecastingresearch.substack.com/p/introducing-the-brier-index.

Li, Z., Wang, Y., Lahib, A. E., Xia, Y.-J., and Pi, X. Simulated ignorance fails: A systematic study of LLM behaviors on forecasting problems before model knowledge cutoff. *arxiv*, 2026. URL https://arxiv.org/abs/2601.13717.

Lightning Rod Labs. Foresight-32b: An LLM-based forecasting system, 2025. URL https://blog.lightningrod.ai/p/using-the-future-to-train-prediction-models.

Liu, Z., Han, P., Yu, H., Li, H., and You, J. Time-R1: Towards comprehensive temporal reasoning in LLMs. *arxiv*, 2025. URL https://arxiv.org/abs/2505.13508.

Metaculus. FutureEval: Continuously updated AI forecasting benchmark, 2026. URL https://www.metaculus.com/futureeval/.

Paleka, D., Goel, S., Geiping, J., and Tramer, F. Pitfalls in evaluating language model forecasters. *arxiv*, 2025. URL https://arxiv.org/abs/2506.00723.

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

Pratt, S., Blumberg, S., Carolino, P. K., and Morris, M. R. Can language models use forecasting strategies? *arxiv*, 2024. URL https://arxiv.org/abs/2406.04446.

Qiu, L., Sha, F., Allen, K., Kim, Y., Linzen, T., and van Steenkiste, S. Bayesian teaching enables probabilistic reasoning in large language models. *Nat. Commun.*, March 2025. URL http://arxiv.org/abs/2503.17523.

Schoenegger, P., Tuminauskaite, I., Park, P. S., and Tetlock, P. E. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *arxiv*, 2024. URL https://arxiv.org/abs/2402.19379.

Schoenegger, P., Jones, C. R., Tetlock, P. E., and Mellers, B. Prompt engineering large language models' forecasting capabilities. *arxiv*, 2025. URL https://arxiv.org/abs/2506.01578.

Turtel, B., Franklin, D., Skotheim, K., Hewitt, L., and Schoenegger, P. Outcome-based reinforcement learning to predict the future. *arxiv*, 2025. URL https://arxiv.org/abs/2505.17989.

Turtel, B., Wilczewski, P., Franklin, D., and Skothiem, K. Future-as-label: Scalable supervision from real-world outcomes. *arxiv*, 2026. URL https://arxiv.org/abs/2601.06336.

Wang, Z., Zhou, X., Yang, Y., Ma, B., Wang, L., Dong, R., and Anwar, A. Beyond inherent cognition biases in LLM-based event forecasting: A multi-cognition agentic framework. In *Findings of EMNLP*, 2025. URL https://aclanthology.org/2025.findings-emnlp.258/.

Yang, Q., Mahns, S., Li, S., Gu, A., Wu, J., and Xu, H. LLM-as-a-prophet: Understanding predictive intelligence with prophet arena. *arxiv*, 2025. URL https://arxiv.org/abs/2510.17638.

Zeng, Z., Liu, J., Chen, S., He, T., Liao, Y., Tian, Y., Wang, J., Wang, Z., Yang, Y., Yin, L., Yin, M., Zhu, Z., Cai, T., Chen, Z., Chen, J., Du, Y., Gao, X., Guo, J., Hu, L., Jiao, J., Li, X., Liu, J., Ni, S., Wen, Z., Zhang, G., Zhang, K., Zhou, X., Blanchet, J., Qiu, X., Wang, M., and Huang, W. FutureX: An advanced live benchmark for LLM agents in future prediction. *arxiv*, 2025. URL https://arxiv.org/abs/2508.11987.

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. In *NeurIPS (Datasets and Benchmarks)*, 2022. URL https://arxiv.org/abs/2206.15474.

A. Additional method details

Belief-state schema. The belief b_t produced at each step is a JSON object with the following fields: a probability estimate $p \in [0, 1]$, a confidence label (low/medium/high), a list of supporting evidence items each with a one-line summary and provenance (URL or tool call), a list of contradicting evidence items, and a list of open questions still to be resolved. The LLM is asked to update each slot incrementally: new evidence is appended, redundant items are merged, and resolved open questions are removed. The probability p is treated as the agent's posterior under the current evidence and is what gets returned on `submit`.

Shrinkage aggregator. With K trials yielding probabilities $p_1, \ldots, p_K$ and a prior π (uniform 0.5 or empirical π_q), let $\ell_k = \text{logit}(p_k)$ and $\bar{\ell} = \frac{1}{K}\sum_k \ell_k$. We define $\hat{\ell} = \alpha\bar{\ell} + (1-\alpha)\,\text{logit}(\pi)$ where α is a per-question shrinkage parameter given by $\alpha = max(f, 1 - cs)$, where $s^2 = \frac{1}{K}\sum_k(\ell_k - \bar{\ell})^2$ is the inter-trial variance, and f, c are hyper-parameters chosen by LOO-CV. This estimator (which is a heuristic approximation to an empirical Bayes estimator) reduces to plain logit-space averaging when α=1 and to the prior when α=0.

Hierarchical Platt scaling. For each raw prediction p from source s we use the calibrated forecast $q = \sigma(a\,\text{logit}(p) + b + \delta_s)$ where a, b are global slope/intercept, and δ_s is a per-source intercept offset. We optimize (a, b, δ_s) wrt binary cross-entropy, adding L2 regularization on $(a - 1, b, \{\delta_s\})$ centered at the identity transform. Regularization strengths are selected by leave-one-out cross-validation on the calibration set. Compared to global Platt, the per-source intercept lets sources with extreme empirical base rates (e.g. Polymarket markets that resolve heavily in one direction) avoid having their well-calibrated extreme predictions shrunk back toward 0.5.

Leakage defense. To make backtesting reliable we layer four defenses against information that post-dates the cutoff d: (1) the search engine's freshness filter constrains results to $\leq d$; (2) a second-pass LLM classifier flags results that mention post-d events; (3) data tools (yfinance, FRED, etc.) clamp returned series to $\leq d$; (4) known resolution-source URLs (e.g. the Polymarket question page itself) are blacklisted. A post-hoc audit on a $\sim$200-question sample finds 1.5% residual leakage, substantially below typical published forecasting evaluations.

B. Additional experimental details

Datasets. Initial development used the test set of 113 binary questions from the Q2 2025 Metaculus AI Benchmark Tournament (AIBQ2). Because AIBQ2 is small, we conducted a much larger study using two ForecastBench tranches (A: 2025-10-26, B: 2025-11-09), each with 100 market and 100 dataset questions, yielding 400 unique questions and 791 binary resolution events. Market sources: Polymarket, Manifold, Metaculus, Rand Forecasting Initiative. Dataset sources: yfinance, FRED, DBnomics, Wikipedia, ACLED.

Mixed-effects analysis. For each method comparison we fit a mixed-effects model with random intercepts per question and report a paired bootstrap CI over the fixed-effect difference, resampled at the question level ($5{,}000$ resamples). This controls for question difficulty, which (in an ANOVA decomposition) accounts for 62% of the total variance in BI across (method, question) pairs — far more than method or LLM choice. All significance markers in Tab. 1 use this procedure.

Cross-LLM ablation. The cross-LLM study covers Gemini-3.1-Pro, Gemini-3.1-Flash, Claude Sonnet-4.6, GPT-5, and Kimi K2.5. Each model is evaluated under five treatments that vary one component at a time: NoBel-mean-uncal (reference baseline), NoBel-mean-hcal (+hier-cal) Bel-mean-hcal (+belief-state) and Bel-shrink-hcal (+shrink-prior). The last line is the full `BLF` system compared to the reference. The per-model effect-size buildup in Fig. 2 reveals several qualitative patterns: (1) for models with poor NoBel-uncal BI (e.g.,Kimi, Flash), the largest single component is hierarchical calibration, which closes most of the headline gap; (2) after calibration, several models (Pro, Kimi) gain additional performance from the belief state; (3) for some models with already-strong NoBel baselines (Sonnet, GPT-5), the marginal contribution of the `BLF` stack is small and not statistically significant; (4) all models benefit from hierarchical shrinkage in the absence of a crowd baseline. Source-type analysis further shows that the `BLF` stack contributes most on market questions (where careful sequential search is required) and least on dataset questions (where a single well-chosen tool call is often sufficient).