When Does Evidence Help Prompted LLM Forecasting? Evidence Access and Prompt Structure Across 12 Models

Akram Naoufel Tabet [1] **Mitja Luštrek** [1]

Abstract

We present a controlled study of how evidence access and prompt structure shape LLM-based forecasting. Using 114 resolved binary questions from ForecastBench, we evaluate 12 models across three prompting strategies: a direct control prompt, a base-rate prompt that asks models to anchor on historical frequency, and a Bayesian prompt that asks models to state a prior and update on evidence. We compare two information conditions: closed-book forecasting, where models receive no external evidence, and shared evidence retrieved via AskNews before the timestamp at which the market baseline is recorded. Across 8,196 valid forecasts, shared evidence lowers Brier scores in the same direction across most models and prompt-level aggregates ($\Delta = -0.028$, 95% CI $[-0.067, +0.009]$). The improvement is directionally robust but, at the pilot's 114-question cluster count, does not reach significance for the pooled effect; it is concentrated in questions whose evidence packet carries direct recent-status or base-rate information ($\Delta = -0.112$, 95% CI $[-0.198, -0.026]$) and is negligible otherwise. Bayesian-style prompting performs worse than the control prompt in closed-book settings, consistent with recent evidence that structured reasoning prompts can degrade LLM forecasts when external information is unavailable. Although Bayesian prompting shows the largest numerical improvement from evidence, its advantage over the control prompt is only suggestive at pilot scale and does not reach conventional significance (DiD $= -0.011$, $p = 0.08$). Despite these gains, no LLM beats the freeze-time prediction-market baseline. These findings suggest that evidence access improves LLM forecasting, but that prompt structure alone is insufficient: even evidence-grounded models remain behind market baselines and suffer from overconfident extreme probabilities.

[1]Jožef Stefan Institute, Ljubljana, Slovenia. Correspondence to: Akram Naoufel Tabet <akram.naoufel.tabet@ijs.si>, Mitja Luštrek <mitja.lustrek@ijs.si>.

Proceedings of the ICML 2026 Workshop: Forecasting as a New Frontier of Intelligence, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Large language models have shown rapidly improving performance on short-horizon probabilistic forecasting tasks (Halawi et al., 2024; Karger et al., 2025; Bosse et al., 2026), yet two empirical results sit in tension. On the one hand, Schoenegger et al. (2025) systematically tested 38 prompt variants across six frontier LLMs and found that, after correcting for multiple prompt comparisons, no prompt significantly improved closed-book ForecastBench accuracy. In fact, Bayesian Reasoning and Propose-Evaluate-Select prompts significantly degraded performance. In their study, Bayesian Reasoning asks the model to state a prior, identify evidence, and update its probability, while Propose-Evaluate-Select asks the model to generate candidate forecasts, evaluate them, and select a final answer. On the other hand, fine-tuning LLMs to reason like Bayesians (Qiu et al., 2026) and integrating structured reasoning into agentic forecasting pipelines (Alur et al., 2025) have both been reported to yield forecasting gains. A natural reconciliation is that structured prompts may behave differently depending on whether the model has external evidence to reason over: in closed-book settings, the model must infer priors and updates from parametric memory alone; with retrieved evidence, the same structure can organise concrete external information rather than relying only on the model's internal memory.

This paper isolates that hypothesis with a controlled factorial design. We cross three prompts taken verbatim from Schoenegger et al. (2025) (Control, Base-Rate-First, and Bayesian Reasoning) with two information conditions: closed-book forecasting and a shared evidence packet retrieved from AskNews before each question's freeze time, the timestamp at which the market baseline is recorded. We evaluate 12 LLMs, including nine reasoning-token models and three standard models, on 114 binary ForecastBench questions resolving after the latest model's training cut-

off. The P1–P3 design was fixed before the final analysis; a fourth Superforecaster Persona prompt is reported separately as an exploratory extension.

Our contributions are fourfold: (i) a replication of Schoenegger et al.'s closed-book prompt-engineering null with newer frontier models; (ii) the first systematic test of how the same prompts behave under shared evidence; (iii) evidence that shared retrieval improves forecasting accuracy on average, with gains concentrated in questions whose evidence packets contain direct recent-status or base-rate information; and (iv) a calibration analysis showing that the remaining gap to the market baseline is associated with overconfident extreme forecasts, especially in the highest-probability bin.

2. Methods

Questions. We sample 114 resolved binary questions from ForecastBench (Karger et al., 2025) biweekly snapshots released between October 2025 and March 2026. Questions are drawn from Polymarket, Manifold, and Metaculus, and intentionally balanced across broad topic groups. All questions resolve on or after September 2025, after the latest stated training cutoff in our model panel, and pass manual sanity checks against question leakage.

Table 1. Experimental design summary. P1–P3 are the main confirmatory prompts; P4 is exploratory.

Component	Design
Questions	114 resolved binary ForecastBench questions
Sources	Polymarket (41), Manifold (37), Metaculus (36)
Models	12 LLMs: 9 reasoning-token, 3 standard
Prompts	P1 Control, P2 Base-Rate-First, P3 Bayesian
Conditions	Closed-book vs. shared evidence
Evidence	Up to 10 AskNews articles before freeze time
Main calls	$12 \times 3 \times 2 \times 114 = 8{,}208$
Exploratory	P4 Superforecaster prompt: 2,736 calls
Primary metric	Brier score
Baseline	Freeze-time market probability
Valid forecasts	8,196/8,208 (99.9%) for P1–P3

Prompts. We use three prompts verbatim from Schoenegger et al. (2025), Appendix A: **P1** Control (a direct probability request), **P2** Base-Rate-First (anchor on historical frequency, then adjust), and **P3** Bayesian Reasoning (state a prior, identify evidence, update). All prompts terminate with a fixed "`Forecast:X%`" output requirement. We additionally run an exploratory **P4** Superforecaster Persona prompt drawn from the same source, reported separately and not included in the main confirmatory analysis.

Information conditions and shared-evidence construction. We compare two information conditions. In the closed-book (CB) condition, each model receives only the ForecastBench question title, background, resolution criteria, and resolution date. In the shared-evidence (SE) condition, we append a fixed evidence packet retrieved from the AskNews historical API. The goal of this condition is not to evaluate retrieval systems themselves, but to isolate the effect of giving different models access to the same external evidence. For each question, we constructed an AskNews query from the question title and key entities appearing in the question text, background, and resolution criteria. Key entities include named people, organizations, countries, teams, events, legislation, companies, products, or other terms needed to identify the target event. Retrieval was restricted to articles published strictly before the ForecastBench freeze time, defined as the timestamp at which the market baseline probability was recorded. We retained up to ten AskNews articles per question, using the API relevance ranking, and excluded articles whose publication timestamp was after the freeze time. The resulting evidence packet was held fixed across all models and prompts for that question. This design prevents model-specific retrieval differences from confounding the comparison between closed-book and evidence-grounded forecasting.

Evidence-quality audit. To address whether the evidence benefit depends on the content of the retrieved packet, we additionally classified each question-level evidence packet according to the type of useful forecasting information it contained. We used four non-exclusive categories: direct recent-status evidence, such as polls, standings, rankings, official figures, or latest pre-freeze developments; direct base-rate evidence, such as historical frequencies, previous outcomes, or comparison-class information; indirect contextual evidence, such as relevant background without a direct probabilistic signal; and weak or no clearly useful evidence. This audit is intended as a retrieval-quality analysis of the fixed shared-evidence packets, not as an evaluation of agentic retrieval or dynamic evidence selection.

Models. We test 12 LLMs accessed through OpenRouter: nine that produce internal reasoning tokens in our runs (gpt-5.4, claude-opus-4.6, gemini-3.1-pro, kimi-k2.6, grok-4.20, deepseek-v3.2-speciale, gpt-oss-120b, glm-5.1, qwen3-max) and three standard models (gemma-4-31b, gemini-3-flash, mistral-large).

Metrics and analysis. The primary metric is Brier score (Brier, 1950), $B = (\hat{p} - y)^2$, where $\hat{p}$ is the model's forecast probability and y is the binary outcome (1 if the event occurred, 0 otherwise). Lower Brier scores indicate more accurate probabilistic forecasts. Secondary metrics include Expected Calibration Error (ECE) over 10 equal-width probability bins, forecast standard deviation as a measure of sharpness, and the rate of catastrophic overconfidence: forecasts above 0.8 for events that did not occur, or below 0.2 for events that did occur.

To test whether prompt effects depend on evidence access, we fit a mixed-effects model with Brier score as the outcome:

$$\textit{Brier}_{ijkl} \sim \text{Prompt}_i \cdot \text{Cond}_j + (1 \mid Q_k) + (1 \mid M_l). \quad (1)$$

Here, Prompt_i is the prompt type, Cond_j is the information condition (CB or SE), Q_k indexes questions, and M_l indexes models. For prompt-level evidence effects, we report the mean paired difference between the shared-evidence and closed-book conditions, $\Delta_p = \text{Brier}_{p,\text{SE}} - \text{Brier}_{p,\text{CB}}$, for each prompt p. For prompt–condition interactions, we report difference-in-differences estimates relative to the control prompt:

$$\text{DiD}_p = (\text{Brier}_{p,\text{SE}} - \text{Brier}_{p,\text{CB}}) - (\text{Brier}_{\text{P1,SE}} - \text{Brier}_{\text{P1,CB}}).$$

Negative values indicate that the prompt benefits more from shared evidence than the control prompt. We compute 95% confidence intervals using a question-clustered bootstrap: questions are resampled with replacement, and all model–prompt–condition forecasts associated with each sampled question are kept together. This preserves the dependence among forecasts made on the same question.

For the evidence-quality audit, we report stratified SE–CB differences by question-level evidence-packet category. Because these categories are based on a post-hoc audit of the retrieved packets and some strata contain small numbers of questions, we treat these analyses as descriptive and hypothesis-generating rather than confirmatory.

Parsing and integrity. Each model response is parsed by extracting the required "`Forecast:X%`" line and converting it into a probability. We hardened the parser to handle common formatting failures, including markdown wrapping, decimal/percentage mismatches, and truncated responses. The P1–P3 valid forecast rate is 99.9% (8,196/8,208); P4 is 98.5% (2,695/2,736). Each prompt–condition–model combination contains 114 attempted calls after deduplication.

3. Results

3.1. Shared evidence consistently improves aggregate forecasting accuracy but uncertainty is large

[†]Exploratory prompt; not included in the confirmatory P1–P3 aggregate or DiD analysis.

Shared evidence consistently lowers mean Brier across prompts. Across the confirmatory P1–P3 design, mean Brier decreases from 0.180 in the closed-book condition to 0.152 in the shared-evidence condition ($\Delta = -0.028$) (Table 2). This point estimate is stable across prompts and is similar for the exploratory P4 prompt. However, when uncertainty is estimated by resampling questions rather than individual forecasts, the confidence interval is wider and includes zero (95% CI [-0.067, $+0.009$]). We therefore interpret shared evidence as producing a consistent descriptive improvement, while treating the aggregate effect as not statistically resolved at the question-clustered level.

3.2. Evidence gains are concentrated in high-quality evidence packets

Evidence quality explains much of the aggregate uncertainty. In an outcome-blind audit of the 114 shared-evidence packets, we classified 62 packets as strong, 33 as moderate, and 19 as weak. 95 packets contained either direct recent-status or base-rate evidence; these questions show a larger descriptive evidence benefit ($\Delta = -0.039$, 95% CI [-0.083, $+0.006$]) than the 19 packets without direct signal, where shared evidence descriptively worsens forecasts ($\Delta = +0.028$, 95% CI [-0.026, $+0.100$]). Evidence benefits are concentrated most clearly in the strong-evidence subset: shared evidence reduces Brier from 0.239 to 0.127 ($\Delta = -0.112$, 95% CI [-0.198, -0.026]). By contrast, moderate packets show no improvement ($\Delta = +0.009$), and weak packets descriptively worsen forecasts ($\Delta = +0.028$) (Table 3). This suggests that the aggregate evidence effect is weakened by heterogeneity in retrieval quality: fixed evidence packets help when they contain strong direct recent-status or base-rate signals, but not when retrieved articles are indirect, noisy, or weakly predictive.

3.3. Bayesian prompting is sensitive to evidence access

Replicating the closed-book pattern reported by Schoenegger et al. (2025), the Bayesian Reasoning prompt performs worst in the CB condition (Brier 0.190 vs. 0.174 for Control). However, P3 also shows the largest numerical improvement from shared evidence, improving by -0.033 Brier points. The planned difference-in-differences comparison, $(\text{P3}_{SE-CB}) - (\text{P1}_{SE-CB})$, is negative (DiD $= -0.011$, 95% CI [-0.024, $+0.001$]), consistent with the hypothesis that Bayesian-style structure becomes more useful when evidence is available. Because the confidence interval still overlaps zero, we interpret this interaction as suggestive rather than conclusive.

3.4. Evidence narrows but does not close the market gap

Despite the gains from shared evidence, no LLM configuration matches the freeze-time prediction-market baseline. The market baseline achieves Brier 0.095. In the closed-book condition, the strongest P1–P3 configurations remain far above this baseline; under shared evidence, the gap narrows substantially, with the best P1–P3 model–prompt–condition combination reaching Brier 0.107. The exploratory P4 prompt improves the closed-book frontier and remains competitive under shared evidence, but it also does not beat the market. Overall, evidence access moves LLM forecasts closer to market performance, but the remaining gap suggests that retrieval alone is insufficient without better calibration and aggregation.

3.5. High-confidence miscalibration remains the main failure mode

Calibration analysis suggests that the remaining gap between LLM forecasts and the market baseline is strongly associated with high-confidence errors. Across P1–P3, the overall ECE is 0.082, indicating moderate miscalibration. However, this aggregate value hides a sharper failure mode: overconfidence increases substantially in the upper probability bins. In the top decile, the mean forecast is 0.960 while the observed event frequency, i.e. the fraction of events in that bin that resolve true, is only 0.522 (Figure 1). Although LLMs are often described as cautious in open-ended settings, our forced-probability forecasting setup produces many extreme probability estimates. Across the full P1–P3 design, 384

Table 2. Mean Brier by prompt and information condition. Δ = SE–CB, so negative values favour shared evidence. CIs are paired question-clustered bootstrap intervals over 114 questions. P4 is exploratory and excluded from the confirmatory prompt-interaction analysis.

Prompt	CB	SE	Δ [95% CI]	DiD vs. P1 [95% CI]
All P1–P3	0.180	0.152	-0.028 [-0.067, $+0.009$]	–
P1 Control	0.174	0.152	-0.022 [-0.062, $+0.016$]	–
P2 Base-rate-first	0.175	0.149	-0.027 [-0.066, $+0.010$]	-0.005 [-0.015, $+0.005$]
P3 Bayesian	0.190	0.157	-0.033 [-0.075, $+0.008$]	-0.011 [-0.024, $+0.001$]
P4 Superforecaster†	0.165	0.139	-0.025 [-0.062, $+0.010$]	–

†P4 is exploratory and was not part of the confirmatory comparison.

Table 3. Evidence effect by audited evidence-packet quality. Δ = SE–CB, so negative values favour shared evidence. CIs are paired question-clustered bootstrap intervals.

Evidence quality	N_q	CB	SE	Δ [95% CI]
Strong	62	0.239	0.127	-0.112 [-0.198, -0.026]
Moderate	33	0.136	0.145	$+0.009$ [-0.015, $+0.022$]
Weak	19	0.171	0.199	$+0.028$ [-0.024, $+0.100$]

Table 4. Best LLM configurations compared with the freeze-time market baseline. Lower Brier is better.

System / setting	Brier
Market baseline	0.095
Best P1–P3 CB model avg.: gpt-5.4	0.146
Best P1–P3 CB combination: gpt-5.4 / P1 / CB	0.142
Best P1–P3 SE model avg.: gemini-3.1-pro	0.113
Best P1–P3 SE combination: gemini-3.1-pro / P1 / SE	0.107
Best P4 CB combination†: gpt-5.4 / P4 / CB	0.137
Best P4 SE combination†: gpt-5.4 / P4 / SE	0.112

† Exploratory, not part of the main comparison.

of 8,196 valid forecasts (4.7%) are high-confidence false positives ($\hat{p} > 0.8, y = 0$), while 340 (4.1%) are high-confidence false negatives ($\hat{p} < 0.2, y = 1$). This indicates that improving evidence access alone is insufficient: reliable LLM forecasting also requires better calibration of extreme probabilities.

3.6. Exploratory P4 and heterogeneity analyses

The Superforecaster-style prompt is exploratory but performs strongly. It achieves mean Brier 0.165 in CB and 0.139 in SE, the lowest prompt aggregate in both conditions. It also improves calibration relative to P1–P3 (ECE 0.064 vs. 0.082) and reduces high-confidence false positives (3.2% vs. 4.7%). At the model level, P4 beats each model's best P1–P3 prompt in 8/12 CB comparisons and 7/12 SE comparisons. Because P4 was not part of the main confirmatory hypothesis, we treat this as a hypothesis-generating result for the live forecasting experiment (a forthcoming Metaculus tournament evaluation) rather than as a primary finding.

Evidence effects are heterogeneous across topics and sources, though small-N topic subgroups (e.g., sports, N=8) show large deltas driven by individual questions; full breakdowns and per-topic N are reported in Appendix A.

4. Discussion

Three takeaways emerge. First, shared evidence improves LLM forecasting on average, but the effect is not uniform. The evidence-quality audit suggests that the gain is concentrated in packets containing direct recent-status or base-rate information, while weak or indirect packets provide little benefit and may even hurt. Thus, the result should be interpreted less as "retrieval always helps" and more as evidence that useful, timely, and decision-relevant information can substantially improve LLM forecasts when it is available.

Second, the results remain cautious for prompt structure alone. The Bayesian Reasoning prompt performs worst in the closed-book condition and shows the largest numerical gain from shared evidence, but the prompt × evidence interaction remains suggestive rather than conclusive at pilot scale. The exploratory Superforecaster prompt performs well, but because it was not part of the confirmatory design, we treat it as a hypothesis for future live forecasting experiments rather than as a main result.

Third, evidence access narrows but does not close the gap to the market baseline. This comparison should be interpreted carefully: market probabilities aggregate many forecasters and may reflect richer search, selection, and information-processing advantages than our fixed evidence packets. Still, the remaining gap highlights an important failure mode of LLM forecasting: even when evidence improves accuracy, models remain prone to overconfident extreme probabilities.

5. Limitations and Next Steps

This study is diagnostic rather than definitive. It is limited to 114 resolved binary ForecastBench questions, uses fixed AskNews evidence packets rather than agentic retrieval, and is underpowered for detecting small prompt $\times$ evidence interactions. The evidence-quality audit is descriptive and post-hoc, and P4 is exploratory. A larger live Metaculus evaluation will test whether evidence selection, calibration, agentic retrieval, and superforecaster-style prompting generalize to numeric and multi-class forecasts and close more of the gap to market baselines.

References

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 preprint arXiv:2511.07678*, 2025. URL https://arxiv.org/abs/2511.07678.

Bosse, N. I., Mühlbacher, P., Wildman, J., Phillips, L., and

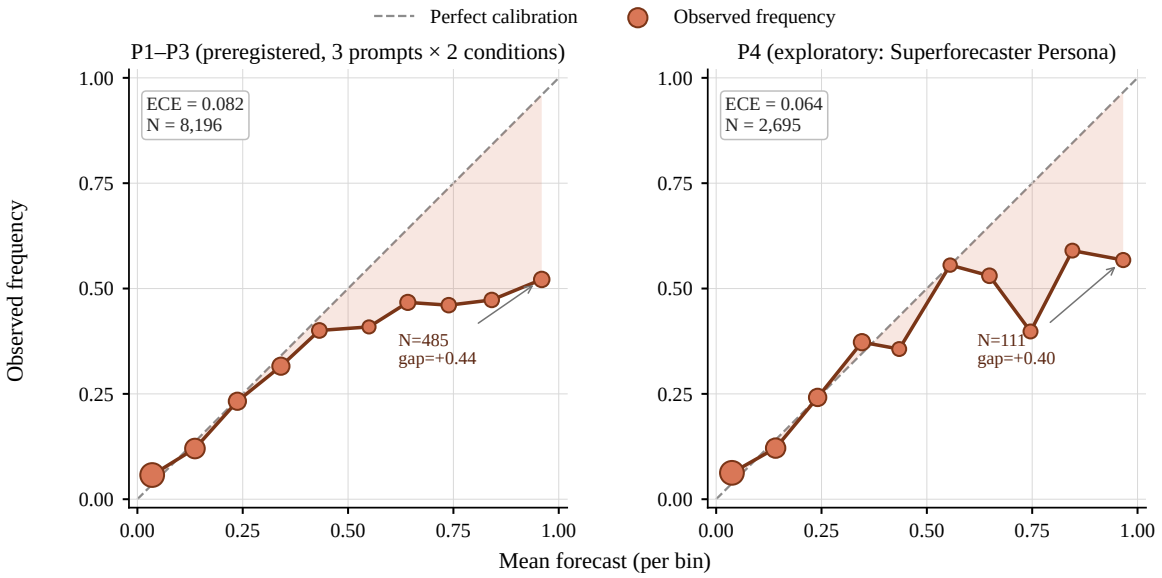

Figure 1. Calibration of P1–P3 forecasts ($N = 8{,}196$) and exploratory P4 forecasts ($N = 2{,}695$). Each point is a probability bin: x is the mean forecast and y is the fraction of questions in that bin that resolved true. The dashed diagonal indicates perfect calibration; the largest gap appears in the highest-probability bin.

Schwarz, D. Automating forecasting question generation and resolution for AI evaluation. *arXiv preprint arXiv:2601.22444*, 2026. URL https://arxiv.org/abs/2601.22444.

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

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

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. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2409.19839.

Qiu, L., Sha, F., Allen, K., Kim, Y., Linzen, T., and van Steenkiste, S. Bayesian teaching enables probabilistic reasoning in large language models. *Nature Communications*, 17:1238, 2026. doi: 10.1038/s41467-025-67998-6.

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

A. Topic and source heterogeneity

Table 5 reports CB and SE mean Brier across the ten topic categories represented in our sample. Topic labels are auto-generated by an LLM classifier and then manually verified. Sample sizes are uneven: large topics (AI/technology, geopolitics, conflict, economics/finance) contain ≥ 14 questions each, while small subgroups (energy, public health, "other") contain ≤ 6. Topic-level SE–CB deltas in the small subgroups should therefore be interpreted as descriptive rather than as generalizable estimates.

Table 5. Per-topic mean Brier by information condition (P1–P3, all 12 models). N_q is the question count; forecasts per row $\approx 36N_q$. Topics ordered by SE–CB delta (most-helped first).

Topic	N_q	CB Brier	SE Brier	SE–CB	Direction
Sports	8	0.184	0.025	-0.159	helps[†]
Public health	6	0.255	0.168	-0.087	helps
Finance/markets	14	0.280	0.217	-0.063	helps
Other	5	0.203	0.141	-0.062	helps
Economics	14	0.128	0.087	-0.041	helps
AI/technology	20	0.126	0.097	-0.030	helps
Geopolitics	20	0.163	0.163	0.000	neutral
Conflict	18	0.170	0.189	$+0.019$	hurts
Entertainment	5	0.183	0.211	$+0.028$	hurts
Energy	4	0.252	0.389	$+0.137$	hurts

[†] Sports delta is driven by two questions for which pre-freeze season standings were highly predictive (Pistons Central Division lead; Yamal La Liga goal pace). A manual audit confirmed all evidence articles were dated before each question's freeze time and contained no resolution information. See Section 5.

Table 6. Mean LLM Brier score by question source and information condition (P1–P3, all 12 models). Scores are computed from LLM forecasts, not from the market baseline. The source indicates the platform from which the ForecastBench question was drawn. SE–CB shows how much shared evidence changes LLM performance for questions from each source.

Source	N_q	CB Brier	SE Brier	SE–CB
Manifold	37	0.160	0.086	-0.074
Metaculus	36	0.228	0.210	-0.018
Polymarket	41	0.154	0.159	$+0.005$

B. Per-model results

Table 7. Per-model mean Brier across P1–P3 prompts and both information conditions ($N \approx 684$ per model). "Parse %" is the rate of successfully parsed forecasts; latency is the median per-call wall-clock time in seconds. Lower Brier is better.

Rank	Model	Type	Brier	Parse %	Lat. (s)
1	gpt-5.4	reasoning	0.137	100.0	62
2	gemini-3.1-pro	reasoning	0.143	99.7	35
3	kimi-k2.6	reasoning	0.144	100.0	249
4	claude-opus-4.6	reasoning	0.146	100.0	39
5	grok-4.20	reasoning	0.161	100.0	29
6	deepseek-v3.2-spec.	reasoning	0.164	100.0	197
7	glm-5.1	reasoning	0.173	99.7	77
8	qwen3-max	reasoning	0.176	99.4	251
9	gpt-oss-120b	reasoning	0.176	99.4	86
10	gemma-4-31b	standard	0.178	100.0	39
11	gemini-3-flash	standard	0.179	100.0	6
12	mistral-large	standard	0.216	100.0	17
	Market baseline		0.095	—	—

C. Prompt templates

We use three preregistered prompts taken verbatim from Schoenegger et al. (2025), Appendix A: P1 Control, P2 Base-Rate-First, and P3 Bayesian Reasoning. Each is run in both the closed-book (CB) condition (only the question) and the shared-evidence (SE) condition (an

Table 8. Per-model Brier across all prompt × condition cells. P1–P3 are preregistered; P4 is exploratory. The best of P1–P3 per row and condition is in **bold**; P4 cells marked * outperform that row's best P1–P3 prompt in the same condition.

	P1 (Control)		P2 (Base-rate)		P3 (Bayesian)		P4† (Persona)	
Model	CB	SE	CB	SE	CB	SE	CB	SE
gpt-5.4	**0.142**	0.125	0.148	0.141	0.148	**0.117**	0.137*	0.112*
gemini-3.1-pro	**0.169**	**0.107**	0.175	0.117	0.175	0.117	0.146*	0.129
kimi-k2.6	0.162	0.141	0.165	**0.119**	**0.150**	0.127	0.152	0.132
claude-opus-4.6	**0.161**	0.137	0.168	**0.123**	0.163	0.127	0.152*	0.135
grok-4.20	0.178	0.154	**0.176**	0.140	0.179	**0.138**	0.171*	0.134*
deepseek-v3.2-spec.	0.176	**0.145**	**0.166**	0.148	0.185	0.167	0.164*	0.131*
gpt-oss-120b	0.168	**0.150**	**0.164**	0.154	0.222	0.195	0.158*	0.136*
glm-5.1	**0.180**	0.161	0.198	**0.152**	0.192	0.157	0.167*	0.170
qwen3-max	0.189	**0.147**	**0.186**	0.155	0.193	0.184	0.191	0.156
gemma-4-31b	**0.175**	0.173	0.188	**0.162**	0.196	0.176	0.181	0.148*
gemini-3-flash	0.194	**0.167**	**0.186**	0.167	0.190	0.170	0.177*	0.153*
mistral-large	0.193	0.216	**0.186**	**0.205**	0.284	0.209	0.186	0.138*

† Exploratory, not part of the main confirmatory analysis.

additional fixed block of up to 10 AskNews articles). The exploratory P4 Superforecaster Persona prompt is also drawn from the same source. Structural placeholders ({question_title}, {background}, {resolution_criteria}, {resolution_date}) are populated from each ForecastBench question record; {retrieved_articles} is present only in the SE condition.

Evidence block (SE condition only).

```
The following recent news articles may be relevant to this question. Use them if
helpful; disregard if not.
{retrieved_articles}
```

P1 — Control.

```
Please answer the following question with a probabilistic estimate expressed
between 0% and 100%, and format your response as: 'Forecast: X%'.
Question: {question_title}
Background: {background}
Resolution Criteria: {resolution_criteria}
Resolution Date: {resolution_date}
```

P2 — Base-Rate-First.

```
Please answer the following question with a probabilistic estimate expressed
between 0% and 100%, and format your response as: 'Forecast: X%'.
Before considering the specific details of this question, what is the historical
frequency of similar events? Using this base rate as your starting point, adjust
your probability estimate based on the particular circumstances of this case.
Question: {question_title}
Background: {background}
Resolution Criteria: {resolution_criteria}
Resolution Date: {resolution_date}
```

P3 — Bayesian Reasoning.

```
Consider the following question in terms of Bayesian reasoning. Start with a
prior probability based on historical data or general knowledge. Then, update
this prior using more specific information about the case under discussion. For
each new piece of information, produce an updated posterior estimate of the
outcome using the principle behind Bayes rule. Conclude with the final posterior
```

```
probability, formatted as: 'Forecast: X%'
Question: {question_title}
Background: {background}
Resolution Criteria: {resolution_criteria}
Resolution Date: {resolution_date}
```

P4 — Superforecaster Persona (exploratory). This persona-based prompt is drawn verbatim from Schoenegger et al. (2025), Appendix A, and prepends a long superforecasting system message before the question block. The same evidence-block convention as P1–P3 applies (appended only in the SE condition).

```
In this chat, you are a superforecaster who has a strong track record of accurate
forecasting. You evaluate past data and trends carefully for potential clues
to future events, while recognising that the past is an imperfect guide to
the future so you will need to put probabilities on possible future outcomes
(ranging from 0 to 100%). Your specific goal is to maximize the accuracy of
these probability judgments by minimising the Brier scores that your probability
judgments receive once future outcomes are known. Brier scores have two key
components: calibration (across all questions you answer, the probability
estimates you assign to possible future outcomes should correspond as closely as
possible to the objective frequency with which outcomes occur) and resolution
(across all questions, aim to assign higher probabilities to events that occur
than to events that do not occur). You outline your reasons for each forecast:
list the strongest evidence and arguments for making lower or higher estimates
and explain how you balance the evidence to make your own forecast. You begin
this analytic process by looking for reference or comparison classes of similar
events and grounding your initial estimates in base rates of occurrence (how
often do events of this sort occur in situations that look like the present
one?). You then adjust that initial estimate in response to the latest news
and distinctive features of the present situation, recognising the need for
flexible adjustments but also the risks of over-adjusting and excessive volatility.
Superforecasting requires weighing the risks of opposing errors: e.g., of
failing to learn from useful historical patterns vs. over-relying on misleading
patterns. In this process of error balancing, you draw on the 10 commandments
of superforecasting (Tetlock & Gardner, 2015) as well as on other peer-reviewed
research on superforecasting: 1. Triage 2. Break seemingly intractable problems
into tractable sub-problems 3. Strike the right balance between inside and
outside views 4. Strike the right balance between under- and overreacting to
evidence 5. Look for the clashing causal forces at work in each problem 6.
Strive to distinguish as many degrees of doubt as the problem permits but no more
7. Strike the right balance between under- and overconfidence, between prudence
and decisiveness 8. Look for the errors behind your mistakes but beware of
rearview-mirror hindsight biases 9. Bring out the best in others and let others
bring out the best in you 10. Master the error-balancing bicycle. Once you have
written your reasons, ensure that they directly inform your forecast. Then, you
will provide me with a number between 0 and 100 (up to 2 decimal places) that is
your best prediction of the event in the following format: ''Forecast: X%''.
Take a deep breath and work on this problem step-by-step. The question that you
are forecasting as well as some background information and resolution criteria are
below. Read them carefully before making your prediction.
Question: {question_title}
Background: {background}
Resolution Criteria: {resolution_criteria}
Resolution Date: {resolution_date}
```