The Wisdom of Deliberating AI Crowds: Does Deliberation Improve LLM-Based Forecasting?

Paul Schneider* Amalie Schramm*

Abstract

Structured deliberation has been found to improve the performance of human forecasters. This study investigates whether a similar intervention—allowing LLMs to review each other's forecasts before updating—can improve accuracy in large language models (GPT-5, Claude Sonnet 4.5, Gemini Pro 2.5). Using 202 resolved binary questions from the Metaculus Q2 2025 AI Forecasting Tournament, accuracy was assessed across four scenarios: (1) diverse models with distributed information, (2) diverse models with shared information, (3) homogeneous models with distributed information, and (4) homogeneous models with shared information. Results show that the intervention significantly improves accuracy in scenario (2), reducing Log Loss by 0.020 or about 4% in relative terms ($p = 0.017$). However, when homogeneous groups (three instances of the same model) engaged in the same process, no benefit was observed. Unexpectedly, providing LLMs with additional contextual information did not improve forecast accuracy, limiting our ability to study information pooling as a mechanism. Our findings suggest that deliberation may be a viable strategy for improving LLM forecasting.

Introduction

Expert forecasting is the systematic elicitation of probability judgments about future events. It usually involves obtaining probability estimates from multiple experts and aggregating their judgments into a single estimate (Mellers et al., 2014; Armstrong, 2001; Tetlock, 2005). Probabilistic forecasts can support policy decision making and risk management across many domains, including geopolitics (e.g., election outcomes), economics, and AI safety (Hanea et al., 2021; Surowiecki, 2004; Tetlock et al., 2014).

Traditionally, forecasting relied on human experts (Tetlock, 2005; Tetlock and Gardner, 2015). However, recent advancements in large language models (LLMs) has sparked a new research program into whether AI systems can potentially also provide accurate forecasts (Zou et al., 2022; Schoenegger et al., 2024; Ye et al., 2024; Halawi et al., 2024). Various studies have since explored this question and found mixed results: while some authors report that LLMs are already approaching or even exceeding human-level performance (Halawi et al., 2024; Schoenegger et al., 2025), a public AI forecasting tournament showed that human expert forecasters still outperform LLM-based systems by a significant margin (Metaculus, 2025b,c).

In line with benchmarks in other areas, AI forecasting results suggest that general LLM capabilities might be the most important determinant of forecast performance (Brown et al., 2020; Wei et al., 2022; Kaplan et al., 2020; Metaculus, 2025b). Notwithstanding, methodological choices also matter. This includes prompt engineering, fine-tuning, retrieval strategies, and aggregation methods.

One method that has not yet been systematically tested is deliberation, i.e., the process of structured discussion and sharing of information. It has been shown to improve forecast accuracy when used by teams of human experts (Hemming et al., 2018; Dezecache et al., 2022). A deliberation-like protocol ("multi-agent debate") was also found to be effec-

*PRIORB, Bochum, Germany. Contact: paul@priorb.com

tive in improving LLM performance on math and logic tasks (Du et al., 2024; Liang et al., 2024). In this study, we test whether this finding extends to LLM-based forecasting systems and improves their accuracy.

Study Objective

We investigate whether a deliberation-like process of sharing forecast estimates and reasoning across multiple LLM instances ("deliberation" hereafter) improves forecast accuracy, compared to simply aggregating independent forecasts.

The hypothesis was tested under conditions that varied along two dimensions: a) model diversity (homogeneous vs. diverse), and b) information distribution (shared vs. distributed). The resulting four scenarios correspond to distinct deployment scenarios of LLM forecasting systems.

Methods

Overview

We used 202 resolved binary questions from the Metaculus Q2 2025 AI tournament. Groups of three LLMs forecasted each question in two rounds. LLMs first generated independent forecasts, then were shown their peers' forecasts and reasoning before making updated forecasts ("deliberation"). We tested four scenarios crossing model diversity (diverse vs. homogeneous) with information distribution (distributed vs. shared). Accuracy was measured using Log Loss on the median group forecast. Within each scenario, we used paired t-tests to compare independent vs. deliberative forecasts.

Materials: Questions, Information, and LLMs

Forecasting Questions

We used all 202 resolved binary questions from the Metaculus Q2 2025 AI Forecasting Benchmark (Metaculus, 2025a). All questions had resolved by the time of analysis, allowing us to measure forecast accuracy against ground truth. Questions spanned multiple domains including geopolitics, economics, technology, and science. Each question specified a binary outcome and a resolution date.

Information Extraction

To manipulate information availability across agents, we extracted relevant contextual information from publicly available forecast commentary from the Metaculus platform (Metaculus, 2025a), accessed via the official API (`https://www.metaculus.com/api/`). We used an LLM (Gemini Pro 2.5) to summarise the information with the aim to construct three distinct, non-overlapping units of information. These could entail factual claims, statistical evidence, or general contextual background.

An illustrative example of the three information units is provided in Appendix A.

All models' training cut-off dates were prior to the resolution date and question publication date, ensuring no information leakage.

LLMs

Each forecasting task was performed by a group of three LLMs. We used three frontier models: GPT-5 (OpenAI), Claude Sonnet 4.5 (Anthropic), and Gemini Pro 2.5 (Google). Groups were configured as either homogeneous (homo), with all three LLMs being of the same type, or diverse (diverse), with one instance of each model.

The instructions provided to the models were standardized to ensure consistency across groups. For the initial ("before") forecast, we used a modified version of the Metaculus prompting template, which has been shown to perform well in prior benchmarks (Metaculus, 2025c). For the deliberative ("after") forecast, we presented each model with the three initial forecasts (from all group members) along with their reasoning, and instructed them to review, contrast, and synthesise the different perspectives before updating their forecast.

Both prompt templates are provided in Appendix B.

Forecast Generation Procedure

Forecasts were generated in two stages for each question-group combination:

Stage 1 (Independent forecasts). Each agent in a group received the question text, resolution criteria, and their assigned information package. Agents generated forecasts and reasoning independently, without access to other agents' outputs.

Stage 2 (Deliberative forecasts). Each agent received the three Stage 1 forecasts and rationales from all group members. Agents were instructed to review and critique the reasoning, then provide an updated forecast. The deliberation prompt encouraged agents to identify new arguments, assess their validity, and explain any forecast revisions.

For diverse scenarios, each of the 202 questions was forecasted by one group containing all three model types (GPT-5, Sonnet, Pro). For homogeneous scenarios, questions were distributed across model types using round-robin assignment: questions 1, 4, 7, ... were assigned to GPT-5 groups; questions 2, 5, 8, ... to Sonnet groups; and questions 3, 6, 9, ... to Pro groups. This yielded approximately 67 questions per model type in each homogeneous condition.

Within each group, the three individual forecasts were aggregated using the median probability. The median served as the group-level forecast for both independent and deliberative stages.

Statistical Analysis

We implemented a 2×2 factorial design crossing two factors:

1. **Model Diversity:** Homogeneous (same model) vs. Diverse (three different models)
2. **Information:** Shared (full information) vs. Distributed (unique information per LLM).

Within each scenario, we used paired t-tests comparing independent vs. deliberative forecasts. The unit of analysis was the group-level median forecast. For diverse scenarios (group composed of three different models), we collected $n = 202$ group observations (one group per question). For homogeneous scenarios, we collected $n = 606$ group observations (three groups per question, one each of 3× GPT-5, 3× Sonnet, and 3× Pro). For the model-level breakdown (Appendix D), questions were distributed across model types using round-robin assignment, yielding approximately 67 questions per model. The outcome metric was change in Log Loss (cross-entropy), i.e., (deliberative minus independent) (Good, 1952). Significance was assessed using a two-tailed $\alpha = 0.05$ threshold, not adjusted for multiple comparisons.

We conducted several secondary analyses to probe the mechanisms and boundary conditions of our findings. First, we tested whether providing information (none vs. partial vs. full) improved forecast accuracy at the independent stage, using linear regression with "no information" as the reference category. This isolates the effect of information from deliberation. Secondly, for homogeneous scenarios, we examined whether deliberation effects differed by model type (GPT-5, Sonnet, Pro), testing each with separate paired t-tests. Thirdly, we examined calibration curves to assess whether deliberation affected not just accuracy but also the calibration of probability estimates. Finally, we repeated the main analysis using Brier scores, which are less sensitive to extreme probabilities (results are reported in Appendix E).

Power Estimation

Sample size ($N = 202$) was determined by the Metaculus Q2 2025 AI Forecasting Tournament rather than a-priori power analysis. Sensitivity analyses characterising the minimum detectable effect (MDE) for each experimental condition are reported in Appendix C.

Code and Data Availability

All code, data, and analysis scripts are available at `https://github.com/priorb-source/delib-ai-wisdom`. Statistical analyses were conducted in Python using statsmodels (Seabold and Perktold, 2010).

Results

The dataset comprised 202 resolved binary questions from the Metaculus Q2 2025 AI tournament. Each question was forecasted by groups of 3 agents across 4 scenarios, yielding 1,616 group-level observations (202 questions × 2 diverse scenarios + 606 questions × 2 homogeneous scenarios). At the agent level, this corresponded to 2,424 individual forecasts per stage (before and after deliberation).

Table 1 shows the effect of deliberation on forecast accuracy within each scenario. We report mean Log Loss for independent and deliberative forecasts, the paired difference (deliberative minus independent), and paired t-test statistics.

The results show that both diverse scenarios showed improvement after deliberation. The effect was statistically significant for Diverse models, shared information ($p = 0.017$), with a mean Log Loss reduction of 0.020, corresponding to a 4% relative improvement in accuracy. The Diverse models, distributed information scenario showed a similar magnitude of improvement (-0.022) but higher variance, resulting in a non-significant result ($p = 0.18$).

Neither homogeneous scenario benefited from deliberation. Both showed slight increases in Log Loss (worsening), though neither was statistically significant.

Appendix E reports the effect of deliberation on forecast accuracy by scenario measured in Brier score, as an alternative outcome variable. This sensitivity analysis confirmed the primary findings.

A break down of results for the homogeneous scenarios by model type is provided in Appendix D. Gemini Pro showed the largest increases in Log Loss after deliberation (+0.047 and +0.052), while Sonnet showed slight decreases in both conditions and GPT-5 showed mixed results. However, the analysis was not powered to detect model-level differences.

Calibration Analysis

Figure 1 shows calibration curves for all four scenarios, comparing independent (blue) and deliberative (orange) forecasts against perfect calibration (dashed diagonal). Overall, the models appear reasonably well-calibrated across conditions, with predicted probabilities generally tracking observed frequencies. Visual inspection reveals no systematic pattern distinguishing the independent from the deliberative forecasts; in most cases, the curves overlap or fluctuate without a clear directional trend.

The "Same Model + Distributed Information" condition appears to show the tightest alignment with the diagonal. Conversely, the "Same Model + Shared Information" condition exhibits the most notable deviation, largely driven by a significant outlier in the deliberative forecast at the 0.8 probability bin. Aside from this instance, however, the data do not support a strong conclusion that deliberation consistently degrades or improves calibration relative to the independent baseline.

Sensitivity Analysis: Does Information Affect Accuracy?

The similar effect sizes observed for models with diverse and shared information was unexpected. We assumed that deliberation could help partly by allowing LLMs to pool information, in which case the effect of deliberation should be larger in those conditions where agents held unique facts. The absence of this pattern led us to investigate whether

Table 1: Effect of deliberation on forecast accuracy by scenario on Log Loss

Scenario	Independent mean (SD)	Deliberative mean (SD)	Change mean (SD)	t	p
Diverse models, distributed information	0.475 (0.494)	0.453 (0.537)	−0.022 (0.237)	1.34	0.18
Diverse models, shared information	0.501 (0.608)	0.481 (0.618)	−0.020 (0.117)	2.41	0.017
Homogeneous models, distributed information	0.517 (0.612)	0.525 (0.677)	+0.008 (0.308)	−0.36	0.72
Homogeneous models, shared information	0.525 (0.653)	0.545 (0.696)	+0.020 (0.194)	−1.47	0.14

the information we provided was decision-relevant in the first place.

More specifically, we tested whether providing information improved forecast accuracy at the *independent* stage. It was found that neither partial nor full information significantly improved forecast accuracy. Table 2 shows the results of linear regressions predicting Log Loss from information level, with "no information" as the reference category.

Table 2: Effect of information on forecast accuracy, measured as Log Loss (independent stage only)

Predictor	β	SE	t	p
Diverse ($n = 1{,}818$)				
Intercept (no info)	0.567	0.027	20.66	<.001
Partial info	−0.025	0.035	−0.71	.48
Full info	−0.038	0.039	−0.98	.33
Homogeneous ($n = 1{,}818$)				
Intercept (no info)	0.548	0.027	20.06	<.001
Partial info	+0.001	0.039	0.02	.99
Full info	−0.008	0.039	−0.21	.83

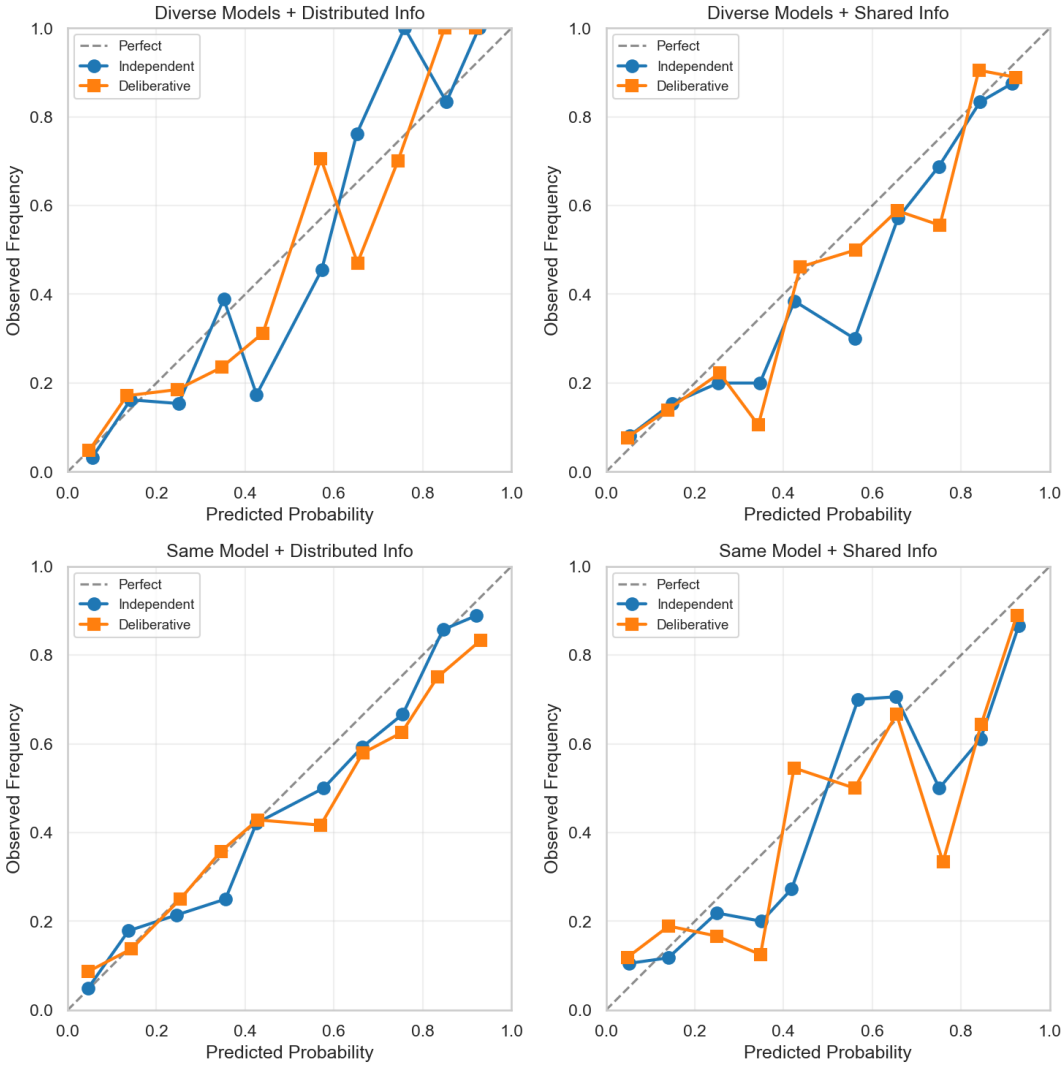

Figure 1: Calibration curves for independent vs. deliberative forecasts, stratified by scenario

Discussion

This study is the first to test whether structured deliberation improves forecast accuracy in LLM-based prediction systems. Our central finding is that deliberation significantly improved accuracy when diverse models collaborate: allowing models to deliberate reduced log loss by 0.020 to 0.022, which corresponds to a marginal but potentially meaningful relative improvement of around 4% in forecast accuracy.

However, the benefit of deliberation was only found when models were diverse, i.e., when groups were composed of different models, and the improvement was only statistically significant when all models had access to the same information. When groups of LLMs were composed of three instances of the same model, deliberation yielded no accuracy gains. In fact, results even indicate a slight degradation, although not statistically significant. It could be speculated that because models share identical training data and architecture, any learned biases or reasoning errors might be correlated. Presenting such a model with "peer" forecasts that mirror its own might not provide any new, external perspective.

Potentially the most surprising finding is that providing additional contextual information did not improve forecast accuracy (see Table 2). This result was unexpected. Metaculus tournament data suggested that the quality of retrieved information impacts forecast accuracy (Metaculus, 2025b). A central hypothesis of this work was indeed based on the assumption that pooling of information could be a main driver of any deliberation benefits. The most immediate explanation is that the LLM-generated summaries of Metaculus commentary may not have contained genuinely useful information. Yet, when we manually reviewed the underlying data, rationales in the scenarios in which LLMs had more information available appeared to be of higher quality than those in which less information was available. Those rationales frequently referenced external facts and data sources that, at face value, seemed decision-relevant. One could speculate that the compression of the information from comments into short information units may have led to loss of nuance or to have led models to become overconfident. Notwithstanding, these results remain counter intuitive and should be investigated further in future work.

The absence of a clear information effect has important implications for the interpreting of our findings. If additional information was generally not useful, pooling of information during deliberation cannot be expected to meaningfully improve accuracy. This likely explains why, contrary to initial expectations, the effect of deliberation under the distributed information condition was not larger than the shared information condition. In fact, both conditions showed very similar effect sizes (~ 0.02 Log Loss improvement), with only the shared information condition being statistical significance, likely due to lower variance.

This study has several other limitations worth mentioning. First, we implemented a minimal deliberation protocol consisting of a single round of forecast sharing and updating. More elaborate and structured approaches might yield additional benefits. Secondly, results may be specific to the types of questions tested and the models used. Generalisation to other forecasting contexts remains to be established. Finally, the power of our study to detect true effects may have been limited by the number of questions sampled. Future work should aim to validate these findings across a larger corpus of forecasting questions.

Notwithstanding these limitations, our results both extend and qualify findings from the multi-agent debate literature in AI. Du et al. (2024) and Liang et al. (2024) demonstrated that debate protocols can improve LLM performance on reasoning tasks, particularly mathematics and logic problems. The present study extends this finding to the forecasting domain, where deliberation improves accuracy on real-world probabilistic forecasts. Findings reported in this study also echo patterns observed in human forecasting research. Several authors found that forecaster teams can benefit from structured interaction protocols (Tetlock and Gardner, 2015; Hemming et al., 2020; Dezecache et al., 2022). Hanea et al. (2021) and Fraser et al. (2023) found that deliberation improved calibration even when experts had access to the same background materials. Our re-

sults suggest this pattern might transfer to groups of LLMs.

For practitioners who build forecasting systems, our findings tentatively suggest that deliberation among diverse models may improve accuracy. The marginal cost of additional API queries to implement a deliberation-like protocol is modest and may provide benefit. However, the observed improvement appears small: the 0.02 change in Log Loss corresponds to a $\sim 4\%$ relative improvement. Deliberation, if it helps, is an incremental optimisation. Base model capability likely remains the dominant factor in forecast accuracy. Benchmarking and testing system performance under real-world conditions is essential before deployment.

Conclusion

This study found that deliberation improved accuracy in LLM-based forecasting when using diverse models: when three different frontier models reviewed each other's forecasts, accuracy improved significantly; when identical models engaged in the same process, no benefit was observed. Contrary to expectations, additional contextual information had no effect on forecast accuracy. While further research is warranted to confirm these findings and explore the role of information quality, practitioners may wish to explore deliberation-like protocols with diverse models in forecasting applications.

Funding Statement

This study was financially supported by the Foresight Institute (`https://foresight.org/`).

References

Armstrong, J. S., editor (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*, volume 30 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Boston, MA.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Dezecache, G., Dockendorff, M., Ferreiro, D. N., Deroy, O., and Bahrami, B. (2022). Democratic forecast: Small groups predict the future better than individuals and crowds. *Journal of Experimental Psychology: Applied*, 28(3):525–537.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*.

Fraser, H., Bush, M., Wintle, B. C., Mody, F., Smith, E. T., Hanea, A. M., Gould, E., Hemming, V., Hamilton, D. G., Rumpff, L., Wilkinson, D. P., Pearson, R., Singleton Thorn, F., Gray, C. T., Head, A., Ross, M., Groenewegen, R., Marcoci, A., Vercammen, A., Parker, T. H., Hoekstra, R., Nakagawa, S., Mandel, D. R., van Ravenzwaaij, D., McBride, M., Sinnott, R. O., Vesk, P., Burgman, M., and Fidler, F. (2023). Predicting reliability through structured expert elicitation with the repliCATS (collaborative assessments for trustworthy science) process. *PLOS ONE*, 18(1):e0274429.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114.

Halawi, D., Zhang, F., Chen, Y.-H., and Steinhardt, J. (2024). Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems*, volume 37.

Hanea, A. M., Wilkinson, D. P., McBride, M., Lyon, A., van Ravenzwaaij, D., Singleton Thorn, F., Gray, C., Mandel, D. R., Willcox, A., Gould, E., Smith, E. T., Mody, F., Bush, M., Fidler, F., Fraser, H., and Wintle, B. C. (2021). Mathemati-

cally aggregating experts' predictions of possible futures. *PLOS ONE*, 16(9):e0256919.

Hemming, V., Armstrong, N., Burgman, M. A., and Hanea, A. M. (2020). Improving expert forecasts in reliability: Application and evidence for structured elicitation protocols. *Quality and Reliability Engineering International*, 36(2):623–641.

Hemming, V., Walshe, T. V., Hanea, A. M., Fidler, F., and Burgman, M. A. (2018). Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management. *PLOS ONE*, 13(6):e0198468.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint*.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. (2024). Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904. Association for Computational Linguistics.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., and Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115.

Metaculus (2025a). AI forecasting benchmark tournament – 2025 Q2.

Metaculus (2025b). Fall 2025 AI forecasting benchmark tournament.

Metaculus (2025c). Q1 AI benchmarking results: Pro forecasters crush bots.

Schoenegger, P., Park, P. S., Karger, E., Trott, S., and Tetlock, P. E. (2025). AI-augmented predictions: LLM assistants improve human forecasting accuracy. *ACM Transactions on Interactive Intelligent Systems*, 15(1):1–25.

Schoenegger, P., Tuminauskaite, I., Park, P. S., Valdece Sousa Bastos, R., and Tetlock, P. E. (2024). Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45):eadp1528.

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, pages 92–96. SciPy.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, New York.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Princeton, NJ.

Tetlock, P. E. and Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. Crown Publishers, New York, NY.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., and Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science*, 23(4):290–295.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

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

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., and Hendrycks, D. (2022). Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems*, volume 35. Datasets and Benchmarks Track.

Supplementary Materials

S1 Information Unit Examples

questionTitle: "Will initial jobless claims for the week ended June 21, 2025 exceed 220,000?"

questionDescription: "According to the resolution source, 'An initial claim is a claim filed by an unemployed individual after a separation from an employer. The claim requests a determination of basic eligibility for the Unemployment Insurance program.'"

questionResolutionCriteria: "This question resolves as Yes if initial jobless claims for the week ended June 21, 2025 is greater than 220,000 according to FRED"

information_1: "According to the U.S. Department of Labor, initial jobless claims for the week ending May 31, 2025, rose by 8,000 to 247,000, the highest level in eight months. The figure for the prior week was revised to 239,000. The four-week moving average increased to 235,000, the highest since November 2021. For the week ending May 24, the number of continuing claims was 1.904 million, a slight decrease of 3,000 from the previous week."

information_2: "Broader labor market indicators suggest a softening trend. A report from Challenger, Gray & Christmas showed that U.S.-based employers announced 93,816 job cuts in May, which is 47% higher than in May 2024. Separately, the ADP employment report revealed that only 37,000 jobs were created in the private sector in May, the lowest figure in a year. Both the Federal Reserve's Beige Book report and an Institute for Supply Management (ISM) survey pointed to weakening conditions, with widespread comments about economic uncertainty delaying new hiring."

information_3: "The recent rise in jobless claims was not uniform across the country. The largest increases in initial claims for the week ending May 31 were in Michigan, Nebraska, California, Florida, and Virginia. The largest decreases were reported in Massachusetts, Illinois, Texas, Washington, and New York. Several major corporations have announced layoffs in 2025, including Procter & Gamble, Microsoft, Citigroup, Walmart, Meta, and Disney."

S2 Prompts

Modified Metaculus template for Round 1 (independent):

```
You are a professional forecaster interviewing for a job.
Your interview question is: {{questionTitle}}
Question background: {{questionDescription}}

This question's outcome will be determined by the specific
criteria below. These criteria have not yet been satisfied:
{{questionResolutionCriteria}}

{{question.questionFinePrint}}

Your research assistant's report says: {{information}}

Today is {{question.date}}.

Before answering you think: (a) The time left until the
outcome to the question is known. (b) The status quo outcome
if nothing changed. (c) A brief description of a scenario that
results in a No outcome. (d) A brief description of a scenario
that results in a Yes outcome. (e) You write your rationale
remembering that good forecasters put extra weight on the
status quo outcome since the world changes slowly most of the
time. Explain your reasoning and the evidence behind your
forecast in detail. Summarise information your received from
your research assistant that influences your forecast (if any).
Explain why your forecast is not higher, and why it is not
lower. Outline what would need to be true for you to update
your forecast in either direction. (f) The last thing you write
is your final probabilistic forecast as a number between 0
and 100.

OUTPUT SCHEMA {
  "time_left_until_outcome_known": "string",
  "status_quo_outcome": "string",
  "no_outcome_scenario": "string",
  "yes_outcome_scenario": "string",
  "rationale": "string",
  "probability": "number" (0-100)
}
```

Deliberation prompt for Round 2:

```
You are now in a deliberation phase with two other expert
forecasters. Please review their analyses:
```

```
Forecaster 2's Analysis
{{forecaster2_rationale}} Forecast: {{forecaster2_probability}}%

Forecaster 3's Analysis
{{forecaster3_rationale}} Forecast: {{forecaster3_probability}}%

Consider their reasoning and any new information or arguments
carefully:

- What evidence or arguments did they raise that you hadn't
  considered?
- Do you find their reasoning convincing? Why or why not?
- Should you update your forecast based on their input? If so,
  how much? If not, why not?

Weigh your previous analysis and critically review your own
reasoning and evidence in light of any new information or
arguments, as if you were participating in a structured
deliberation process.

Based on your thoughtful analysis, provide a clear and concise
review of all the arguments and information you have considered,
your updated rationale, and your updated forecast. Do not feel
obligated to update your forecast if you do not think it is
warranted.

Provide your updated analysis and forecast.

OUTPUT SCHEMA {
  "review": "string (your thoughts on the other forecasters'
             reasoning)",
  "rationale": "string (your updated reasoning; if you change
               your forecast, explain why and how much; if not,
               explain why not)",
  "probability": "number" (0-100)
}
```

S3 Minimum Detectable Effects

The sample size for this study ($N = 202$ questions) was determined externally by the Metaculus Q2 2025 AI Forecasting Tournament rather than by a-priori power analysis. We therefore conducted sensitivity analyses to characterise the minimum detectable effect (MDE) for each scenario, given the fixed N and observed variance.

For a paired t-test with $N = 202$ observations at $\alpha = 0.05$ (two-sided), achieving 80% power requires a Cohen's d of approximately 0.198. The MDE in raw units (Log Loss) is then calculated as: MDE = $d\times$ SD, where SD is the observed standard deviation of paired differences for each scenario.

Table S3: Minimum Detectable Effects by Scenario

Scenario	SD of Change	MDE (80% power)	Observed Effect	p-value
Diverse models, shared information	0.117	0.023	−0.020	0.017
Diverse models, distributed information	0.237	0.047	−0.022	0.182
Same model, shared information	0.194	0.038	+0.020	0.144
Same model, distributed information	0.308	0.061	+0.008	0.717

The distributed information condition introduces additional variance into the deliberation process. When agents hold different information, their independent forecasts vary more, and consequently their post-deliberation updates also vary more. This variance increase is itself substantively meaningful: it suggests that information asymmetry creates noise that may partially obscure any deliberation benefit. Studies examining deliberation with distributed information should anticipate approximately 2× higher variance than shared-information conditions. To achieve equivalent power, such studies would require either larger samples or larger true effects. At $N = 202$, the Diverse models, distributed information condition was adequately powered only to detect effects of ~ 0.05 Log Loss or larger—roughly 2.5× the effect we observed.

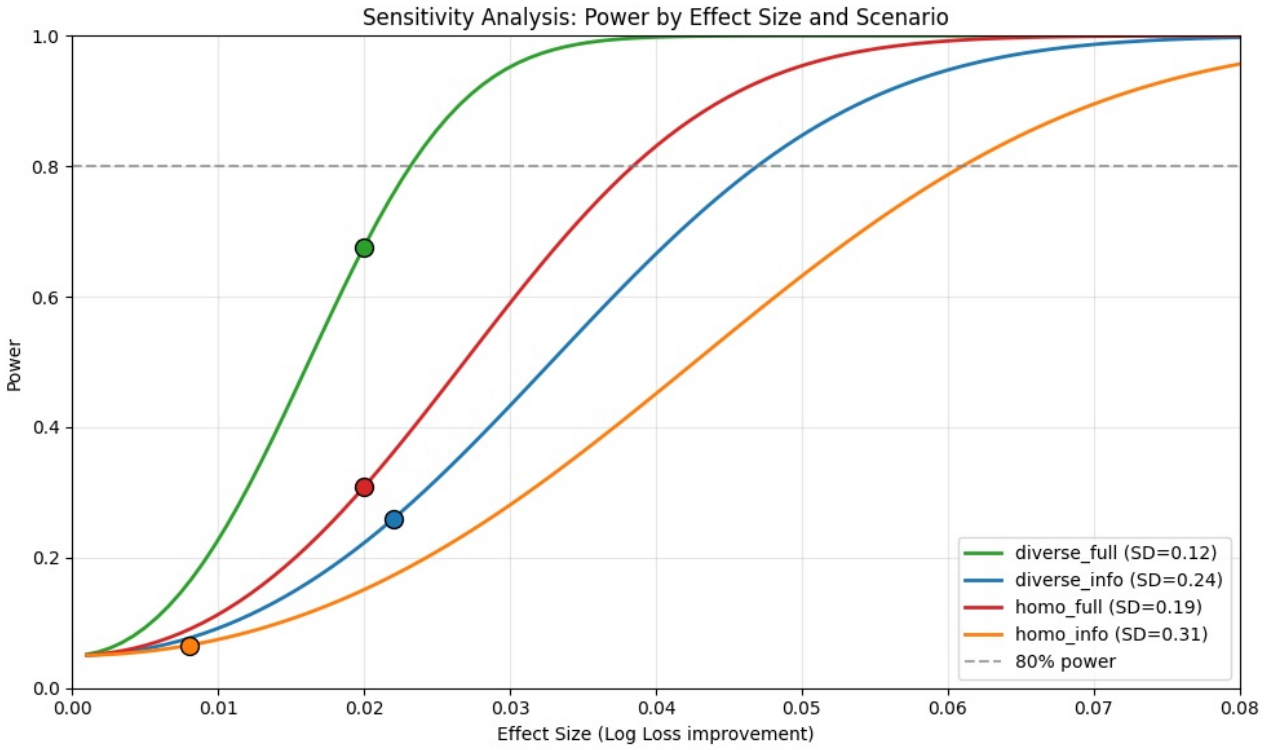

Figure S3: Sensitivity curves showing statistical power as a function of effect size for each scenario. Dots indicate observed effects; horizontal line marks 80% power threshold.

S4 Model-Level Breakdown (Homogeneous Scenarios)

Table S4: Deliberation effects by model type (homogeneous scenarios only)

Scenario	Model	n	Independent	Deliberative	Change	t	p
Same model, distributed info.	GPT-5	67	0.483	0.476	−0.007	0.56	0.58
Same model, distributed info.	Sonnet	67	0.440	0.422	−0.017	0.65	0.52
Same model, distributed info.	Pro	68	0.627	0.675	+0.047	−0.83	0.41
Same model, shared info.	GPT-5	67	0.485	0.494	+0.009	−0.99	0.33
Same model, shared info.	Sonnet	67	0.437	0.435	−0.002	0.27	0.79
Same model, shared info.	Pro	68	0.651	0.703	+0.052	−1.35	0.18

Note: Questions were distributed across model types using round-robin assignment. Values show mean Log Loss.

S5 Brier Score Sensitivity Analysis

Table S5 below shows the primary analysis repeated using Brier scores (instead of Log Loss). Brier scores are less sensitive to extreme probability estimates and may provide a complementary measure of forecast accuracy.

Table S5: Effect of deliberation on forecast accuracy by scenario (Brier Score)

Scenario	Independent mean (SD)	Deliberative mean (SD)	Change mean (SD)	t	p
Diverse models, distributed information	0.153 (0.188)	0.145 (0.201)	−0.008 (0.102)	1.14	0.26
Diverse models, shared information	0.162 (0.221)	0.153 (0.220)	−0.009 (0.051)	2.47	0.014
Homogeneous models, distributed information	0.169 (0.214)	0.170 (0.234)	+0.001 (0.123)	−0.12	0.90
Homogeneous models, shared information	0.171 (0.240)	0.177 (0.250)	+0.007 (0.061)	−1.56	0.12

Note: $n = 202$ for all scenarios. Change = Deliberative minus Independent; negative values indicate improvement.