Diversity is the Strength of the AI Crowd

Anonymous Authors[1]

Abstract

Top AI forecasting systems are approaching superforecaster-level accuracy on future world events, but still rely primarily on off-the-shelf LLMs combined with forecasting-specific context gathering and scaffolding. We study how to improve this recipe through ensembling: given a fixed number of samples, which off-the-shelf model forecasts should be combined to maximize accuracy? On binary questions from the Metaculus AI Benchmark, we find that individual accuracy is not enough: many frontier LLMs make highly correlated predictions, limiting the value of additional forecasts from the same or similar models. Instead, the strongest ensembles combine accurate but diverse forecasters, with models such as *Grok 4* contributing disproportionately because their predictions are less correlated with other frontier LLMs. These results suggest that the strength of the AI crowd comes not from sampling more forecasts indiscriminately, but from combining forecasts across models with complementary errors, motivating forecasting systems that explicitly optimize for both model quality and diversity.

1. Introduction

Top-performing AI forecasting systems are now approaching skilled human forecasters on tournament-style questions about future world events (Halawi et al., 2024; Schoenegger et al., 2024).

The prevailing recipe in current forecasting tournaments wraps an off-the-shelf large language model in a forecasting-specific pipeline that handles, for example, retrieval, structured prompting, and multi-sample aggregation. Practitioners building such systems face a cost/time constraint the literature rarely addresses head-on: given a fixed number of forecasts per question, which model forecasts should be combined to maximise performance?

A natural starting place would be to draw all samples from the strongest single model. We show this answer is wrong, for a direct reason: samples from well-performing (and indeed less well-performing) frontier LLMs are highly correlated with each other[1]. A second draw from a single model, whose distribution closely tracks the first, contributes only a fraction of the information that one draw from a similarly skilled but more diverse model would contribute. Classical ensemble theory makes this precise: an averaging ensemble's error decomposes into a member-accuracy term and a diversity term (Krogh & Vedelsby, 1994; Wood et al., 2023), so the strength of a multi-LLM crowd comes from sampling models with *complementary* errors, not from sampling more indiscriminately.

We evaluate this claim on binary questions from the Metaculus AI Benchmark Q2 2025 tournament, with five forecasters representing contemporary frontier LLMs and one fine-tuned variant of *gpt-oss-120b*. We brute-force-evaluate every valid sample allocation up to budget $B{=}5$, and report three findings. First, the strongest off-the-shelf frontier LLMs make tightly correlated predictions; *Grok 4* and the fine-tuned model sit further from this cluster on a pairwise Jensen–Shannon-divergence diversity matrix. Second, the optimal fixed-budget allocation never concentrates all samples on a single model: at $B{=}5$ it splits the budget across the fine-tuned model, *Gemini 3 Pro*, *GPT-5*, and *Grok 4*. Third, *Grok 4* is the *least-replaceable* member of the optimum despite ranking third in solo accuracy: its outsize contribution to the ensemble comes from its low correlation with the rest of the pool, not from its solo skill. [2]

2. Related Work

AI forecasting. Top AI forecasting systems now approach skilled human forecasters on tournament-style questions (Halawi et al., 2024; Schoenegger et al., 2024). The standard recipe is shared: an off-the-shelf LLM wrapped in retrieval-augmented reasoning, with multi-sample aggre-

[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.

[1] Correlation here is between the predicted probabilities; the reasoning traces that produce them may differ.

[2] A non-archival version of these findings appears as a technical blog post; the venue is redacted to preserve anonymity for review.

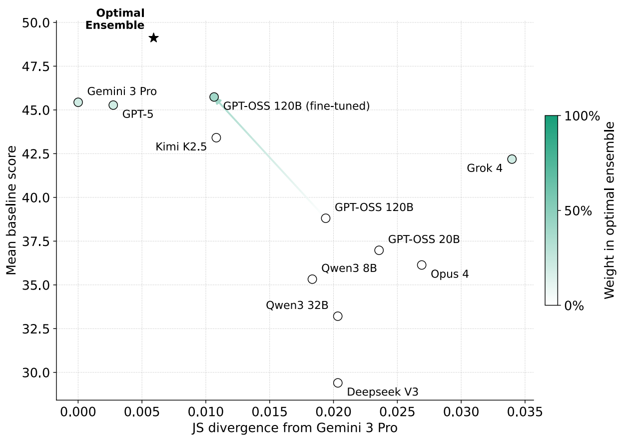

Figure 1. **Diversity pays off.** Mean baseline score on binary event questions from the Metaculus AI Benchmark Q2 2025 against the test-averaged Jensen–Shannon divergence from *Gemini 3 Pro* for our five evaluated models. Marker shade indicates each model's weight in the optimal B=5 ensemble $a_5^\star = (\text{FT}: 2, \text{Gem}: 1, \text{GPT}: 1, \text{Grok}: 1, \text{Kimi}: 0)$. See §5.1 for discussion.

gation, evaluated on dynamic benchmarks such as ForecastBench (Karger et al., 2025) and FutureX (Zeng et al., 2025) that draw resolution-unknown questions to avoid data leakage. Two axes of improvement have been pursued. The first strengthens the per-question pipeline via retrieval-augmented reasoning (Halawi et al., 2024), structured belief updating (Murphy, 2026), or task-specific fine-tuning (Turtel et al., 2025b;a). The second ensembles across multiple LLMs: aggregating 12 frontier models matches a tournament of ∼900 human forecasters (Schoenegger et al., 2024). Within this second axis, prior work has implicitly assumed uniform allocation across models; we instead ask which non-uniform allocation maximises ensemble accuracy under a fixed sample budget.

Ensemble theory and LLM ensembling. Classical ensemble learning gives a clean characterisation of when ensembling helps: under squared loss the error of an averaging ensemble decomposes as *average member error minus a non-negative diversity term* (Krogh & Vedelsby, 1994; Brown et al., 2005); the same decomposition extends to broader loss families, including the log-score we use, via the recent unified theory of Wood et al. (2023). The implication that ensembles benefit from *both* accurate *and* decorrelated members is old; what this paper adds is an empirical characterisation of the trade-off in an LLM-crowd setting, where members are heterogeneous frontier LLMs rather than resamples of a single base learner. Existing LLM-ensemble work focuses on text-generation outputs, using ranking-based or generative fusion (Jiang et al., 2023), or studies weight-space averaging within a model family (Wortsman et al., 2022). We instead ensemble probabilistic forecasts and quantify how much of an ensemble's gain comes from per-model accuracy versus per-pair diversity.

3. Preliminaries

Forecasting questions. A *binary forecasting question* q is a statement whose truth is revealed at a known resolution time, paired with a ground-truth outcome $y \in \{0, 1\}$. We write a test set as $\mathcal{D} = \{(q_i, y_i)\}_{i=1}^N$.

Forecasters and ensembles. A *forecaster* π_m for model $m \in \mathcal{M}$ is the full prediction pipeline: a model wrapped in the scaffolding described in §4, not the raw LLM call. Each forecaster is a stochastic map from a question to a probability of the affirmative outcome; $\hat{p}_{m,i}^{(r)} \in [0, 1]$ denotes the r-th sample drawn from π_m on question q_i. Sample-level stochasticity comes from nonzero sampling temperature, retrieval variability, and inference-kernel non-determinism that cannot generally be switched off (He & Thinking Machines Lab, 2025); we treat all such variability as part of π_m. We write $\bar{p}_{m,i} = \frac{1}{R_m} \sum_{r=1}^{R_m} \hat{p}_{m,i}^{(r)}$ for the per-question average over R_m independent samples. A *weighted ensemble* of M models with weights $\omega = (\omega_1, \ldots, \omega_M) \in \Delta^{M-1}$ predicts

$$\hat{p}_i^\omega \;=\; \sum_{m=1}^{M} \omega_m \, \bar{p}_{m,i}, \tag{1}$$

followed by a small clipping operation that bounds each side away from 0 and 1 by a fixed mass $\varepsilon = 0.05$; see Appendix B for the operator details. We treat this aggregation operator as fixed throughout: the question we study is which models contribute the weight, not how the contributions are combined. A *discrete team* of total sample budget B is the special case of integer weights $\omega_m = c_m/B$ with $c_m \in \mathbb{Z}_{\geq 0}$ and $\sum_m c_m = B$, which we use in §4 to define replaceability.

Scoring and diversity. We evaluate ensembles with the Metaculus *baseline score*, a rescaled log score under which a uniform 50% prediction earns 0 points and perfect foresight earns 100 (Metaculus); the score is strictly proper (Gneiting & Raftery, 2007). We report per-question baseline points (BP) and mean BP over the test set. To quantify how much two forecasters disagree on a question, we use the symmetric Jensen–Shannon (JS) divergence between the two Bernoulli predictive distributions induced by their probabilities. Writing $\bar{p} = \frac{1}{2}(\hat{p}_a + \hat{p}_b)$ for the mixture distribution, the JS divergence is

$$\mathrm{JS}(\hat{p}_a, \hat{p}_b) \;=\; \tfrac{1}{2}\,\mathrm{KL}(\hat{p}_a \,\|\, \bar{p}) + \tfrac{1}{2}\,\mathrm{KL}(\hat{p}_b \,\|\, \bar{p})\,. \tag{2}$$

We average JS across the test set to give a per-pair diversity score $\mathrm{JS}(m, m')$; for a model m relative to a set S we write $\mathrm{JS}(m, S) = \frac{1}{|S|} \sum_{m' \in S} \mathrm{JS}(m, m')$.

4. Methods

Models. We evaluate four off-the-shelf frontier LLMs (*Gemini 3 Pro*, *GPT-5*, *Grok 4*, and *Kimi K2.5*) and one fine-tuned model, *FT-gpt-oss-120b*: a fine-tuned *gpt-oss-120b* trained with reinforcement learning on a training set of forecasting questions disjoint from our test set. Each model is wrapped in the same two-stage scaffolding pipeline: a retrieval phase that gathers question-specific evidence, followed by a structured prediction phase in which the model emits a probability for the binary question. We use each model's default sampling temperature. Between-run variation comes from sampling stochasticity and inference-kernel non-determinism (§3). We draw $R_m = 5$ independent samples per model per question.

Data. Our test set is the binary subset of the Metaculus AI Benchmark Q2 2025 tournament, $N{=}113$ binary event questions resolved between May and July 2025. The full question set can be accessed here[3]. We run the research phase upfront to collect information that would have available at prediction time, and pair the output with the question to produce static prompts. The knowledge cutoff of every evaluated model precedes the tournament's start, and retrieval is limited to sources available before each question's resolution date. A sample of questions is given in Appendix C.

Continuous-weight sweeps. We characterise the ensemble in two complementary ways. For Figure 1 we report each model's solo baseline score against its test-averaged Jensen–Shannon divergence from *Gemini 3 Pro*, with marker shade encoding its weight in the discrete-team optimum $a_5^\star$ (defined below). For Figure 2 we grid the 2-simplex of three-model mixing weights with step 0.05 and score each $(\omega_1, \omega_2, \omega_3)$ point under Equation 1 on the test set.

Replaceability. At $B{=}5$ over the five-model pool with per-model cap $c_m \leq 3$, exactly 101 allocations are valid (out of 126 unconstrained). We enumerate all of them, score each with Equation 1 under the integer weights $\omega_m = c_m/B$, and identify the optimal team $a_5^\star = \arg\max_a S(a)$, where $S(a) = \frac{1}{N}\sum_i \mathrm{bs}(\hat{p}_i^a, y_i)$ is the mean baseline score. The *replaceability* of a model m in $a_5^\star$ is the score gap to the best team that *excludes* m,

$$\Delta_m \;=\; S(a_5^\star) \;-\; \max_{a:\, c_m=0} S(a), \tag{3}$$

where the maximum is over the same allocation universe with the additional constraint $c_m = 0$. A large positive Δ_m identifies m as *least-replaceable*: dropping it from the optimum costs the most relative to the best alternative team. By construction $\Delta_m = 0$ for any model not in $a_5^\star$.

[3]The link is omitted to preserve anonymity for review and will be provided in the camera-ready version.

5. Results

5.1. Pairwise mixing with a frontier anchor

Figure 1 plots each model's mean baseline score against its test-averaged Jensen–Shannon divergence from *Gemini 3 Pro*, with marker shade encoding the model's weight in the optimal $B{=}5$ team $a_5^\star$. Most frontier LLMs (*GPT-5*, *Kimi K2.5*) cluster at low JS divergence from *Gemini 3 Pro*, scoring well but contributing little diversity. *Grok 4* and *FT-gpt-oss-120b* sit further out on the JS axis, scoring comparably with the frontier cluster while correlating less, and together with *Gemini 3 Pro* and *GPT-5* they make up the four models that earn a slot in $a_5^\star$.

5.2. Three-way mixing

The pairwise sweep tracks each model against a single anchor and misses joint effects. Figure 2 (left) reports the three-way heatmap over *GPT-5*, *Gemini 3 Pro*, and *FT-gpt-oss-120b*, gridded across the 2-simplex with step 0.05. The optimum lies near $(\omega_{\mathrm{FT}}, \omega_{\mathrm{Gem}}, \omega_{\mathrm{GPT}}) = (0.56,\, 0.26,\, 0.18)$: the fine-tuned model receives the largest weight. The heatmap is asymmetric: moving along the *GPT-5–Gemini 3 Pro* edge (no FT) gains at most ≈ 0.6 BP over the better single-model corner, whereas moving toward the FT vertex picks up ≈ 2.7 BP. The right panel of Figure 2 substitutes *Grok 4* for *GPT-5* and shows the same qualitative asymmetry.

5.3. Replaceability under a $B{=}5$ budget

Switching from continuous weights to integer sample budgets, we brute-force every $B{=}5$ allocation over the five-model pool (101 valid allocations under the per-model cap $c_m \leq 3$). The optimal team is $a_5^\star = (\mathrm{FT}{:}\,2,\ \mathrm{Gem}{:}\,1,\ \mathrm{GPT}{:}\,1,\ \mathrm{Grok}{:}\,1)$. Figure 3 reports the replaceability Δ_m (Equation 3) for each member of $a_5^\star$: the score gap from $a_5^\star$ to the best team that *excludes* m. *Grok 4* is the largest contributor by a wide margin ($\Delta_{\mathrm{Grok}} \approx 1.7$ BP), with the fine-tuned model second ($\Delta_{\mathrm{FT}} \approx 0.6$ BP). *GPT-5* and *Gemini 3 Pro* contribute < 0.1 BP each, indicating that the two frontier-cluster members are largely interchangeable within the optimal team. *Grok 4* ranks only third in solo baseline score, so its outsize ensemble contribution comes from its low correlation with the frontier cluster (§5.1), which is exactly the quantity classical ensemble theory predicts should govern the diversity term of an ensemble's error decomposition (Krogh & Vedelsby, 1994; Wood et al., 2023). We make this connection explicit in Appendix A.

6. Discussion

The pattern across our three results is consistent with the diversity term in the classical ambiguity decomposi-

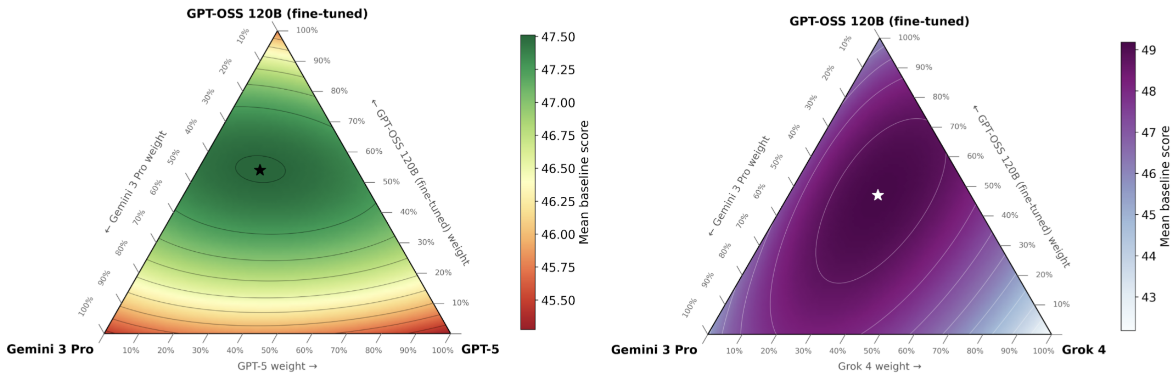

Figure 2. **Three-way ensembles.** Mean baseline score over the three-model weighted ensemble simplex. *Left: FT-gpt-oss-120b*, *Gemini 3 Pro*, and *GPT-5*; the optimum ($\star$) lies at $(\omega_{\mathrm{FT}}, \omega_{\mathrm{Gem}}, \omega_{\mathrm{GPT}}) \approx (0.56, 0.26, 0.18)$. *Right: Grok 4* substitutes for *GPT-5*; the optimum lies at $(\omega_{\mathrm{FT}}, \omega_{\mathrm{Gem}}, \omega_{\mathrm{Grok}}) \approx (0.48, 0.26, 0.26)$. In both, the fine-tuned model receives about half the weight.

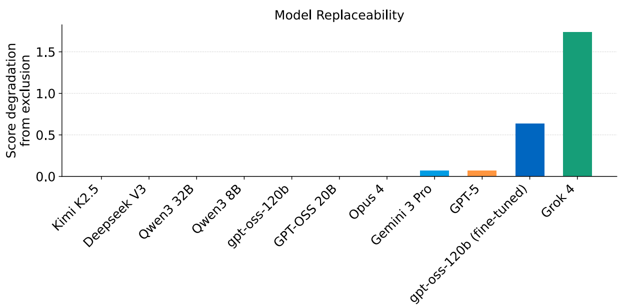

Figure 3. **When selecting an ensemble of frontier and open-source models, *Grok 4* and fine-tuned *gpt-oss-120b* are the least replaceable.** Model replaceability is defined as the reduction in score incurred when removing a model from the optimal ensemble (Eq. 3). By definition, if a model is not included in the optimal ensemble, there is no cost to removing it.

tion (Krogh & Vedelsby, 1994; Wood et al., 2023): at a fixed sample budget the marginal value of an additional forecast decomposes into a solo-accuracy contribution and a diversity contribution. In our pool the five solo scores cluster within a few BP of each other (§5.1), so it is the diversity contribution that differentiates additions, and the most-decorrelated model in the pool, *Grok 4*, is the least-replaceable member of the optimal team (§5.3) despite ranking only third in solo accuracy. For practitioners building forecasting systems, this means the natural strategy of stacking samples from the strongest single model is suboptimal relative to mixing in a decorrelated alternative.

For model developers the implication is more speculative but worth flagging: training pipelines that incidentally decorrelate via distinct pre-training (*Grok 4* is plausibly one such case in our pool) or deliberately via task-specific fine-tuning (*FT-gpt-oss-120b*) carry ensemble-level value that is not visible in a head-to-head solo accuracy comparison. A model that is third in solo accuracy can still be first in marginal ensemble contribution.

7. Limitations

Our study uses $N{=}113$ binary event questions from a single tournament, caps independent runs per model at $R_m = 3$, ignores per-sample cost differences across models, and uses one replaceability definition (leave-one-out re-optimisation) among several plausible choices.

8. Conclusion

Across pairwise, three-way, and discrete-team analyses on the Metaculus AI Benchmark Q2 2025 binary set, we have shown that the strength of the AI crowd lies in combining forecasters that are both accurate and decorrelated.

Impact Statement

This paper studies how to ensemble off-the-shelf language-model forecasters. Improvements in automated forecasting could inform decision-making across the economy and in government (Tetlock & Gardner, 2015; Karger et al., 2025); the specific contribution here, identifying which models contribute most to a forecasting ensemble, is intended to make existing forecasting systems more accurate, not to change the kinds of decisions those systems are used for. We see no specific ethical or societal harms beyond those already discussed in the broader AI forecasting literature.

References

Brown, G., Wyatt, J., Harris, R., and Yao, X. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2402.18563.

He, H. and Thinking Machines Lab. Defeating nondeterminism in LLM inference. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/, 2025. Connectionism research blog.

Jiang, D., Ren, X., and Lin, B. Y. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Annual Meeting of the Association for Computational Linguistics*, 2023.

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*, 2025. arXiv:2409.19839.

Krogh, A. and Vedelsby, J. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, 1994.

Metaculus. Metaculus scoring documentation: Baseline score. https://www.metaculus.com/help/scores-faq/. Accessed: 2026-05-13.

Murphy, K. P. Agentic forecasting using sequential Bayesian updating of linguistic beliefs. *arXiv preprint arXiv:2604.18576*, 2026.

Schoenegger, P., Tuminauskaite, I., Park, P. S., Bastos, R. V. S., and Tetlock, P. E. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45), 2024. doi: 10.1126/sciadv.adp1528.

Tetlock, P. E. and Gardner, D. *Superforecasting: The art and science of prediction*. Crown, 2015.

Turtel, B., Franklin, D., Skotheim, K., Hewitt, L., and Schoenegger, P. Outcome-based reinforcement learning to predict the future. *arXiv preprint arXiv:2505.17989*, 2025a.

Turtel, B. et al. LLMs can teach themselves to better predict the future. *arXiv preprint arXiv:2502.05253*, 2025b.

Wood, D., Mu, T., Webb, A. M., Reeve, H. W. J., Luján, M., and Brown, G. A unified theory of diversity in ensemble learning. *Journal of Machine Learning Research*, 24 (359):1–49, 2023.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., and Schmidt, L. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022.

Zeng, Z. et al. FutureX: An advanced live benchmark for LLM agents in future prediction. *arXiv preprint arXiv:2508.11987*, 2025.

A. Ambiguity decomposition of the ensemble's gain

We make explicit the connection between the empirical replaceability ranking of Section 5.3 and the *ambiguity decomposition* of an averaging ensemble. For a single binary question with outcome $y \in \{0,1\}$, member predictions $\hat{p}_1, \ldots, \hat{p}_B$, and uniform ensemble $\bar{p} = \frac{1}{B}\sum_b \hat{p}_b$, Krogh and Vedelsby's identity (Krogh & Vedelsby, 1994) gives the squared error of the ensemble as the average squared error of its members minus a non-negative diversity term,

$$(\bar{p} - y)^2 \;=\; \frac{1}{B}\sum_{b=1}^{B}(\hat{p}_b - y)^2 \;-\; \frac{1}{B}\sum_{b=1}^{B}(\hat{p}_b - \bar{p})^2. \tag{4}$$

This identity is exact under squared loss. Wood et al. (2023) generalise it to arbitrary Bregman divergences, including the log-score behind the Metaculus baseline metric we report: the ensemble strictly improves on the average member exactly to the extent that members disagree, regardless of the choice of Bregman loss. The Δ_m ranking of Section 5.3 is the empirical counterpart of the per-member diversity term: a model whose removal forces the next-best team to be substantially worse is one that contributed disproportionately to the diversity term of the ensemble it was part of.

B. Aggregation operator and haircut parameters

The ensemble probability $\hat{p}_i^{\omega}$ used throughout (Eq. 1) is followed by a small clipping operation that bounds the affirmative-outcome probability away from $\{0,1\}$. Concretely, the raw weighted mean $\bar{p} \in [0,1]$ is replaced by

$$\hat{p} \;=\; \max\!\big(\varepsilon,\, \min(1-\varepsilon,\, \bar{p})\big),$$

where $\varepsilon = 0.05$ is the per-side clipping mass. This matches the binary special case of the projection used inside the forecasting pipeline that supplied our predictions and is required for the log-score to be finite on resolution. All results reported in the main body are under this fixed aggregation operator; results under the unclipped mean and under larger ε are qualitatively unchanged but shift mean BP by a small additive constant.

C. Question sample

To give the reader a sense of the kinds of questions in the test set (their topics, framings, and resolution timing), we list three examples drawn from the $N{=}113$ binary event questions of the Metaculus AI Benchmark Q2 2025 tournament.

- *Will Haley Stevens announce her candidacy for US Senator from Michigan before May 1, 2025?* Resolution date 2025-05-01. Resolved: yes.
- *Before July 1, 2025, will the government of Greenland officially announce a date for an independence referendum?* Resolution date 2025-07-01. Resolved: no.
- *Before July 1, 2025, will Discord announce that it is planning an IPO?* Resolution date 2025-07-01. Resolved: no.