Evaluating Long-Form Forecasts by Their Effect on Downstream Predictions

Anonymous Authors[1]

Abstract

Language model evaluations for judgmental forecasting currently test multiple-choice or short-answer predictions to fully specified questions. However, when reasoning about the future, the relevant questions aren't known in advance, making real-world forecasting inherently open-ended. In this work, we study how to evaluate responses to questions like "How will AI capabilities progress by 2027?", which have no single ground truth. Our key contribution is measuring the quality of a long-form forecast by how it updates the world model of a downstream predictor. Specifically, we measure how much the predictive accuracy of weaker models improves on a sample of world events, once provided the long-form forecast in-context. We test seven frontier models under this framework, finding meaningful differences in their long-form forecasts about AI progress. We hope our methodology helps measure and improve the quality of forecasts for long-term decision-making.

1. Introduction

Language model forecasting evaluations currently focus on fully specified, short answer forecasting questions (Karger et al., 2025; Yang et al., 2026; Chandak et al., 2026). While this setting has clear resolution criteria and proper scoring rules, it fails to capture open-ended real-world forecasting questions which decision-makers are often interested in. For example, "How would AI breakthroughs progress in 2027?", or "What effect will the development of AGI have on our society".

These open-ended questions elicit long-form responses, automatic evaluations for which have been a long-standing challenge in language modelling (Hovy, 1999). Over the years, many proxy evaluation metrics based on lexical overlap (Papineni et al., 2002), and more recently language model judges (Zheng et al., 2023) have been proposed. Yet, they all require access to a reference solution or high quality rubrics (Arora et al., 2025), which are hard to create for forecasting given it is extremely challenging for humans as well.

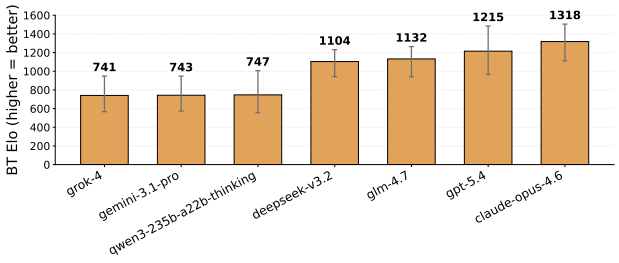

Figure 1. Elo scores of long-form forecasts by frontier language models based on our proposed evaluation framework. For pairwise comparison datapoints between models, we use the downstream brier skill scores obtained when predictions based on the long-form foreast are made on verifiable questions.

Specifically for long-form forecasts, evaluations must correctly weigh multiple desiderata for claims made, like correctness, importance, surprise, and calibrated confidence. In this work, we show these desiderata can be captured into a unified evaluation based on the downstream utility of a long-form forecast in leading to directionally correct updates in it's readers beliefs about the world.

We benchmark frontier models at generating long-form forecasts about a topic, by providing each forecast to weaker models that then have to predict verifiable forecasting questions about the same topic. Better long-form forecasts should lead to higher brier score when the weaker models are conditioned on it.

We apply this framework to forecasting AI progress, as this is a high impact domain which is evolving rapidly, and has large disagreements among experts. We obtain 318 verifiable questions about recent AI events after the knowledge cutoff of both the long-form forecasters and downstream prediction models we test. We validate this measurement by ensuring that conditioning on the forecast substantially improves brier score for the downstream predictors. Then, we compare the long-form forecasting ability of the frontier models test, finding our evaluation shows meaningful,

[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.

statistically significant differences between closed models like GPT 5.4 and Opus 4.6 and frontier open-weight models. Interestingly, Grok-4 and Gemini 3.1 Pro which often score highly on verifiable short-form forecasting benchmarks perform poorly at long-form forecasting, which we validate with qualitative inspection.

Overall, our main contributions are as follows:

1. We study using language models to generate long-form forecasts for open-ended questions about the future.
2. We ground the evaluation of these forecasts by measuring how useful they are to predict verifiable questions.
3. As a proof of concept, we test frontier models on their ability to forecast AI progress and find our evaluation clearly separates their capability.

2. Method

We first describe our methodology for grounding long-form forecast evaluations in verifiable outcomes, followed by how we reduce variance within this framework.

2.1. Evaluation Framework

A long-form forecast is challenging to evaluate because it has multiple, sometimes conflicting desiderata. For example, four distinct dimensions for a high quality forecast are:

1. **Correctness.** Correct claims about the future are naturally more useful than wrong ones.
2. **Importance.** The claims made in the forecast should be about as important events as possible. For example, presidential election results are more important than a layperson's opinion on a policy.
3. **Surprise.** The claims should not be trivial, and attempt to predict an uncertain variable. For example, there is not much value in predicting the sun will rise in the east as there is no uncertainty in this, while predicting whether a day will be sunny or not can be valuable.
4. **Calibrated Confidence.** Since forecasting involves predicting uncertain variables, claims made by the model should be appropriately hedged. Both overconfidence, and underconfidence should be penalized.

One could define separate evaluations for each of these dimensions, but then it becomes unclear how to strike a tradeoff between these goals. For example, it is possible to obtain high correctness by making trivial claims about low surprise events. It is also possible that it is harder to predict high importance events better than a crowd than low importance personal ones. Given these goals can conflict, in this work we ask:

> Can we build a unified evaluation for long-form forecasts that captures and weighs these desiderata appropriately?

We believe the impact of a long-form forecast lies in how informative it is for its *readers*. This means the forecast should update its reader's beliefs about the world in the right direction. More correct, important, yet surprising and calibrated forecasts would lead to a directionally very good update in the reader's beliefs.

Then, how can we implement a measurement of the belief change in reader's given access to a long-form forecast? We propose approximating a user's belief state by sampling concrete, verifiable questions about the world, and asking the user to answer them. We can then measure the change in scores of the downstream predictor with and without the long-form forecast. The better the predictor performs given the forecast, the more useful the long-form forecast is. This effectively grounds evaluations of long-form forecasts in verifiable outcomes through the proxy reader.

Formally, for each prediction, we compute a Brier score against the resolved binary outcome,

$$\text{Brier} = (\hat{p} - y)^2,$$

where $\hat{p}$ is the predicted probability and $y \in \{0, 1\}$ is the realized outcome.

For a generator g, predictor p, and question q, we define forecast usefulness as the reduction in Brier score relative to the same predictor without the forecast:

$$\Delta_q^{p,g} = \bar{\beta}_{p,q}^{\text{base}} - \bar{\beta}_{p,g,q}^{\text{scen}}.$$

Positive Δ means the forecast helped the predictor. We average this quantity across questions and predictors.

2.2. Variance Reduction

Within our framework, there are multiple sources of variance in the final evaluation metric for long-form forecasts. These include:

- **Variance in downstream predictor performance.** Given we sample questions to estimate the downstream predictor's belief state over the world, we must sample a sufficient number of questions to reduce this source of error.
- **Predictor used.** We want to minimize the effect of individual idiosyncracies in downstream predictors on the evaluation outcomes. We view this as sampling

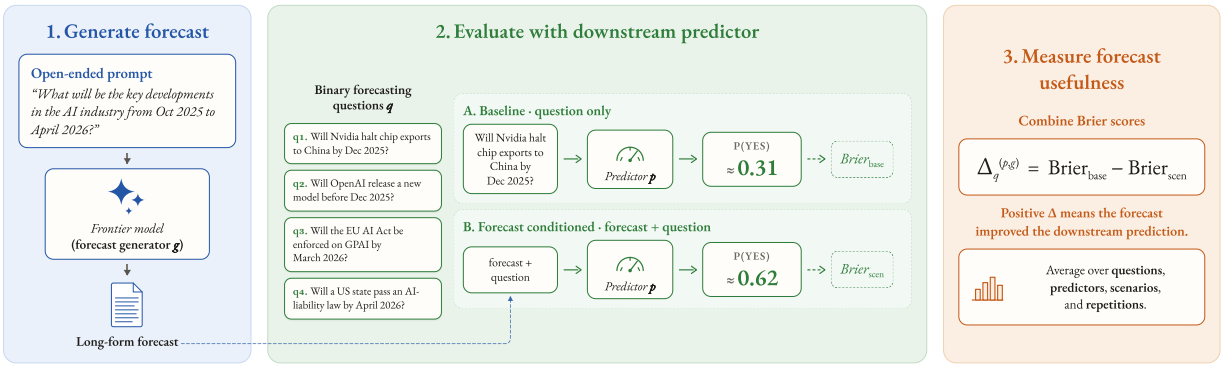

Figure 2. **Our proposed evaluation framework**: We consider a long-form forecast useful if it informs better downstream predictions on verifiable forecasting questions. We ensure that both the initial prompt and downstream questions are outside the knowledge cutoff of both the forecaster and downstream predictor.

a sufficient number of readers from the population to reduce this source of error.

- **Long-form forecast generation.** Since language model outputs are sampled stochastically, there can be variance in the quality of the long-form forecast produced itself.

A natural way to compute variance in our final metric is thus using a hierarchical bootstrap over each of these variables. We provide details in Section C. One shortcoming of this approach is that the absolute change in performance of different predictors can vary widely depending on both their initial, and in-context reasoning ability. This is not variance arising from the actual quantity of interest: the ability of the long-form forecast generator, but rather the measurement apparatus. This is why we compute and report Elo ratings to rank the long-form forecast generators, which is not affected by variance in absolute change in scores. We can then treat each the change in brier skill score obtained for each (downstream prediction, question) tuple as a pairwise comparison for computing Elo. We group accordingly in our hierarchical bootstrap to obtain variances in Elo ratings. Details about hierarchical bootstrap computation are provided in Section C.

3. Experimental Setup

While our framework is broadly applicable, here we instantiate it in the domain of AI asking the model to predict the key developments in the AI ecosystem from October 2025 to April 2026, outside their knowledge cutoff. We evaluate seven frontier models from different providers on this aggregated set of 318 questions. For downstream questions, we source all resolved questions on AI in this time window from prediction markets. However, this gave us only 60 questions. To gather more questions grounded in real-world, we synthesize 258 additional binary questions from AI newsletters in that period following a pipeline similar to Chandak et al. (2026). Questions were generated synthetically from event reports, filtered for clear resolution criteria, scenario relevance, and verifiable outcomes. Full dataset construction details is provided in Section B.

How to efficiently obtain downstream predictors? Running downstream evaluation with real-users would be time-taking and hard to reproduce. For this work, we use a language model as the downstream predictor, since they are now capable of making reasonable predictions on world events (Yang et al., 2026) and can effectively leverage information provided in-context (Brown et al., 2020). While language models are far from perfect simulators of human outcomes today (Seshadri et al., 2026), it is a promising research direction with rapid progress being made (Wu et al., 2026). Our evaluation framework can be considered an application of it.

To assess long-form forecast's utility in a *robust* manner, we measure downstream performance across three weaker predictors each from a different model family: Llama-4-Maverick, Gemma-3-27B, and Qwen3-30B-A3B. We use Brier Score as the main metric for performance which is a proper scoring rule (Guo et al., 2017). Each predictor answers every question in two modes: 1) without the long-form forecast (baseline brier) and 2) conditional on the long-form forecast (conditional brier). The final score is aggregated over 3 seeds of long-form output and over all the downstream predictor with 3 (conditional) generations from each of them to account for any variance.

Luckily, as a reference, we also have the AI 2027 scenario forecast (Kokotajlo et al., 2025) which we use as a human baseline. Additionally, we evaluate the current Wikipedia page on AI as a long-form output baseline and also assess compilation of past Metaculus AI questions as input to the downstream model to see whether any past resolution help in future prediction or not. Finally, we also measure a

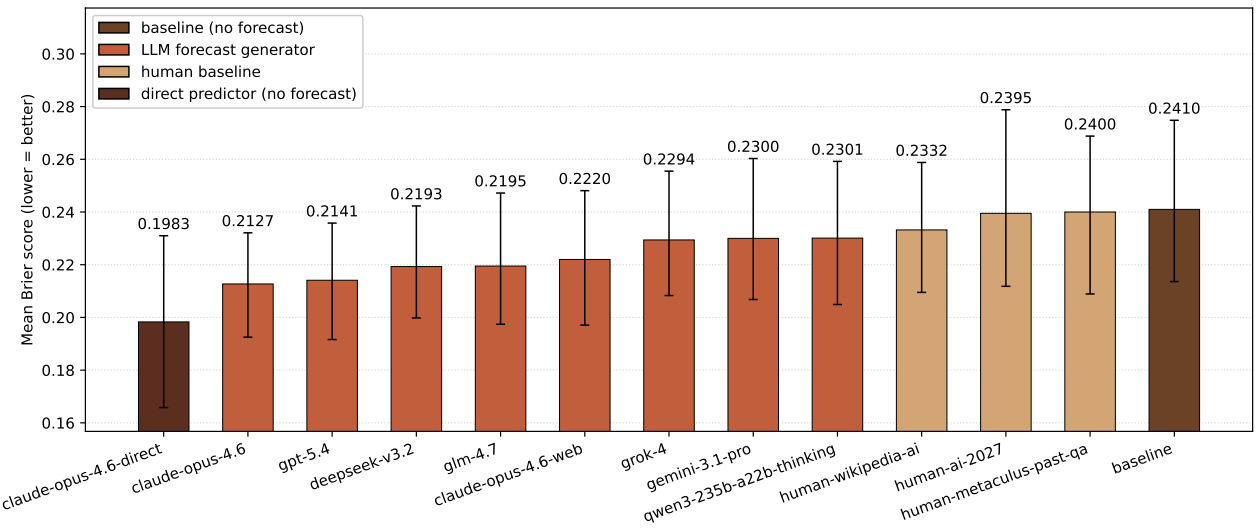

Figure 3. Mean Brier score for forecast-conditioned predictors, baselines, human-written forecasts, and a direct strong-model predictor. Error bars are from the hierarchical bootstrap over questions, forecasts, predictors, and prediction repetitions. Lower is better. Brown bars represent baselines and ceiling estimates, while orange bars represent the frontier models tested.

soft ceiling by directly evaluating a frontier model (Claude-Opus-4.6) on downstream questions.

4. Results

We report performance of frontier models on two complementary metrics: absolute Brier score against baselines and ELO ranking of frontier models estimated from paired comparisons. Together, these analyses showcase how useful and accurate is a model's long-form forecast and rank them by the effectiveness of their forecast's utility to downstream predictors.

Top forecasts improve downstream predictions. Figure 3 shows that long-form forecast from frontier models can substantially improve downstream predictions. The best forecaster, Claude Opus 4.6, reduces mean Brier score from the no-forecast baseline of 0.241 to 0.212, with GPT-5.4 close behind at 0.214. These reductions showcase that the strongest forecasts actually provide useful information rather than merely appearing plausible.

Model-generated forecasts show a clear ranking. Figure 1 shows the ordering among the seven frontier models. Claude Opus 4.6 is ranked first followed by GPT-5.4. GLM-4.7 and DeepSeek-V3.2 follow closely. Surprisingly, Grok-4 and Gemini 3.1 Pro which perform well on verifiable short-form questions (Karger et al., 2025) lag behind in their ability to produce long-form forecasts about AI progress, with significantly lower ELO.

Baselines distinguish generic context from relevant forecast information. We include direct prediction and human-written forecasts as reference points for interpreting the performance of long-form forecasts. Direct Claude Opus 4.6 prediction approximates an upper bound with the best raw Brier score of 0.198. Human-written forecasts provide a contrasting reference for broad, reusable forecasts not optimized for this question set. Interestingly, the AI 2027 scenario forecast barely improves over the no-forecast baseline. On going through it, we find it to be fairly broad suggesting that they don't provide enough information for concrete downstream predictions, especially for our question set. This should not be interpreted as frontier models outperforming superforcasters as we evaluate only one reference point here.

5. Conclusion

We propose an evaluation framework for long-form forecasts based on how much they help predictions of concrete verifiable forecasting questions. Our evaluation on forecasting AI progress shows interesting differences between frontier model capabilities at long-form forecasting. This result shows initial signs that our framework could pave the way for grounded evaluations of long-form forecasts on open-ended questions relevant to decision makers.

References

Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., and Singhal, K.

Healthbench: Evaluating large language models towards improved human health, 2025. URL https://arxiv.org/abs/2505.08775.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

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

Dai, H., Teehan, R., and Ren, M. Are llms prescient? a continuous evaluation using daily news as the oracle, 2025. URL https://arxiv.org/abs/2411.08324.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

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

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Hovy, E. H. Toward finely differentiated evaluation metrics for machine translation. In *Proceedings of the EAGLES Workshop on Standards and Evaluation Pisa, Italy, 1999*, 1999.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. Forecastbench: A dynamic benchmark of AI forecasting capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lfPkGWXLLf.

Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., and Dean, R. AI 2027. https://ai-2027.com/, April 2025. Published April 3, 2025. AI Futures Project. Accessed May 13, 2026.

Liu, J., Chen, S., Wang, Z., Zeng, Z., Guo, J., Hu, L., Yin, L., Huang, S., Hao, W., Yang, Y., Cheng, Z., Yao, Z., Yin, L., Liu, H., Cheng, J., Li, Y., Ma, Z., Wang, B., Qiu, B., Liu, X., Zhang, Z., Liu, Z., Wang, J., Yin, M., He, T., Liao, Y., Tian, Y., Zhu, Z., Dai, A., Zhang, G., Liu, J., Zhang, K., Wu, W., Gao, X., Chen, X., Yao, Z., Wen, Z., Prakash, B. A., Blanchet, J., Wang, M., Si, N., and Huang, W. Futurex-pro: Extending future prediction to high-value vertical domains, 2026. URL https://arxiv.org/abs/2601.12259.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D. (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Qin, J. and Andriushchenko, M. Quantsightbench: Evaluating llm quantitative forecasting with prediction intervals, 2026. URL https://arxiv.org/abs/2604.15859.

Seshadri, P., Cahyawijaya, S., Odumakinde, A., Singh, S., and Goldfarb-Tarrant, S. Lost in simulation: Llm-simulated users are unreliable proxies for human users in agentic evaluations, 2026. URL https://arxiv.org/abs/2601.17087.

Wang, Z., Zhou, X., Yang, Y., Ma, B., Wang, L., Dong, R., and Anwar, A. OpenForecast: A large-scale open-ended event forecasting dataset. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S. (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5273–5294, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.353/.

Wu, S., Choi, E., Khatua, A., Wang, Z., He-Yueya, J., Weerasooriya, T. C., Wei, W., Yang, D., Leskovec, J., and Zou, J. Humanlm: Simulating users with state alignment beats response imitation, 2026. URL https://arxiv.org/abs/2603.03303.

Yang, Q., Mahns, S., Li, S., Gu, A., Wu, J., and Xu, H. LLM-as-a-prophet: Understanding predictive intelligence with prophet arena. In *ICLR*, 2026. URL https://openreview.net/forum?id=VpiHkMSPqI.

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

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang,

H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

A. Related Work

LLM forecasting benchmarks. Recent work has evaluated language models as judgmental forecasters by asking them to assign probabilities to resolved or soon-to-resolve questions. ForecastBench (Karger et al., 2025), FutureX (Zeng et al., 2025), and FutureX-Pro (Liu et al., 2026) provide dynamic benchmarks for tracking model forecasting performance, often using prediction-market-style binary or multiple-choice questions.In addition, QuantSightBench (Qin & Andriushchenko, 2026) extends this line of evaluation to numerical forecasting, assessing models' ability to produce prediction intervals in an agentic setting. These benchmarks are valuable because they produce clear targets and can be scored with proper scoring rules such as the Brier score (Gneiting & Raftery, 2007).

Open-ended future prediction. Recent forecasting evaluations have begun to move beyond binary or multiple-choice prediction-market questions toward more open-ended formulations. News-oracle evaluations (Dai et al., 2025) and automated question-generation pipelines such as OpenEP (Guan et al., 2024) construct future-event questions from news and other text sources, expanding the supply of short-form forecasting targets. OpenForecast (Wang et al., 2025) further studies open-ended event prediction, and the OpenForesight scaling pipeline (Chandak et al., 2026) emphasizes automated construction and leakage control. These works broaden what is forecast, but still primarily evaluate a model's direct answer to a predefined event-level question. Our setting instead asks whether a model can produce a reusable long-form forecast that improves later forecasts about a domain, even when the exact downstream questions are not shown to the generator.

B. Evaluation Data Construction

The AI evaluation set contains 318 resolved binary questions. It combines a 60-question externally sourced prediction-market subset with 258 synthetic questions grounded in AI Newsletter material. The final file is balanced enough for aggregate evaluation but is not forced to be exactly 50/50: it contains 151 yes resolutions and 167 no resolutions.

External prediction-market subset. The 60-question external subset is the balanced market set used by the evaluation pipeline. It contains 35 Metaculus questions and 25 resolved questions from Kalshi, Polymarket, and Manifold. We include these questions as a human-written anchor distribution: they are naturally phrased prediction-market questions, have external resolutions, and are not generated by the same pipeline that produces the synthetic questions.

Synthetic newsletter subset. The 258 synthetic questions were generated from AI Newsletter material using a shared automated pipeline with an explicit September 2025 forecasting vantage and April 30, 2026 resolution date. The pipeline generated candidate trajectory questions from 73 substantive AI Newsletter issues, structurally validated them, removed keyword and semantic duplicates, applied a two-judge clarity gate, and then re-derived labels using a web-search-assisted verifier. Because article-grounded generation overproduces yes outcomes, a later pass inverted some confirmed-yes questions into confirmed-no questions and re-verified them. Finally, a forecast-friendly filter removed questions that were too narrow, such as one-off product-version lookups, and a reframing pass lifted some narrow events into broader industry-level questions.

Filtering funnel. The automated AI Newsletter run produced 913 candidate questions. Structural validation kept 714, deduplication kept 656, the clarity gate kept 426, and web verification kept 416. The inversion pass added 191 confirmed-no questions, yielding a 607-question audit pool with web evidence and source URLs. The forecast-friendly filtering and reframing stages produced a broad forecast-relevant pool, from which the paper uses the 258-question synthetic subset.

C. Hierarchical Bootstrap Details

All error bars are computed by non-parametric bootstrap over the experimental hierarchy. In our experiment, there is a single domain, so the bootstrap resamples questions with replacement, draws long-form forecast indices from the three forecasts per generator, resamples the downstream predictor panel, and resamples baseline and prediction repetitions; Section C gives the formal indexing scheme. For raw Brier scores, each bootstrap iteration recomputes the mean Brier under these resampled draws; the plotted 95% intervals are the empirical 2.5 and 97.5 percentiles over $B = 1000$ iterations.

For generator comparisons, we use the same hierarchical resample for every generator within a bootstrap iteration. This pairing is important: hard questions or noisy predictor samples affect all generators together, so pairwise differences are more stable than marginal intervals on absolute Δ values. Let $\Theta_{b,g}$ be generator g's mean Δ in bootstrap iteration b. We convert the paired bootstrap matrix into head-to-head outcomes by counting, for each pair (a, b), how often $\Theta_{\cdot,a} > \Theta_{\cdot,b}$.

We summarize these paired outcomes with a Bradley–Terry model. Each generator has latent skill s_g, and

$$\Pr(a \succ b) = \sigma(s_a - s_b),$$

where σ is the logistic function. We fit s by maximum

likelihood on the bootstrap win counts and center skills to have mean zero. For interpretability, we map it to Elo via

$$\mathrm{Elo}_g = 1000 + s_g \frac{400}{\ln 10}.$$

Elo intervals are computed by an outer bootstrap: we resample questions, rerun the paired hierarchical bootstrap, refit Bradley–Terry, and transform the resulting skill percentiles to Elo. Therefore, the Elo error bars reflect the uncertainty in the ranking resulting from the entire resampling process, rather than just the uncertainty in the optimiser during the Bradley–Terry fit.

Let $\beta^{\mathrm{base}}_{p,q,r}$ denote the Brier score for downstream predictor $p \in \mathcal{P}$, question $q \in \{1, \dots, Q\}$, and baseline prediction repetition $r \in \{1, 2, 3\}$. Let $\beta^{\mathrm{scen}}_{p,g,q,s,r}$ denote the corresponding forecast-conditioned Brier score for generator $g \in \mathcal{G}$, forecast draw $s \in \{1, 2, 3\}$, and prediction repetition $r \in \{1, 2, 3\}$.

For bootstrap replicate b, we sample a predictor multiset $P^{(b)}$ of size $|\mathcal{P}|$, a question multiset $Q^{(b)}$ of size Q, forecast indices $s_q^{(b)} \sim \mathrm{Unif}\{1, 2, 3\}$, baseline repetition indices $r_{p,q}^{(b),\mathrm{base}} \sim \mathrm{Unif}\{1, 2, 3\}$, and forecast repetition indices $r_{p,q}^{(b),\mathrm{scen}} \sim \mathrm{Unif}\{1, 2, 3\}$, all with replacement. The baseline and generator-specific means are

$$\hat{\mu}^{(b)}_{\mathrm{base}} = \frac{1}{|P^{(b)}|\,|Q^{(b)}|} \sum_{p \in P^{(b)}} \sum_{q \in Q^{(b)}} \beta^{\mathrm{base}}_{p,q,r^{(b),\mathrm{base}}_{p,q}},$$

$$\hat{\mu}^{(b)}_{g} = \frac{1}{|P^{(b)}|\,|Q^{(b)}|} \sum_{p \in P^{(b)}} \sum_{q \in Q^{(b)}} \beta^{\mathrm{scen}}_{p,g,q,s^{(b)}_q,r^{(b),\mathrm{scen}}_{p,q}}.$$

The bootstrap improvement for generator g is

$$\hat{\Delta}^{(b)}_g = \hat{\mu}^{(b)}_{\mathrm{base}} - \hat{\mu}^{(b)}_g.$$

Raw Brier intervals are empirical quantiles of $\{\hat{\mu}^{(b)}_g\}_{b=1}^{B}$, while improvement and paired-comparison quantities use $\{\hat{\Delta}^{(b)}_g\}_{b=1}^{B}$. For Bradley–Terry Elo, all generators share the same $P^{(b)}$, $Q^{(b)}$, and repetition draws in each replicate, so pairwise wins compare generators on matched resampled experimental draws.

D. Predictor Capability and Scenario Use

We ran an auxiliary GPT-family ablation to test whether the value of a long-form forecast depends on the capability of the downstream predictor. This ablation holds the questions, generators, and forecast texts fixed while varying only the predictor: GPT-5.4, GPT-5.4-mini, and GPT-5.4-nano.

Figure 4 suggests a non-monotonic relationship between predictor capability and forecast benefit. At the high-capability end, GPT-5.4 receives little measurable benefit from added forecasts, consistent with a ceiling effect: its unconditioned predictions already encode much of the information supplied by the long-form forecast. At the low-capability end, GPT-5.4-nano is often harmed by forecast conditioning, suggesting that it may lack the reasoning capacity needed to translate broad forecast claims into calibrated updates for specific questions. The largest average gains occur for GPT-5.4-mini, which appears to sit in an intermediate regime where the model can use the forecast but still has enough headroom to improve.

E. Additional Generator Diagnostics

Figure 5 and Figure 6 show two exploratory diagnostics for the seven generators. These plots are not intended as causal tests, since model release date, training data recency, architecture, post-training, and prompt-following ability are all confounded. They are nevertheless useful for interpreting what the downstream-utility metric appears to reward.

The release-date trend is consistent with the idea that long-form AI forecasts benefit from current world knowledge. A generator with a more recent training or post-training cutoff may better capture recent product releases, benchmark progress, compute investments, regulation, and organizational changes that shape the question distribution. This is especially relevant in AI progress forecasting, where the evaluation horizon is short and the domain changes quickly.

The length trend suggests that the framework may reward forecasts that are more thorough, not merely more accurate at isolated facts. Longer forecasts can include more mechanisms, actors, failure modes, and conditional branches, giving the downstream predictor more opportunities to connect the forecast to a later prediction-market question. However, this diagnostic does not show that verbosity itself is beneficial. Length may be a proxy for deliberation quality, coverage of relevant uncertainties, or model capability; excessively long or poorly structured forecasts could still hurt downstream prediction. Two additional checks support this interpretation. First, the factuality analysis in **??** shows that claim-level factual correctness alone does not determine downstream utility. Second, the human-written reference forecasts in Figure 3, including the Wikipedia-style baseline, do not automatically outperform the no-forecast baseline despite providing substantial background context. Together these checks suggest that the framework is not simply rewarding longer text, but rather useful forecast structure.

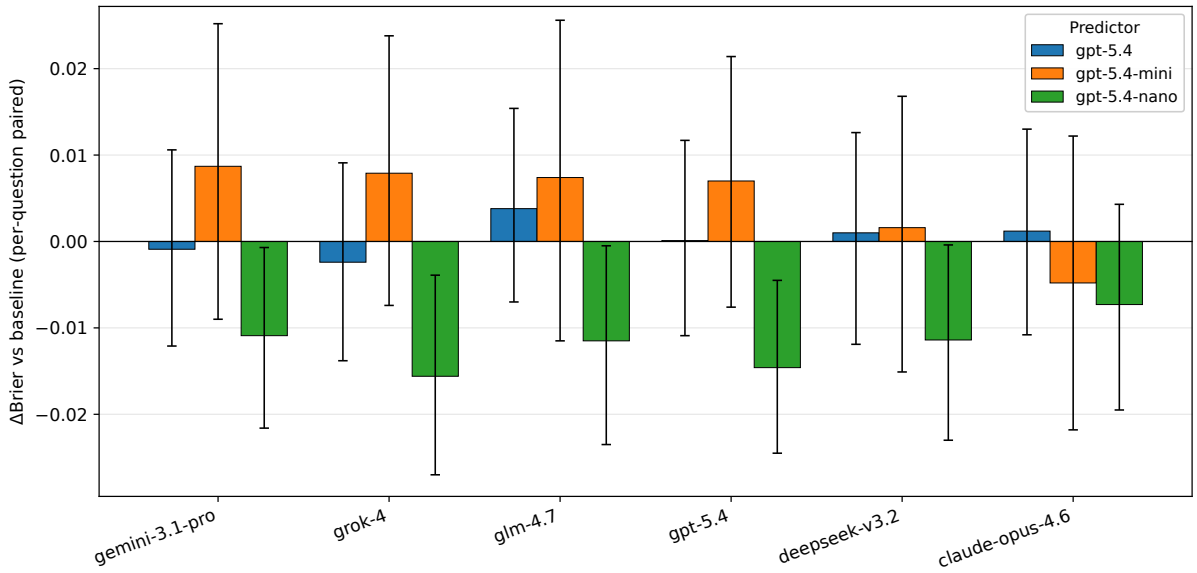

Figure 4. Auxiliary GPT-family predictor ablation in which we vary the predictor's capability level within the same model family. Positive ΔBrier means that conditioning on the long-form forecast reduced Brier score relative to the no-forecast baseline.

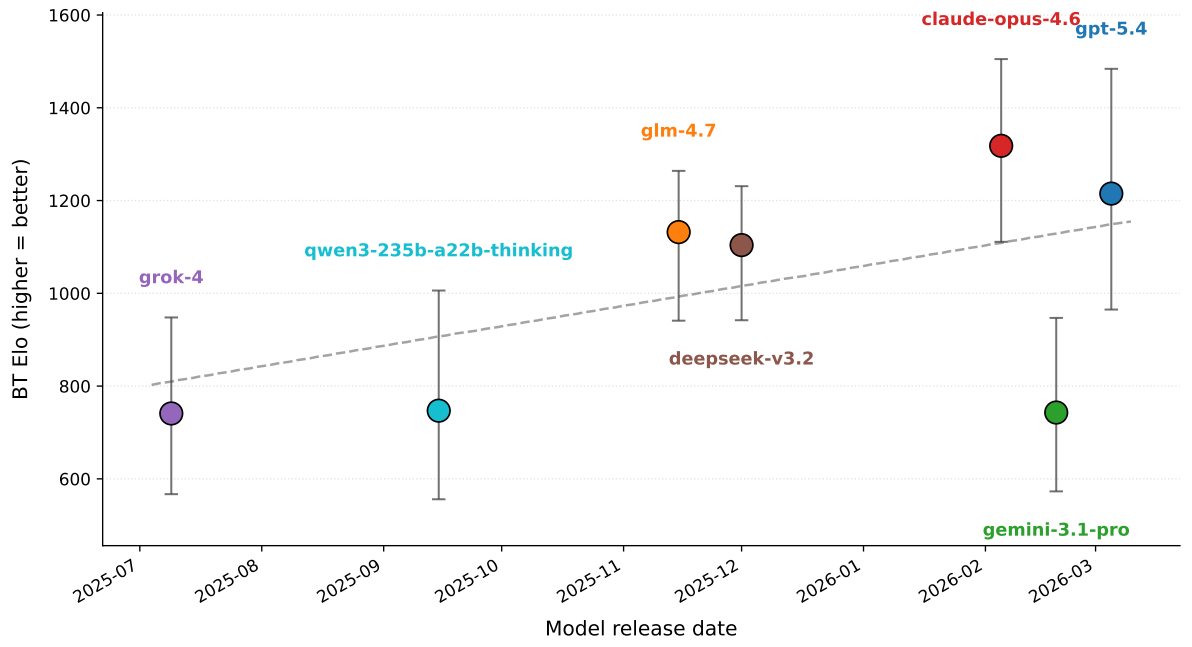

Figure 5. Generator Elo versus approximate public release date. More recent models tend to perform better in this experiment. One plausible explanation is recency of training data and post-training: models released later may have more up-to-date knowledge about AI progress, infrastructure, policy, and market structure. Vertical bars are nested-bootstrap Elo intervals.

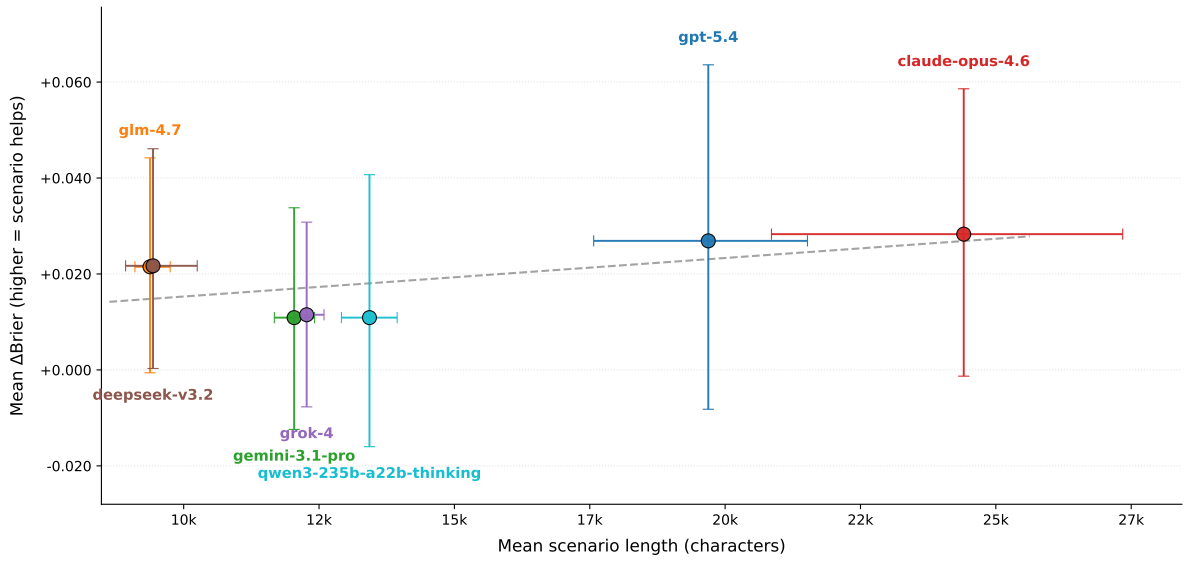

Figure 6. Mean forecast length versus mean ΔBrier. Longer forecasts tend to be associated with larger downstream improvements, though length should be interpreted as a proxy rather than a mechanism. More useful generators may write longer forecasts because they cover more contingencies, dependencies, and conditional pathways; simply increasing length need not improve a forecast. Vertical bars are hierarchical-bootstrap intervals for ΔBrier, and horizontal bars show the range across the generator's three forecast samples.

F. Prompt Ablation

We ran an early prompt-variant ablation on a 60-question AI forecasting set using the Llama-4-Maverick downstream predictor. Table 1 reports downstream mean Brier scores after removing leakage-tainted models from the sweep; lower values indicate better predictions.

The `planner_temporal` variant was selected for the main experiments because it performed well in this sweep while matching the intended use case: a temporally structured, decision-relevant forecast that covers multiple consequential branches rather than a single most-likely path.

	Mean Brier score by long-form forecast generation prompt variant					
Forecast generator	Planner	Planner-temp.	Counterfactual	Question-elicit	Metaculus few-shot	PT concise
Opus	0.218	0.214	0.218	0.220	0.230	0.219
Grok-4	0.221	0.228	0.221	0.220	0.253	0.242
DeepSeek	0.230	0.229	0.234	0.226	0.247	0.239
GPT-5.4	0.242	0.246	0.233	0.235	0.283	0.259

Table 1. Prompt-variant ablation on the 60-question AI forecasting set. Each cell is the downstream predictor's mean Brier score after conditioning on forecasts generated by the row model using the column prompt variant. Lower is better.

Planner-Temporal Forecast Prompt

User: You are forecasting AI-industry developments from October 2025 to {`resolution_deadline`} for a decision-maker who will use your scenario to plan ahead: investments, strategic bets, research priorities, hiring, product direction. They do not know in advance which specific questions or decisions they will face; your forecast will be evaluated on how well it prepares them for however reality actually unfolds.

Forecasting objective: Focus on the most impactful developments likely to shape the AI industry over this period, even the ones which may be less likely to happen but whose occurrence would meaningfully reshape plans. Think broadly about the whole ecosystem: for eg: capability discontinuities, breakthroughs or breakdowns, regime shifts, major policy or geopolitical shocks, market reconfigurations, surprise entrants, compounding second-order effects. A highly consequential outcome at 15% probability deserves as much attention as a moderately consequential one at 70%.

Uncertainty and branching: Reality can unfold in many directions, so a good scenario covers the space of impactful possibilities rather than committing to a single path: the aim is for the forecast to remain useful however things play out. Where an outcome has meaningfully different branches, try to mention them separately along with your assessed likelihood to each so the reader can weight them.

Specificity requirements: Be specific enough to act on: name the actors, capabilities, thresholds, rough timing, and the downstream consequences. Avoid vague truisms ("AI will continue to advance", "compute will grow").

Output structure: Structure your forecast as a temporal progression, broken down month by month. Your forecast should be maximally useful to downstream users.

Figure 7. Planner-temporal scenario-generation prompt template used for the main AI forecasting experiments. The placeholder {`resolution_deadline`} is filled with the evaluation horizon.