Reaching the frontier of AI forecasting with reinforcement learning

Scott Jeen [1] **Matthew Aitchison** [1] **Max Clark** [1] **Toby Shevlane** [1] **Ben Day** [1]

Abstract

Top AI forecasting systems perform similarly to skilled humans using off-the-shelf LLMs that aren't necessarily trained for the task. We ask whether this recipe can be improved by fine-tuning models specifically for judgmental forecasting. To answer this question, we fine-tune gpt-oss-120b with reinforcement learning on roughly 10,000 binary forecasting questions, rewarding probabilities assigned to realized outcomes. We find that fine-tuning elevates gpt-oss-120b from below frontier LLM performance to marginally above them on held-out Metaculus AI Benchmark questions.

1. Introduction

Top AI forecasting systems are approaching skilled-human accuracy on questions about geopolitics and current affairs (Karger et al., 2025; Metaculus, 2025). This is exciting because scalable, automated forecasting could significantly improve the quality of decision-making in both public and private sectors.

To date, the most successful recipe in forecasting tournaments has been to combine an off-the-shelf LLM (like Gemini 3 or GPT-5.4) with forecasting-specific context-gathering. These models, to our knowledge, have not been explicitly trained for forecasting. Can we improve the recipe by replacing them with models fine-tuned specifically for forecasting?

In this paper, we introduce a method for fine-tuning LLMs with reinforcement learning (RL) for forecasting (Sutton et al., 1998), and show it significantly improves the forecasting performance of gpt-oss-120b. Using the LLM fine-tuning library Tinker (Thinking Machines Lab, 2025), we fine-tune the model on around 10,000 binary questions of the form "Will [event] occur before [date]?". We reward the model for placing greater probability on the correct real-world outcome.

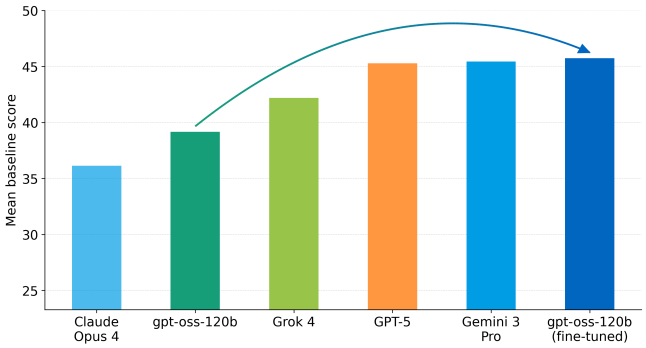

Figure 1. **RL fine-tuning makes gpt-oss-120b competitive with the frontier LLMs on questions from the Metaculus AI Benchmark Q2 2025**. Naively predicting 50% on every question would get a score of 0, and perfect foresight would get a score of 100, per the construction of the Metaculus "baseline score". Naively predicting 18.8% on every question (the rate at which the equivalent questions resolved "yes" in the previous tournament, Q1 2025) yields a score of 22.3, which we use to truncate the Y-axis. Fine-tuning improves gpt-oss-120b's score from 38.6 to 45.8, on par with the best general models.

In a head-to-head contest, the fine-tuned model achieves marginally superior performance to the frontier LLMs (Figure 1), despite much lower initial performance. We find that providing forecast-specific context increases the gains from fine-tuning.

2. Related Work

AI forecasting. Early AI forecasting models substantially underperformed human crowds (Schoenegger & Park, 2023; Zou et al., 2022). Several lines of work have since narrowed this gap. At the system level, Halawi et al. (2024) introduced a two-stage architecture pairing a retrieval-and-summarisation phase with a reasoning phase, approaching human-crowd accuracy on a held-out set; Murphy (2026) extends this direction with a Bayesian linguistic forecaster (BLF) that maintains a structured belief state across iterative tool-use steps. A second line is ensembling: Schoenegger et al. (2024) showed that aggregating predictions across a diverse set of LLMs recovers a "wisdom of the crowd" effect that rivals human aggregates. Benchmarking has matured in

[1]Mantic Technologies, London, UK. Correspondence to: Scott Jeen <scott@mantic.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

parallel, with static benchmarks (Zou et al., 2022; Jin et al., 2021) joined by dynamic ones–ForecastBench (Karger et al., 2025), FutureX (Zeng et al., 2025), and the live Metaculus AI tournaments–that mitigate data leakage by drawing questions whose outcomes are unknown at submission. Recent technical reports (Alur et al., 2025) and tournament results (Metaculus, 2025) indicate that the best AI systems now meet or exceed median tournament forecasters, though elite human forecasters remain ahead.

Fine-tuning LLMs for forecasting. Most published forecasting systems use off-the-shelf LLMs as the prediction model, with gains driven by retrieval, prompting, and ensembling rather than weight updates. The exceptions are recent. Halawi et al. (2024) supervised-fine-tune GPT-4-Base on reasoning traces from a stronger predictor, showing that distilling good forecasting rationales transfers measurable accuracy. Turtel et al. (2025a) instead use self-play: a model generates pairs of reasoning trajectories on questions resolving after its knowledge cutoff, pairs are ranked by distance to the actual outcome, and the model is fine-tuned with DPO — improving the forecasting accuracy of Phi-4 14B and DeepSeek-R1 14B by 7–10% and bringing them to GPT-4o level. Closest to our setup is Turtel et al. (2025b), who train a 14B reasoning model with RL on Polymarket-derived questions. Methodological choices in that work–omitting per-question standard-deviation normalisation in GRPO, scaffolding around malformed outputs–echo broader findings on RLVR stability (Liu et al., 2025) and inform our setup. Our contribution extends the RL-for-forecasting line by training a larger open model (gpt-oss-120b), and demonstrating that the fine-tuned model is competitive with frontier LLMs conditioned on informative research.

3. Preliminaries

Forecasting questions. We consider binary forecasting questions of the form *"Will event E occur before date τ?"*, each paired with a ground-truth outcome $y \in \{0, 1\}$ revealed at resolution time. Let $\mathcal{D} = \{(q_i, c_i, y_i)\}_{i=1}^{N}$ denote a dataset of N such questions, where q_i is the question text, c_i is a context produced by a research phase (collected search results and summaries) prepared in advance, and y_i is the resolved outcome.

Prediction model. A prediction policy π_θ, parameterized by an LLM with weights θ, takes (q_i, c_i) as input and emits a chain of thought followed by structured tool calls that specify a mixture model over the event time T:

$$p_\theta(T \mid q_i, c_i) = \sum_{k=1}^{K} w_k \, f_k(T; \phi_k), \qquad \sum_{k=1}^{K} w_k = 1, \quad (1)$$

where the number of components K, the component parameters $\{\phi_k\}$, and the mixture weights $\{w_k\}$ are all selected by the policy. An illustrative mixture model is provided in Figure 2. The implied probability of a "yes" resolution is the CDF evaluated at the question's deadline:

$$\hat{p}_\theta(q_i, c_i) = \int_{-\infty}^{\tau_i} p_\theta(T \mid q_i, c_i) \, dT. \quad (2)$$

This formulation allows the model to express and weight different beliefs about the timing of the event than would be permitted by a standard Bernoulli distribution. We also find that it improves the performance of the base model prior to finetuning.

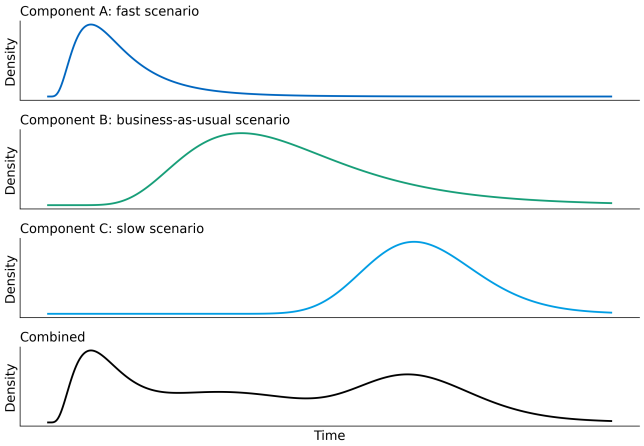

Figure 2. **Illustrative mixture model.** The LLM selects: the number of components in the mixture, their parameters, and their respective weights. The LLM is prompted to select components capturing different scenarios that could lead to the event occurring. The final prediction is a weighted combination of the components.

Objective. We optimize θ to maximize the expected reward under the Brier score, which is strictly proper and bounded in $[0, 1]$:

$$r(\hat{p}, y) = 1 - (\hat{p} - y)^2. \quad (3)$$

The training objective is

$$\mathcal{J}(\theta) = \mathbb{E}_{(q,c,y)\sim\mathcal{D}} \, \mathbb{E}_{\hat{p}\sim\pi_\theta(\cdot|q,c)} \big[r(\hat{p}, y) \big], \quad (4)$$

which we optimize with a policy gradient estimator using GRPO-style group-relative advantage normalization.

Evaluation. At test time we report the Metaculus *baseline score*, a rescaled log score where 0 corresponds to a uniform 50% prediction and 100 corresponds to perfect foresight on each question. The full list of questions can be accessed here[1].

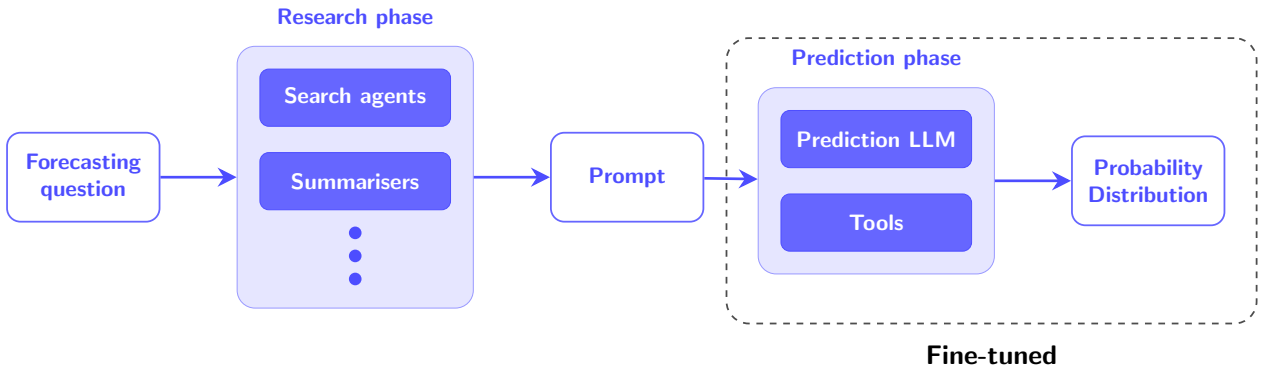

Figure 3. **Full forecasting pipeline.** The research phase takes the forecasting question as input and performs deep research to collect information relevant to the question which goes into the prompt for the prediction LLM. The prediction LLM outputs chain-of-thought reasoning and specifies a probability distribution using specialized tools. In this paper we run the research phase upfront for each question in the train and test sets, store the research as prompts and hold them fixed. We are concerned only with fine-tuning the model used in the prediction phase.

4. Methods

Dataset generation. We generated the training set using a three-stage LLM pipeline similar to existing work (Bosse et al., 2026; Turtel et al., 2026). Concretely, we first sample a news article on some date between gpt-oss-120b's knowledge cutoff and the date on which we run the generator. Then we prompt an LLM to generate a forecasting question conditioned on the news article. Finally, we equip a second LLM with web search and prompt it to resolve the question. We use this setup to generate approximately 10,000 binary event questions.

Before training, we run a backtest-compliant deep research phase that collects information that would have been available to the model at prediction time for each question. We store the output of the process as a static prompt paired with each question. The research phase is not rolled into the fine-tuning loop, as visualised in Figure 3.

We test on unseen questions from the Q2 2025 Metaculus AI Benchmark Tournament. We compare models whose knowledge cutoff is before this tournament's start date. We run the same deep research phase described above for each question.

We evaluate using the baseline score, following the Metaculus platform (Metaculus, 2026). This is log scoring, i.e. $\ln p(\text{outcome})$, rescaled such that 100 is the maximum possible score and 0 is the score for a uniform prediction (in our setting, 50%).

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

4.1. RL details.

Infrastructure and base model. We run all experiments on Tinker (Thinking Machines Lab, 2025). We train gpt-oss-120b (Agarwal et al., 2025) for its strong initial forecasting performance–second only to Kimi K2.5 (Team et al., 2026)–while being cheaper and faster. We fine-tune with LoRA (Hu et al., 2022) at rank 32, following Schulman & Lab (2025).

Algorithm. We optimize $\mathcal{J}(\theta)$ from Section 3 with a standard policy gradient algorithm using GRPO-style group-relative advantage normalization (Shao et al., 2024) (without dividing by the standard deviation (Liu et al., 2025)). The Brier score is strictly monotonic in the predicted probability for a fixed outcome, so rollouts in a group almost always receive distinct rewards. This lets us train with a small group size of 8 without breaking ties or injecting variance. We use a batch size of 64; larger batch sizes destabilized training.

Choice of reward. We use the Brier score as the reward, which is strictly proper (Gneiting & Raftery, 2007). The log score is also strictly proper but led to less stable training, which we attribute to its unboundedness producing higher-variance gradient estimates.

KL regularization. We add a KL penalty against a reference policy π_{ref}, initialized to the base policy:

$$\mathcal{J}_\beta(\theta) = \mathcal{J}(\theta) - \beta\, \mathbb{E}_{(q,c)\sim\mathcal{D}}\big[\mathrm{KL}(\pi_\theta(\cdot \mid q,c) \,\|\, \pi_{\text{ref}}(\cdot \mid q,c))\big]. \quad (5)$$

We find the KL penalty stabilises training and maintains calibration. When the test set performance plateaus, we

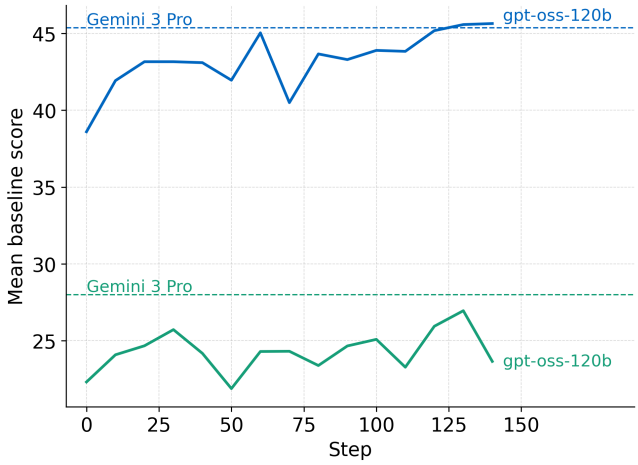

Figure 4. **Test set baseline score of gpt-oss-120b with (blue) and without (green) research and tools.** In the model-only setup, test set performance improves, but never reaches the initial score of gpt-oss-120b with research and tools. With research and tools, gpt-oss-120b climbs 7 points through training and marginally exceeds the performance of Gemini 3 Pro. Training continued for further steps but performance no longer improved.

perform a KL reset following (Liu et al., 2026) and continue training.

5. Results and Discussion

Fine-tuning closes the gap to frontier models. Figure 1 reports the headline result on the held-out Metaculus AI Benchmark Q2 2025 questions. The base gpt-oss-120b scores 38.6, below every frontier model except for Claude Opus 4, which it beats (36.2); after RL fine-tuning the same model reaches 45.8, marginally above Gemini 3 Pro (45.5) and GPT-5 (45.3) and well above Grok 4 (42.3). The 7.2-point gain from fine-tuning exceeds the gap between any two adjacent frontier models we tested, suggesting that targeted on-task training is a more reliable lever than choice of base model in this score range.

The research and tooling scaffold are important. Figure 4 decomposes this gain by training the same base model with and without our pre-generated research context and the mixture-distribution tools from Section 3. The full setup (blue) climbs from 38.6 to 45.8 over roughly 130 steps. The model-only setup (green) starts near 22 and ends near 27 — never approaching even the *initial* score of the scaffolded run. Two things follow: the bulk of absolute performance comes from the research pipeline and mixture-model tooling rather than the weights (Halawi et al., 2024), and the gains from RL are themselves larger when the scaffold is in place (7.2 vs. ∼3 points). We interpret the latter as the scaffold giving the policy a denser space of useful actions.

Table 1. **Expected Calibration Error (ECE) across models**. Lower is better; the best result is shown in bold.

Model	ECE
Grok 4	0.1092
gpt-oss-120b	0.1091
gpt-oss-120b (fine-tuned)	0.0903
GPT-5	0.0865
Gemini 3 Pro	0.0862
Claude Opus 4	**0.0601**

Calibration. Table 1 reports the Expected Calibration Error (ECE) across the same models. Fine-tuning improves gpt-oss-120b from 0.1091 to 0.0903, but it remains noticeably worse than the best-calibrated model in our comparison, Claude Opus 4 (0.0601). And neither is especially well-calibrated in absolute terms: an ECE of 0.09 means predicted probabilities deviate from empirical frequencies by roughly 9 percentage points on average, which would be considered substantial for high-stakes forecasting. The fine-tuned model's calibration gain comes "for free" from optimising a proper score (Gneiting & Raftery, 2007); closing the remaining gap to Claude Opus 4 would likely require a dedicated calibration loss or post-hoc recalibration.

Gains from fine-tuning plateau. The training curve flattens after roughly step 130, and further training (including a KL reset, Section 4.1) yielded no improvement. There are two candidate explanations: dataset saturation, since ∼10,000 questions of fairly uniform difficulty may exhaust the available signal; or a ceiling imposed by the frozen research context, since no policy improvement can recover information the prompt does not contain. The much smaller gains in the model-only ablation are consistent with the latter. Folding retrieval into the training loop is the natural way to test this.

6. Conclusions

RL fine-tuning on roughly 10,000 binary questions lifts gpt-oss-120b from 38.6 to 45.8 mean baseline points on the Metaculus AI Benchmark Q2 2025, matching Gemini 3 Pro and GPT-5, while reducing ECE by 17% without any explicit calibration term. The deep-research context and mixture-distribution output are important: removing them roughly halves the gain from RL. The most consequential extension is bringing the research phase into the training loop so that retrieval is optimised end-to-end with the forecast; scaling to stronger base models and to numerical and multiple-choice question formats are natural follow-ons.

Impact Statement

This work improves the forecasting accuracy of open-weight LLMs on questions about geopolitics, economics, and current affairs, with both benefits and risks worth naming. Accurate, well-calibrated forecasts are a public good: decision-makers in government, public health, humanitarian response, and journalism rely on probabilistic judgements about uncertain events, and open-weight forecasters lower the cost of high-quality forecasting and make it auditable in ways closed APIs are not. The risks are largely about misuse — a model that forecasts geopolitical events well is also useful to actors trying to front-run markets, time disinformation campaigns, or anticipate the responses of states and institutions — and we mitigate this only weakly, since the base model is already public and we do not think withholding the methodology would meaningfully slow capable actors. We also flag that our training data is English-language news, so performance outside that distribution is uncharacterised, and that confident forecasts can crowd out human judgement; downstream users should treat outputs as one input among several rather than as ground truth.

References

Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Alur, R., Stadie, B. C., Kang, D., Chen, R., McManus, M., Rickert, M., Lee, T., Federici, M., Zhu, R., Fogerty, D., et al. Aia forecaster: Technical report. *arXiv preprint arXiv:2511.07678*, 2025.

Bosse, N. I., Mühlbacher, P., Wildman, J., Phillips, L., and Schwarz, D. Automating forecasting question generation and resolution for ai evaluation. *arXiv preprint arXiv:2601.22444*, 2026.

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. *Advances in Neural Information Processing Systems*, 37: 50426–50468, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Jin, W., Khanna, R., Kim, S., Lee, D.-H., Morstatter, F., Galstyan, A., and Ren, X. Forecastqa: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4636–4650, 2021.

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*, volume 2025, pp. 93943–93980, 2025.

Liu, M., Diao, S., Lu, X., Hu, J., Dong, X., Choi, Y., Kautz, J., and Dong, Y. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *Advances in Neural Information Processing Systems*, 38: 17998–18031, 2026.

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Metaculus. Metaculus cup summer 2025, 2025. URL https://www.metaculus.com/tournament/metaculus-cup-summer-2025/. Accessed: 2026-05-13.

Metaculus. Scores faq, 2026. URL https://www.metaculus.com/help/scores-faq/. Accessed: 2026-05-12.

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

Schoenegger, P. and Park, P. S. Large language model prediction capabilities: Evidence from a real-world forecasting tournament. *arXiv preprint arXiv:2310.13014*, 2023.

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):eadp1528, 2024.

Schulman, J. and Lab, T. M. Lora without regret. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250929. https://thinkingmachines.ai/blog/lora/.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sutton, R. S., Barto, A. G., et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S., Cao, Y., Charles, Y., Che, H., Chen, C., Chen, G., et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Thinking Machines Lab. Tinker, 2025. URL https://thinkingmachines.ai/tinker/.

Turtel, B., Franklin, D., and Schoenegger, P. Llms can teach themselves to better predict the future. *arXiv preprint arXiv:2502.05253*, 2025a.

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

Turtel, B., Wilczewski, P., Franklin, D., and Skothiem, K. Future-as-label: Scalable supervision from real-world outcomes. *arXiv preprint arXiv:2601.06336*, 2026.

Zeng, Z., Liu, J., Chen, S., He, T., Liao, Y., Tian, Y., Wang, J., Wang, Z., Yang, Y., Yin, L., et al. Futurex: An advanced live benchmark for llm agents in future prediction. *arXiv preprint arXiv:2508.11987*, 2025.

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., and Hendrycks, D. Forecasting future world events with neural networks. *Advances in Neural Information Processing Systems*, 35: 27293–27305, 2022.