Scaling Open-Ended Reasoning To Predict the Future

Nikhil Chandak[1,3*] **Shashwat Goel**[1,2*]
Ameya Prabhu[3,4†] **Moritz Hardt**[1,3†] **Jonas Geiping**[1,2,3†]

[1]Max Planck Institute for Intelligent Systems [2]ELLIS Institute Tübingen
[3]Tübingen AI Center [4]University of Tübingen

Blog Code Dataset and Models

Abstract

High-stakes decision making involves reasoning under uncertainty about the future. In this work, we train language models to make predictions on open-ended forecasting questions. To scale up training data, we synthesize novel forecasting questions from global events reported in daily news, using a fully automated, careful curation recipe. We train the Qwen3 thinking models on our dataset, `OpenForesight`. To prevent leakage of future information during training and evaluation, we use an offline news corpus, both for data generation and retrieval in our forecasting system. Guided by a small validation set, we show the benefits of retrieval, and an improved reward function for reinforcement learning (RL). Once we obtain our final forecasting system, we perform held-out testing between May to August 2025. Our specialized model, `OpenForecaster8B`, matches much larger proprietary models, with our training improving the accuracy, calibration, and consistency of predictions. We find calibration improvements from forecasting training generalize across popular benchmarks. We open-source all our models, code, and data to make research on language model forecasting broadly accessible.

1 Introduction

Every day, people navigate decisions under uncertainty, due to incomplete evidence or competing hypotheses. The highest-stakes choices are inherently forward-looking: governments set policy while anticipating macroeconomic and geopolitical shifts; investors allocate capital amid market and regulatory uncertainty; individuals choose careers as technologies evolve; and scientists pursue research directions in search of the next breakthrough. Decades of work (Tetlock et al., 2014) on human forecasting shows that while prediction is hard and skill varies widely, it is possible to train humans to become better forecasters. In fact, some "superforecasters" consistently outperform peers. While there is a ceiling to predictability in social systems (Franklin, 1999), we do not yet know where that ceiling lies in the real world.

If trained at scale for forecasting world events, Large Language Models (LLMs) may enjoy structural advantages over humans: they can ingest and synthesize vast, heterogeneous corpora across thousands of topics; and update predictions rapidly as new information arrives. Just like language models now show superhuman reasoning on some exam-style math and coding problems (OpenAI, 2025), in the future, language model forecasters may be able to come up with possibilities that humans miss. So in this work, we study:

How can we train language models to better forecast open-ended questions?

Scaling training data for forecasting. As forecasting world events is hard for humans, detailed and correct reasoning traces for forecasting are difficult to obtain. Fortunately, recent success in Reinforcement Learning (RL) for language models enables training with

*Equal contribution †Equal co-supervision

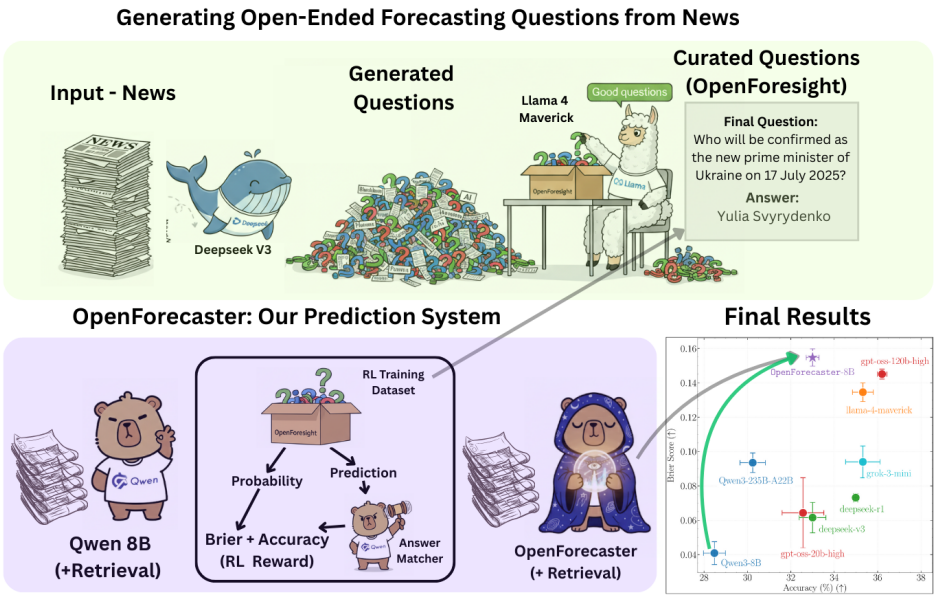

Figure 1: A summary of our methodology for training language model forecasters.

just the eventual outcome of the question (Guo et al., 2025). Further, the static knowledge cutoff of LLMs enables a unique opportunity: events that resolve after the cutoff are in the future for the model. Even then, sourcing questions at scale for training forecasting abilities has a few key challenges. First, waiting for events to resolve is too slow as a feedback loop for training. Second, prediction markets–the primary source for existing forecasting questions–mostly consist of binary yes or no questions. As there is a 50% chance of success on these questions even with incorrect reasoning, they make for noisy rewards.

Instead, we use global news, which covers a large number of salient events every day, to synthesize open-ended forecasting questions like "Who will be confirmed as the new prime minister of Ukraine on 17 July 2025?". Our recipe for creating training data is entirely automated and scalable, with one language model extracting events from news articles to generate questions, and a different model filtering and rewriting questions to avoid leaking future information. For this work, we use this recipe with 250,000 articles up till April 2025, to create `OpenForesight`, a dataset of $\sim$ 50,000 open-ended forecasting questions for training. To grade responses for open-ended questions, we use model-based *answer matching* (Chandak et al., 2025) consistent with frontier benchmarks like Humanity's Last Exam (Phan et al., 2025).

Ensuring we truly improve forecasting. We take extensive measures to avoid the leakage of future information during training and evaluation. First, we do not use online search engines for sourcing news, as they have unreliable date cutoffs due to dynamic updates to documents and search ranking (Paleka et al., 2025a). Instead, we use the CommonCrawl News corpus, which provides static, monthly snapshots of global news. Second, we only train on events until April 2025, which is when the Qwen3 model weights we train were released. Finally, we do not observe performance on the test set until the very end. Our test set is composed of diverse news sources, different from the ones used in training and validation, to ensure we are not just learning distributional biases of the training data.

Validating design choices for LLM Forecasting Systems. We start from Qwen3 (Yang et al., 2025) 4B and 8B models with thinking enabled. We perform all ablations on a small validation set. We use dense retrieval with the Qwen3-8B Embedding model to provide forecasters relevant chunks from our offline news corpus. Despite a cautious approach of only retrieving articles until *one month* before the question resolution date to avoid leakage, the retrieved information leads to large improvements. Then, we train language models using RL with GRPO. For the reward function, we propose combining accuracy, and an adaptation of the brier score for open-ended responses (Damani et al., 2025). Ablations show rewarding accuracy alone hurts calibration, while optimizing only the brier score hurts exploration on hard questions. Our final methodology is illustrated in Figure 1.

Final results. In Section 6, we report results on our held-out test set of open-ended forecasting questions from May to August 2025, and FutureX (Zeng et al., 2025), an external forecasting benchmark. RL training on `OpenForesight` makes the predictions of our specialized 8B model competitive with much larger proprietary models in both accuracy and calibration. We also observe large improvements on consistency evaluations for long-term predictions (Paleka et al., 2025b). Finally, we find calibration from our forecasting training generalizes to multiple out of distribution benchmarks.

By providing rigorous probabilistic predictions, open-ended forecasting systems could transform policy making, corporate planning, and financial risk management (Tetlock, 2017). To promote forecasting research, we open-source our dataset, code, and models.

2 Related Work

Forecasting World Events. Much prior work in Machine Learning and Statistics has focused on forecasting numeric or time-series data (Box & Jenkins, 1976) in diverse domains like weather (Richardson, 1922), econometrics (Tinbergen, 1939) or finance (Cowles, 1933). Instead, our work focuses on the prediction of discrete world events, with both questions and answers described in natural language, also called *judgemental forecasting* (Tetlock & Gardner, 2016). In the rest of our paper, we refer to this as *forecasting* for brevity. In prior work on evaluating language models for forecasting (Zou et al., 2022; Karger et al., 2024), questions are primarily sourced from prediction markets like Metaculus, Manifold, and Polymarket. Prediction markets provide a platform for online participants to register predictions on questions like "Will Donald Trump win the US Presidential Election in 2024?", which mostly have binary, yes or no, outcomes and have rapidly grown in popularity over the last few years.

Evaluating LLMs for Forecasting. New information (before the event resolves) benefits forecasting. Thus, LLM forecasting work (Zou et al., 2022; Halawi et al., 2024) provides relevant retrieved articles to models (Lewis et al., 2020) often obtained via web-search APIs. Paleka et al. (2025a) discuss pitfalls of LLM forecasting evaluations, including leakage of outcomes from online search in backtests, and distributional biases of prediction market questions. To avoid these issues, we use static, monthly snapshots of global news for retrieval and creating questions. Jin et al. (2021) ask humans to create forecasting questions, while Dai et al. (2024) try to automate this process with LLMs. However, their questions pre-define a few outcomes to choose from. Guan et al. (2024); Wang et al. (2025) evaluate models on open-ended forecasts, but we go a step further by showing how to train models for this task.

Reinforcement Learning for LLMs. Shao et al. (2024) proposed *Group Relative Policy Optimization* (GRPO), an RL algorithm that only uses outcome rewards. This approach has been highly successful in training LLMs to *reason* about well-specified coding (Jain et al., 2024) and exam-style questions across domains (Phan et al., 2025). Instead, forecasting requires LLMs to reason about uncertainty. Halawi et al. (2024) proposed training language models for forecasting by Supervised Finetuning (SFT) on chain of thought traces that lead to brier scores better than the prediction market aggregate. In the same setting of binary forecasting questions, Turtel et al. (2025a) optimize brier scores using GRPO, while Damani et al. (2025) extend it to short answer questions in other domains. We depart from these works in showing how to synthesise large-scale forecasting training data from daily news, to train models at open-ended reasoning about the future.

3 Open-Ended Forecasting

Motivation. The forecasting task we study is *open-ended* in two key ways:

1. It allows expressing arbitrary natural language forecasting questions
2. It may not have a structured outcome set, unlike numeric or categorical predictions. This differentiates it from both time-series forecasting, and prediction markets.