FORECASTBENCH: A DYNAMIC BENCHMARK OF AI FORECASTING CAPABILITIES

Ezra Karger *
Forecasting Research Institute
Federal Reserve Bank of Chicago
ezra@forecastingresearch.org

Houtan Bastani *
Forecasting Research Institute
houtan@forecastingresearch.org

Chen Yueh-Han *
New York University
yc7592@nyu.edu

Zachary Jacobs
Forecasting Research Institute
zach@forecastingresearch.org

Danny Halawi
University of California, Berkeley
dhalawi@berkeley.edu

Fred Zhang
University of California, Berkeley
z0@berkeley.edu

Philip E. Tetlock
Forecasting Research Institute
University of Pennsylvania
tetlock@wharton.upenn.edu

ABSTRACT

Forecasts of future events are essential inputs into informed decision-making. Machine learning (ML) systems have the potential to deliver forecasts at scale, but there is no framework for evaluating the accuracy of ML systems on a standardized set of forecasting questions. To address this gap, we introduce **ForecastBench**: a dynamic benchmark that evaluates the accuracy of ML systems on an automatically generated and regularly updated set of 1,000 forecasting questions. To avoid any possibility of data leakage, ForecastBench is comprised solely of questions about future events that have no known answer at the time of submission. We quantify the capabilities of current ML systems by collecting forecasts from expert (human) forecasters, the general public, and LLMs on a random subset of questions from the benchmark ($N = 200$). While LLMs have achieved super-human performance on many benchmarks, they perform less well here: expert forecasters outperform the top-performing LLM (p-value < 0.001). We display system and human scores in a public leaderboard at www.forecastbench.org.

1 INTRODUCTION

Forecasting the future is a challenging but important task (Armstrong, 2001; Tetlock and Gardner, 2015). Economic forecasts influence investment and hiring decisions (Christensen et al., 2018). And forecasts of the Covid-19 pandemic in early 2020 prompted local lockdowns to slow the spread of the virus (Adam, 2020). However, human forecasting is often expensive, time-consuming, applicable only in specific domains, and subject to concerns about human biases. Motivated by these limitations, recent work explores the use of machine learning (ML) models, especially large language models (LLMs), in automated forecasting (Fluri et al., 2024; Halawi et al., 2024; Hendrycks et al., 2021; Phan et al., 2024; Pratt et al., 2024; Yan et al., 2024; Zou et al., 2022).

*Equal contribution.
Correspondence to forecastbench@forecastingresearch.org.
The views expressed in this paper do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve system.

To assess LLMs' forecasting capabilities, prior work built static benchmarks of questions where the answer was known (resolved) *after* the knowledge cut-offs of the LLMs they test (Halawi et al., 2024; Yan et al., 2024; Zou et al., 2022). For example, "Will a nuclear weapon be detonated in 2023," though resolved on Jan 1, 2024, is a valid out-of-sample question for testing a model with a known knowledge cut-off before the end of 2023.

This static evaluation methodology comes with three key drawbacks. First, as the knowledge cut-offs of frontier models are updated, static benchmarks become obsolete, necessitating further data-sourcing. This makes it difficult to continuously track and compare the top models in the field. Second, knowledge cut-offs are usually estimated using the time range of pre-training data. Such estimates may not be accurate, and post-training may inject further post-cutoff knowledge into the model, contaminating the test sets. Lastly, model developers face financial incentives to exaggerate their accuracy on benchmarks and claim that their models are state-of-the-art performers. While some fudging is easily identified, other subtle benchmark manipulation or overfitting is harder to catch, and some studies show significant evidence of benchmark contamination and/or memorization by popular LLM models (Elazar et al., 2024; Li et al., 2023; Roberts et al., 2023).

To address these drawbacks, we introduce **ForecastBench**, a dynamic benchmark that is continuously updated with new questions about future events. Our automated system gathers new questions from nine sources on a daily basis. These sources include prediction markets, forecasting platforms, and real-world time series. We regularly elicit predictions on these questions from both LLMs and human forecasters. As they resolve, we update a public leaderboard to show participant performance. This process makes ForecastBench an accurate real-time benchmark of forecasting ability.

Our initial benchmark consists of 1,000 standardized forecasting questions randomly sampled from a much larger real-time question bank. To establish baseline levels of performance, we record predictions from dozens of LLMs on the initial set, using methods like retrieval-augmentation to boost performance (Halawi et al., 2024; Lewis et al., 2020). We independently elicit predictions from two groups of human forecasters: the general public and seasoned forecasters (superforecasters) (Tetlock and Gardner, 2015) who have consistently performed well in competitive forecasting tournaments. As questions resolve, we rate the submissions of registered models and the human comparison groups, updating our public leaderboard.

Our initial results indicate that state-of-the-art models, such as Claude-3.5 Sonnet and GPT-4 Turbo, perform only roughly as well as a simple median of forecasts from a survey of humans with no (or minimal) forecasting experience, even when the LLMs are augmented with news retrieval and prompt engineering and when they have access to forecasts from a human crowd (on market-based questions). The models also perform significantly worse than the median forecast of superforecasters.

These findings leave significant room for researchers to improve AI-based forecasting systems using innovative approaches, such as developing methods for continuously updating models with current events and enhancing LLMs to reason over extended time frames. To support progress in this area, we publish an auxiliary dataset of LLM and human forecasts, rationales, and accuracy for use in future LLM fine-tuning and testing.

1.1 Related work

Automated forecasting ML systems that make accurate forecasts can help inform human decision-making (Hendrycks et al., 2021; Schoenegger et al., 2024a). Recent work studies the use of LLMs for automated forecasting (Halawi et al., 2024; Jin et al., 2021; Pratt et al., 2024; Yan et al., 2024; Zou et al., 2022). These papers all build static benchmarks of resolved questions. A recent paper from Halawi et al. (2024) uses questions resolved between June 2023 and January 2024. Unfortunately, the latest LLMs have knowledge cut-offs past this point. This fact motivates our work to build a dynamically updating benchmark that can accurately evaluate frontier model performance.

In addition, Schoenegger and Park (2023) and Abolghasemi et al. (2023) compare the accuracy of GPT-4 and other LLMs to human forecasters. Schoenegger et al. (2024b) found that an ensemble of 12 LLMs rivaled human forecasts in a forecasting tournament with a small number of questions, limiting the study's statistical power. Unlike our work, that tournament was run only once and is no longer updated. Also, our much larger question set allows us to make precise statistical claims about the performance of LLMs relative to each other and to humans.

Finally, recent work focuses on the use of LLMs and transformer-based systems in statistical time-series forecasting (Das et al., 2024; Dooley et al., 2023; Goswami et al., 2024; Gruver et al., 2023; Jin et al., 2024; Nie et al., 2023; Rasul et al., 2023; Woo et al., 2024), but many of the most important forecasting questions do not have well-defined time series that can be used in standard statistical forecasting models (e.g., what is the probability that a nuclear weapon will be used offensively in the next ten years?). Our benchmark is more general, and evaluates forecasting performance across questions with and without underlying time series and historical baseline data.

Language model evaluation Evaluating highly capable LLMs is a challenging task—with many models saturating key benchmarks soon after their release (Maslej et al., 2023; Owen, 2024) and with benchmarks potentially leaked to models' training data (Balloccu et al., 2024; Jacovi et al., 2023; Jiang et al., 2024b; Magar and Schwartz, 2022; Sainz et al., 2023; Xu et al., 2024a;b; Zhang et al., 2024). Our dynamic forecasting benchmark avoids both of these issues. First, automated forecasting is highly unsaturated: Halawi et al. (2024) showed that under simple zero-shot evaluation, frontier models such as GPT-4 are much less accurate than aggregates of human predictions. Second, our benchmark is dynamic. The final resolution of each question is only determined in the future and cannot be leaked in any training data (by construction). This eliminates concerns of contamination.

2 PRELIMINARIES

Forecasting A forecasting question asks for a probabilistic prediction of a future event. A forecaster may assign probabilities to potential outcomes of the event. Forecasting platforms, including prediction markets, host a wide range of questions. Many questions are of public interest, such as "Will a Democrat win the 2028 US presidential election?"

Metrics For binary questions, we use the Brier score as the performance metric, defined as $(f-o)^2$, where $f \in [0,1]$ is the probabilistic forecast and $o \in \{0,1\}$ is the outcome. Lower Brier scores are better, and a score of 0.25 is associated with the uninformed forecast of 0.5. Brier scores are strictly proper, incentivizing truthful reporting from participants.

Models We evaluate 17 LLMs on our initial benchmark: GPT-3.5-Turbo-Instruct (Brown et al., 2020), GPT-4 (OpenAI, 2023), GPT-4o, Llama-2-70B (Touvron et al., 2023), Llama-3-7B, Llama-3-70B, Mistral-7B (Jiang et al., 2023), Mixtral-8x7B (Jiang et al., 2024a), Mixtral-8x22B, Mistral-Large, Qwen1.5-110B-Chat (Bai et al., 2023), Claude-2.1 (Anthropic, 2023), Claude-3-Haiku, Claude-3.5-Sonnet, Claude-3-Opus (Anthropic, 2024), Gemini 1.5 Flash, and Gemini 1.5 Pro (Gemini Team, 2023). We outline the various baselines we run with these models in Section 5.

3 BENCHMARK, LEADERBOARD, AND DATASETS

We maintain a question bank with a growing set of forecasting questions. Continuously adding questions to the question bank allows it to stay relevant and ensures that we have a large selection of unresolved questions to sample from.

Every night, our automated system updates the question bank with new questions and resolution values. We drop invalid, low-quality, and resolved questions, categorizing the remaining questions by topic. The process is fully automated, as detailed in Section 3.1.

Every two weeks, we sample from the question bank and release question sets for those interested in submitting their forecasts to the benchmark. We also survey a standard set of LLM-based models to allow for comparisons of performance over time. Submitted forecasts are resolved continuously with daily updates to our leaderboard. We provide the resulting data on forecast questions and submissions to researchers. See Section 3.2 for details. Finally, we discuss the question resolution procedure in Section 3.3 and our leaderboard design in Section 3.4.

3.1 QUESTION BANK

In an automated process that runs nightly at `0:00` UTC, questions are added to the question bank, resolution values are updated, and metadata generated. Details follow.