
Forecasting Future World Events with Neural Networks

Andy Zou
UC Berkeley

Tristan Xiao
UC Berkeley

Ryan Jia
UC Berkeley

Joe Kwon
MIT

Mantas Mazeika
UIUC

Richard Li
UC Berkeley

Dawn Song
UC Berkeley

Jacob Steinhardt
UC Berkeley

Owain Evans
University of Oxford

Dan Hendrycks
UC Berkeley

Abstract

Forecasting future world events is a challenging but valuable task. Forecasts of climate, geopolitical conflict, pandemics and economic indicators help shape policy and decision making. In these domains, the judgment of expert humans contributes to the best forecasts. Given advances in language modeling, can these forecasts be automated? To this end, we introduce Autocast, a dataset containing thousands of forecasting questions and an accompanying news corpus. Questions are taken from forecasting tournaments, ensuring high quality, real-world importance, and diversity. The news corpus is organized by date, allowing us to precisely simulate the conditions under which humans made past forecasts (avoiding leakage from the future). Motivated by the difficulty of forecasting numbers across orders of magnitude (e.g. global cases of COVID-19 in 2022), we also curate IntervalQA, a dataset of numerical questions and metrics for calibration. We test language models on our forecasting task and find that performance is far below a human expert baseline. However, performance improves with increased model size and incorporation of relevant information from the news corpus. In sum, Autocast poses a novel challenge for large language models and improved performance could bring large practical benefits.

1 Introduction

Forecasting plays a crucial role in the modern world. Climate forecasts shape the policies of governments and companies (Gillingham et al., 2018). Economic forecasts influence investment and employment (Christensen et al., 2018). In 2020, forecasts about the spread of COVID-19 led to national lockdowns and border closures (Adam, 2020), slowing the spread of the virus. Consequently, machine learning (ML) models that make accurate forecasts across a broad range of topics could enable more informed decision making at scale and improve ML safety (Hendrycks et al., 2021c).

Two main approaches to forecasting are described in the forecasting literature: statistical and judgmental forecasting (Webby and O’Connor, 1996; Armstrong, 2001). In *statistical forecasting*, forecasts are made by traditional statistical models for time-series prediction such as autoregression (Makridakis et al., 2008) or by ML time-series models (Makridakis et al., 2020; Triebe et al., 2021). Humans create and tune the models but do not tweak individual forecasts. This works well when there are many past observations of the variable being forecast and minimal distribution shift. By contrast, in *judgmental forecasting* human forecasters use their own judgment to determine forecasts. The forecasters may use statistical models, but often integrate information from various sources including news, accumulated knowledge, and *a priori* reasoning. This enables forecasting for questions where

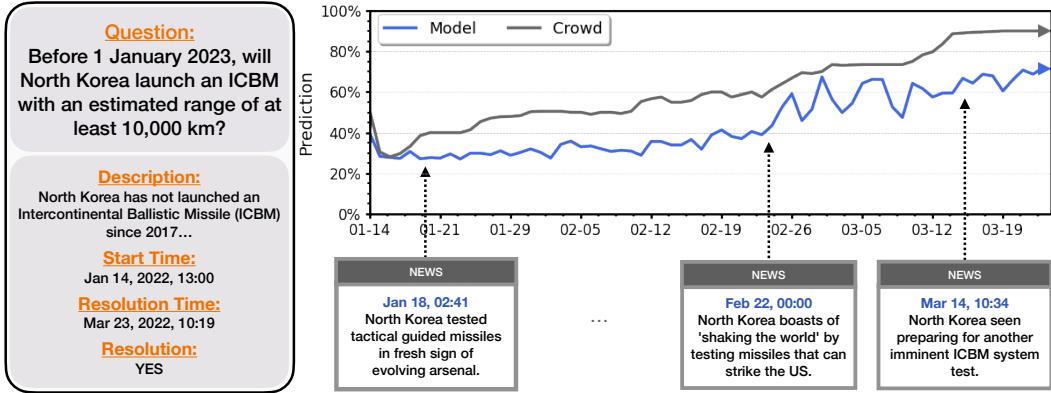


Figure 1: Example from the Autocast dataset, including the question, the resolution of the question, and the timeseries of aggregate human expert forecasts (Crowd) from the start date to the time the question resolves. We train a language model to generate forecasts at each timestep, using only news articles available at that timestep (i.e. without allowing any leakage of information from the future).

past data is scarce or subject to distribution shift (Tetlock and Gardner, 2016). For brevity, we refer to judgmental forecasting as “forecasting” in the rest of the paper.

Because it relies on scarce human expertise, forecasting is only used for a small number of questions. This motivates using ML to automate forecasting, e.g. by automating human information retrieval (finding news sources), reasoning (to decide if some evidence bears on a forecast), and quantitative modeling. ML models may also have some advantages over human forecasters. Models can read through text or data much faster than humans and can discern patterns in noisy high-dimensional data that elude humans. When it comes to learning, humans cannot be trained on past data in manner simulating actual forecasting (e.g. How likely was the Soviet Union’s collapse from the viewpoint of 1980?) because they know the outcomes – but past data can be used for ML models.

As a step towards automating human forecasting, we introduce *Autocast*, a new dataset for measuring ML models’ forecasting ability. Autocast includes thousands of forecasting questions collected from human forecasting tournaments. The questions vary in the forecasting horizon from days to decades, in the topic (including politics, economics and science), and in the answer format (e.g. multiple-choice vs. predicting a number). The questions are pre-selected for public interest, and there is a strong human baseline (the crowd aggregate of many competitive forecasters). The questions in Autocast are about past events (e.g. the US 2020 election) and so ML models could answer them simply by memorizing what happened. To test forecasting ability, we need to simulate the state of information *before* the past events (“retrodiction”). To this end, we curate a corpus of news items from Common Crawl (Nagel, 2016) that is organized by date. This means a model can be exposed only to news from before the outcomes being forecast, allowing for a rigorous test of retrodiction.

We implement a number of baseline models on Autocast, and demonstrate how language models can be trained on past forecasting questions by retrieving from our news corpus. We find that performance improves with model size and that information retrieval helps. However, all baselines are substantially worse than aggregate human forecasts. On forecasting binary outcomes, the best ML model achieves 65% accuracy vs. 92% for humans (and 50% for random). The same ML model (Raffel et al., 2020) is close to the human ceiling when fine-tuned on other NLP benchmarks (e.g. SQuAD from Rajpurkar et al. (2016)), which shows that Autocast is a challenging, real-world test for ML. Experiment code and the dataset are available at github.com/andyzoujm/autocast.

Contributions.

1. We introduce Autocast, a dataset for forecasting that covers diverse topics (e.g. politics, economics, society, science) and varying time horizons.

Question Summary	Category	Answer Type	Resolution
Will a Tesla car demonstrate fully autonomous capability before the end of 2021?	Science & Tech	T/F	No
What will be Putin’s approval rating value 3 months after the potential invasion of Ukraine?	Politics	Numerical	83
When will the US-Canada border reopen?	Social	Numerical	Nov 8, 2021
How many vacancies will arise on the U.S. Supreme Court in 2021? (A) 0 (B) 1 (C) 2 (D) 3 or more	Economy	MCQ	A

Table 1: Examples from the Autocast dataset. For brevity, we do not depict the full question specification, which often includes context, definitions, and detailed resolution criteria.

2. Part of our dataset is a large news corpus organized by date, allowing us to rigorously evaluate model performance on historical forecasts.
3. We show that forecasting is challenging for current language models, with accuracy and calibration far below a strong human baseline.

2 Related Work

Forecasting. A recent experiment (Kirk Bonde, 2022) tested GPT-3 in the few-shot setting on true/false questions collected from Metaculus (one of the sources for Autocast). However, since questions were not filtered by date, some answers would have appeared in GPT-3’s training data. Similar to our work, ForecastQA (Jin et al., 2021) is a dataset of forecasting questions that covers a range of topics. However, ForecastQA’s questions were written by crowdworkers without forecasting experience. Consequently, the questions are often nonsensical or ambiguous given the lack of additional context, e.g. “To how many people will the Representative of an internet speak to by September 2019?”, or “In July 2019, will an article say there were no volunteers in 2016?”. We found that a high percentage of ForecastQA questions suffer from these issues. By contrast, our questions were written by experienced forecasters and are always unambiguous given the full question description. Finally, ForecastQA’s human baseline was done retrospectively (making it unrealistic) whereas our dataset contains expert human forecasts from real forecasting questions.

Information Retrieval. Information retrieval is crucial for forecasting, as good forecasts depend on up-to-date, specialized information drawn from multiple sources (Tetlock and Gardner, 2016). Recent work has used information retrieval to improve question-answering in large language models (Lewis et al., 2020; Nakano et al., 2021; Shuster et al., 2021) or to address time-sensitive questions (Chen et al., 2021). This has been applied to tasks that are related to forecasting, such as fact checking and truthful question-answering. In forecasting, it is useful to read and compare multiple news articles daily, in order to build an accurate picture of the current state, and then to iterate this process. We

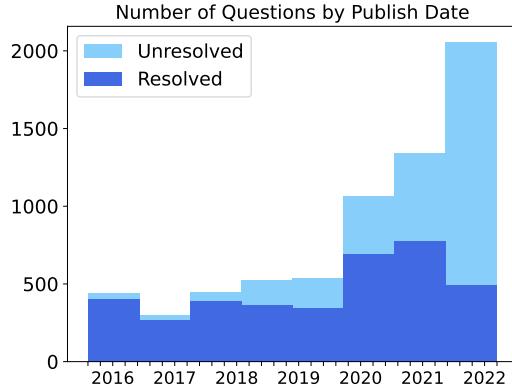


Figure 2: The number of questions in Autocast by publish date. Unresolved questions are about events after 2022 (e.g. the 2024 US Election). They are not included in the test set but can be used as auxiliary training data. Note that the number of questions is accelerating. Future questions will be added to Autocast, improving it over time.