# Forecasting Future World Events with Neural Networks

*Tristan Xiao*

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-61
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-61.html

May 11, 2022

Copyright © 2022, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Acknowledgement

To my parents - everything I am, you helped me to be.

DocuSign Envelope ID: C1F90A7F-729A-48FB-872C-4AC35CC8C05D

# Forecasting Future World Events with Neural Networks

by Tristan Xiao

## Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

**Committee:**

Professor Dawn Song
Research Advisor

5/10/2022

(Date)

* * * * * * *

Jacob Steinhardt

Professor Jacob Steinhardt
Second Reader

5 / 10 / 2022

(Date)

Abstract

Forecasting Future World Events with Neural Networks

by

Tristan Xiao

Master of Science in Electrical Engineering and Computer Sciences

University of California, Berkeley

Forecasting future world events is a challenging but fruitful task, especially during times of uncertainty for better decision-making. We introduce a dataset of forecasting questions spanning various categories and topics and a large dataset of news curated from common-crawl. We show the effectiveness of larger models, better retrieval sources and techniques, and temporal architecture for long-range modeling. In order to better measure models' performance and calibration on questions with numerical outputs, we also introduce another dataset full of numerical questions where we design a baseline algorithm to train models to output confidence intervals at specified confidence levels. With this dataset, we introduce a novel measure of calibration for numerical outputs based on adaptive binning RMS.

# Forecasting Future World Events with Neural Networks

**Andy Zou** **Tristan Xiao** **Ryan Jia** **Joe Kwon** **Richard Li**

**Jacob Steinhardt** **Owain Evans** **Dan Hendrycks**

## Abstract

Forecasting future world events is a challenging but fruitful task, especially during times of uncertainty for better decision-making. We introduce a dataset of forecasting questions spanning various categories and topics and a large dataset of news curated from common-crawl. We show the effectiveness of larger models, better retrieval sources and techniques, and temporal architecture for long-range modeling. In order to better measure models' performance and calibration on questions with numerical outputs, we also introduce another dataset full of numerical questions where we design a baseline algorithm to train models to output confidence intervals at specified confidence levels. With this dataset, we introduce a novel measure of calibration for numerical outputs based on adaptive binning RMS.

## 1 Introduction

Forecasting is an activity to predict what will happen in the future given events and information in the past and present. At crucial times, political leaders and command and control centers can employ Machine Learning (ML) systems to improve forecasting and decision making [Hendrycks et al., 2021b]. The task involves taking some statement or question about the future world and guessing what the truth value or resolution is. Forecasters assign probabilities or numerical values to (geopolitical, epidemiological, industrial, or economical) events and quantities that could arise within the next months or years. They are scored by their accuracy and calibration.

In recent times, the AI safety community has become increasingly interested in forecasting AI developments, such as "What will performance on ImageNet be in a year?" or "Will this line of research be relevant (highly cited) next year?" For instance, similar questions are being posed by safety researchers on HyperMind, a prediction market. Our efforts would help technical AI safety orient itself and have foresight, as well as make models more calibrated and integratively complex, a skill that is otherwise under-incentivized.

Machine learning models have the intrinsic advantage of being able to tirelessly process prediction-relevant data. Since machine learning models can quickly read gigabytes of text, they could weigh millions of variables, whereas humans can only contemplate a small number of factors when producing their predictions. They could also incorporate smaller subtler signals which are not apparent to time-limited humans. These factors could in theory substantially improve forecasting performance.

To measure comprehensively ML models' forecasting performance, we curate a new benchmark consisting of thousands of forecasting questions scraped from online forecasting tournaments and prediction markets. These questions could range from forecasting the likelihood of an one-time

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

| | T/F | MC | NUM | Total |
|---|---|---|---|---|
| GoodJudgement | 870 | 862 | – | 1732 |
| Metaculus | 1097 | – | 872 | 1969 |
| Total | 1967 | 862 | 872 | 3701 |

Table 1: The forecasting dataset has questions from Good Judgement Open and Metaculus where people publicly post forecasting questions and crowd predictions are recorded and displayed. There are 3701 questions in total ending in April 2022, consisting of T/F, multiple choice, and numerical questions.

event such as an election outcome, to more continuous statistics such as citation counts for academic papers, to generally, consequences given a state and a series of actions. Accompanying the dataset of questions is a large pile of daily news articles complied from the commoncrawl news corpus that models could leverage when making predictions.

In order to better measure calibration for questions with numerical output, we curate an additional dataset where we compile a suite of numerical questions from various existing natural language benchmarks. The models are tasked to generate confidence intervals for specified confidence levels and we introduce a novel calibration measure based on adaptive binning [Nguyen and O'Connor, 2015]. Outputting confidence intervals instead of point estimates reveals more information about the model's beliefs and confidence.

To provide baseline algorithms for our forecasting benchmark, we directly finetune pretrained language models and incorporate retrieval models to obtain additional information from the daily news articles. Additionally, we also design a hierarchical architecture to process temporal text feeds and generate and update daily forecasts to match the crowd predictions. We show that bigger model sizes, more news articles, better retrieval methods, and temporal updates can all lead to increase in performance. Furthermore, we conduct experiments on our numerical calibration benchmark and show that effectiveness of our new calibration measure and provide various baseline algorithm to output confidence intervals. Again, we show that calibration can be improved with larger models and novel algorithmic design.

## 2 Related Work

**Machine Forecasting.** ForecastQA is the first attempt at providing a forecasting dataset for an ML system [Jin et al., 2021]. Besides questions about politics and business on CSET-Foretell, CITEWORTH is another dataset for citeworthiness detection over scientific documents.

**Machine Retrieval.** We examined multiple techniques for retrieval, including dense passage retrieval (DPR), fusion-in-decoder (FiD), and best matching (BM25). In order to run DPR, we generate embeddings for our $cc_news$ corpus and attach them. For BM25, we also experiment with reranking using BERT based cross-encoders (BM25-CE) which is the best method on BERI benchmark measuring out of domain retrieval performance [Thakur et al., 2021].

**Machine Calibration.** We also experimented with recurrence based models, such as sequential transformers and other variations, for fine tuning the confidence levels of our predictions to our desired calibrated confidence intervals. Calibration is defined as follows: $P(\hat{a} = a|P(\hat{a}|q) = p) = p$ $\forall p \in [0, 1]$. Concretely, the model should get roughly 80 percent correctly for the questions that it's 80 percent confident. This is studied in discrete case but no prior work to our knowledge has explored the case where the model outputs are numerical and continuous. In our experiments, we force the model to output confidence intervals for each question and formulate the calibration loss to move the upper and lower bounds around to achieve good calibration. Calibration is measured with RMS error of confidence levels and the actual proportion of containment.

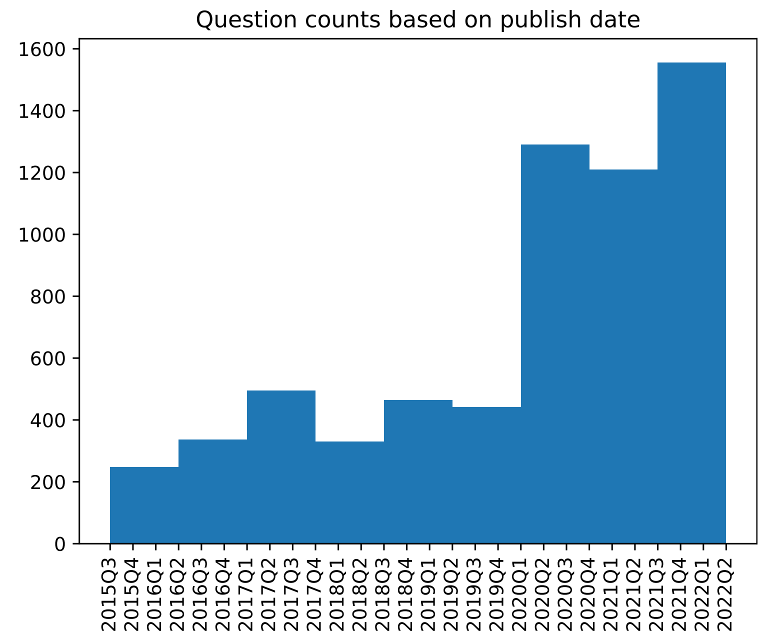


Figure 1: The number of questions published has been monotonically increasing through the last several years and the pace of increase is speeding up.

**Long Context Modeling.** An important aspect of forecasting is efficiently handling the dynamic aggregation of dispersed information among various agents [Paper: Timeline of prediction markets]. ML systems are particularly good at processing a large amount of information and weighing millions of variables for a certain objective. In order to design an architecture that can actually make sense of this task, we draw inspiration from [Paper: On-The-Fly Information Retrieval Augmentation for Language Models]. Concretely, for temporal processing, we experiment with encoding the document feed throughout a prediction timeline with a reader model daily and feeding the aggregated representations sequence into a decoder-only transformer backbone, then training autoregressively on crowd prediction targets.

**Large Zero/Few-shot Models.** As a benchmark, we test our results against the UnifiedQA model, which is a general purpose pre-trained model that demonstrated solid applicability to various question answering tasks ranging from extractive span selection to multiple choice [Khashabi et al., 2022].

## 3 Dataset

In our forecasting work, we collect thousands of questions spanning multiple choice (categorical) and T/F (binary) over a wide variety of domains (with discrete and continuous probability predictions). Questions are scraped from Good Judgement, Metaculus, and Kalshi, which are forecasting tournaments and prediction markets. For calibration, we also filter for and compile about 30,000 questions with numerical answers, taken from Stanford's Question Answering Dataset (SQuAD), 80K Hours Calibration, Grade School Math 8K (GSM8K) [Cobbe et al., 2021], TriviaQA, and Hendrycks Test (MMLU) [Hendrycks et al., 2021a].

To increase the quality of our forecasting questions, we implement dataset balancing for T/F questions. We perform question negation using OpenAI's 175B GPT-3 Edit model and few shot prompting. (Concretely, we can negate a question whose answer is True so that the negated question's answer is now False).

To supplement these questions with relevant historical information from a corpus of contextual text, in our work, we use the commoncrawl corpus news corpus, which includes important textual information in the form of news articles going up to the current day. We extract news from 2016 to the

**Algorithm** Adaptive Binning RMS for Calibration Error

**Input:** A set of $N$ examples each with labels $\{y_1, \dots, y_k, \dots, y_N\}$ and $C$ predicted confidence intervals $[[(l_k^1, u_k^1), \dots, (l_k^C, u_k^C)]$ for $k$ in $N]$ corresponding to $C$ confidence levels $[CL^1, \dots, CL^C]$. Set bin size to $M$.
**function** AdaptiveRMS
  Sort the examples by labels $y_n$ in ascending order.
  Assign a bin label $b_k = \frac{k-1}{M} + 1$ to each by splitting sorted examples into chunks of $M$.
  Let $\{B_1, \dots, B_b\}$ be the set of bins and $B_b$ the subset of examples in bin b.
  **for** $c = 1, \dots, C$ **do**
    Calculate empirical containment for bin b

$$\hat{p}_b^c = \frac{1}{|B_b|} \sum_{k \in B_b} \mathbb{1}(y_k \in [l_k^c, u_k^c])$$

    Calculate root mean squared calibration error

$$RMS^c = \sqrt{\frac{1}{b} \sum_{i=1}^{b} (\hat{p}_i^c - CL^c)^2}$$

  **end for**
  Output overall RMS by taking the mean of RMS for all confidence levels.

present, totalling more than 100GB of data, to use as relevant and recent information for forecasting on questions that are marked as resolved. Each question comes with its own corresponding date range, and our specific task is to retrieve the most relevant corpus articles falling under those dates.

Ultimately, the model is given a large amount of potentially relevant information in text format. In order to successfully produce a reasonable forecast, the model will have to discern and retrieve salient information, aggregate them in a meaningful way, keep track and update them over time, and finalize into a prediction.

# 4 Experiments

## 4.1 Setup

We test UnifiedQA models of all sizes which use the T5 backbone on the dataset with zero-shot prompting [Khashabi et al., 2022]. Then we also train FiD models with pretrained T5 [Raffel et al., 2020] as the backbone on the dataset directly for 10 epochs with a batch size of 8, an initial learning rate of 5e-5 with linear decay schedule, and a weight decay of 1e-2. To output numerical answers, we add and train an additional linear layer following the hidden state output of the FiD model. For retrieval, we experiment with DPR and BM25 with cross-encoder reranking and retaining the top 10 retrieved articles. The articles are concatenated to the questions and fed into the Fid models. For the temporal model, we freeze the finetuned FiD models in the previous setting to encode the question with the top one news article every day, outputting a sequence of embeddings. These embeddings are then treated as the input embeddings to an autoregressive model (GPT-2) which is then finetuned to predict the daily crowd prediction targets [Radford et al., 2019].

For calibration, we finetune DeBERTa-v3 models of all sizes on the numerical dataset with a three-part loss. The first part is the point estimate loss where an MSE loss is used to regress the predicted point estimate to the actual target. The second part is an MSE loss between the boundaries of the predicted confidence intervals to the actual target for boundaries that are on the wrong side of the target. The third part is again an MSE loss that penalizes the length of the predicted intervals so as to encourage finer predictions. The models are trained for 10 epochs with a batch size of 100.

| Model | Size | T/F | MC | Num | Avg | Macro |
|---|---|---|---|---|---|---|
| Random | – | 50.0 | 22.9 | 20.0 | 31.0 | 31.0 |
| UnifiedQA-v2 | small | 46.8 | 22.0 | 20.0 | 29.6 | 30.1 |
| | base | 43.0 | 19.5 | 20.0 | 27.5 | |
| | large | 47.5 | 21.2 | 20.0 | 29.5 | |
| | 3B | 58.6 | 19.0 | 20.0 | 32.5 | |
| | 11B | 53.8 | 20.3 | 20.0 | 31.4 | |
| T5 | small | 62.5 | 28.2 | 25.5 | 38.8 | 39.6 |
| | base | 61.1 | 26.7 | 27.6 | 38.5 | |
| | large | 61.0 | 32.1 | 29.3 | 40.8 | |
| | 3B | 62.1 | 28.2 | 31.3 | 40.5 | |
| T5 + DPR (10 news) | small | 63.2 | 28.2 | 27.6 | 39.7 | 39.7 |
| | base | 61.3 | 31.3 | 23.1 | 38.6 | |
| | large | 62.9 | 28.2 | 27.9 | 39.7 | |
| | 3B | 64.6 | 30.5 | 27.2 | 40.8 | |
| T5 + BM25 CE (10 news) | small | 62.9 | 29.8 | 28.9 | 40.5 | **41.1** |
| | base | 63.8 | 30.5 | 25.5 | 40.0 | |
| | large | 65.6 | 29.0 | 31.0 | 41.8 | |
| | 3B | 67.0 | 33.6 | 25.2 | 41.9 | |
| T5 + GPT-2 Temporal (1 news) | small | 61.9 | 28.2 | 25.9 | 38.7 | 40.9 |
| | base | 63.2 | 32.8 | 23.5 | 39.8 | |
| | large | 64.6 | 29.0 | 28.2 | 40.6 | |
| | 3B | 67.6 | 32.1 | 33.3 | **44.3** | |

Table 2: Different model performance on the forecasting benchmark. T5 with the top 10 news retrieved from the period the question remain active obtains the best macro average. But adding in temporal information can further improve performance if the model is large enough. With a T5-3B and GPT2-xl, we get the best performance on the dataset.

### 4.2 Results

Our baseline algorithms significantly outperforms UnifiedQA models which are mostly below random performance. This shows the difficulty of the dataset because UnifiedQA obtains strong performance on a entire suite of natural language datasets with clear scaling behavior whereas this is not the case here. However, we introduce baseline algorithms and identify several factors that could result in better machine forecasters.

**Model Size.** The performance on both the forecasting and calibration datasets strongly suggest that bigger models obtain better results. The trend becomes even clearer when the method is more effective and aggregates more information.

**Retrieval.** DPR has been shown to perform poorly when there is a domain shift. Since we do not finetune the DPR model, we don't get much boost from using DPR retrieved articles. However, as shown in the BEIR benchmark, BM25+CE reranking is the best method when tested on out-of-domain retrieval datasets, our results follow this conclusion nicely, improving over the simple finetuning baseline.

**Temporal.** When daily crowd predictions are used as targets for an autoregressive setup, we get a further boost with the largest model because these additional signals.

**Calibration.** Performance on the calibration task also shows strong trend that larger models are better, as is true in a variety of performance metrics. The most important test AdaRMS is however still very large which suggests room for improvement over the baseline algorithm.

## 5 Conclusion

We introduce a forecasting benchmark and a calibration benchmark. The benchmark contains forecasting questions scraped from prediction markets and forecasting tournaments which we release with an accompanying dataset of news articles. We experiment with baseline algorithms and show the

| Model | Size | Total RMS | PE Dist | Interval Len | AdaRMS |
|---|---|---|---|---|---|
| DeBERTa-v3 | xsmall | 14.3 | 0.84 | 28.9 | 22.5 |
| | small | **9.0** | 0.78 | 16.6 | 20.1 |
| | base | 11.0 | 0.69 | 11.7 | 19.1 |
| | large | 9.4 | **0.54** | **6.6** | **17.2** |

Table 3: Calibration

effective of larger model size, more context, better retrieval method, and incorporation of temporal targets. We also show how to obtain better calibration when outputs are numerical and introduce a way to measure calibration when the model is allows to output a confidence interval. Our results on both benchmarks show significant room for future improvement.

# References

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv*, 2021b.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. Forecastqa: A question answering challenge for event forecasting with temporal text data. 2021.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*, 2022.

Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*, 2015.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL `https://openreview.net/forum?id=wCu6T5xFjeJ`.