Event Prediction Based on Deep Learning and Neural Networks

Pengfei Li
Boston University
lpf00@bu.edu

Siyi DU
Boston University
dsharon@bu.edu

Abstract

A challenge of event prediction is proposed by the given paper. The paper highlights the critical role that forecasting plays in global decision-making, with examples such as the impact of COVID-19 predictions on national lockdowns and the influence of economic forecasts on interest rates. While human experts traditionally make these forecasts, AI systems offer the potential to process vast amounts of data and improve the accuracy of predictions. The goal of our project is to build a machine-learning model that makes accurate and calibrated forecasts. We implemented several different models for this objective, evaluated the result using beir scores, and did a comparison between models. Our latest submission scored 86.8 and ranked 24 on the leaderboard.

1 Introduction

Precisely predicting forthcoming global events is a challenging yet crucial task that can significantly impact policymaking and decision-making in a wide range of areas, including climate, geopolitical conflicts, pandemics, and economic indicators. Despite relying on their expertise to create accurate projections in these fields, human specialists' ability to do so has recently come into question due to advancements in language modeling.

The author of the given paper introduced Autocast, which contains thousands of forecasting questions obtained from real-world forecasting contests, ensuring that the data is diverse, of high quality, and relevant to real-world scenarios. Accompanying this dataset is a news corpus organized by date, allowing us to precisely replicate the conditions under which humans made past forecasts, without any leakage from the future.

Our research indicates that language models' forecasting ability is far inferior to that of human experts. However, by increasing the model size and integrating essential information from the news corpus, we have seen improvements in performance. Autocast represents an exciting challenge for large language models, and enhancing their forecasting abilities can lead to better decision-making in numerous domains. At the same time, We find several drawbacks to the proposed datasets and objective settings.

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

Preprint. Under review.

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. The author of Autocast designed an architecture for this purpose (albeit with limits on article length and time horizon), drawing inspiration from Wang and McAllester (2020).

Calibration Calibration is important in forecasting (Tetlock and Gardner, 2016). Even expert forecasters will be highly uncertain about some outcomes of interest. Such forecasts will be more useful in the form of calibrated probabilities than as point estimates. Thus forecasters are evaluated with proper scoring rules, which incentivize calibration. There is an extensive literature on improving the calibration of deep learning models (Guo et al., 2017; Nguyen and O'Connor, 2015; Lin et al., 2022; Minderer et al., 2021; Kull et al., 2019b), mostly for classification with a fixed set of classes. One part of Autocast requires models to forecast continuous quantities varying over multiple orders of magnitude, which has not been explored in prior work.

Truthful question-answering Current language model soften generate falsehoods when answering questions (Shuster et al., 2021; Lin et al., 2021), and they also achieve poor calibration when giving probabilistic answers (Hendrycks et al., 2021a) to human knowledge questions. However, for questions with a known ground truth answer, we expect models to improve as a result of scale, fine-tuning, and information-retrieval from reliable sources (Bai et al., 2022; Nakano et al., 2021; Hadfield-Menell et al., 2016; Turner et al., 2020; Wainwright and Eckersley, 2019). Yet humans also want models to give calibrated and truthful answers to questions that are too difficult or costly for us to answer ourselves (Irving et al., 2018; Evans et al., 2021; Leike et al., 2017; Hendrycks et al., 2021d; Reddy et al., 2020; Nahian et al., 2021). Forecasting is useful for this purpose. Forecasting questions are challenging but eventually become easy to evaluate. By contrast, it may be difficult for humans to evaluate superior answers to open problems in fundamental philosophy or science.

ChatGPT (GPT3.5) ChatGPT is a cutting-edge language model trained by OpenAI using the GPT-3.5 architecture. It has been designed to be highly versatile and adaptable, able to engage in natural language conversations on a wide range of topics. ChatGPT has been extensively trained on a massive corpus of textual data, allowing it to generate high-quality responses that are both coherent and relevant to the input provided. Its ability to understand and interpret human language makes it an invaluable tool for a wide range of applications, from virtual assistants to language translation services. Overall, ChatGPT represents a major step forward in the development of artificial intelligence, and is likely to have a profound impact on the way we interact with technology in the years to come.

3 Problem formulation

Event prediction is a challenging task. In our experiment, we adopted the Autocast dataset as the objective of our event prediction project. The questions in Autocast cover a very wide variety of topics. The questions are divided into five main categories: Economy, Politics, Science, Social, and Other. Each category contains numerous subcategories for a total of 44 subcategories ranging from foreign policy to AI.

At the moment the result of present language models is far more interior than human experts in terms of prediction accuracy. Our project seeks to try new methods on the given dataset and find insight along the process.

4 Methods

4.1 UnifiedQA

We tried UnifiedQA on our dataset at first. UnifiedQA does not rely on pre-training or fine-tuning on specific tasks or datasets. Instead, it is trained on a diverse set of question-answering datasets, which allows it to generalize to new tasks and domains with minimal additional training. This makes UnifiedQA highly versatile and adaptable, able to provide high-quality responses to a wide range of natural language queries.

4.2 chatGPT(GPT3.5)

ChatGPT is a cutting-edge language model developed by OpenAI, based on the GPT (Generative Pre-trained Transformer) architecture. At its core, ChatGPT consists of a stack of transformer encoder-decoder blocks, with each block containing multi-head attention and feed-forward neural networks. The transformer architecture allows the model to effectively capture and process contextual dependencies in the input text, allowing it to generate highly relevant and coherent responses. Additionally, ChatGPT leverages pre-training on a massive corpus of textual data, enabling it to learn and generalize to a wide range of language tasks.

4.3 chatGPT(fine-tuned)

ChatGPT provides open API for fine-tuning the model. We created a fine-tuned dataset for chatGPT using Autocast training data(Figure 1). In order to cut the cost of training, we didn't apply news retrieval to this task. Instead, we only take the background and question of the data as questions. Otherwise, the computational cost would be unaffordable. The process crashed every time training on my local machine (32GB unified memory).

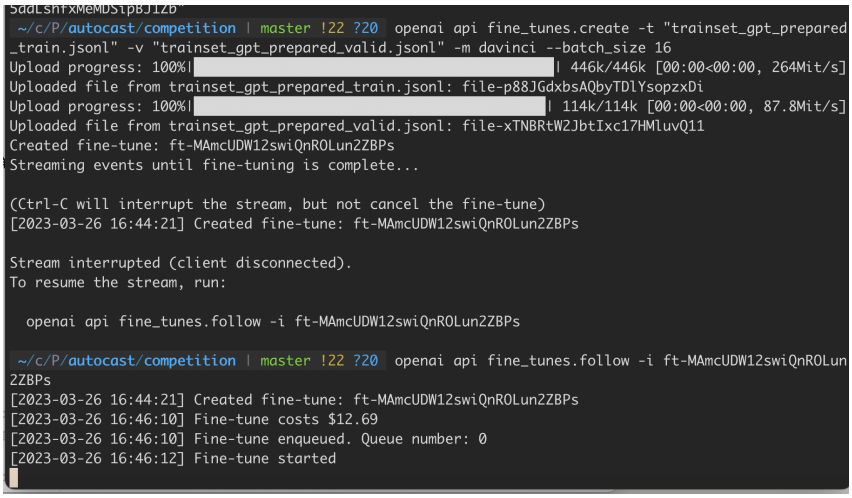

Figure 1: Fine-tuning chatGPT

4.4 Retrival based method

The method from the given paper is a retrieval based method. The author used ccnews as an article base to find related articles of a question. We did try to reproduce the experiment. But again, the computational cost is unaffordable for training a large language model on our own. The memory requirement is hard to reach(Figure 2).

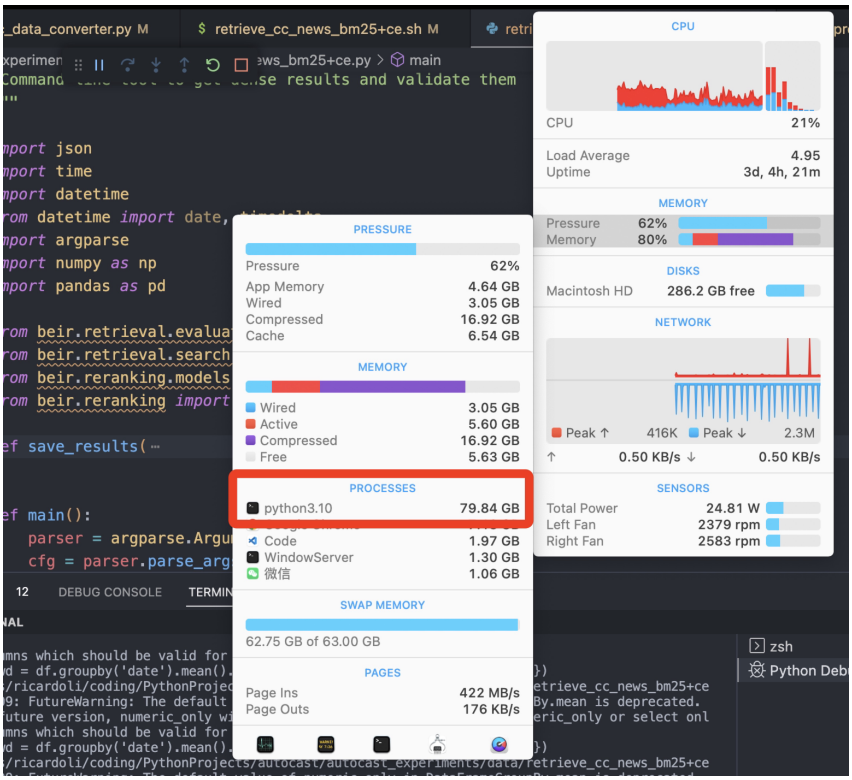

Figure 2: Memory usage of reproducing the proposed algorithm

Table 1: Result

Name	combined	TF	MCQ	Numerical
unifiedQA	148.3	47.0	78.0	23.3
chatGPT	151.5	55.6	73.3	22.7
chatGPT(fine-tuned)	111.2	49.5	39.1	22.7
random	86.8	25.0	39.1	22.7

5 Results

We had in all 23 tries in this experiment submitted. After tuning the parameter in our code, they received 148.3, 151.5, and 111.2 respectively. All of them are worse than the random answer given in the paper provided(Table 1).

The reason for this happening, We believe is because of the difficulty of this problem. It is generally hard for an LLM to analyze this chaotic world at present, while the random result given by the author can produce a result of 1/n, which is pretty good. Besides the problem difficulty, the dataset used in this paper is all confusing(Figure 3). As there are questions in pair collected to this dataset. I think this is very misleading. Settings like this need more consideration and clarification.

Overall, We believe it is still too early to adopt LLM to do event prediction.

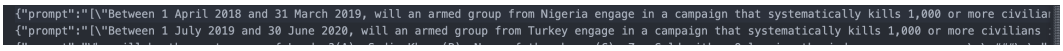

Figure 3: questions in pair

6 Technical depth and innovation

As the potential of large-scale language models becomes increasingly apparent, researchers are exploring a variety of exciting applications. One such application, as noted by the author, is event prediction. In our own project, we built upon the author's experiment by using a larger and more powerful language model for both training and testing. Through our work, we discovered certain laws and limitations that apply to language models.

7 Architecture and design

The model we used in our project are all transformer based. A transformer architecture has two segments, an encoder and a decoder, that work on input and target sequences, respectively. The encoder calculates attention vectors to determine important parts of the input, while the decoder computes embedding vectors for each target output and uses masked attention to predict the next word. GPT models use data compression to convert words into numerical representations and improve accuracy by calculating conditional probabilities. They perform well in "few shots" settings and require minimal examples to produce relevant responses.

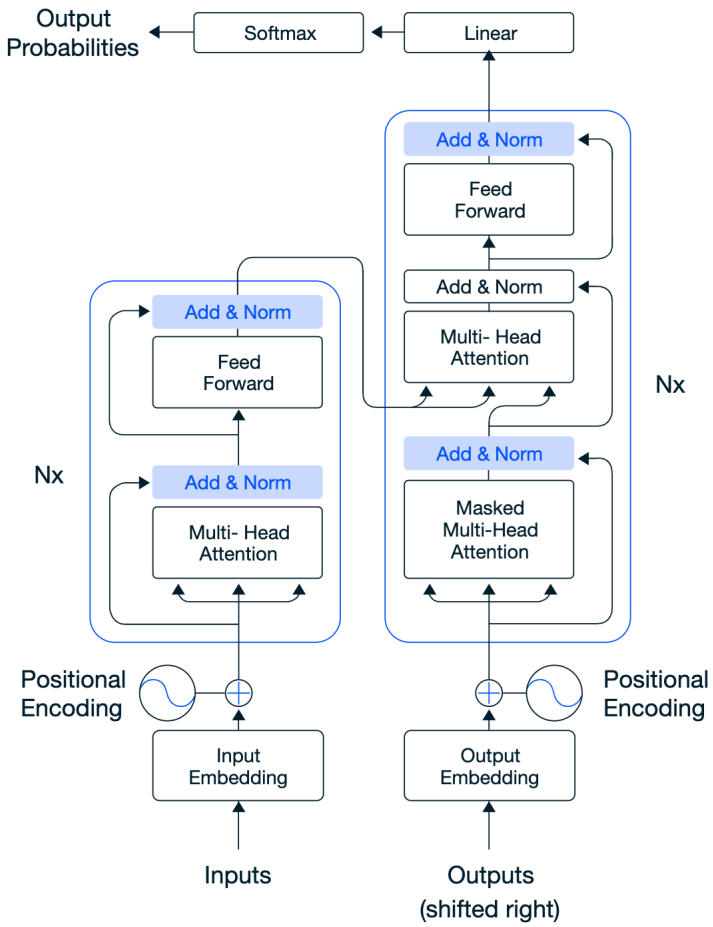

Figure 4: Transformer architecture

8 Github

https://github.com/R1cardoo/eventPrediction

9 Reference

David Adam. Modelling the pandemic the simulations driving the world's response to covid-19. Nature, 580(7803):316–318, 2020.

Jon Scott Armstrong. Principles of forecasting: a handbook for researchers and practitioners, volume 30. Springer, 2001.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

Peter Christensen, Kenneth Gillingham, and William Nordhaus. Uncertainty in forecasts of long-run economic growth. Proceedings of the National Academy of Sciences, 115(21):5409–5414, 2018.

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.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. Truthful ai: Developing and governing ai that does not lie. arXiv preprint arXiv:2110.06674, 2021.

Kenneth Gillingham, William Nordhaus, David Anthoff, Geoffrey Blanford, Valentina Bosetti, Peter Christensen, Haewon McJeon, and John Reilly. Modeling uncertainty in integrated assessment of climate change: A multimodel comparison. Journal of the Association of Environmental and Resource Economists, 5(4):791–826, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In International Conference on Machine Learning, pages 1321–1330. PMLR, 2017.

Dylan Hadfield-Menell, S. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. In NIPS, 2016.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. news-please: A generic news crawler and extractor. In Proceedings of the 15th International Symposium of Information Science, pages 218–223, March 2017. doi: 10.5281/zenodo.4120316.

Sven Ove Hansson. Fallacies of risk. Journal of Risk Research, 2004.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. CoRR, abs/2006.03654, 2020.

James Hedlund. Risky business: safety regulations, risk compensation, and individual behavior. Injury Prevention, 2000.

Dan Hendrycks and Mantas Mazeika. X-risk analysis for ai research. arXiv preprint arXiv:2206.05862, 2022.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. Proceedings of the International Conference on Learning Representations, 2019.

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, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874, 2021b.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. arXiv preprint, 2021c.

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

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. arXiv preprint arXiv:1805.00899, 2018.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In EACL, 2021.

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. 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), pages 4636–4650. Association for Computational Linguistics, August 2021. doi: 10.18653/v1/2021.acl-long.357.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. CoRR, abs/1705.03551, 2017.

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

Mathias Kirk Bonde. Getting gpt-3 to predict metaculus questions, 2022. https://www.lesswrong.com/posts/c3cQgBN3v2Cxpe2kc/getting-gpt-3-to-predict-metaculus-questions, Last accessed on 2022-06-08.

Meelis Kull, Miquel Perelló-Nieto, Markus Kängsepp, Telmo de Menezes e Silva Filho, Hao Song, and Peter A. Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In NeurIPS, 2019a.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. Advances in neural information processing systems, 32, 2019b.

J. Leike, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and S. Legg. Ai safety gridworlds. ArXiv, abs/1711.09883, 2017.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33: 9459–9474, 2020.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958, 2021.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. arXiv preprint arXiv:2205.14334, 2022.

Spyros Makridakis, Steven C Wheelwright, and Rob J Hyndman. Forecasting methods and applications. John wiley sons, 2008.

Spyros Makridakis, Rob J Hyndman, and Fotios Petropoulos. Forecasting in social settings: The state of the art. International Journal of Forecasting, 36(1):15–28, 2020.

Barbara Mellers, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, S Emlen Metz, Lyle Ungar, Michael M Bishop, Michael Horowitz, Ed Merkle, and Philip Tetlock. The psychology of intelligence analysis: drivers of prediction accuracy in world politics. Journal of experimental psychology: applied, 21 (1):1, 2015.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. Advances in Neural Information Processing Systems, 34, 2021.

Sebastian Nagel. Common crawl news dataset, 2016. URL https://data.commoncrawl.org/ crawl-data/CC-NEWS/index.html.

Md Sultan Al Nahian, Spencer Frazier, Brent Harrison, and Mark Riedl. Training value-aligned reinforcement learning agents using a normative prior. arXiv preprint arXiv:2104.09469, 2021.