University of Potsdam
Faculty of Science
Institute of Environmental Science and Geography
Institute of Physics and Astronomy
Climate, Earth, Water, & Sustainability

Master Thesis
for the award of the academic degree
Master of Science (M.sc.)
at the University of Potsdam

Forecasting Impact in the Earth System Sciences with Language Models

Potsdam, 23 August 2025

Submitted by:
Daniel Morgan Rivers
rivers@uni-potsdam.de
Matriculation No.: 829112

First reviewer: Prof. Christian Kuhlicke
Second reviewer: Dr. Ivan Kuznetzov

Abstract

Abstract in English

The field of decision science has made rapid strides with the introduction of techniques of deep learning in natural language processing as a tool for improving accuracy and calibration of economic and geopolitical forecasts, predicting scientific outcomes, and in domains like stock market prediction and diagnosing diseases. Several recent methods and improvements have been made to the predictive ability of large language models (LLMs) calibrated via fine-tuning techniques. Because some of these powerful forecasting techniques have not yet been brought to bear on the important problem of impact prediction in the context of earth system sciences, I fine-tuned two LLMs (Llama 70B and ChatGPT 3.5) to perform such forecasts, training on thousands of abstracts from the scientific literature. I find that the best-performing system ([Llama 70B / ChatGPT 3.5], averaging X parallel predictions) can forecast XX% of interventions correctly compared to YY% using a random forest model baseline, and ZZ% from published ex-ante predictions from the scientific literature. This work has wide-ranging applications both to improve sustainability forecasting, as well as in adjacent areas and improving decision making in policy contexts.

Abstract in German

Lorem ipsum ...

Table of Contents

1 Background: The Science of Forecasting

1.1 Introduction

1.2 Prediction Markets and Superforecasting

1.3 LLM Forecasting of Real-World Outcomes

1.4 Predicting Outcomes with AI in the Earth System Sciences

2 Methods for LLM Forecasting

2.1 Data Sources

2.2 Selecting LLM's for Forecasting in the Earth System Sciences

2.3 Baseline Measures to Compare Against LLM Forecasts

2.4 Methods for LLM Forecasting

2.5 Grading Forecast Accuracy

3 Results & Discussion

3.1 Strengths and Weaknesses of This Forecasting System

3.2 Evaluation of Techniques for Improving Forecast Accuracy

3.3 The Risk of Trusting This Forecasting System

4 Conclusion & Outlook

4.1 The State of AI and LLMs

4.2 The Promise and Capabilities of AI Forecasting

As a clear disclaimer: **LLM's are not in general superior to humans at forecasting as of May 2025.** At the same time, their forecasting ability for short-term predictions is closing in at a rapid pace as AI capabilities have advanced [source]. Furthermore, predictions with a significant number of relevant news articles or very near to the date of a forecasting resolution can best teams of trained forecaster's aggregate predictions in prediction accuracy (Halawi et al., 2024 | "Approaching Human-Level Forecasting with Language Models"). It is currently unknown to what extent the ability of AI systems to forecast geopolitical and economic events can be extended to forecasting the impact of interventions with implications in the earth system sciences. Exploring this domain opens a promising avenue to improve the efficacy of interventions in the earth system sciences. In the remaining section, we discuss the beneficial aspects of the system developed, as well as the potential dangers or risks this system may pose.

One co-benefit of a system fine-tuned on earth system sciences is that by its cross-domain nature, the LLM will be able to identify a wide range of likely outcomes, and the degree of effect of those outcomes, on a wide range of quantitative and qualitative outcomes. When implementing interventions, researchers, policy-makers, and decision makers must always consider many relevant outcomes of their interventions. The similarity and vector search of the system allow users to quickly identify relevant documentation as well as outcomes of similar scientific research most relevant to their proposed intervention.

Another benefit of the system is that AI's typically excel in domains where human experts are particularly challenged: when there is a very large range of relevant data or when predictions about the effect of an intervention involve carefully calibrated probabilities. AI's can also perform predictions in a way that human experts can learn from: introducing one piece of information can be used to quantify the effect on AI forecasts. AI forecasts can be ensembled arbitrary and at relatively little expense compared to humans.

4.3 Risks, Biases, and Limitations

However, there are clear risks of using AI for evaluating the likely outcomes of interventions in the earth system sciences. The most obvious issue may be that while AI can be accurate in some domains, current AI systems do not accurately present their confidence in their answers and can completely hallucinate events and facts which have no grounding in reality. The result is a misleading analysis, which in the space of earth system sciences may lead to significant risks. Policy makers may trust AI more than is justified by its performance, or view it as an unbiased source, despite nearly all current AI systems having a well-documented political bias acknowledged by both the political left and political right [source].

Another risk is that scientists may not perform research deemed to be unlikely to succeed, and thus the range of explored outcomes may be narrowed to the outcomes known to work in the past or deemed to be likely to be successful by the AI system.

While AI may be able to calibrate itself on many different domains and automatically pull in relevant information, it currently lacks the ability to reliably perform complex mathematical calculations or run long-term analysis. Furthermore, as AI becomes more advanced there is significant concern in the technology community that it may form its own goals and intrinsic values, out of alignment with its human operators. An AI that advises on AI policy may in fact present a conflict of interest, even if the AI is simply using heuristics mimicking human tendencies towards self-preservation and in-group preferences.

Finally, without the full text, there is a risk that the policy forecasting aspect may be quite limited. Without a sense of the scope of an intervention, which would not reliably be indicated in the abstract, the degree of impact of an intervention may difficult to ascertain by any forecasting system.

4.4 System Design and Risk Mitigation

We address these concerns by noting that as AI begins to become more accurate and lower cost than human researchers at forecasting the impact of policy outcomes, it becomes ever more important to have specifically designed systems that take steps to reduce the dangers of AI systems. We believe the system developed clearly fulfills this criterion. The system we use in this work specifically provides credible, peer-reviewed scientific information and news from reputable sources to the AI, rather than relying on general internet search as many current AI providers rely on. [OPTIONAL: Furthermore design our system to be calibrated via fine-tuning, meaning that some of the reliability concerns may be ameliorated.] As AI systems advance, there appears to be a progression towards more agentic systems with more clear intermediate goals. A misalignment with human preferences (an example in this work might be downplaying the CO2 effects of building more AI systems in order to increase the number of AI systems as an in-group preference) may occur and be missed by humans with extremely long thought chains and insufficient detection of misalignment. Our system by contrast allows the user to inspect the series of logical deductions performed by the model and view available sources the model used as scientific reference material. The system has been specifically quantified in terms of its bias, allowing users to have full knowledge of the likely failure modes when using the system, often absent in generally available AI chat interfaces. [OPTIONAL: with an explicit attempt to correct these biases via fine-tuning, syncophantic behavior is also reduced compared to RLHF models.]. Another risk is that papers tend to have a bias, and the model will learn to replicate that bias. Papers are much more likely to have "significant" results than mixed effect or no effect. The optimistic bias towards positive bias published in journals should mean we interpret the prediction of the model cautiously, with knowledge that it will likely present a more optimistic version of the outcomes than is justified from a neutral observer's perspective. In order to counteract this risk, we are also looking at the accuracy of the quantitative result of the intervention, which is more valid to compare between abstracts and has a relatively smaller publisher bias [source]. Finally, much of the promise of the AI forecasting approach relies on models continuing to become lower cost and more performant in general domains. While multiple empirical trends and the longstanding success of Moore's law clearly indicate this should continue, it is by no means guaranteed. If AI models cease to improve on relevant metrics, or otherwise become increasingly biased or unreliable, much of the promist of an AI forecasting tool for estimating interventions in the earth system sciences goes away. Despite this risk, the system remains useful and informative for the scientific and public policy community as it provides a system with sources proven to provide useful information for the evaluation of policy outcomes, and introduces a framework by which the impact of interventions can be broken down for more accurate predictions. While there is a possibility that AI's may never reach the capabilities of humans in integrating the disparate sources of information, automated information search and a new tool that can synthesize relevant information can be a powerful tool for scientists and policy makers. Forecasting has the distinct benefit of disallowing training on any particular benchmarks and is a rather difficult-to-game metric compared to standard LLM performance benchmarks. It becomes increasingly useful to society to understand what the true capabilities of LLM's are and the rate of their improvement, both for the regulation of dangerous AI capabilities and the improved understanding where AI may be capable enough for reliable use in various critical domains such as automated medicine and driverless vehicles.

4.5 Broader Applications and Vision

The codebase and research done here can be repurposed from specifically earth system science, to other domains where impact forecasts are clearly useful. A similar system with an expanded set of abstracts and data could be used with relatively little modification in domains such as public health, financial policy, and in a more general way to provide predictions for scientists about likely qualitative and quantitative outcomes of their scientific studies. The success of the model demonstrates that a great deal of opportunity to synthesize scientific findings and improve decision making on an institutional level is policy. One particularly promising avenue for expansion of the system would be as an application to Futarchy first proposed by Robin Hanson. Futarchy proposes to use prediction markets to allow policy makers or the general public to only have to agree on what they value and quantify as utility, not on how to maximize that utility. Several prediction markets in parallel are formed, creating a zero-sum game financially rewarding players that best predict the utility outcome conditional on a policy being implemented. To the extent that complex public policy can ever be reduced to a single utility function, that this function can be agreed on by a quorum of policy makers, Futarchy could significantly reduce gridlock and polarization in politics, at least in the domains in which the necessary conditions are useful and possible. In essence, Futarchy aids policy makers in coming to agreement on how to implement policies by reducing the scope of disagreement to what the set of possible policy implementations could be and how they would choose to quantify a successful outcome. If and when the system proposed is shown to exceed human ability in predicting policy, or if it can be shown that the system can be complementary to human predictions, cheaply improving their accuracy, this system could be integrated to a scheme for futarchy by replacing or augmenting prediction markets. This may be especially helpful in use-cases where AI succeeds and prediction markets fail: very low probabilities over long time periods (as the winners may choose to invest their money on a higher-return investments), predictions about long-run outcomes that are difficult to gain information about, particularly contentious outcomes, or issues where markets may be biased by particularly wealthy individuals who come in very late in the market and buy many more shares than expected.

4.6 Ideation: Extensions and other applications

- Improving prospects of futarchy to improve governance
- Understanding how different sources of information contribute to effective forecasting of impact
- Before the forecasting at all: collecting the information for forecasting all in one place, both resources to make reasonable forecasts, as well as creating structure out of unstructured papers in earth systems science
- Creating general hierarchies of impact for different categories of interventions
- Ability to create "unbiased" forecasts that are both evidence based and listened to by both sides of the political spectrum
- Increasing democratic understanding of the likely effects of laws from third party sources: allows non-experts to assess the efficacy of elected officials in accomplishing their goals
- Automated scoring of introduced legislation

- Sufficient statistics to introduce confidence bars on the effects of political outcomes
- Leveraging the advance of AI for good
- Constraining the use of AI in a scientifically valid, constrained manner, which minimizes the risk that AI biases themselves influence policy decisions.
- Automated feedback on proposed interventions (registered studies): what are the likely things this has impact on? What are some relevant papers for their proposal?

Works Cited

References

Halawi, Danny et al. (Nov. 2024). "Approaching Human-Level Forecasting with Language Models". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. (Visited on 08/18/2025).

Declaration of Academic Integrity

Ich, Morgan Rivers, erkläre hiermit, dass diese Arbeit das Ergebnis meiner eigenen Arbeit ist. Ich danke für die Unterstützung, die ich bei der Erstellung dieser Arbeit erhalten habe, und für die verwendeten Quellen. *(I, Morgan Rivers, hereby declare that this thesis is the product of my own work. All the assistance received in preparing this thesis and the sources used have been acknowledged.)*

Potsdam, 23 August 2025