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, 27 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

Background The earth system sciences concern the complex interaction between biological, chemical, physical, and anthropogenic processes. A broad goal of the earth system sciences is to model and accurately predict the outcomes of interventions with regard to the environment and its impact on humans. Much of the progress in earth system science has been on linking these complex phenomena into large models, such as integrated assessment models (IAMs), computable general equilibrium models (CGEs), or agent-based models (ABMs). While many attempts have been made to model specific subsystems within the earth system, such as the carbon cycle, environmental and economic linkages, or understanding human impacts in the climate-water-food nexus, there have been few attempts to create a comprehensive model which can predict quantitative or qualitative outcomes of a wide range of cross-domain interventions in the earth system which could be described in natural language.

In particular, the earth system is a “complex system” - characterized by difficult-to-predict, emergent phenomena, and both positive and negative feedback loops.

Thus far, models in the earth system sciences have largely relied on mechanistic, theoretically-based models of the underlying complex systems they analyzed. However, this is not the only way to predict outcomes -

“Machine Learning” (ML) outcomes , while lacking the rigorous mechanistic underlying processes characterizing IAMs, CGEs, and ABMS, have recently been shown to perform better than human or other computer modelling techniques to model complex systems in domains such as language modelling (Brown et al., 2020 | “Language Models Are Few-Shot Learners”), protein folding (Jumper et al., 2021 | “Highly Accurate Protein Structure Prediction with AlphaFold”), biodiversity protection (Silvestro et al., 2022 | “Improving Biodiversity Protection through Artificial Intelligence”), and weather forecasting (Lam et al., 2023 | “Learning Skillful Medium-Range Global Weather Forecasting”).

The collective failure of the scientific community to model complex outcomes in the earth system has severe implications. For example, work from (Stechemesser et al., 2024 | “Climate Policies That Achieved Major Emission Reductions: Global Evidence from Two Decades”) has demonstrated that out of 1500 policies between 1998 and 2022, only 68 had statistically significant causal effect to reduce country emissions with a 99% or higher confidence. Furthermore, they find that more than four times the effort witnessed so far in emissions reductions from implementing more successful policies in line with past reductions would have to be exerted to close the emissions gap to remain below 2 degrees C in global temperature rise. Broadly, their findings support the claim that even when climate policy is implemented, it is largely ineffective, and in the future it will need to be much more effective to avoid dangerous levels of CO_2 concentrations. In terms of biodiversity, achieving sustainability cannot be met by current trajectories, and goals for 2030 and beyond may only be achieved through transformative changes across economic, social, political and technological factors (Watson et al., 2019 | *The Global Assessment Report on BIODIVERSITY AND ECOSYSTEM SERVICES*). As of 2022 pollution remains responsible for approximately 9 million deaths per year, corresponding to one in six deaths worldwide (Fuller et al., 2022 | "Pollution and Health: A Progress Update").

While much scientific effort has been expended on understanding underlying systems, much less effort has been directly focused on predicting which specific policies, if enacted, would realistically improve outcomes on the indicators of interest in the earth system sciences. Meanwhile, examples exist in the literature where regulation can greatly reduce or even eliminate environmental problems - the Montreal protocol has met with great success in closing the hole in the ozone layer [CITE IF KEEP THIS SENTENCE!].

IAMs have shown promise in modelling outcomes of specific policies, with the disadvantage that they are harder to use and set up, require a high computational power and expertise to use effectively, and are not rigorously benchmarked on large databases of existing interventions and associated outcomes. For any user-defined policy package (for example, introducing efficient clean-burning cookstoves in India), GAINS can calculate the reduction in emissions (PM-2.5, NO_x, CO_2, etc), the improvement in ambient air quality, and the health impacts such as lives saved from lower PM-2.5 exposure (, 2011 | "Cost-Effective Control of Air Quality and Greenhouse Gases in Europe: Modeling and Policy Applications"). Other IAMs include the MIT Emissions Prediction and Policy Analysis (EPPA) model, which requires manually entering assumptions of the effects of policies into models of the world economy, calculates the implications on health and runs a CGE to estimate the economic effects (, | *The MIT Emissions Prediction and Policy Analysis (EPPA) Model: Version 4 | MIT CS3*).

In the domain of biodiversity, an ML-based framework called CAPTAIN uses a reinforcement learning (RL) agent coupled with a spatially explicit ecosystem simulation to statistically learn which areas to protect over time in order to maximize species survival under budget constraints, to maximize cost-effectiveness in protecting biodiversity (Silvestro et al., 2022 | "Improving Biodiversity Protection through Artificial Intelligence"). Other techniques used to predict outcomes of interventions include linear optimisation combined with econometric theory, such as the Open Source Energy Modelling System (OSeMOSYS). OSeMOSYS simulates energy production and consumption under policy constraints including a model of the energy grid. By incorporating physical and known constraints, such models have the potential to predict outcomes of policy interventions over longer time horizons (, 2011 | "OSeMOSYS: The Open Source Energy Modeling

System: An Introduction to Its Ethos, Structure and Development"). An even more fine-grained, bottom up approach of modelling intervention outcomes is possible. For example, combining bio-economic farm optimization models with ABMs, researchers have modelled evolution of pesticide-related risks for the country of Switzerland (Dueri and Mack, 2024 | "Modeling the Implications of Policy Reforms on Pesticide Risk for Switzerland").

Despite all these examples, in many relevant sub-domains, such as climate policy, ex ante analysis of mitigation action and of mitigation plans is limited (Intergovernmental Panel On Climate Change (Ipcc), 2023 | "Mitigation and Development Pathways in the Near to Mid-term"). Given the overwhelming complexity of the earth system, and the corresponding failures to properly model many of the system components in the earth system and especially how they interact with human interventions, complementing mechanistic understanding and prediction with ML approaches is urgently needed.

Proposal We set out to predicting near-term, future states in a wide array of different contexts. One method that has shown a great deal of promise in such domains is "judgemental forecasting", which allows experts in the skill of forecasting generally, to use tools including fermi estimates, intuition, and information gathering to make a calibrated prediction on the likelihood of a given outcome (Halawi et al., 2024 | "Approaching Human-Level Forecasting with Language Models"). This can be contrasted with "statistical forecasting" which typically uses time-series prediction methods or purely quantitative approaches.

This thesis proposes the use of Large Language Models (LLMs) to implement judgemental forecasting to predict how effective interventions will be in the earth system. By splitting records of the effectiveness of thousands of interventions in the earth system from the scientific literature into an intervention and an outcome, I will use language models to mimic the reasoning and data gathering skills of trained forecasters, hoping to replicate the success at using judgemental forecasting from language models in adjacent domains. Ultimately, the goal is to learn whether it is possible to compliment a scientifically founded prediction for the effectiveness of a given intervention with the reasoning produced from a system involving a system with LLMs that are specifically trained for the task at its core. Given the difficulty of field testing ideas, policymakers often rely on expert forecasts on how an intervention will meet its intended goals to select which interventions will be implemented. Replacing or augmenting that advisory role could greatly improve decision making in this context (Hewitt et al., | "Predicting Results of Social Science Experiments Using Large Language Models").

I will briefly review current progress in specific domains of outcome prediction in the earth system sciences, and then discuss progress with LLMs in adjacent domains. To my knowledge, there has been no attempt at predicting real-world outcomes of interventions in the earth system sciences while also rigorously quantifying the skill of such a system.

Within the domain of LLM use, there has been some progress. A recent tool called "climsight" summarizes and aggregates information about climate adaptation and mitigation (Koldunov and Jung, 2024 | "Local Climate Services for All, Courtesy of Large Language Models"), but stops short of making predictions towards adaptation. Machine learning and LLMs have been used to collect over 80,000 articles about climate adaptation and provide analysis about which areas of implementation are lacking and point out gaps in attention towards promising categories of policies.

Limited work has also been done using LLMs such as ChatGPT-4 to serve as data sources for policy deliberation and multi-criteria assessment of climate and sustainability interventions, finding GPT-4 is in rough agreement with the policy rankings of human experts for the expected outcomes (Bina et al., 2025 | "On Large Language Models as Data Sources for Policy Deliberation on Climate Change and Sustainability"). However, very little is done to improve on GPT4's abilities, the assessment was made on only a few dozen generic policy examples, and no attempt was made to compare outcomes between these policies and real-world outcomes. Despite these limitations, the findings are promising. For multiple criteria decision making (MCDM), GPT-4 provided a useful collaborative starting point, eased the process of considering multiple criteria effectively, and aided policy deliberation on climate change and sustainability.

One study with the bottom-up technique of simulating individual respondents using GPT4 according to their demographic profiles (while matching demographic data in the US), found that the "social policy" papers had a relatively low correlation with prediction accuracy at an accuracy of 0.64 compared to an average of about 0.9 compared to other studies (Hewitt et al., | "Predicting Results of Social Science Experiments Using Large Language Models"). Although the methodology may lead to differing outcomes (simulating individual profiles versus the direction of this thesis which prompts the LLM to directly reason out the answer), this may hint that public policy may be more difficult to predict than other scientific results. In the soical science prediction work, they separately analyzed 346 treatment effects estimated from the responses of over one million participants, with hundreds of ex-ante predictions made before the outcomes from experts. The interventions included surveys that simulated the effect of informational content which promoted pro-democratic attitudes, encouraged respondents to increase beneficial choiced with respect to climate change, and increase their vaccination rates. LLM predictions were more accurate for survey experiments than field experiments (79% vs 64% accurate respectively). LLMs also matched or exceeded expert prediction accuracy.

1.2 Prediction Markets and Superforecasting

In recent years, significant progress has been made on accurate near-term forecasting outside of specific domains. The most promising approaches appear to be a mix of prediction markets, and specialized, trained experts known as "superforecasters" (Tetlock and Gardner, 2015 | *Superforecasting: The Art and Science of Prediction*). Prediction markets have gained recent prominence in the domain of geopolitical forecasting, with significant volumes of transactions on predicting future geopolitical outcomes with a broad purview, including election results, the outcomes of treaties, or whether a regime will topple. Predictions are typically more accurate than a guess of 50% chance on a given outcome only within approximately one year time horizon, with the accuracy notably improving as the event reaches only a few months until the question is closed (Tetlock and Gardner, 2015 | *Superforecasting: The Art and Science of Prediction*) TODO: CHECK+REFINE THIS STATEMENT. In a broad range of complex, human-involved outcomes, prediction markets are superior to expert analysis. In the words of the economist Robin Hanson, "racetrack market odds improve on the prediction of racetrack experts; orange juice commodity futures improve on government weather forecasts; stocks fingered the guilty firm in the Challenger crash long before the official NASA panel; Oscar markets beat columnist forecasts; gas demand markets beat gas demand experts; betting markets beat Hewlett Packard official printer sale forecasts; and betting markets beat Eli Lily official drug trial forecasts." (Hanson, 2013 | "Shall We Vote on Values, But Bet on Beliefs?").

However, prediction markets have demonstrated that a smaller subset of forecasters in the market, known as "superforecasters", are statistically much better forecasters than the prediction market, and ensembling these forecasters and letting them exchange information among themselves leads to higher accuracy predictions than prediction markets alone (Mellers et al., 2015 | "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions"). Due to the relatively high expense and human effort required to organize superforecasters, the efforts have been largely focused on specific geopolitical and economic questions, some of which may fall under the domain of intervention impact in the earth system sciences, although most point to broad trends where information may be gathered from the news and informal internet searches, and deep expertise in any single domain would not be required for an accurate forecast. In fact, in terms of "calibration", superforecasters usually beat domain experts in their own fields by maintaining a broad sense of good judgement and cultivating a trained skill at accurately estimating prediction probabilities, rather than overly relying on a single strategy (such as econometric analysis, or specific statistical methods).

1.3 LLM Forecasting of Real-World Outcomes

Methods and Capabilities Assembling superforecasters to predict the efficacy of interventions such as laws, specific interventions such as introduction of cleaner burning ovens, or regulations on air quality would be prohibitively costly. Instead, this work focuses on the mimicking of techniques known to be effective for superforecasters with LLMs.

As might be expected given the lack of world knowledge and complex reasoning abilities of LLMs, simply replacing a crowd of humans with a crowd of untrained LLMs does not generally outperform the crowd average (Abolghasemi, Ganbold, and Rotaru, 2025 | "Humans vs. Large Language Models: Judgmental Forecasting in an Era of Advanced AI") (Schoenegger and Park, 2023 | *Large Language Model Prediction Capabilities: Evidence from a Real-World Forecasting Tournament*) CHECK THESE REFS.

While the base models appear to underperform compared to crowds of humans, more recent work on the question has shown that increasing model reasoning ability increases the forecasting accuracy (Yan, Q., Seraj, R., He, J., Meng, L., and Sylvain, T., 2024 | "AutoCast++: Enhancing World Event Prediction with Zero-shot Ranking-based Context Retrieval"), and that with proper techniques and careful prompting, LLMs will approach or sometimes exceed accuracy of superforecasters on questions with a high degree of context and with proper ensembling and fine-tuning of the LLM system (Halawi et al., 2024 | "Approaching Human-Level Forecasting with Language Models"). (, | *Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy | Science Advances*) (Abolghasemi, Ganbold, and Rotaru, 2025 | "Humans vs. Large Language Models: Judgmental Forecasting in an Era of Advanced AI") (Yan, Q., Seraj, R., He, J., Meng, L., and Sylvain, T., 2024 | "AutoCast++: Enhancing World Event Prediction with Zero-shot Ranking-based Context Retrieval"). TODO: CHECK THIS STATEMENT

The best results are achieved by capitalizing on the broad world-knowledge of LLMs and the augmentation of their knowledge in high-news or near-term contexts (Halawi et al., 2024 | "Approaching Human-Level Forecasting with Language Models"). A few crucial improvements to the base-level prediction ability can be applied to approach superforecasting level calibration and accuracy. These improvements are:

1. Fine-tuning the LLMs to replicate the format of good forecasts, using hundreds or thousands of correct forecasts as the fine-tuning dataset (or in some cases, directly fine-tuning on existing content in the target area (Wen et al., 2025 | *Predicting Empirical AI Research Outcomes with Language Models*))
2. Have the LLM integrate relevant and timely information into the context to improve the forecast (such as NewsCatcher and Google News API)

3. Have the LLM split questions into sub-questions before being used to query news API

4. Prompting techniques: Have the LLM think step-by-step, rephrase the question to improve comprehension, and reason over chains of crafted prompts to ensure sufficient reasoning effort has gone into the answer

5. Reduce error rates by aggregating the final predictions ("Wisdom of the crowd")

TODO: Would be good to go through and see which of the cited papers have adopted these techniques, and which were successful.

In one similar work, the technique of Chain of Thought (CoT) has been used to improve the reasoning abilities of GPT4 in predicting the outcome of 1261 conclusions from 276 papers which analyze the real-world outcomes of field experiments in the social sciences. While not specifically investigating outcomes with relevance in the earth system sciences, they do investigate the prediction ability for the impact of educational incentives, household finance behavior, healthcare enrollment, and financial planning. Remarkably, over the 1261 outcomes, 78% were predicted accurately by the system (Chen, Hu, and Y. Lu, 2025 | *Predicting Field Experiments with Large Language Models*).

Lastly, a recent study found that LLMs can correctly predict outcomes in scientific domains such as predicting results of papers in neuroscience, not just in geopolitical or economic forecast contexts (X et al., 2025 | "Large Language Models Surpass Human Experts in Predicting Neuroscience Results"). Notably, this result used the raw probabilities generated by the language model rather than explicit reasoning, and for this reason was able to use very small language models compared to GPT4 as was used in most other studies. Because language models work by assigning a probability of each word, multiplying the probabilities of all the words multiplied in the entire abstract allows researchers to compare the multiplied probability of the real abstract to the multiplied probability of the fabricated abstract directly, without having the language model generating any text involving reasoning or chain of thought.

Another study found a similar result with regards to publications in the domain of AI algorithms, finding their system beating human experts in predicting the ability of an AI algorithm to improve on the state of the art performance in AI models (Wen et al., 2025 | *Predicting Empirical AI Research Outcomes with Language Models*). In this domain, the researchers use a sophisticated framework with RAG and fine-tuning, and still find a good result.

While much cheaper than prediction markets or IAMs, LLMs are also more expense than simpler ML models. When attempting to forecast whether results and effect sizes replicate in social sciences, simple neural network classifiers trained on millions of scientific abstracts and hundreds of full texts, the unordered semantic vectors of the words in the abstracts of the papers, combined with statistical were sufficient to approach prediction market level accuracy of approximately 70% accuracy in predicting which paper results would replicate, despite lacking fundamental logical relationships between words in the text or any deeper language comprehension of the methods of the abstracts (Yang, Youyou, and Uzzi, 2020 | "Estimating the Deep Replicability of Scientific Findings Using Human and Artificial Intelligence"). This finding mirrors that of the neuroscience study (X et al., 2025 | "Large Language Models Surpass Human Experts in Predicting Neuroscience Results") which finds that reasoning ability is not strictly to predict the outcomes in neuroscience abstracts.

Insofar as identifying whether results from social science papers will replicate is a similar task as forecasting the impact of an intervention in the earth system sciences, we can be encouraged that statistical and categorical aspects of the interventions should be sufficient to identify the likely success of real-world outcomes, and remain skeptical that LLMs are strictly necessary to rival humans at predicting categorical outcomes, where ML may be sufficient. However, ML models certainly have a lower upper bound in potential accuracy than a full reasoning model, and computational resources are not so restricted that LLMs could not be used. Furthermore, ML models using simple semantic vectors cannot produce free-form predictions of outcomes like LLMs, limiting the flexibility of their application in real-world use-cases.

If LLMs are able to approach or surpass human ability in predicting unpublished results in complex domains of predicting which techniques in improving state of the art AI system performance, predicting the outcomes of neuroscience papers, predicting social science replicability, or predicting geopolitical events such as election results, then it stands to reason that they may be able to predict the outcomes of interventions in the earth system sciences. While geopolitical forecasting may not be amenable to scientific techniques, neuroscience and AI algorithm improvements certainly are - yet LLMs still beat human experts in these domains. Furthermore, LLM systems are far simpler to use, and far less costly to run and maintain than IAMs, CGEs, or ABMs, while having the benefit of human-interpretable reasoning and the ability to be extremely flexible as to their domain of application. Finally, given their low cost to use, LLMs can often be used as starting points or augmentation to expert judgement in ex-ante outcome prediction, rather than being the sole source of judgement about expected intervention outcomes, and the collaboration often bests forecast accuracy when compared to expert forecasts or LLM forecasts alone (Schoenegger, Park, et al., 2025 | "AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy")(Schoenegger, Jones, et al., 2025 | *Prompt Engineering Large Language Models' Forecasting Capabilities*).

Limitations The use of LLMs to inform decision making for promising interventions in the earth system comes with several downsides. Notably, LLMs do not reason like humans, and are prone to "hallucinations" where facts are fabricated. These hallucinations can be either factual fabrications attributed to external source material, or false statements which come intrinsically from the model (Huang et al., 2025 | "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions"). For the purposes of probabilistic reasoning, LLM's are not typically skilled at ensuring probabilities sum to 100%, or related quantitative skilled, even after fine-tuning on the task of probability predictions (Lyu et al., 2025 | "Calibrating Large Language Models with Sample Consistency"). As mentioned previously, LLMs are more computationally costly than other machine learning methods. There are also issues (which we will leave for the Conclusion & Outlook section) with overly trusting LLMs, false beliefs from users of LLMs that they are less biased than humans or not biased at all, and issues with AI safety, if LLMs begin to replace or distort, rather than augment, human decision making.

Furthermore, the majority of work thus far has focused on classification, or fixed categories. At best, assigning a numerical score to a list of fixed objectives (Bina et al., 2025 | "On Large Language Models as Data Sources for Policy Deliberation on Climate Change and Sustainability"). Open-ended future event prediction will be increasingly necessary for specific event prediction which cannot be easily quantified into a series of rankings or clear outcome categories. Some of the most important outcomes of interventions are the unexpected effects and learnings from the work, which cannot be captured by rigid outcome category schemes. Past work has used LLMs such as GPT4 to evaluate free-form event prediction on Accuracy, Completeness, Relevance (how pertinent the prediction is to the actual outcomes), Specificity (not overly broad nor vague), and Reasonableness (logical coherence and believability of the prediction) (Guan et al., 2024 | *OpenEP: Open-Ended Future Event Prediction*). However, the work finds that accurately predicting future events in open-ended settings is challenging for existing LLMs, as predictions are often incomplete, underspecified, irrelevant, or illogical.

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

Data Filtering In order to ensure sufficient numbers of interventions are collected, openalex was used to collect published works (Priem, Piwowar, and Orr, 2022 | *OpenAlex:*

A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts). openalex used LLM categorization to sort the papers into thousands of "topics" (, | *OpenAlex Topic Classification Whitepaper.Docx*). Regex filtering on the abstract and restricting the publication year to 2021 or later was first used to identify papers which were likely to contain relevant interventions in the abstracts (See Figure XY).

Figure XY: The keyword query for openalex ("evaluate the program" OR "causal impact" OR "empirical result" OR "case study" OR "life cycle" OR "evaluate the effectiveness" OR "insights from" OR "first report" OR "quasi-experimental" OR "estimate the impact" OR "to control for" OR "impact evaluation" OR "adoption of" OR "percent of" OR "project was" OR "the project" OR "efficacy of" OR "the beneficiaries" OR "electricity use" OR "this evaluation" OR "of improved" OR "randomly selected" OR "intervention" OR "highest values" OR "achieved the" OR "gave the" OR "conducted" OR "RCBD" OR "RCT" OR "randomiz" OR "randomis" OR "difference-in-difference" OR "score matching" OR "triple difference" OR "event study" OR "before-after design" OR "pre-post design" OR "two-stage least squares" OR "2SLS" OR "propensity score" OR "inverse probability weighting" OR "matching estimator" OR "kernel matching" OR "coarsened exact matching" OR "Mahalanobis matching" OR "entropy balancing" OR "doubly robust estimation" OR "RDD" OR "RKD" OR "interrupted time series" OR "interrupted time-series analysis" OR "synthetic control") AND ("law" OR "ban" OR "tax" OR "subsid" OR "policies" OR "policy" OR "regulation" OR "regulatory" OR "instrument" OR "levy" OR "grant" OR "tariff" OR "mandate" OR "ordinance" OR "directive" OR "beneficiary" OR "statute" OR "campaign" OR "rollout" OR "legislation" OR "pilot" or "treatment" OR "program" OR "programme") NOT ("simulation" OR "2018" OR "2017" OR "2016" OR "2015" OR "2014" OR "2012" OR "2011" OR "2010" OR "2009" OR "2008" OR "2007" OR "2006" OR "2005" OR "2004" OR "2003" OR "2002" OR "2001" OR "2000" OR "1999" OR "1998" OR "1997" OR "1996" OR "1995" OR "1994" OR "1993" OR "1992" OR "1991" OR "1990" OR "1985" OR "1980" OR "1975" OR "1970" OR "1960")

After a preliminary filtering of approximately 600 topics within the earth system sciences, 135 of these topics were chosen as containing a high proportion of papers analyzing outcomes of interventions in the earth system, which do not reference interventions before the model cutoff date.

Next, GPT o4-mini was used to categorize the topics (see Figure 1).

After these filtering stages, 1741 abstracts remained.

Outcome and Intervention Classification GPT-o4-mini was used to extract descriptions of interventions from abstracts (See Figure XZ (a)). In addition, a series of outcome categories were defined to allow intercomparison between the intervention and the result (See Figure XZ (b)).

After a preliminary filtering of approximately 600 topics within the earth system sciences, *Below is the title and abstract of a scientific article. You will score the degree to which the abstract of the paper below reports the result(s) of some program(s) or policy/policies, conditional on the following criteria: 1. The abstract must provide quantitative or qualitative information informing decision makers, or other members of government civil society the impact of the program or policy. 2. The abstract cannot only describe theoretical work to simply improve models or scientific understanding, it must evaluate a program or policy. 3. The abstract below must provide at least one qualitative or quantitative statement regarding the extent to which the specific program or policy achieved relevant outcomes, including the overall success of the policy. On the first line of your response, provide a best guess for the four digit year the evaluated program(s) or policy/policies took effect. Use background knowledge or context clues if no date is mentioned. You may assume surveys and interventions implemented by the authors occurred two years prior to publication, if the date is not mentioned. If there is no possibility to estimate the timeframe, return N/A. On the second line, provide a score between 1 and 10 for the degree to which this abstract fits the criteria, where 1 is no fit, and 10 is a perfect fit for the criteria.*

Figure 1: The GPT4o-mini query for scoring the degree of fit of an abstract.

```
Figure XY: The keyword query for openalex ("evaluate the program" OR "causal
impact" OR "empirical result" OR "case study" OR "life cycle" OR "evaluate the
effectiveness" OR "insights from" OR "first report" OR "quasi-experimental" OR
"estimate the impact" OR "to control for" OR "impact evaluation" OR "adoption
of" OR "percent of" OR "project was" OR "the project" OR "efficacy of" OR "the
beneficiaries" OR "electricity use" OR "this evaluation" OR "of improved" OR
"randomly selected" OR "intervention" OR "highest values" OR "achieved the"
OR "gave the" OR "conducted" OR "RCBD" OR "RCT" OR "randomiz" OR "randomis" OR
"difference-in-difference" OR "score matching" OR "triple difference" OR "event
study" OR "before-after design" OR "pre-post design" OR "two-stage least squares"
OR "2SLS" OR "propensity score" OR "inverse probability weighting" OR "matching
estimator" OR "kernel matching" OR "coarsened exact matching" OR "Mahalanobis
matching" OR "entropy balancing" OR "doubly robust estimation" OR "RDD" OR "RKD"
OR "interrupted time series" OR "interrupted time-series analysis" OR "synthetic
control") AND ("law" OR "ban" OR "tax" OR "subsid" OR "policies" OR "policy" OR
"regulation" OR "regulatory" OR "instrument" OR "levy" OR "grant" OR "tariff"
OR "mandate" OR "ordinance" OR "directive" OR "beneficiary" OR "statute" OR
"campaign" OR "rollout" OR "legislation" OR "pilot" or "treatment" OR "program"
OR "programme") NOT ("simulation" OR "2018" OR "2017" OR "2016" OR "2015" OR "2014"
OR "2012" OR "2011" OR "2010" OR "2009" OR "2008" OR "2007" OR "2006" OR "2005"
OR "2004" OR "2003" OR "2002" OR "2001" OR "2000" OR "1999" OR "1998" OR "1997"
OR "1996" OR "1995" OR "1994" OR "1993" OR "1992" OR "1991" OR "1990" OR "1985" OR
"1980" OR "1975" OR "1970" OR "1960")
```

Figure 2: Keyword query used for the OpenAlex search to identify policy- and program-evaluation abstracts.

135 of these topics were chosen as containing a high proportion of papers analyzing outcomes of interventions in the earth system, which do not reference interventions before the model cutoff date.

Next, GPT o4-mini was used to categorize the topics (see Figure 3).

```
Figure XX: The gpt o4-mini query Below is the title and abstract of a scientific
article.  You will score the degree to which the abstract of the paper below
reports the result(s) of some program(s) or policy/policies, conditional on the
following criteria:  1.  The abstract must provide quantitative or qualitative
information informing decision makers, or other members of government civil
society the impact of the program or policy.  2.  The abstract cannot only describe
theoretical work to simply improve models or scientific understanding, it must
evaluate a program or policy.  3.  The abstract below must provide at least one
qualitative or quantitative statement regarding the extent to which the specific
program or policy achieved relevant outcomes, including the overall success of
the policy.  On the first line of your response, provide a best guess for the
four digit year the evaluated program(s) or policy/policies took effect.  Use
background knowledge or context clues if no date is mentioned.  You may assume
surveys and interventions implemented by the authors occurred two years prior to
publication, if the date is not mentioned.  If there is no possibility to estimate
the timeframe, return N/A. On the second line, provide a score between 1 and 10 for
the degree to which this abstract fits the criteria, where 1 is no fit, and 10 is a
perfect fit for the criteria.
```

Figure 3: Scoring prompt given to GPT-o4-mini for identifying abstracts that evaluate programs or policies and report outcomes.

After these filtering stages, 1741 abstracts remained.

Outcome and Intervention Classification GPT-o4-mini was used to extract descriptions of interventions from abstracts (See Figure 4 (a)). In addition, a series of outcome categories were defined to allow intercomparison between the intervention and the result (See Figure 4 (b)).

Outcome Grading Next, each outcome was evaluated in multiple ways.

First, a grading scheme was identified as a useful taxonomy, which could allow easy comparison between the predictions and the test set:

1. **Very significant**: Substantial improvement with robust evidence
2. **Significant**: Noticeable improvement with moderate evidence
3. **Neutral/mixed results**: Some improvement but limited or unclear
4. **No effect**: No discernible impact
5. **Outcome was worsened**: Negative impact

```
(a) What is the intervention that is
described in the abstract?  This is
an abstract for an impact evaluation
report about an intervention in a
developing country.  Do not mention
the outcomes or analysis method, only
describe the intervention with as much
detail as is present in the abstract.
Ensure to include any contextual
information about what was done and
where in your response.  If nothing is
said about the intervention write:  No
Intervention Described.
```

(a) Intervention extraction prompt.

```
(b) What does the abstract say
regarding the

            outcome_category

outcome?  Be sure to include
relevant quantitative or categorical
information where present.  If nothing
is said about the outcome write:  No
Information.
```

(b) Outcome extraction prompt.

Figure 4: Prompts used to extract (a) intervention descriptions and (b) outcome statements from abstracts.

Grading for free-form prediction was also allowed, whereby o4-mini was used to directly compare a free-form prediction of the outcome, to the outcome described in the abstract.

Outcome Forecasting Next, the outcome itself was forecasted. At this stage, we used the forecast techniques (scratchpad, RAG, fine-tuning, ensembling) to produce a single most accurage prediction.

Forecasting context was restricted to RAG context obtained, the GPT-generated intervention description, and the name of the outcome metric.

For each forecast, a qualitative free-form forecast is generated. Subsequently, a grade on the 5-point scale is also generated.

Forecast Evaluation

The last stage of forecasting was the evaluation of the forecast itself. I compare forecasted grades against actual outcome grades using the Root Mean Square Error (RMSE), Accuracy (fraction of exact match), Macro-F1 score, and the Brier score (binary classification where "positive" ≥ 0.75)

The forecast is also compared against two baselines: - Most-common grade baseline - Most-similar abstract and outcome (using vector similarity search)

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 AI Scientist-like idea

Extending the system for searching for high-impact policies is possible, rather than simply using the fine-tuned model for forecasting. While use of reasoning models outside of the domain in which they are trained for often reduces their performance, it still may be possible to re-train the model for these use cases. For instance, the model could be prompted to generate many policy options for a given country to reduce CO2 emissions, and each idea could have the emissions reduction forecasted. Seeding the model with many similar policies and suggesting that it think of a wide range of options may allow for consideration of a wide range of policy options. Next, only the ideas which are forecasted to have high emissions would be suggested to the user of the system. Such a system would be similar to the "AI Scientist" released by Google which iteratively generates new hypotheses and reasons over the hypotheses to discover better scientific theories behind biological phenomena (C. Lu et al., 2024 | *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery*).

4.7 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

Abolghasemi, Mahdi, Odkhishig Ganbold, and Kristian Rotaru (Apr. 2025). “Humans vs. Large Language Models: Judgmental Forecasting in an Era of Advanced AI”. In: *International Journal of Forecasting* 41.2. “Schoenegger and Park (2023) and Abolghasemi et al. (2023) evaluated GPT-4 and other LLMs on forecasting tournaments and found that they underperform the human crowd. This observation is in line with ours in Section 3.3. Unlike us, they make little efforts to improve these LMs on forecasting.” (Halawi et al., 2024, p. 3), pp. 631–648. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2024.07.003. (Visited on 08/19/2025).

Bina, Rachel et al. (Feb. 2025). “On Large Language Models as Data Sources for Policy Deliberation on Climate Change and Sustainability”. In: Hence, we conclude that GPT-4 can be used as a credible input, even starting point, for subsequent deliberation processes on climate and sustainability policies.
multiple criteria decision making
(MCDM) models for comparative assessment of climate and sustainability policies
in my work I extend past the purely non-technical to include technical assessment
Other as-
pects, particularly those focused on how a policy could impact well-being and quality of life in
the community, perhaps should not be assessed by policy makers except through consultation with
representatives of the general public. These well-being aspects of climate and sustainability policies
are our main, but not exclusive, concern in this study. WHAT KIND OF POLICY:
The third kind of justification occurs when the policy in question fails or comes out uncertain
on a strict cost–benefit basis but is judged to have sufficient co-benefits to overcome a cost–benefit
shortcoming (Boyd et al., 2022; Creutzig et al., 2022; Dagnachew and Hof, 2022; Finn and Brockway,
2023; Karlsson et al., 2020; Sharifi, 2021). An example of such a policy might be a ban on single-use
plastic bags (yielding co-benefits of reduced litter in the streets and reduced landfill material) or a
ban on gasoline-powered leaf blowers (yielding co-benefits of cleaner air, reduced ecological damage,
and reduced noise).

The present study centers upon this third kind of co-benefit-based warrant for climate and
sustainability policies, and within it on factors that contribute to quality-of-life (alias well-being).
Certain aspects of climate and sustainability plans, such as costs, emissions, degree of protection etc., are technical by nature and best evaluated by climate scientists, city planners, engineers,
and area experts in general, or obtained through careful assessment of the literature. Other aspects, particularly those focused on how a policy could impact well-being and quality of life in
the community, perhaps should not be assessed by policy makers except through consultation with
representatives of the general public. These well-being aspects of climate and sustainability policies
are our main, but not exclusive, concern in this study.
Our study presumes that climate and sustainability policies are usefully—or even necessarily—
compared across multiple evaluation criteria and that these criteria are inevitably in conflict, even
limiting our attention to quality-of-life and well-being criteria. No policy is always and everywhere
the best, and trade-offs are inevitable STRATEGY:
Broke into ~10 categories of policy assessment
e.g.
Moral considerations
Economy
Improving and creating destinations
etc
and scored from 1-10 on all. DOI: `10.2139/ssrn.5123359`. (Visited on 08/18/2025).

Brown, Tom et al. (2020). “Language Models Are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. show that language models improve as they scale -> they will keep getting better ”Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches.”
Fine-Tuning (FT) - updates the weights of a pre-trained model by training on thousands of supervised labels specific to the desired task. The main advantage of fine-tuning is strong performance on many benchmarks. The main disadvantages are the need for a new large dataset for every task, the potential for poor generalization out-of-

distribution [MPL19], and the potential to exploit spurious features of the training data [GSL+18 , NK19]. We focus on task-agnostic performance, leaving fine-tuning for future work. Curran Associates, Inc., pp. 1877–1901.

Chen, Yaoyu, Yuheng Hu, and Yingda Lu (May 2025). *Predicting Field Experiments with Large Language Models.* For example, our framework achieves nearly 100% prediction accuracy on 71% of conclusions while it completely fails to
arXiv:2504.01167v3 [cs.CY] 21 May 2025
SLW, August 6th, 2025, Toronto
Yaoyu Chen1 Yuheng Hu1 Yingda Lu1
1The Department of Information and Decision Sciences, College of Business Administration, University of Illinois at Chicago
{ychen563, yuhenghu, yingdalu}@uic.com
predict 18% of the conclusions with close to 0% accuracy. DOI: `10.48550/arXiv.2504.01167`. arXiv: `2504.01167 [cs]`. (Visited on 08/19/2025).

"Cost-Effective Control of Air Quality and Greenhouse Gases in Europe" (Dec. 2011). "Cost-Effective Control of Air Quality and Greenhouse Gases in Europe: Modeling and Policy Applications". In: *Environmental Modelling & Software* 26.12, pp. 1489–1501. ISSN: 1364-8152. DOI: `10.1016/j.envsoft.2011.07.012`. (Visited on 08/24/2025).

Dueri, Sibylle and Gabriele Mack (June 2024). "Modeling the Implications of Policy Reforms on Pesticide Risk for Switzerland". In: *The Science of the Total Environment* 928, p. 172436. ISSN: 1879-1026. DOI: `10.1016/j.scitotenv.2024.172436`.

Fuller, Richard et al. (June 2022). "Pollution and Health: A Progress Update". In: *The Lancet Planetary Health* 6.6, e535–e547. ISSN: 2542-5196. DOI: `10.1016/S2542-5196(22)00090-0`. (Visited on 08/24/2025).

Guan, Yong et al. (Aug. 2024). *OpenEP: Open-Ended Future Event Prediction.* Experiment results indicate that accurately predicting future events in open-ended settings is challenging for existing LLMs. DOI: `10.48550/arXiv.2408.06578`. arXiv: `2408.06578 [cs]`. (Visited on 08/18/2025).

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).

Hanson, Robin (2013). "Shall We Vote on Values, But Bet on Beliefs?" In: *Journal of Political Philosophy* 21.2. Futarchy, pp. 151–178. ISSN: 1467-9760. DOI: `10.1111/jopp.12008`. (Visited on 08/19/2025).

Hewitt, Luke et al. (n.d.). "Predicting Results of Social Science Experiments Using Large Language Models". In: (). "LLM-derived predictions remained highly accurate for studies that could not have been in the LLM training data given they were not published prior to the LLM training data cutoff date." (Hewitt et al., p. 7).

Huang, Lei et al. (Mar. 2025). "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Transactions on*

Information Systems 43.2, pp. 1–55. ISSN: 1046-8188, 1558-2868. DOI: 10.1145/3703155. (Visited on 08/20/2025).

"Mitigation and Development Pathways in the Near to Mid-term" (Aug. 2023). In: *Climate Change 2022 - Mitigation of Climate Change*. Ed. by Intergovernmental Panel On Climate Change (Ipcc). 1st ed. Cambridge University Press, pp. 409–502. ISBN: 978-1-009-15792-6. DOI: 10.1017/9781009157926.006. (Visited on 08/24/2025).

Jumper, John et al. (Aug. 2021). "Highly Accurate Protein Structure Prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. (Visited on 08/24/2025).

Koldunov, Nikolay and Thomas Jung (Jan. 2024). "Local Climate Services for All, Courtesy of Large Language Models". In: *Communications Earth & Environment* 5.1, p. 13. ISSN: 2662-4435. DOI: 10.1038/s43247-023-01199-1. (Visited on 08/24/2025).

Lam, Remi et al. (Dec. 2023). "Learning Skillful Medium-Range Global Weather Forecasting". In: *Science*. DOI: 10.1126/science.adi2336. (Visited on 08/24/2025).

Lu, Chris et al. (Sept. 2024). *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery*. DOI: 10.48550/arXiv.2408.06292. arXiv: 2408.06292 [cs]. (Visited on 08/19/2025).

Lyu, Qing et al. (Apr. 2025). "Calibrating Large Language Models with Sample Consistency". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 39.18, pp. 19260–19268. ISSN: 2374-3468. DOI: 10.1609/aaai.v39i18.34120. (Visited on 08/24/2025).

Mellers, Barbara et al. (2015). "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions". In: *Perspectives on Psychological Science* 10.3, pp. 267–281. ISSN: 1745-6924. DOI: 10.1177/1745691615577794.

OpenAlex Topic Classification Whitepaper.Docx (2025). https://docs.google.com/document/d/1bDopkhu

(Visited on 08/19/2025).

"OSeMOSYS" (Oct. 2011). "OSeMOSYS: The Open Source Energy Modeling System: An Introduction to Its Ethos, Structure and Development". In: *Energy Policy* 39.10, pp. 5850–5870. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2011.06.033. (Visited on 08/24/2025).

Priem, Jason, Heather Piwowar, and Richard Orr (June 2022). *OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts*. Comment: Submitted to the 26th International Conference on Science, Technology and Innovation Indicators (STI 2022). DOI: 10.48550/arXiv.2205.01833. arXiv: 2205.01833 [cs]. (Visited on 08/19/2025).

Schoenegger, Philipp, Cameron R. Jones, et al. (June 2025). *Prompt Engineering Large Language Models' Forecasting Capabilities*. DOI: 10.48550/arXiv.2506.01578. arXiv: 2506.01578 [cs]. (Visited on 08/21/2025).

Schoenegger, Philipp and Peter S. Park (Oct. 2023). *Large Language Model Prediction Capabilities: Evidence from a Real-World Forecasting Tournament*. "Schoenegger and

Park (2023) and Abolghasemi et al. (2023) evaluated GPT-4 and other LLMs on forecasting tournaments and found that they underperform the human crowd. This observation is in line with ours in Section 3.3. Unlike us, they make little efforts to improve these LMs on forecasting." (Halawi et al., 2024, p. 3) Comment: 13 pages, six visualizations (four figures, two tables). DOI: 10.48550/arXiv.2310.13014. arXiv: 2310.13014 [cs]. (Visited on 08/19/2025).

Schoenegger, Philipp, Peter S. Park, et al. (Mar. 2025). "AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy". In: *ACM Transactions on Interactive Intelligent Systems* 15.1, pp. 1–25. ISSN: 2160-6455, 2160-6463. DOI: 10.1145/3707649. (Visited on 08/21/2025).

Silvestro, Daniele et al. (May 2022). "Improving Biodiversity Protection through Artificial Intelligence". In: *Nature Sustainability* 5.5, pp. 415–424. ISSN: 2398-9629. DOI: 10.1038/s41893-022-00851-6. (Visited on 08/24/2025).

Stechemesser, Annika et al. (Aug. 2024). "Climate Policies That Achieved Major Emission Reductions: Global Evidence from Two Decades". In: *Science (New York, N.Y.)* 385.6711, pp. 884–892. ISSN: 1095-9203. DOI: 10.1126/science.adl6547.

Tetlock, Philip E. and Dan Gardner (2015). *Superforecasting: The Art and Science of Prediction.* Superforecasting: The Art and Science of Prediction. New York, NY, US: Crown Publishers/Random House, p. 340. ISBN: 978-0-8041-3669-3 978-0-8041-3670-9.

The MIT Emissions Prediction and Policy Analysis (EPPA) Model (2025). *The MIT Emissions Prediction and Policy Analysis (EPPA) Model: Version 4 | MIT CS3.* https://cs3.mit.edu/publication/14578. "Existing efforts I could identify focus on manually entering assumptions of the effects of policies into models of the world economy, such as the The MIT Emissions Prediction and Policy Analysis (EPPA) Model "

EPPA is a recursive-dynamic multi-regional general equilibrium model of the world economy, which is built on the GTAP dataset and additional data for the greenhouse gas and urban gas emissions. (Visited on 08/19/2025).

Watson, Robert T et al. (2019). *The Global Assessment Report on BIODIVERSITY AND ECOSYSTEM SERVICES.* Tech. rep. Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES).

Wen, Jiaxin et al. (June 2025). *Predicting Empirical AI Research Outcomes with Language Models.* from devon, shows empirical results can be predicted. DOI: 10.48550/arXiv.2506.00794. arXiv: 2506.00794 [cs]. (Visited on 08/18/2025).

Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy | Science Advances (2025). https://www.science.org/doi/10.1126/sciadv.adp1528. (Visited on 08/21/2025).

X, Luo et al. (Feb. 2025). "Large Language Models Surpass Human Experts in Predicting Neuroscience Results". In: *Nature human behaviour* 9.2. braingpt

"LLM training data memorization analysis. One concern regarding LLMs outperforming human experts on BrainBench is the possibility that LLMs were exposed to the original abstracts during their pre-training. If LLMs have simply memorized the training data, they would naturally assign lower perplexity scores to the correct abstracts. To address this concern, we employed a common method from the literature to determine whether a given text is part of LLM's training data22,33. This method involves calculating the zlib entropy and the perplexity ratio (equation (3)) of a text sequence to infer its membership status:" (X et al., 2025, p. 312). ISSN: 2397-3374. DOI: 10.1038/s41562-024-02046-9. (Visited on 08/18/2025).

Yan, Q., Seraj, R., He, J., Meng, L., and Sylvain, T. (2024). "AutoCast++: Enhancing World Event Prediction with Zero-shot Ranking-based Context Retrieval". In: *International Conference on Learning Representations (ICLR)*. "In a competition with a large prize pool, no machine learning system was able to approach the performance of human forecasters on Autocast (Zou et al., 2022). The knowledge cut-offs of the latest LMs have moved past 2022, necessitating more recent data. In this work, we source questions in 2023–2024, enabling us to apply recent LMs. Yan et al. (2024) built a retrieval system that led to improved accuracy on Autocast. They trained a Fusion-in-Decoder model to directly predict the final (binary) resolution" (Halawi et al., 2024, p. 2). (Visited on 08/19/2025).

Yang, Yang, Wu Youyou, and Brian Uzzi (May 2020). "Estimating the Deep Replicability of Scientific Findings Using Human and Artificial Intelligence". In: *Proceedings of the National Academy of Sciences* 117.20, pp. 10762–10768. DOI: 10.1073/pnas.1909046117. (Visited on 08/24/2025).

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, 27 August 2025