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 Earth System Science Intervention Outcomes with Language Models

Potsdam, 25 September 2025

Submitted by:
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-tune three LLMs (Llama 70B, ChatGPT-3.5, [SOME MODERN LLM I CHOOSE]) 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 / [SOME MODERN LLM I CHOOSE], averaging X parallel predictions) can forecast XX% of interventions correctly compared to YY% using the most-similar-abstract baseline. Through the use of conformal prediction, I show that QQ% of outcome prediction forecasts can be made using the system with >90% confidence in a true positive outcome prediction. This work has wide-ranging applications both to improve sustainability forecasting, as well as in adjacent areas and improving decision making in policy contexts. I also release the largest extant structured database of RR thousand interventions and associated outcomes in the context of Earth system sciences.

Abstract auf Deutsch

Will do, once abstract is finalized

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 the best prior computational approaches in several complex-system 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!]. Despite many examples of other computer models which have some success (see section XXYY), 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 predict 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 expert forecasters 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 when applied to the Earth-human system encompassing sustainability, the environment, clean energy production, pollution, and other human $\leftrightarrow$ nature interaction contexts ("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 use language models to mimic the reasoning and data gathering skills of trained forecasters, in an attempt to replicate the success at using judgemental forecasting from language models in geopolitical forecasting to the adjacent domain of forecasting Earth system science intervention outcomes. Ultimately, the goal is to learn whether it is possible to complement a scientifically founded prediction for the effectiveness of a given intervention with a system with LLMs that are specifically

trained for the task at their core. Given the difficulty of field testing ideas, policymakers and funding agencies 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 event outcome prediction in the context of 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 context of Earth system sciences while also rigorously quantifying the skill of such a system.

In the process of training the LLM, it was necessary to collect and label a large volume of interventions and associated outcomes along a wide range of metrics in the context of Earth system sciences. Accordingly, in tandem with the open source LLM forecasting system, I also release the largest extant structured database of interventions and associated outcomes in the context of Earth system sciences. The database contains intervention descriptions, quantitative and qualitative outcomes identified with each intervention, and further statistical information about intervention categories and other statistical trends described in Section 3.1.

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 (GPT-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 GPT-4'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.

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. Prediction accuracy is typically above-chance hundreds of days before resolution and steadily improves as deadlines approach. Predictions are typically above-chance within approximately one year time horizon, with the accuracy notably improving as the event reaches question resolution: one study finds a Brier score of approximately 0.2-0.3 for geopolitical and economic questions within about 3 months before resolution using a large constructed prediction market, dropping close to a Brier score of about 0.75 within a day or two of the question resolution (Tetlock and Gardner, 2015 | *Superforecasting: The Art and Science of Prediction*). 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"). One source shows superforecasters saw the correct outcome with a 60% probability approximately 300 days out (significantly earlier than prediction markets), and 75% probability 250 days ahead of the outcome (Tetlock and Gardner, 2015 | *Superforecasting: The Art and Science of Prediction*).

Due to the relatively high expense and human effort required to organize superforecasting tournaments (the gold standard for event prediction), they 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) (Tetlock and Gardner, 2015 | *Superforecasting: The Art and Science of Prediction*).

1.3 LLM Forecasting of Outcomes in the Earth System Sciences

1.3.1 Other Computer Modelling Methods

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), Greenhouse Gas and Air Pollution Interactions and Synergies (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").

1.3.2 Methods and Capabilities

Implementing the gold standard prediction method - superforecaster tournaments - to predict the efficacy of interventions such as new environmental laws in low and middle income countries (LMIC), specific interventions such as introduction of cleaner burning ovens, or regulations on air quality would be worthwhile, but also costly and logistically challenging given the very large number of annual interventions over wide geographic regions. Even if such a tournament were to be ran, ML methods to estimate the outcomes could be complementary and increase the accuracy for such a tournament. This work focuses on the mimicking of techniques known to be effective for tournaments of superforecasters with LLMs, both to aid expert forecasters and grantmakers, and to provide direct, useful predictions for those without access to expert knowledge. While there has been no attempt at predicting real-world outcomes of interventions in the context of Earth system sciences while also rigorously quantifying the skill of such a system, much encouraging progress has been made in closely adjacent domains which I will survey below.

If using LLMs to directly output probabilities or yes/no answers to forecasting questions, the base models appear to underperform compared to crowds of humans (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*). In such a context, 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 assemblages 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").

In a recent study, a RAG+fine-tuned LLM system was sufficiently more skilled than the human crowd to reliably earn a profit on Polymarket event predictions (Turtel, Franklin, and Schoenegger, 2025 | *LLMs Can Teach Themselves to Better Predict the Future*),

providing a real-world example of the prediction skill of such systems against humans.

Despite these findings, it has been argued that utilization of the direct probabilities in complex domains may be more accurate, if the prediction is a function of "many noisy intertwined signals across subfields", in which case methods such as CoT may reduce the power of "intuition" available to the model (X et al., 2025 | "Large Language Models Surpass Human Experts in Predicting Neuroscience Results").

In general however, 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"). Along these lines, several improvements to the base-level prediction ability can be applied to approach superforecasting level calibration and accuracy. These include:

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
3. Have the LLM split questions into sub-questions before being used to query RAG system
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 ensembling the final predictions ("Wisdom of the crowd")

In one similar work, the technique of Chain of Thought (CoT) has been used to improve the reasoning abilities of GPT-4 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*).

In terms of social intervention outcome prediction, another study separately analyzed 346 treatment effects estimated from the responses of over one million participants, with hundreds of ex-ante predictions made from experts before the outcomes were known (Hewitt et al., | "Predicting Results of Social Science Experiments Using Large Language

Models"). The study adopted a bottom-up technique of simulating how individual respondents would respond to surveys and field experiments using GPT-4 according to their demographic profiles, specifically mimicking demographic profiles in the USA. The interventions included surveys that simulated the effect of informational content which promoted pro-democratic attitudes, encouraged respondents to increase beneficial choices with respect to climate change, and increase their vaccination rates. Notably GPT-4 matched or exceeded expert prediction accuracy in this domain. Interestingly, GPT-4 predictions were more accurate for survey experiments than field experiments (79% vs 64% accurate respectively).

Another recent study found that LLMs can correctly predict outcomes in scientific domains such as predicting results of papers in neuroscience (X et al., 2025 | "Large Language Models Surpass Human Experts in Predicting Neuroscience Results"). 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 GPT-4 as was used in most other studies. Because language models work by assigning a probability of each token (typically some commonly occurring part of a 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 generate any text involving reasoning or CoT.

This capability could be related to the surprising ability of language models to perform direct time-series even in zero-shot settings. The findings relate to a wide range of domains (energy, traffic, weather, retail, health), and show that RLHF reduces performance in such domains (Ghasemloo and Moradi, 2025 | *Informed Forecasting: Leveraging Auxiliary Knowledge to Boost LLM Performance on Time Series Forecasting*).

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.

Insofar as identifying whether results from social science papers will replicate is a similar task as forecasting the impact of an intervention in the context of 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, insofar as reasoning is required for forecasting in complex domains, non-reasoning ML models have a lower upper bound in potential accuracy than a full reasoning model, and regardless computational resources are not so restricted that LLMs could not be used in the context of Earth system sciences. 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.

A very different result was found in the context of ex-post impact evaluations of interventions in developing countries. One study determined that from a large collection of existing ex-post evaluations of outcomes of similar interventions, there is a very high variability in the effect sizes even from the same intervention (Vivalt, 2020 | "How Much Can We Generalize From Impact Evaluations?"). They find that a mixed model, rather than a random **TODO: A lot more to write here, once complete reading the study** whether an intervention improves or makes worse We may infer from (Vivalt, 2020 | "How Much Can We Generalize From Impact Evaluations?") that there is both the possibility that by taking into account contextual heterogeneity between interventions, prediction could be greatly improved compared to a statistical baseline, and the risk for LLMs that quantitative predictability is simply very low in general (because no other studies I found collected as many quantitative results and compared them). A similar result was found in the case study of education where "parameter heterogeneity" was found to be driven by economy- or institution-wide contextual factors, rather than specific characteristics of the intervention itself (Pritchett and Sandefur, 2013 | "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix - Working Paper 336").

TODO: rewrite in own words after reading study: "An inference about another study will have the correct sign about 61% of the time. If trying to predict the treatment effect of a similar study using only the mean treatment effect in an intervention-outcome combination, the median ratio of the MSE to that mean is 2.49 across intervention-outcome combinations. Only about 6% of total variance can be attributed to sampling variance. Modelling the variation with a mixed model can help a little, but not a lot....only about 6% of the observed variation in study results can be attributed to sampling variance. I find about 20% of the remaining variance could be explained using a single best-fitting explanatory variable. However, this statistic obscures a lot of heterogeneity, with the median decrease being about 10% among the intervention-outcomes for which this comparison was made. In a separate study by the same author, they a clear overall publication bias using the "caliper" test, but it remains somewhat smaller than other social sciences. The "caliper" test counts the number of papers nudging their results to be just over the 5% significance boundary for reporting a significant result. There is a clear statistical "nudging" going on, but in development, it appears to be primarily for very small shifts in results, and not particularly common when considering wider bands around the 5% threshold. RCT studies fared significantly better than non-RCT, with statistically lower bias (Vivalt, 2019 | "Specification Searching and Significance Inflation Across Time, Methods and Disciplines"). The biases in development programs were found to be much smaller than those previously observed in other social sciences (such as in (Gerber and Malhotra,

2008 | "Publication Bias in Empirical Sociological Research")). Another analysis finds: " "For ease of exposition we begin by comparing results that are insignificant at the 5 percent level to results that are significant at the 5 percent level. In the IV lit- erature, a result that is statistically insignificant is only 21.4 percent as likely to be published as a significant one. Said differently, a significant IV result is almost 5 times more likely to be published than an insignificant IV result. In the DID liter- ature, a statistically significant result is 4.2 times more likely to be published. For RCTs, a significant result is only 1.9 times more likely to be published. For RDDs, a significant result is 2.8 times more likely to be published. All of these estimates are statistically significant at the 1 percent level."

" We find that the ratio of tests just above and below 1.96 is only 1.10 in economics in comparison to over 2 for political science and sociology. This result provides strong evidence that the extent of p-hacking is much smaller in economics (at least when using these inference methods) than in other disciplines" (Brodeur, Cook, and Heyes, 2020 | "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics")

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, the impact of informational field campaigns, 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 context of 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 producing 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 has been found to produce a higher forecast accuracy than 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*).

1.3.3 Limitations

As might be expected given the absence of real-world experience and limited reasoning abilities of LLMs, simply replacing a crowd of humans with a crowd of untrained LLMs does not generally outperform the crowd average, especially where unpredictability and volatility of the question require strong reasoning abilities and good judgement to integrate relevant

information into forecasts (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*). Therefore, moderate-to-high complexity in the forecasting framework surrounding the LLM is required for a well-performing system, limiting this work's reproducibility and increasing software maintenance costs, and making it more difficult to produce useful forecasting systems.

It remains an open question whether forecasting systems can reproduce the success in other domains, with at least one study indicating forecasting in the context of Earth system sciences may be especially challenging. The study previously mentioned, with the bottom-up technique of simulating how individual respondents would respond to surveys and field experiments using GPT-4 according to their demographic profiles, 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 in their work, as compared to versus the approach of this thesis, which prompts the LLM to directly reason out the answer), this may hint that public policy and similar domains may be more difficult to predict than other scientific results.

The use of LLMs to inform decision making for outcome prediction in the context of Earth system sciences comes also 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, LLMs 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 ML 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 either 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 GPT-4 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.

While much cheaper than prediction markets or IAMs, LLMs are also more computationally expensive 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 explicit reasoning through CoT is not strictly required to predict the outcomes in neuroscience abstracts. It is an open question in the context of Earth system sciences whether LLMs are necessary, where maybe simpler ML techniques could be sufficient in many use-cases, although we leave it for future work.

There are also several limitations in extending best-performing or fine-tuned LLM forecasting systems to real-world use cases.

One issue is that the interventions in the literature are highly skewed toward a narrow range of topics, meaning the best performing system may succeed by being a specialist, rather than a generalist. For example, China has a much larger number of evaluations than other countries in the dataset, meaning that an LLM system may devote resources to becoming skilled at predicting Chinese development context, rather than development as a whole.

Model skill may not transfer when releasing a model into a real-world domain where the predicted outcome is truly in the future. Model cutoff dates are often not truly leakage free - some training, such as Reinforcement Learning from Human Feedback (RLHF) can introduce coarse details about events occurring after the model cutoff date. The system prompt (which cannot be directly inspected in closed-source LLMs such as GPT-3.5) can also contain unintended information leakage, and post-resolution documents in search results can further leak hints or the outcome itself(Paleka et al., 2025 | *Pitfalls*

in Evaluating Language Model Forecasters).

Even if there is no leakage, ranking forecasting skill using single scoring metrics can be misleading - each evaluation metric has its own issues (Paleka et al., 2025 | *Pitfalls in Evaluating Language Model Forecasters*) (See Table 1 and section 2.6). Therefore what may appear to be the best combination of accuracy-improving techniques and the best selection of base LLM may not in fact be the same outside of the test and validation sets. Language models themselves contain both political and stereotype biases which can bleed into both the rationales and the probabilities a system outputs (Nadeem, Bethke, and Reddy, 2021 | "StereoSet: Measuring Stereotypical Bias in Pretrained Language Models") (Bang et al., 2024 | "Measuring Political Bias in Large Language Models: What Is Said and How It Is Said").

Language models also don't always report their true reasoning - even if they reason something through scratchpads or CoT, the true reasons behind the answer may differ significantly. This can make using free-form reasoning for forecasts unreliable (Turpin et al., 2023 | "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting").

2 Methods for LLM Forecasting

2.1 Selecting LLMs for Forecasting in the Earth System Sciences

Multiple studies have measured zero-shot LLM forecasting capability against the base model performance, and found better general ability base models tend to perform better on forecasting tasks (Halawi et al., 2024 | "Approaching Human-Level Forecasting with Language Models") (Karger et al., 2024 | "ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities"): In one study with dozens of base models and a dynamically updating benchmark on prediction market forecasting questions, an inverse linear relationship was found between the human preference of a model's answer (in terms of an ELO score) and the Brier score, and similarly a log-linear inverse relationship between the compute used to train the model and the Brier score (Karger et al., 2024 | "ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities").

In order to guard against leakage of information from the training, we select ChatGPT-3.5 and Llama 70B (Touvron et al., 2023 | *Llama 2: Open Foundation and Fine-Tuned Chat Models*) as our base models due to their strong performance and early training cutoff date (approximately the beginning of 2022 for both models). NOTE: IN THE CASE THAT TRAINING DATA LEAKAGE IS NOT SIGNIFICANT We also run the YET TO DETERMINED - DEEPSEEK? OPENAI OSS? as an example of a stronger, more recently trained model in order to establish whether base model performance correlates with forecasting skill in the context of Earth system sciences.

Llama 70B is notable as it is a strong open source model with a relatively early training cutoff date of 2022, allowing us to inspect more directly the system prompt, the direct probabilities of sets of output tokens (the "logits"), and the degree of memorization of the training via use of the zlib entropy and the perplexity ratio (See Section 3.3 for more details).

2.2 Baseline Measures to Compare Against LLM Forecasts

Several baselines are required in order to justify as expensive and complex a scheme as fine-tuning an LLM for forecasting. We choose three simple baseline methods, in order to ensure our predictions are significantly better than the baseline methods of emission reduction predictions.

Prediction baseline: Same outcome as most similar policy This baseline technique provides a sanity check that more sophisticated methods are worthwhile. By selecting the most similar policy (defined YET TO BE DETERMINED), we can see if more complex techniques are being mislead by their usage of large amounts of less relevant data. While

other ML methods would be a useful baseline as well, training additional ML models is time-consuming, and the results are not as readily interpretable as other baselines. I leave the development of simpler machine learning methods for future work.

Prediction baseline: Zero-shot LLM In order to ensure the system is an improvement, to show that the methods used to improve accuracy are indeed increasing accuracy above simply a single generated prediction by a non-finetuned language model such as ChatGPT.

2.3 Data Sources

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*).

2.4 Data Filtering

After the OpenAlex topic filtering to restrict the results to XX thousand from the original X.X million abstracts within the selected OpenAlex topics Regular Expression (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 1).

After these filtering stages, XX thousand abstracts remained.

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 2).

After these filtering stages, 1741 abstracts remained with a score of 6 or higher from GPT-o4-mini, indicating they are high quality, good matches for the intervention-prediction split with appropriate years for the intervention.

GPT-o4-mini was used to extract descriptions of interventions from abstracts (See Figure 3 (a)). In addition, a series of outcome categories were defined to allow intercomparison between the intervention and the result (See Figure 3 (b)).

```
("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 1: Regex query searching the words within abstract used for the OpenAlex search to reduce policy- and program-evaluation abstracts which evaluated a post-2022 intervention. The query also filtered abstracts to those published after 2022. Further filtering using GPT-o4-mini was required to reduce false positive matches.

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:

- 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.
- The abstract cannot only describe theoretical work to simply improve models or scientific understanding, it must evaluate a program or policy.
- 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.

example_abstract

Figure 2: The GPT-o4-mini query for scoring the degree of fit of an abstract.

What is the intervention that is described in the abstract? This is an abstract for an impact evaluation report about an intervention in a LMIC. 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.

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

2.5 Techniques for Improving Forecasting Skill

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

We proceed to discuss how each technique for improving the composite forecasting skill metric was implemented.

Scratchpad More details will come, once I am sure exactly how I plan to implement the methods.

RAG More details will come, once I am sure exactly how I plan to implement the methods.

Ensembling More details will come, once I am sure exactly how I plan to implement the methods.

Fine-tuning The Llama 70B model was fine-tuned using past paper results and pairings collected before the model cutoff date. In past work, even though data were in the training data, fine-tuning significantly improved prediction performance (Wen et al., 2025 | *Predicting Empirical AI Research Outcomes with Language Models*).

NOTE: MORE DETAILS ON FINE-TUNING TO COME, ONCE ESTABLISH THAT I REALLY HAVE TIME TO DO THIS

2.6 Scoring Rules

We combine Brier score, calibration, and accuracy into a composite forecasting skill metric to attempt to mitigate the various issues in the individual metrics. In general, we only proceed with adding an accuracy boosting feature if it does not worsen any individual metric. We also utilize a held-out test set to ensure the validation metrics remain similarly performant in the final dataset. In event forecasting, scoring rules are typically used to quantify forecaster skills. A scoring rule is "proper" if the forecaster maximizes the expected score for an observation drawn from the distribution F if they issue the probabilistic forecast F, rather than G $\neq$ F (Gneiting and Raftery, 2007 | "Strictly Proper Scoring Rules, Prediction, and Estimation"). Brier score and log-loss are both strictly proper, but accuracy is not, limiting its comparability to other domains. However, accuracy has the advantage of intuitive simplicity.

Each form of evaluating the quality of forecasts has its own limitations.

Table 1: Caution when ranking forecasting systems using single metrics: each evaluation metric has its own issues (Paleka et al., 2025 | *Pitfalls in Evaluating Language Model Forecasters*).

Metric	**Method**	**Equation**	**Pitfalls when using for comparing forecasting skill**
Calibration	Consider all questions where the forecaster predicts a probability close to p; compare predicted and observed frequencies across bins.	$\Pr(Y = 1 \mid \hat{p} = p) \approx p$.	Can penalize useful forecasting; depends on binning; see text below.
Accuracy	Percent of correct classifications after thresholding probabilities (e.g., predict "Yes" if $p \geq 0.5$).	$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i = y_i]$.	Rises with class imbalance and fewer options; not comparable across outcomes with different base rates.
Brier score	Mean squared error between predicted probability and outcome (lower is better). A brier score is strictly proper. More outcome categories will raise the brier score (as the correct outcome is more difficult to predict)	$\text{Br} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$.	Overweights mid-probability discrimination relative to rare-event skill; see text below.
Logarithmic score	Strictly proper scoring rule for multi-category outcomes (often reported as negative log-loss).	$\ell(\mathbf{p}, y) = \log p_y$	Extremely sensitive to overconfident errors; undefined at $p_y = 0$ without clipping; scale depends on log base.

Calibration issues Calibration can penalize useful forecasting. For example, guessing a card suit and rank in a 52-card deck. A base-rate forecaster that predicts 1/52 for every card is perfectly calibrated. A more discerning forecaster assigns 10% to five "front-runners" and 0.5% to the remaining 47. If their ranking is genuinely informative so that the true card lies among the top five in 60% of rounds, the observed success rates are about 12% in the 10% bin and 0.8% in the 0.5% bin so calibration looks worse, yet this forecaster is far more useful.

Brier score issues Out of 120 "Good" or "Bad" outcomes, 60 are mid-probability with a 40% base rate (half 60% "Good" events and half are 20% "Good" events). The remaining 60 are rare, events with a 2% "Good" rate.

- Forecaster A is perfect on rare events but predicts 0.4 on all mid-prob questions, yielding an average Brier of 0.12.
- Forecaster B is baseline on rare events (predicts 0.02) but discriminates mid-prob questions (predicts 0.6 for 60% cases and 0.2 for 20% cases), yielding an average Brier of ≈ 0.1.

  Despite being useless on rare events, B is ranked better overall.

Logarithmic score issues

- **Extreme penalties:** overconfident mistakes with $p_y \approx 0$ dominate the average.
- **Undefined at the boundaries:** $\log 0$ is undefined, so implementations clip $p_y \in [\epsilon, 1-\epsilon]$ and results depend on ϵ.
- **Unit dependence:** the scale changes with the log base, complicating comparisons across papers.
- **Dataset mix sensitivity:** comparisons can be distorted when outcome prevalence differs across evaluation sets.

2.7 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

Grading for free-form prediction was also allowed, whereby GPT-o4-mini was used to directly compare a free-form prediction of the outcome, to the outcome described in the abstract. NOTE: FURTHER DETAILS WILL COME UPON IMPLEMENTING THIS METHODOLOGY

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

3 Results & Discussion

3.1 Database of Evaluations

In addition to producing a useful LLM forecasting system, this work has also (HOPEFULLY) produced the largest extent collection of intervention outcomes in the context of Earth system sciences. This database of abstracts is shared publicly on zenodo at `https://zenodo.org/records/XXYYZZ`.

One large existing collection can be found in (Vivalt, 2020 | "How Much Can We Generalize From Impact Evaluations?"), with 15,024 estimates from 635 papers on 20 types of interventions in international development. Notably, this work has catalogued 1,932 quantitative results from 307 separate papers over approximately 70 categories, when restricting attention to only those results that can be compared with results from another paper on the same intervention-outcome. A majority of papers were found to be assessing the same outcome, so only the latest results were chosen for the quantitative analysis. Only a small percentage of quantitative outcomes were among the 70 categories.

Here, we will analyze which categories of interventions perform above average.

Additionally, we will analyze the relationship between stated grade and extracted quantitative outcomes.

3.2 Strengths and Weaknesses of This Forecasting System

To be analyzed once results are available.

3.3 Evaluation of Techniques for Improving Forecast Accuracy

To be analyzed once results are available.

3.4 The Risk of Trusting This Forecasting System

Even if a forecasting system performs well on the test set, several well-known biases and failure modes can inflate apparent skill. First, there are many subtle pitfalls in evaluating forecasting systems discussed in Section 1.3.3, which can inflate their expected abilities. Second, published abstracts themselves are not neutral evidence: they often overemphasize the significance of effect sizes, which can systematically bias both human and model judgments when training or evaluating on abstract-level text (Duyx et al., 2019 | "The Strong Focus on Positive Results in Abstracts May Cause Bias in Systematic Reviews: A Case Study on Abstract Reporting Bias"). A large study of findings in economics determine an effect size overestimation factor due solely to publication bias of 1.62 in Medicine, Environmental 1.78, Psychology 1.39, and Economics 2.16 (F et al., 2024 | "Footprint of Publication Selection Bias on Meta-Analyses in Medicine, Environmental Sciences, Psychology, and Economics"). In general, including unpublished working papers, while perhaps reducing rigor, may allow for significantly reduced effect size inflation especially in the field of economics.

Some researchers claim that most published research findings are false. In our case flexibility in designing reported outcomes and analytical modes increase the chances that the study was "gamed" to report unrepresentative significance on that particular metric (Ioannidis, 2005 | "Why Most Published Research Findings Are False"). Cases where large financial payouts are required for a significant result are also more likely to lead to false findings (Ioannidis, 2005 | "Why Most Published Research Findings Are False"). The targeted selection of RCTs in our work increases the chance that the discovered outcomes are true, but many uses of RCTs are insufficient - especially if underpowered (Ioannidis, 2005 | "Why Most Published Research Findings Are False"). We should accordingly more heavily weight outcomes from RCTs with higher effect sizes and large sample sizes. Ideally, outcomes are also from pre-registered studies that commit to the research and analysis methododolgies before reporting the results.

Furthermore, people often ascribe objectivity to algorithmic outputs and therefore overweight automated advice leading to "automation bias" (, 2019 | "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment")

4 Conclusion & Outlook (NOTE: CURRENTLY LOW PRIORITY)

4.1 The State of AI and LLMs

- Timelines for AI surpassing human ability in forecasting (, | *AI4Research: A Survey of Artificial Intelligence for Scientific Research*)(Lee et al., 2025 | *Advancing Event Forecasting through Massive Training of Large Language Models: Challenges, Solutions, and Broader Impacts*)
- Scaling up language models improves few-shot and task-agnostic performance (Brown et al., 2020 | "Language Models Are Few-Shot Learners")
- AI safety and regulation

4.2 Extensions of This Work

- Applications in health, policy, law, economics, advancing future scientific progress
- strategic warning applications (Knack and Balakrishnan, | "The State of AI for Strategic Warning")
- Applications to improve personal and organizational decision making
- Futarchy (Arel, 2024 | "Designing Artificial Wisdom: Decision Forecasting AI & Futarchy") (Lizka, 2021 | "Summary and Takeaways: Hanson's "Shall We Vote on Values, But Bet on Beliefs?"") (Hanson, 2013 | "Shall We Vote on Values, But Bet on Beliefs?")
- Applications to reduce gridlock and polarization in the political domain

4.3 Ways that the Current Forecasting Technique Could Be Improved

Comparing differing reasoning trajectories allows the use of reinforcement learning techniques to further improve upon AI forecasting, without additional externally derived training data (Turtel, Franklin, and Schoenegger, 2025 | *LLMs Can Teach Themselves to Better Predict the Future*). This allows much smaller models to best larger model reasoning capabilities.

4.4 The Promise and Capabilities of AI Forecasting

As a clear disclaimer: **LLMs 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 AIs 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. AIs 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.

4.5 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.

4.6 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. 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 CO_2 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. With an explicit attempt to correct these biases via fine-tuning, sycophantic 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 promise 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 AIs 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. In real-world forecasting, the true answers are genuinely unknown at the time of prediction unlike in other benchmark tasks where answers could be memorized from the training data (Schoenegger and Park, 2023 | *Large Language Model Prediction Capabilities: Evidence from a Real-World Forecasting Tournament*). It will be increasingly useful to society to understand what the true capabilities of LLMs 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.

4.7 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.

4.8 AI Scientist 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 CO_2 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*).

4.9 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, pp. 631–648. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2024.07.003. (Visited on 08/19/2025).

AI4Research: A Survey of Artificial Intelligence for Scientific Research (2025). https://arxiv.org/html/250 (Visited on 08/19/2025).

"Algorithm Appreciation" (Mar. 2019). "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment". In: *Organizational Behavior and Human Decision Processes* 151, pp. 90–103. ISSN: 0749-5978. DOI: 10.1016/j.obhdp.2018.12.005. (Visited on 08/31/2025).

Arel, Jordan (July 2024). "Designing Artificial Wisdom: Decision Forecasting AI & Futarchy". In: (visited on 08/20/2025).

Bang, Yejin et al. (Aug. 2024). "Measuring Political Bias in Large Language Models: What Is Said and How It Is Said". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11142–11159. DOI: 10.18653/v1/2024.acl-long.600. (Visited on 08/31/2025).

Bina, Rachel et al. (Feb. 2025). "On Large Language Models as Data Sources for Policy Deliberation on Climate Change and Sustainability". In: DOI: 10.2139/ssrn.5123359. (Visited on 08/18/2025).

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (Nov. 2020). "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics". In: *American Economic Review* 110.11, pp. 3634–3660. ISSN: 0002-8282. DOI: 10.1257/aer.20190687. (Visited on 09/20/2025).

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. Curran Associates, Inc., pp. 1877–1901.

Chen, Yaoyu, Yuheng Hu, and Yingda Lu (May 2025). *Predicting Field Experiments with Large Language Models*. 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.

Duyx, Bram et al. (Dec. 2019). "The Strong Focus on Positive Results in Abstracts May Cause Bias in Systematic Reviews: A Case Study on Abstract Reporting Bias". In: *Systematic Reviews* 8.1, pp. 1–8. ISSN: 2046-4053. DOI: 10.1186/s13643-019-1082-9. (Visited on 08/31/2025).

F, Bartoš et al. (May 2024). "Footprint of Publication Selection Bias on Meta-Analyses in Medicine, Environmental Sciences, Psychology, and Economics". In: *Research synthesis methods* 15.3. ISSN: 1759-2887. DOI: 10.1002/jrsm.1703. (Visited on 09/20/2025).

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).

Gerber, Alan S. and Neil Malhotra (Aug. 2008). "Publication Bias in Empirical Sociological Research". In: *Sociological Methods & Research*. DOI: 10.1177/0049124108318973. (Visited on 09/20/2025).

Ghasemloo, Mohammadmahdi and Alireza Moradi (Aug. 2025). *Informed Forecasting: Leveraging Auxiliary Knowledge to Boost LLM Performance on Time Series Forecasting*. DOI: 10.48550/arXiv.2505.10213. arXiv: 2505.10213 [cs]. (Visited on 08/21/2025).

Gneiting, Tilmann and Adrian E Raftery (Mar. 2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477, pp. 359–378. ISSN: 0162-1459. DOI: 10.1198/016214506000001437. (Visited on 08/19/2025).

Guan, Yong et al. (Aug. 2024). *OpenEP: Open-Ended Future Event Prediction*. 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, 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: ().

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).

Ioannidis, John P. A. (Aug. 2005). "Why Most Published Research Findings Are False". In: *PLOS Medicine* 2.8, e124. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.0020124. (Visited on 08/31/2025).

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).

Karger, Ezra et al. (Oct. 2024). "ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities". In: *The Thirteenth International Conference on Learning Representations*. (Visited on 08/28/2025).

Knack, Anna and Nandita Balakrishnan (2025). "The State of AI for Strategic Warning". In: (). (Visited on 08/22/2025).

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).

Lee, Sang-Woo et al. (July 2025). *Advancing Event Forecasting through Massive Training of Large Language Models: Challenges, Solutions, and Broader Impacts*. DOI: 10.48550/arXiv.2507.19477. arXiv: 2507.19477 [cs]. (Visited on 08/21/2025).

Lizka (Aug. 2021). "Summary and Takeaways: Hanson's "Shall We Vote on Values, But Bet on Beliefs?"" In: (visited on 08/20/2025).

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.

Nadeem, Moin, Anna Bethke, and Siva Reddy (Aug. 2021). "StereoSet: Measuring Stereotypical Bias in Pretrained Language Models". 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)*, pp. 5356–5371. DOI: 10.18653/v1/2021.acl-long.416. (Visited on 08/31/2025).

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).

Paleka, Daniel et al. (May 2025). *Pitfalls in Evaluating Language Model Forecasters*. DOI: 10.48550/arXiv.2506.00723. arXiv: 2506.00723 [cs]. (Visited on 08/21/2025).

Priem, Jason, Heather Piwowar, and Richard Orr (June 2022). *OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts*. DOI: 10.48550/arXiv.2205.01833. arXiv: 2205.01833 [cs]. (Visited on 08/19/2025).

Pritchett, Lant and Justin Sandefur (Aug. 2013). "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix - Working Paper 336". In: (visited on 09/20/2025).

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*. 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. (Visited on 08/19/2025).

Touvron, Hugo et al. (July 2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. DOI: 10.48550/arXiv.2307.09288. arXiv: 2307.09288 [cs]. (Visited on 08/19/2025).

Turpin, Miles et al. (Nov. 2023). "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting". In: *Thirty-Seventh Conference on Neural Information Processing Systems*. (Visited on 08/31/2025).

Turtel, Benjamin, Danny Franklin, and Philipp Schoenegger (Feb. 2025). *LLMs Can Teach Themselves to Better Predict the Future*. DOI: 10.48550/arXiv.2502.05253. arXiv: 2502.05253 [cs]. (Visited on 08/29/2025).

Vivalt, Eva (Aug. 2019). "Specification Searching and Significance Inflation Across Time, Methods and Disciplines". In: *Oxford Bulletin of Economics and Statistics* 81.4, pp. 797–816. ISSN: 1468-0084. DOI: 10.1111/obes.12289. (Visited on 09/20/2025).

— (Dec. 2020). "How Much Can We Generalize From Impact Evaluations?" In: *Journal of the European Economic Association* 18.6, pp. 3045–3089. ISSN: 1542-4766. DOI: 10.1093/jeea/jvaa019. (Visited on 09/18/2025).

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*. 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. 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)*. (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).

Erklärung zur akademischen Integrität / Declaration of Academic Integrity

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und nur mit den angegebenen Quellen und Hilfsmitteln (z. B. Nachschlagewerke oder Internet) angefertigt habe. Alle Stellen der Arbeit, die ich aus diesen Quellen und Hilfsmitteln dem Wortlaut oder dem Sinne nach entnommen habe, sind kenntlich gemacht und im Literaturverzeichnis aufgeführt. Weiterhin versichere ich, dass weder ich noch andere diese Arbeit weder in der vorliegenden noch in einer mehr oder weniger abgewandelten Form als Leistungsnachweise in einer anderen Veranstaltung bereits verwendet haben oder noch verwenden werden. Die Arbeit wurde noch nicht veröffentlicht oder einer anderen Prüfungsbehörde vorgelegt. */ I hereby certify under penalty of law that I have prepared this thesis independently and only using the cited sources and resources (e.g., reference works or the internet). All passages of the thesis that I have taken from these sources and resources, either verbatim or in spirit, are cited and listed in the bibliography. Furthermore, I certify that neither I nor anyone else has used or will use this thesis, either in its present form or in a more or less modified form, as evidence in another course. This thesis has not yet been published or submitted to another examining authority.*

Potsdam, 25 September 2025