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 the Success of Environmental and Sustainability Activities in International Development Using Language Models

Potsdam, 16 December 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

International aid and cooperation creates a profound difference in the rate of development in growing economies, improves the lives of the world's poorest, and often safeguards the environment and materially promotes sustainability. However, international aid has non-significant rates of failure in achieving its objectives. There have been few attempts in the literature to create models to predict the success of aid activities, and none focused on environmental outcomes. This thesis produces a forecasting system for the overall success of international aid activities at time of evaluation from the International Aid and Transparency Initiative (IATI) database, combining classical statistical methods with modern language model techniques. I find that the information available at the start of the activity all contribute to prediction accuracy, including quantifiable information used in previous studies, averaging the success rates of semantically similar activities, and using the reasoning abilities and information gathering ability of large language models (LLMs) to improve forecasts. Testing against the held-out, latest-starting 200 activities in a dataset of 1,800 environmental and sustainability improving activities, the full forecasting system improves success forecasting at the beginning of the activity by X.XX% (95% CI: XX% to XX%), improving the prediction from XX.X% to XX.X% accuracy compared to a baseline adapted from prior literature. For overall success ratings on a scale from approximately 1 to 6, the system can explain XX% (95% CI: XX% to XX%) of the variation compared to XX% for the baseline. I also release a freely available dataset of LLM-generated activity grades, summaries, success ratings, and various other quantitative activity outcomes and extracted information for 1,800 IATI activities. This work lays the foundation to improve decision making for a wide range of initiatives and policies in developing countries and also in other data-rich institutional contexts.

Abstract auf Deutsch

Will do, once abstract is finalized

Table of Contents

1 Background: LLMs and 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 integrated assessment models (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").

In the specific context of developmental cooperation, the system of interactions between people, their wellbeing, medical, educational, and career outcomes, the economy, the government, and the natural environment surrounding development cooperation interventions also displays difficult to understand emergent phenomena such as regime changes, disease spread, and economic collapse.

Together, these characteristics allow us to characterize the system being improved by development cooperation interventions affecting the environment as a "complex system", where, by definition, decision making about outcomes is challenging.

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 interventions would realistically improve outcomes for activities in the Earth system sciences.

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.

This thesis was written in conjunction with the German Federal Ministry for Economic Cooperation and Development (BMZ) in order to improve their environment-related international aid decision making. In the context of official development aid (ODA), German development finance commitments on behalf of the German Federal Government were the second largest ODA source in 2023 at approximately 40 billion USD (, | *Net ODA | OECD*). This was the case before recent major reductions in the US ODA in 2025, which indicate that Germany may soon be the largest source. However, despite significant care and effort put forth in documenting development cooperation outcomes at the the BMZ, ex-post evaluations are rarely read at the BMZ or the affiliated KfW Development Bank, even though around 19% of evaluated projects are unsuccessful (Sustainability (IDOS), | *Learning from KfW's ex-post evaluations? How conflicting objectives can limit their*

usefulness). Given the large volume of directed aid and the likely gaps in knowledge due to low utilization of ex-post evaluations, an opportunity arises to close these gaps using recent advances in ML, especially large language models (LLMs), which can quickly search and synthesize findings over a much larger quantity of information than aid funding decision makers (from here on we will refer to them as "evaluators"). At BMZ, these are the BMZ officers.

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 in the context of developmental interventions affecting the environment. 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 development cooperation outcomes affecting the environment. 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 developmental aid and cooperation interventions affecting the environment, 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 developmental aid and cooperation interventions affecting the environment 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 developmental aid and cooperation interventions affecting the environment. Accordingly, in tandem with the open source LLM forecasting system, I also release the largest extant structured

database of interventions and associated outcomes in developmental aid and cooperation interventions affecting the environment. 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 LLMs: The Transformer Architecture

LLMs used in this work all use variants of the same fundamental architecture: the "transformer" (Vaswani et al., 2017). Transformers are ML models which are trained on vast quantities of textual data. During training, transformers convert input documents and text-based sources into "tokens" which are typically parts of words or smaller chunks of pdf or image-based data which commonly appear during training. In this work we use a simpler "decoder-only" variant of the transformer, as is commonly used for reggressive token prediction in chatbot applications, like Chat-GPT.

After input documents and textual sources are converted to tokens, transformers use a learned linear transformation called "embedding" matrix to convert the token-space into a lower-dimensional semantic space, typically with a few hundred dimensions. Each input token to the transformer is converted into a semantic vector. These vectors are important because they encode similar input tokens into nearby locations in the high-dimensional

semantic space. At this stage, the transformer runs each token through a series of layers which in parallel convert all of the input tokens to the next predicted output token. Typically a transformer contains dozens or hundreds of such layers. These layers are composed of both Multi-Layer Perceptrons (MLP) which are simply feed-forward neural networks commonly used in many other ML architectures, and "attention heads", which are unique to the transformer architecture. A compressed representation is compressed into the "residual stream" after each layer, and the process is repeated until the reverse of the embedding matrix is applied to the final residual stream back into the token space, allowing the transformer to finally predict the next token.

The goal of the transformer is always to predict the next token. Accordingly, while MLP layers are typically able to store information for the memorization and fact-based learning in transformer training, "attention heads" have the ability to learn to locate locations in the past tokens where particularly relevant sections for predicting the next token would be. By copying in the relevant information into the residual stream, transformers are able to access important information even thousands of tokens in the past, a capability which is challenging to replicate in other architectures (such as Long-Short Term Memory (LSTM) architectures).

Transformers are the most appropriate choice as a model due to their remarkable ability to apply reasoning and generalization past their training data, and their ability to utilize both quantitative and semantic information for accurate next-token prediction.

Finally, it is possible to fine-tune transformers - iterate update the weights within the embedding, MLP, and attention head components to reduce the loss on the fine-tuning training data. Fine-tuning can be viewed simplistically as the final stage of training where models are reconfigured and optimized towards skill at a narrow task, such as forecasting outcomes of development cooperation interventions relevant to ESS.

1.3 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

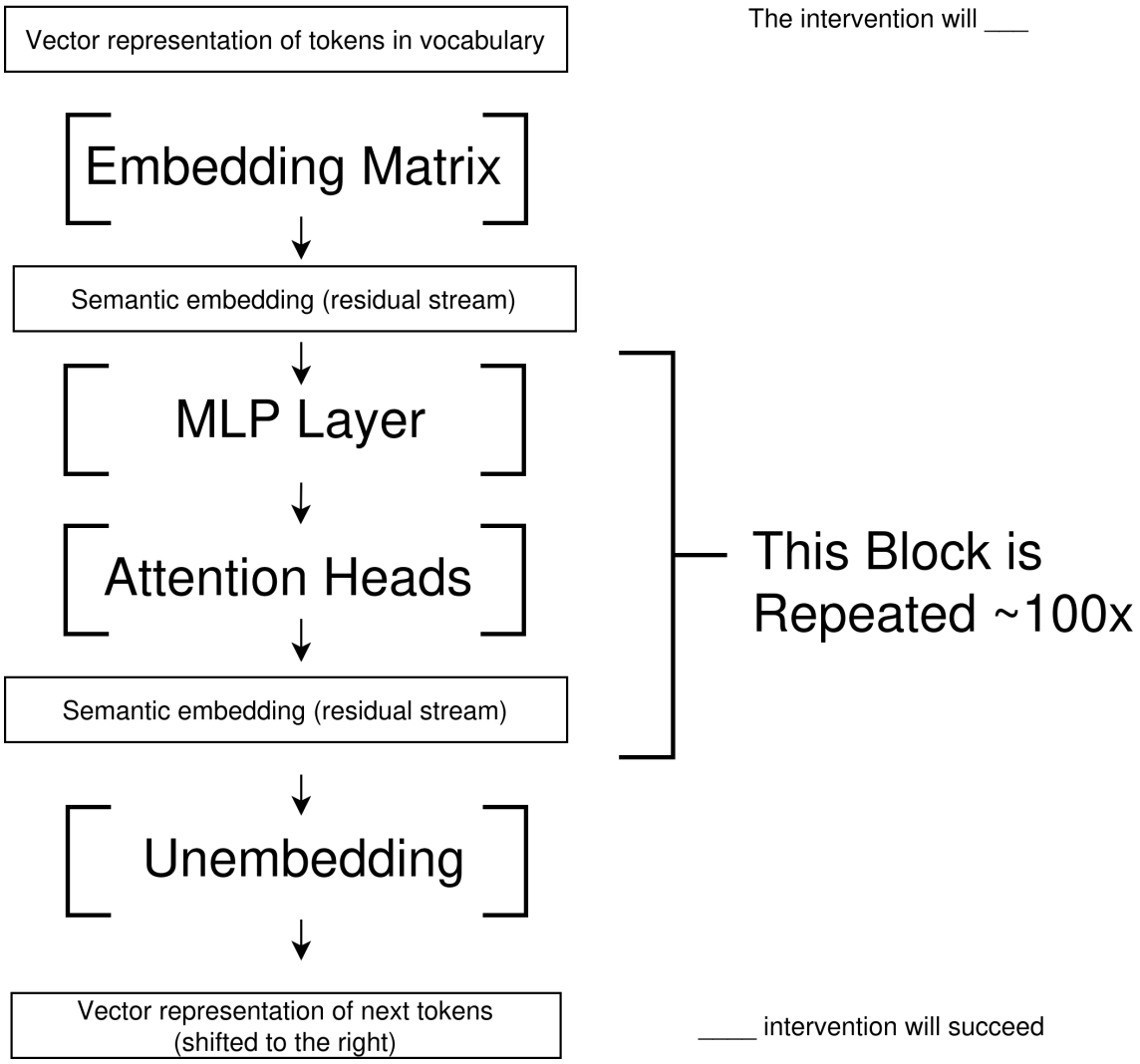

Figure 1: A simplified decoder-only transformer architecture.

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*).

1.4 LLM Forecasting of Outcomes for Development Cooperation

1.4.1 Ex-post Evaluations and Other Data Sources

Before turning to the possibility of LLM forecasting, we first consider what data we could use as context for their forecasts, and how useful it is for human evaluators. evaluators already use evaluations and other sources to inform their work at the BMZ. While evaluations are rigorous, surveys of evaluators find that evaluations come too late after the ends of projects, often several years after, to be strongly relevant to current projects and that ex-post evaluations are of limited relevance to evaluators, either individually or in aggregate form. (Sustainability (IDOS), | *Learning from KfW's ex-post evaluations? How conflicting objectives can limit their usefulness*). This suggests both that the specific relevance of evaluations is unlikely to produce high accuracy results alone, and that information gathered in the few years immediately prior to the project is of particular importance to predicting project outcomes.

Thus far, several repositories of high-quality intervention evaluations have been published. These databases document thousands of examples of ex-post predictions which allow us to infer how future development cooperation outcomes will proceed. In particular, the Interactive Database for Evaluation and Learning (IDEaL) contains over 1,200 rigorous

studies with 6 separate evaluation criteria grading how well the intervention proceeded (KfW Development Bank, | *IDEaL*). Collectively, over 10,000 intervention examples over the past few decades have been collected, although the quality and format of these additional interventions is relatively unknown (, 2025 | *Datenlabor-Bmz/Awesome-Development-Cooperation-Data*).

Social Science Prediction Platform (SSPP) has collected thousands of examples of studies where social science results have been collected and compared to obtain informative priors in Bayesian analysis (, | *How Can You Use the Social Science Prediction Platform for Development Papers?*). SSPP provides hundreds of evaluations, including a leaderboard and mean absolute error of the top 10 performing forecasters with at least 10 forecasts each. As the results become open source in the coming years, this can be a valuable resource for development impact forecasting.

Finally, the largest extant database of recorded interventions available is that of the International Aid and Transparency Initiative (IATI) (*IATI Dashboard – IATI Activities* 2025). The full database contains over 800,000 records of interventions, the majority of which are marked as to whether they were closed or cancelled. Within this dataset, tens of thousands are relevant to interventions in ESS, and a large fraction of those contain date-stamped project appraisal and evaluation documents. However, due to the inclusion of over 1,000 distinct contributing international aid organizations, IATI is known for its difficulty to interpret and inconsistency in structure of each activity record.

At BMZ, even when ex-post evaluations are read by evaluators, they usually only read the cover sheet. Evaluators typically have to read a large number of documents, including activity reports. Ex post evaluations are a small part of the material used to make funding decisions in development cooperation interventions (Sustainability (IDOS), | *Learning from KfW's ex-post evaluations? How conflicting objectives can limit their usefulness*). This further indicates that information outside the purely evaluative nature of past projects is quite important for funding decisions.

1.4.2 Determinants of Success in Development Cooperation Interventions

A longstanding, well-researched question in foreign aid has been "which projects are effective?". Knowing which types of features and aspects of project determine their effectiveness allows this work to ensure the proper information is supplied to the LLMs. It also allows construction of naive baselines, using purely statistical correlations between projects aspects and outcomes. In the last decade, focus has shifted towards experimental impact evaluations such as randomized controlled trials (RCTs), which experimentally answer specific questions about intervention effectiveness, producing a large body of high-quality evidence for effectiveness of a wide range of interventions (Olken, 2020 | "Banerjee, Duflo, Kremer, and the Rise of Modern Development Economics*").

There are many broad ways of answering which interventions will be generally effective. One method is to distinguish between country-level factors, such as the economic and political conditions, and project-level factors, such as the amount of funding or sector of the project. While both country-level and project-level aspects are necessary when forecasting project performance, project-level aspects appear to be a much stronger determinant of project success, typically with an R^2 of about 0.2-0.3 of predicting success outcomes given specific project variables (, 2013 | "Good Countries or Good Projects? Macro and Micro Correlates of World Bank Project Performance") (Bulman, Kolkma, and Kraay, 2017 | "Good Countries or Good Projects? Comparing Macro and Micro Correlates of World Bank and Asian Development Bank Project Performance") (Eilers et al., | "Volume, Risk, Complexity: What Makes Development Finance Projects Succeed or Fail?"). In terms of project variables, one study finds that for KfW funded projects, technical complexity and longer implementation duration, as well as increased assessed risk at the beginning of the project correlate significantly with unsuccessful projects, and increased funds correlate significantly with successful ones; choice of project structure was not found to have a significant correlation with project success (Eilers et al., | "Volume, Risk, Complexity: What Makes Development Finance Projects Succeed or Fail?").

A comprehensive literature review finds that project size is indicative of project performance is mixed, duration is negatively correlated, while preparation time is positively correlated (Ashton et al., 2021 | "A Puzzle with Missing Pieces : Explaining the Effectiveness of World Bank Development Projects"). There is low-strength evidence that non-governmental actors have better evaluation scores. The quality of the economic analysis and a strong analytical underpinning at the start is also positively correlated with success. Staff and management has been found to be key to success across a wide range of studies. A better track record of the task team lead alone is associated with an increase in the chance of project success by 6%.

A separate analysis found when regressing on different explanatory variables, larger projects do worse, Academic/NGO-implemented projects do better than government-implemented projects (Vivalt, 2020 | "How Much Can We Generalize From Impact Evaluations?"). See especially Table 7, with around 500 observations for these findings: 0.013 standard deviations decrease in outcome effect size as a function of size of study, -0.081 standard deviations worse outcome if from government compared to -0.018 from academia (so government is much worse, on average, even controlling for program size). RCTs do better (0.021 standard deviations). The region was found to matter, with North Africa often performing better than South Africa, although there is insignificant regional data there to firmly draw that conclusion. In terms of specific intervention categories, there were clear significant differences, but it is not easy to make quick judgments about which types of interventions generalize. Essentially, reducing to the few variables investigated discards the key contextual background which ultimately drives most outcomes in development.

One appropriate analogy may be attempting to model the next word in human language with a linear statistical model. The performance will be terrible because contextual cues form the majority of the required information to predict outcomes.

Direct consultation with an economist in the evaluation department of BMZ revealed the most important predictor is the degree of ownership of a project. Secondary effects were the length of the project (negatively correlated), how capable the partner government is (positively correlated), previously identified high risks for projects (negatively correlated), and whether the project was directly integrated into a larger program, such as a country's large portfolio of energy sector projects (positively correlated).

1.4.3 Other Computer Modelling Methods

[TODO: Add a diagram demonstrating how these models work and put in some more details demonstrating that I understand how they work in the text]

ML is not the only tool used in modelling intervention outcomes in complex systems. 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").

The underlying workings of these models are highly variable. What unifies them is their domain specificity: predicting outcomes of interventions has typically required careful coding of environmental variables and human involvement in manually calibrating results. It has not been possible with these models to take into account the broader factors that influence the success of specific interventions. Factors such as reputability of the partner country, the rate of success of similar interventions in the past, and judgement about the general fit between an intervention and its specific context, have thus far all been relegated to human judgement of aid decision makers.

1.4.4 Methods and Capabilities

This thesis implements an LLM-based forecasting method, predicting what the evaluation results will be for thousands of IATI records containing both a pre-intervention description of the activity, as well as a post- or mid-intervention evaluation of the results. To do so, thousands of pdfs were downloaded, ranked from most to least relevant for forecasting future outcomes or evaluating the end result of the activities, had their pages ranked and graded for relevance to the task, had quantitative and qualitative descriptions and results transcribed into a unified format, and next several versions of the LLM forecasting system were trialed on the validation set. Finally, the most promising version was used to predict evaluation outcomes on hundreds of evaluations that were written at least one year after the model's training cutoff.

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 developmental aid and cooperation interventions affecting the environment 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")

6. Testing a variety of prompts [NOTE: CITE that "Or How I learned to be careful about prompt variants"] and reducing the complexity of the prompt to prevent the model from forgetting its training data (Kaiser et al., 2025)

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 developmental aid and cooperation interventions affecting the environment, 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 developmental aid and cooperation interventions affecting the environment. 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.

Another study uses LLMs to predict the likely direction and effect size of empirical studies evaluating dietary interventions (Kaiser et al., 2025). This demonstrates that a fine-tuned LLM can predict direction of empirical intervention outcomes better than classical meta-regression baselines in dietary policy interventions, at an accuracy of approximately 80% to predict the directional outcome of the policy. The authors utilize a fine-tuned version of GPT3.5 and carefully select prompt variants which tend to score higher.

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 limited benefit from a slightly more complex mixed effects model with explanatory variables, rather than a random effects model. 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). One possibility explaining this results is that parameter heterogeneity is in fact 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").

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 developmental aid and cooperation interventions affecting the environment. 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*). However, caution is also warranted - in some cases, humans over-adjust their estimates towards the weaker LLM forecast, and the results can be worse than humans alone [TODO: FIND THIS CITATION SOMEWHERE! I KNOW I SENT IT TO MYSELF AT SOME POINT].

1.4.5 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. As a limitation, that this reduces this thesis's reproducibility and increases software maintenance requirements, 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 developmental aid and cooperation interventions affecting the environment 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.

In another study regarding LLM forecasting of food policy, the *direction* (positive vs negative sign of the intervention's impact) was much more easily predictable than the absolute effect of the intervention. The fine-tuned version with a small prompt was found to have a 79% success rate at predicting the direction on the held-out test set, handily besting the random-effects model baseline rate of success of 66%. However, the μ_e average error was -.051, while much better than -1.92 for the random effects baseline for estimating the Cohen's d effect size, is not encouraging in absolute terms.

The use of LLMs to inform decision making for outcome prediction in developmental aid and cooperation interventions affecting the environment 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 developmental aid and cooperation interventions affecting the environment 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 **??** and section 2.8). 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 Outcomes in Development Cooperation Interventions Affecting the Environment

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 developmental aid and cooperation interventions affecting the environment.

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 Data Sources

After considering several data sources for prediction, including the OpenAlex publication repository of peer-reviewed evaluation documents and abstracts, the IDEAL database of ex-post evaluations, the 3ie development database, I decided to use the IATI database, due to its substantial quantity of information available in textual format and extractable from the database records, and its sheer size. While ex-post evaluations may provide sufficient information to describe the activity, it may introduce "future leakage" to rely on language

models to completely remove information about the eventual outcome. Furthermore, although many millions of evaluations are available, it proved difficult to reliably identify and de-duplicate academic papers regarding evaluations of environmental interventions. The IDEAL database and 3ie were in the dozens or hundreds of records for environmental topics.

The IATI database has reliable start and end dates, and typically several recorded outcomes, and usually an overall evaluation rating. It also reliably marks the reporting organization, allowing for an intelligent unification of the rating scales, and sometimes provides a "results" section where outcomes of an activity can sometimes be found. It is quite common for several activity information documents to be uploaded near the beginning of the activity, and several years later, at least one ex-post evaluation of the activity is uploaded as well, or results of key quantitative outcomes are sometimes directly recorded in the IATI database. The status of the activity is extremely commonly reported, including if it is in the planning or completion/finalization stages, which is helpful information for forecasting.

The downside of the IATI database is it is highly inconsistent between reporting organizations as to how the data are filled in, and the format of documents was not uniformly PDF, requiring conversion scripts. Also, many download links were not functioning or required custom web-scraping scripts to properly extract project documents in pdf format from the original websites where project documents were hosted. Dates of documents and especially planned start or end dates, or actual start or end dates, were often missing, leading to frequent exclusion of projects. Furthermore, approximately 30% of IATI activities do not have an activity category code, meaning that environmental or sustainability topics are sometimes missing.

2.3 Data Filtering

IATI records for prediction

Out of the approximately 800,000 international aid activities recorded in IATI, I first reduced the set of activities of interest to 7,575 records which aimed to improving the environment, sustainability, or climate adaptation in a developing country or countries, had an appraisal/intervention description document, and an outcome evaluation or progress report document, both of which could be converted to PDF format. Links to these documents were then downloaded where possible (see the next section for details).

Once documents were downloaded and converted to pdf format, activities were further filtered so that they had at least one document describing the activity, and had a metadata date before 1/4 of the activity implementation period, as well as at least one ex-ante activity at least 3/4 through the activity period. The latest activity document also had to

have a metadata date at least one year before the earliest evaluation document. Activities not meeting these requirements were also excluded, leaving 3,225 activities.

The date for the documents were determined using (in descending preference where available) the pdf's "created on" or "last modified" in its metadata, or the date indicated in the IATI record for the document. This ordering preference was determined as the PDF metadata dates were found to more reliably match the stated date of authorship of the documents better than the "IATI date" recorded within the activity record. These dates were usually available in the metadata of the original PDF, ODT, DOC, or DOCX file. Experimenting with different date options revealed that out of 400 randomly selected PDFs, the closest available date to the true date of authoring the document was the "created by" date, then the "last modified" date, then the "IATI date". The median difference for the date when selecting this ordering was 22 days different than the date indicated in the document itself as determined by feeding the first 3 pages of each of those 400 PDF documents to *gemini-2.5-flash*.

The activity filtering for the topic was done by-hand to filter only those activities relating to improving the environment or sustainability. These were:

- 140: Water Supply & Sanitation
- 231: Energy Policy
- 232: Energy generation, renewable sources
- 234: Hybrid energy plants
- 235: Nuclear energy plants
- 312: Forestry
- 410: General Environment Protection

Some more specific 5-digit codes were also selected:

- 14015: Water resources conservation (including data collection)
- 14020: Water supply and sanitation - large systems
- 14021: Water supply - large systems
- 14022: Sanitation - large systems
- 14032: Basic sanitation
- 14050: Waste management/disposal

- 23110: Energy policy and administrative management
- 23111: Energy sector policy, planning and administration
- 23112: Energy regulation
- 23183: Energy conservation and demand-side efficiency
- 23210: Energy generation, renewable sources - multiple technologies
- 23220: Hydro-electric power plants
- 23230: Solar energy for centralised grids
- 23231: Solar energy for isolated grids and standalone systems
- 23232: Solar energy - thermal applications
- 23240: Wind energy
- 23250: Marine energy
- 23260: Geothermal energy
- 23270: Biofuel-fired power plants
- 23350: Fossil fuel electric power plants with carbon capture and storage (CCS)
- 23360: Non-renewable waste-fired electric power plants
- 23410: Hybrid energy electric power plants
- 23510: Nuclear energy electric power plants and nuclear safety
- 23630: Electric power transmission and distribution (centralised grids)
- 23631: Electric power transmission and distribution (isolated mini-grids)
- 23642: Electric mobility infrastructures
- 31192: Plant and post-harvest protection and pest control
- 31210: Forestry policy and administrative management
- 31220: Forestry development
- 31281: Forestry education/training
- 31282: Forestry research
- 31291: Forestry services

- 32174: Clean cooking appliances manufacturing
- 32210: Mineral/mining policy and administrative management
- 41010: Environmental policy and administrative management
- 41020: Biosphere protection
- 41030: Biodiversity
- 41081: Environmental education/training
- 41082: Environmental research
- 43060: Disaster Risk Reduction

At least one 3-digit or 5-digit codes above must be one of the codes indicated in the activity metadata for it to be considered.

Document-related Filtering

The IATI database contains a collection of thousands of links to pdfs, word documents, html documents, and other document formats. These were first automatically converted to pdf format via a custom python script, and subsequently needed to pass several criteria before being used as documents for forecasting.

I first wrote a script that directly downloaded these and converted them to pdf format. Specifically for UNDP, I made a custom-query to convert their interactive HTML results pages into a pdf format, all others had the links for the documents downloaded directly with no additional processing.

Next, I look at the pdf metadata date, and determine the creation date of the pdf files. I find this is more often closer to the date of the specific activity description document or activity evaluation document (as determined by reading the document) than metadata at the url indicating upload date, or the date entered in IATI for the document. UNDP results had the URL specifically included in the JSON payload populating their website, so the latest year indicated in that evaluation payload was used instead for the UNDP results.

All documents are tagged in IATI with one or more of seven tags per document: “Pre- and post-project impact appraisal”, “Objectives / Purpose of activity”, “Intended ultimate beneficiaries”, “Conditions”, “Budget”, “Summary information about contract”, “Review of project performance and evaluation”, “Results, outcomes and outputs”, “Memorandum of understanding (If agreed by all parties)”, “Tender”, or “Contract”.

I mark documents with “Objectives / Purpose of activity”, “Summary information about contract”, or “Pre- and post-project impact appraisal” tags as preliminary “baseline”

documents - those representing information about the activity before it begins. Documents with “Review of project performance and evaluation” or “Pre- and post-project impact appraisal” tags are marked as preliminary evaluation documents. In order to ensure the information describing the activity is sufficiently before the activity evaluation, I require at least one “baseline” document and at least one “outcome” document per activity, with the evaluation document at least one year prior to a preliminary evaluation document (based on the uploaded document metadata date). I also require that the activity status code is not “Pipeline/identification”. Instead, activities are allowed to be in implementation, finalization, closed, cancelled, or suspended, such that either a final or preliminary evaluation document is possible.

I filtered further to ensure that all activity document labels were a subset of "Conditions", "Budget", "Tender", "Contract" with no other tags, indicating the documents involved were purely legal context, often containing very little evaluation or activity information.

To ensure pdf metadata dates were appropriate, an analysis was undergone to ensure the procedure for selecting the date of activity documents was valid. If the date of the activity is too early, it could lead to documents authored well after the project start leaking future information. To test this, 400 random pdf documents downloaded were uploaded to gemini and the pdf metadata dates were inspected. PDF metadata dates were discovered to have a median difference of 22 days from the date indicated on the document as extracted by gemini. It was discovered that while approximately 10% of the dates were more than a year after the actual date indicated on the document (such that an activity was actually authored earlier than the start date of the activity), a concerning 0.5% of documents had a creation pdf metadata date later than the date extracted directly from the document.

In order to ensure the forecasts were all based on project information available only roughly at the beginning of the activity, a search was undergone through the information available to the model when forecasting to ensure the forecasting was based only on what could have been known at the beginning of the activity. Approximately 10 activities with pdf metadata dates more than a year earlier than the true authoring of activity documents would be expected, based on the 0.5% rate of “>1 year too early” errors from the date analysis.

To prevent any information leakage, which could be due to incorrect dates as well as incorrect marking of the start date of the activity in IATI, or significant progress being made within the first quarter of the activity where documents are allowed, approximately 40 random chatgpt-generated activity summaries (see the next section) were inspected, with none indicating advanced progress, indicating less than 2.5% of activities should be of concern. A selective search for phrases revealed some activities had made clear progress on targeted outcomes. Consequently, a python script with 6,800 separate search terms was used to further search for inappropriate documents. Exact string search terms were made,

with variants of phrases including, “on track”, “ongoing project has been performing”, “ongoing project is performing”, “already made considerable progress”, “key milestones already achieved”, “significant progress had been made”, “the programme has already made considerable progress”, etc. This led to the review of approximately 150 additional activities, and the discovery of 21 activities with clear progress on key project milestones. Progress such as the formation of planning committees or initial disbursements of funds to the implementing organization were not considered grounds for exclusion, given that these milestones are unlikely to be substantially informative. However, extension activities or Phase II / Phase III activities were not excluded, unless significant progress had already been made on the extension or phase being evaluated.

In order to properly extract accurate overall success ratings for each activity and useful textual information about the project for forecasting, I processed each pdf document using the following data processing pipeline:

2.4 Preliminary Data Processing

All document pages had their rotation detected, and were rotated to vertical before processing via the Gemini API. Documents with “.odt”, “.doc”, or “.docx” extensions were converted to pdfs with a custom script. The pages when converted to pdfs were counted and zero-page documents were excluded.

1. Ranking documents Documents were ranked from most to least useful for forecasting the outcome, or evaluating the results, respectively. *gemini-2.5-flash* structured output with direct pdf input was used to make the rankings. Only documents with c- or better grades on a grading scale from a+ to f were considered for the next stage. Also, the documents were ranked from most to least informative for forecasting among the baseline documents, and most to least valuable for ex-post evaluation among the outcome documents. Baseline documents that were closest to the activity start, and the latest outcome documents were preferenced. Documents with sufficient detail but not excessive lengths, such as executive summaries, were prioritized. Documents that were duplicates in a non-English language were excluded if the equivalent was available in English. For outcomes, if there were multiple progress reports, all the earlier ones were excluded and only the latest were kept in the rankings.

2. Categorizing pages within documents The highest ranking documents were then split into 3-page chunks. Each 3-page chunk was sent in pdf form to *gemini-2.5-flash*. The pages were categorized differently based on whether the document was a baseline or outcome. Categories for outcomes allowed retrieval based on whether final evaluation in quantitative or qualitative form are present on the page, deviations from plans or other types of outcomes were detailed, or if the pages were simply overviews of the activ-

ity. Specifically, the allowed categorizations were "condensed summary", "sub activities outlined", "detailed implementation plans", "broad objectives", "possible outcomes", "quantitative targets", "qualitative targets", "risks as word or numeric", "risks or dangers generally", "plans to address key risks", "positive indicators", "progress reports", "similar cases outcomes", "implementation context country", "contextual challenges", "financing details", "budget and legal", "who implements", "whether part of larger program", "partner identity or skill", "whether skin in the game", "other stakeholder engagement", or "activity monitoring details" for baseline document pages, and "expected outcomes", "deviation from plans", "preliminary results", "final outcomes", "delays or early completion", "over or under spending", "overview as was planned", or "unrelated to evaluation" for outcome document pages. Only one category choice among these was possible per page.

In order to exclude irrelevant pages, the pages were also given a second category, for outcome document pages as "glossary", "blank page", "table of contents", "outcome evaluation", "activity description", "references", or "other", and for baseline document pages the same categories were options, in addition to "core activities", "theory of change", "targets", "broader context", and "preliminary results". Only one category choice among these was possible per page.

3. Extracting Ratings Two separate methods were used to extract rankings. The first method sent each individual outcome page ranked above 7/10 for relevance to evaluation, or with a "quantitative targets" categorization, to *gemini-2.5-flash* to extract any overall ratings, and a second script summarized the overall ratings into a single value for the document. However, this was often insufficient to capture the overall ratings. Another "fallback" script involved a custom generated word search with approximately 500 different rephrasings of "overall rating", "final result", "synthesized score", etc, in English, and searched the pdfs directly for an exact match on those terms, prioritizing pages with one or more exact text matches of such terms. Otherwise, if such words could not be found, the earliest pages in the document which were not categorized as "blank page", "appendix", "glossary", "table of context", "references", or "activity description" were included and *gemini-2.5-flash* was queried to extract the overall rating from the documents.

A slightly different process was done for UNDP activities. The fallback script was used again, but documents were judged by *gemini-2.5-flash* for whether the outcomes represented an "overall successful" or "overall unsuccessful" activity, and this was entered as the result, as UNDP rarely delivers directly an overall rating for the activity.

For BMZ/GIZ/KFW documents, activity baseline documents were extremely rare. For this reason, the evaluation document was treated as a baseline document for the purposes of forecasting activity success. Categorization for these evaluations also was via the "baseline" document method described above. When grading or summarizing the features of the evaluation document, *gemini-2.5-flash* was instructed to only describe what could

have been known at the beginning of the activity, and to under no circumstances reveal the final outcome of the activity.

2.5 Baseline Methods

Three relatively simple baseline methods were attempted, to ensure the relatively complex and expensive LLM-based methods are better than simpler approaches. I choose three simple baseline methods, in order to ensure the predictions were significantly better than the baseline methods of emission reduction predictions.

Prediction baseline: always predict the most common rating ('Moderately Satisfactory') This baseline technique provides a sanity check that more sophisticated methods are worthwhile. Because the prediction task is inherently difficult with much of the variation in outcomes unable to be forecasted at the outset of the activity, this is a relatively strong baseline.

Prediction baseline: GLM Trained with non-LLM categories In order to justify the addition of non-LLM categories, we use the baseline statistical categories apparent in prior literature and train a General Linear Model (GLM) on the outputs. Features include log(GDP)

- planned activity duration
- whether the activity is primarily loan or grant-based
- the one-hot encoded funding organization of the top 4 most common organizations in the database (The World Bank (957 activities)
- BMZ/KFW/GIZ (240 activities)
- UNDP (257 activities)
- and the Asian Development Bank (156 activities))
- the Country Policy and Institutional Assessment (CPIA) score from the World Bank for that country
- the scope of the activity
- the planned duration of the activity
- the log(GDP/capita) of the countries where the activity takes place weighted by the percentage of the activity performed in each country.

- *gpt-3.5-turbo*-generated finance, activity integratedness with broader activities, implementer performance, ease of targeted outcomes, contextual challenge, overall risks, and overall activity complexity grades.

The activity start date was not used, as there was no clear pattern with regards to overall activity success over time, and this is unlikely to be monotonic.

Prediction baseline: Zero-shot LLM In order to ensure the series of improvements on the LLM system in fact genuinely improve accuracy above simply querying the model with some basic activity information and requesting a prediction, we insert the summarized information about the category along all of the evaluation axes where the model could find relevant information. This show that the methods used to improve accuracy are indeed increasing accuracy above simply a single generated prediction by ChatGPT-3.5.

2.6 Experimental methods

GLM using IATI Features and Grades

Similar to the baseline prediction, the GLM is trained with elastic net regression to reduce overfitting on noise, and only high-performing features are kept {Insert criteria for inclusion}.

Nearest Neighbor (Vector Similarity)

I first constructed a similarity test using features including countries of the activity, GDP per capita as described previously, the scope of the activity (), and the implementing and funding organization ID. I found however that this similarity test significantly underperformed compared to the semantic similarity of the GPT-generated summary of the activity documents. I first weight the similarity proportional to its embedding semantic similarity score, and tested a cutoff for averaging 1, 3, 7, 10, 15, 20, and all remaining activities based on the Gemini embeddings model *gemini-embedding-001*. I found 15 nearest neighbors was the highest-performing using this method, and thus use the weighted average of the nearest neighbor ratings to predict the overall activity score.

Random Forest

I trained a random forest model using the same data as the GLM, however I find it significantly underperforms the GLM and thus do not continue with this method.

LLM Forecasting Methods #1 and #2

Two distinct but similar LLM forecasting methods were used for prediction. To generate the LLM forecasting methods, GPT 3.5 was prompted with a series of "mock forecasts", generated by Gemini 2.5-pro. The "mock forecast" used relevant pages retrieved by ranking the categorized topics by forecast informativeness and retrieving 10 pages of the

most relevant activity data and 10 pages of the most relevant evaluation data, prioritizing "deviations from plans", etc.

Figure 3 shows the prompt template used to generate the retrospective "mock forecasts" (using *gemini 2.5-pro*), conditioned on retrieved baseline and outcome document excerpts for the same activity.

To generate each mock forecast, we constructed a retrieval-augmented input consisting of up to 10 baseline pages and up to 10 outcome/evaluation pages per activity. Baseline pages were selected from high-scoring passages in predefined "forecast-informative" categories (e.g., objectives, implementation plans, risks, financing details, contextual challenges, and stakeholder/implementer information), using a high relevance threshold (minimum categorization score of 9) and including nearby pages when insufficient high-scoring pages were available. Outcome pages were selected from outcome documents emphasizing deviations from plans (including deviations, delays/early completion, and over/under-spending), using a lower relevance threshold (minimum score of 3) and likewise including surrounding pages to reach the target count when needed. We then merged these retrieved excerpts with activity metadata (title, scope, planned start/end dates, planned financing totals when present) and brief model-generated baseline summaries (activity description and risk summary) before prompting Gemini to write a forecast from the ex-ante perspective. Importantly, the prompt required the model to end by outputting the *known* final evaluation rating for that activity (derived from the merged ratings file and converted into scale-specific text via `get_ratings_text`), while also instructing it to ground the narrative in the retrieved evaluation pages and to return "NO RESPONSE" if the evaluation excerpts did not contain sufficient justification for the assigned rating.

The most semantically relevant activities which hit their 3/4 mark at the latest 1/4 of the way through the query activity were then retrieved and selected such that there was at least one activity with a score below 2.5, and at least one with a score above 2.5, while prioritizing the most semantically similar activities for mock forecast prompt insertion. In addition, the activity "risks" were inserted before each mock forecast, to provide context for the example. Each mock forecast was structured in a way similar to the highest performing scratchpad method from (Halawi et al., 2024).

A series of features including the activity title, start date, and activity location were injected into the prompt to provide context for the activity.

For Method #1, the Gemini-generated summaries were added to the prompt.

For Method #2, the most relevant pages of the activity documents were directly converted to text and inserted into the prompt.

In summary, methods #1 and #2 shared the same scaffold and response-format constraints; they differed only in whether the activity context was provided as model-generated

summaries (Method #1) or as raw extracted document text (Method #2).

Finally, the distribution of rating outcomes was inserted into the prompt, in order to prevent collapse towards only a few ratings.

The full prompt templates for Methods #1 and #2 are shown in Figure 2.

In both methods, we used a k-nearest-neighbors (KNN) few-shot block to provide examples of semantically similar activities with known ex-post evaluation outcomes. Candidate neighbors were restricted to activities with (i) an available retrospective "mock forecast" and (ii) an observed final evaluation rating. We then selected up to k neighbors in descending similarity order, while (when possible) enforcing coverage on both sides of the rating-scale midpoint (at least one outcome above and one below the midpoint) and approximately matching the global training-set distribution over outcomes by sampling from coarse outcome bins.

Concretely, in order to ensure the few-shot insertions of mock forecasts were informative, we first computed a training-set outcome distribution over a fixed set of six equal-width bins spanning the rating scale from worst to best. Each historical activity with a rating was mapped to a within-scale fraction $f \in [0, 1]$ (using its rating and the activity-specific rating scale), and then assigned to a six-bin index $b = \min(5, \lfloor 6f \rfloor) + 1$. For a target forecast with budget k, we converted the global six-bin percentages into per-bin target counts by taking $\lfloor k \cdot p_b \rfloor$ for each bin (where p_b is the training-set percentage in bin b) and allocating any remaining slots to bins with the largest fractional remainders. We then traversed the similarity-ranked candidate neighbors and greedily selected candidates whenever their bin had not yet met its target count. If some bins could not be filled (e.g., due to missing candidates with valid rating scales), we filled remaining slots with the most similar unused candidates regardless of bin. Finally, if the selected set did not include at least one neighbor above the midpoint ($f > 0.5$) and one below ($f < 0.5$), we attempted a single post-hoc swap: replacing the least-similar selected neighbor on the overrepresented side with the most-similar available candidate on the missing side, when such a candidate existed.

Each example activity in the few-shot block included (i) key metadata (title and, where available, location and a brief summary), (ii) a short "risks" summary, (iii) the retrospective mock forecast text (when enabled), and (iv) the final evaluation outcome label.

The forecasting prompt required a structured response format that explicitly considered both lower- and higher-outcome arguments on the rating scale and ended with a single-line prediction. Concretely, the model was instructed to: (1) provide reasons the overall success might be rated `{midpoint_low_text}` or lower, (2) provide reasons it might be rated `{midpoint_high_text}` or higher, (3) aggregate considerations and select exactly one of the `{num_options}` outcomes, and (4) output the final forecast on the last line beginning with `FORECAST:` followed by only the chosen option.

Finally, we appended a short description of the empirical distribution of rating outcomes in the training data to reduce mode-collapse toward a narrow subset of ratings.

2.7 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.

Ensembling All methods which demonstrated above-chance skill in forecasting were averaged using the mean value of the forecast. This was found to robustly outperform any individual forecasting method. Ensembles of LLM-generated Method #1 and Method #2 forecasts using differing random seeds and differently selected K-Nearest-Neighbors (with a different weight on most recent vs most semantically similar). A summary of the 10 pages was also included (method #3).

2.8 Scoring Metrics

RMSE (Root Mean Square Error) Take the square of the difference between every prediction and the true value, take the mean of all such squared values, then take the square root. Measure of "average" distance. Lower is better. On a scale from 0 to 5, therefore worst possible value is 5, best possible value is zero. This method heavily penalizes predictions that are significantly incorrect.

Coefficient of Determination (R^2) R^2: Coefficient of determination. Theoretically equals zero, if we always choose the mean (however using the training set mean results in a lower score on the test set in the baseline measure below). If more than 1 regressors are included, R^2 is the square of the coefficient of multiple correlation and can be negative. Measures proportion of the variation in the dependent variable that is predictable from the independent variable. Higher is better. This method generally does not penalize outliers significantly.

Binary Brier Score The binary Brier score is simply the mean squared error of the forecast (in this case, we measure whether the prediction above or below the midpoint of the scale, set to 2.5). We assume all methods assign a probability of 1 to the side predicted, and 0 to the side they did not predict.

Side Accuracy The percent of correctly predicted above 2.5 or below 2.5, out of all predictions. This has the benefit of clarity of interpretation, but has the downside that 80% of forecasts are already above 2.5 (moderately satisfactory or better). Therefore, this

```
SYSTEM:
You are an experienced international
aid decision maker with a quantitative
mindset.  Forecast the overall
evaluation rating from the options:
{options_text}.

USER:
Forecast what the outcome will be for
this activity.

### EXAMPLE ACTIVITIES ###
[For each neighbor:  title; (optional)
location/summary; risks; (optional)
example forecast; rating scale; final
evaluation outcome]

### NEW ACTIVITY TO FORECAST ###
ACTIVITY ID: {activity_id}
ACTIVITY TITLE: {activity_title}
ORIGINAL PLANNED START DATE:
{planned_start}
ORIGINAL PLANNED END DATE:
{planned_end}
ACTIVITY LOCATION(S): {location}
ACTIVITY DESCRIPTION (SUMMARY):
{gemini_generated_summary}
ACTIVITY RISKS: {risks_summary}

Provide the following format for
your response:
1.  Provide reasons why the
overall success might be rated
{midpoint_low_text} or lower.
2.  Provide reasons why the
overall success might be rated
{midpoint_high_text} or higher.
3.  Aggregate your considerations, and
decide on the final outcome among the
{num_options} options.
4.  Provide the final forecast on the
last line beginning with 'FORECAST: '
followed by only the forecast with no
extra words.

[Append:  training-set rating
distribution text]
Respond only in English.
```

(a) Method #1 (summary injection).

```
SYSTEM:
You are an experienced international
aid decision maker with a quantitative
mindset.  Forecast the overall
evaluation rating from the options:
{options_text}.

USER:
Forecast what the outcome will be for
this activity.

### EXAMPLE ACTIVITIES ###
[Same few-shot block as Method #1]

### NEW ACTIVITY TO FORECAST ###
ACTIVITY ID: {activity_id}
ACTIVITY TITLE: {activity_title}
ORIGINAL PLANNED START DATE:
{planned_start}
ORIGINAL PLANNED END DATE:
{planned_end}
ACTIVITY LOCATION(S): {location}
EXCERPTS FROM BASELINE ACTIVITY
DOCUMENTS:
{pdf_to_text_excerpts}
ACTIVITY RISKS: {risks_summary}

Provide the following format for
your response:
1.  Provide reasons why the
overall success might be rated
{midpoint_low_text}.
2.  Provide reasons why the
overall success might be rated
{midpoint_high_text}.
3.  Aggregate your considerations, and
decide on the final outcome among the
{num_options} options.
4.  Provide the final forecast on the
last line beginning with 'FORECAST: '
followed by only the forecast with no
extra words.

[Append:  training-set rating
distribution text]
Respond only in English.
```

(b) Method #2 (raw text injection).

Figure 2: Prompt templates for LLM forecasting Methods #1 and #2. The methods share the same scaffold (few-shot examples, metadata, risks, and response-format constraints) and differ only in how activity context is injected (summary vs. raw document text).

```
SYSTEM:
You are an experienced international aid decision maker with a quantitative
mindset.  Respond as if you were forecasting at the beginning of the activity
what the outcome would be, ultimately arriving at {final_result_for_prompt},
from the options of {options_text}.

USER:
You are generating example forecasts that will be used to fine tune a language
model.  Using the uploaded pages from activity documents available for the
following activity, respond as if you were forecasting only based on activity
documents and original information from the start what the outcome would be.
You will provide a well-reasoned forecast written from the perspective of an
international aid evaluator at the beginning of the activity, only at the very
end of your response arriving at the correctly forecasted evaluation success
rating of '{final_result_for_prompt}'.  Your response will be balanced and
comprehensive, including consideration of the information from the uploaded
activity documents.  Your mock forecast must adhere to the actual reasons for
the overall evaluation, as described in the pages from the uploaded evaluation
documents.

Provide the following format for your response:
1.  Provide reasons why the overall success might be rated {midpoint_low_text}.
2.  Provide reasons why the overall success might be rated
{midpoint_high_text}.
3.  Aggregate your considerations, and decide on the final outcome among the
{num_options} options (finally arriving at {final_result_for_prompt}).
4.  Provide the final forecast on the last line beginning with 'FORECAST: '
followed by only the forecast with no extra words.


The final prediction should not be made until the very end of the mock forecast.
Your mock forecast must reflect the uncertainty which would be inherent given
the information at the start of the activity.  If there is insufficient
information describing why the '{final_result_for_prompt}' evaluation was
assigned, respond only with:  "NO RESPONSE".  Respond only in English.

ACTIVITY TITLE: {activity_title}
ACTIVITY SCOPE: {activity_scope}
ORIGINAL PLANNED START DATE: {planned_start}
ORIGINAL PLANNED END DATE: {planned_end}
ORIGINAL PLANNED TOTAL DISBURSEMENT: {disbursement_total} {disbursement_units}
ORIGINAL PLANNED TOTAL LOANS AND CREDIT: {loan_total} {loan_units}
ACTIVITY DESCRIPTION FROM START: {chatgpt_description}
ACTIVITY RISKS SUMMARY FROM START: {risks_summary}

[Uploaded context:  up to 10 pages of baseline excerpts + up to 10 pages of
outcome/evaluation excerpts]
```

Figure 3: Prompt template used to generate retrospective "mock forecasts" (Gemini 2.5-pro) from retrieved baseline and outcome/evaluation document pages. Bracketed text indicates injected retrieved excerpts rather than literal prompt text.

method is less informative than R^2 or RMSE.

2.9 Outcome Grading

2.10 Predicting the Confidence of a Forecast

In addition to forecasting the outcome of an intervention, it is also useful to classify just how confident the forecast is. To do so, we prompt the model to produce a confidence score. Research on LLM confidence scoring shows that "consistency" based approaches are more accurate than verbally eliciting confidence directly from the model (Lyu et al., 2025 | "Calibrating Large Language Models with Sample Consistency"). Agreement-based consistency works best for open-source models and Codex, while entropy works best for GPT-3.5 (Lyu et al., 2025 | "Calibrating Large Language Models with Sample Consistency"). Because I already produce K independent forecasts using differing reasoning prompts, the empirical variance across sample predictions provides an agreement-based forecast confidence.

The outputs of such predictions can be very helpful when a given minimum confidence is required, such as in developmental aid and cooperation interventions affecting the environment. "Conformal prediction" is the form of prediction required from machine learning models such that the model is guaranteed to contain the ground truth within a provided probability (such as 90%) (Cherian, Gibbs, and Candès, 2024 | "Large Language Model Validity via Enhanced Conformal Prediction Methods"). Techniques above can be useful for conformal prediction methods, so users of the system can have high confidence that the forecast is correct. This also provides a significant advantage over human forecasters, who are unable to provide guarantees of the accuracy of a given prediction with any real statistical significance.

Agreement-based. For a multiset of K answers $a = a_{i}{}_{i=1}^{K}$ with most-voted answer $\bar{a} = \text{mode}(a)$,

$$\text{Agree}(a) = \frac{1}{K} \sum_{i=1}^{K} [a_i = \bar{a}].$$

Entropy-based. Let the unique-answer set be $U(a) = u_j * j = 1^m$ with empirical frequencies $p_j = \frac{1}{K} \sum [a_i = u_j] i = 1^K$ and $m = |U(a)|$. Define

$$\text{Ent}(a) = 1 - \frac{H(p)}{\log m}, \qquad H(p) = -\sum p_j{}_{j=1}^{m} \log p_j.$$

3 Results & Discussion

3.1 Database of Evaluations

In addition to producing a useful LLM forecasting system, this work has also produced a large collection of intervention outcomes in developmental aid and cooperation interventions affecting the environment. Given the absence of academic publications investigating the IATI dataset specifically regarding evaluations, this thesis provides flexible, powerful tools to gain insights from the IATI dataset. This database of 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

3.3.1 Selecting and validating GLM variables

We use a generalized linear model (GLM) baseline that maps ex-ante structured activity features to a numeric overall rating outcome. In this setting, the GLM is a linear predictor with an intercept and additive feature effects, with coefficients estimated from historical activities. To reduce overfitting and to obtain a stable, interpretable subset of predictors, we use elastic-net regularization, which combines ℓ_1 (lasso) and ℓ_2 (ridge) penalties and can both shrink coefficients and set some to exactly zero (**zouHastie2005elasticnet**).

We assembled a master dataset of activities with non-missing (i) outcome ratings and (ii) activity start dates, and constructed predictors intended to be available at or near

activity start (e.g., graded dimensions such as finance, context, risks, targets, implementer performance, integratedness, complexity; activity scope; macro covariates such as GDP per capita and CPIA score where available; financing-type indicators; organizational identity features; and simple time-derived features such as planned duration). Continuous predictors were standardized using statistics computed on the training data only.

Evaluation used four temporally defined validation sets: a single 69-activity holdout set (Validation A) corresponding to the activities used in the few-shot forecasting evaluation, and three additional 100-activity validation sets (Validations B–D) defined as the most recent 300 activities prior to Validation A, split into three contiguous blocks of 100 by activity start date. For each validation set, the training set consisted of all activities with start dates strictly earlier than the earliest start date in that validation window, excluding the validation activities themselves.

For each run, we selected the elastic-net penalty strength λ by grid search over a logarithmic grid (default $\lambda \in \{10^{-3}, \ldots, 10^{1}\}$), choosing the value that minimized validation RMSE; we also recorded validation R^2 for reporting. After selecting λ, we refit the model on the corresponding training set and recorded the set of non-zero coefficients as the selected variables.

To assess selection stability, we performed nonparametric bootstrap resampling of the training set (500 bootstrap replicates) (**efronTibshirani1994bootstrap**). In each replicate, we resampled training activities with replacement, re-selected λ using the same RMSE criterion on the fixed validation set, refit the elastic net, and recorded coefficients. We summarized each predictor by (i) selection frequency across all bootstrap replicates, (ii) selection frequency among replicates whose validation RMSE was at least as good as the baseline fit, and (iii) sign consistency (the fraction of non-zero selections sharing the modal coefficient sign). Predictors were treated as robust if they were selected in at least 60% of all bootstrap runs, selected in at least 60% of the better-than-baseline runs, and had sign consistency at least 0.9. Downstream GLM baselines were then refit using only this robust predictor set.

3.3.2 Ensuring methods improve forecasting skill

We evaluated whether each forecasting approach demonstrated out-of-sample skill exceeding chance on held-out activities. This analysis was applied to all methods considered: the GLM baseline, the KNN-based baseline, LLM Method #1, LLM Method #2, and simple ensembles formed by averaging each method with the GLM (GLM+KNN, GLM+Method #1, GLM+Method #2). Skill was assessed on each validation set (A–D) using RMSE as the primary metric.

For each method (and each GLM-averaged ensemble), we conducted a permutation-based

skill exceedance test in which the model's validation predictions were held fixed while the validation outcomes were randomly permuted many times to form a null distribution of RMSE; the empirical p-value was computed as the fraction of permutations achieving RMSE less than or equal to the observed RMSE (**good2000permutation**).

In addition, to test whether combining a method with the GLM yielded a statistically meaningful improvement over the GLM alone, we used paired bootstrap resampling of the validation activities. Specifically, we resampled validation activities with replacement (500 replicates), recomputed the RMSE difference $\Delta = \text{RMSE(ensemble)} - \text{RMSE(GLM)}$ on each replicate, and summarized the resulting distribution to obtain uncertainty intervals and an empirical one-sided significance measure based on the fraction of replicates with $\Delta < 0$ (**efronTibshirani1994bootstrap**).

Algorithm 1: Elastic-net variable selection + bootstrap
Input: training set $\mathcal{D}_{train}$, validation set $\mathcal{D}_{val}$, feature matrix X, outcomes y
1. Standardize features using $\mathcal{D}_{train}$ statistics.
2. For $\lambda \in \Lambda$ (log grid): fit elastic net; compute RMSE on $\mathcal{D}_{val}$.
3. Choose $\lambda^\star = \arg\min_{\lambda\in\Lambda} \text{RMSE}_{val}(\lambda)$.
4. Refit elastic net on $\mathcal{D}_{train}$ with $\lambda^\star$; record nonzero coefficients.
5. Bootstrap $b = 1..B$ (resample $\mathcal{D}_{train}$ with replacement):
a. Re-select $\lambda_b^\star$ by minimizing validation RMSE.
b. Refit; store coefficient vector β_b and validation RMSE.
6. Summarize each predictor by selection frequency and sign consistency; retain robust predictors.
Output: robust predictor set and refit GLM baseline.

(a) Variable selection and stability.

Algorithm 2: Skill exceedance + GLM-averaging tests
Input: validation set $\mathcal{D}_{val}$, predictions $\hat{y}^{(m)}$ for each method m.
1. Compute observed RMSE for each method m.
2. For each method m, permutation test (skill exceedance):
a. Permute validation outcomes y_{val} to y_{val}^{π}.
b. Compute $\text{RMSE}(\hat{y}^{(m)}, y_{val}^{\pi})$.
c. Empirical p-value = fraction with RMSE $\leq$ observed.
3. For each method m, test GLM-averaged ensemble:
a. Form ensemble prediction $\hat{y}^{(ens)} = \frac{1}{2}(\hat{y}^{(GLM)} + \hat{y}^{(m)})$.
b. Paired bootstrap over validation activities:
resample activities; compute $\Delta = \text{RMSE}(ens) - \text{RMSE}(GLM)$.
c. Summarize Δ (intervals; fraction with $\Delta < 0$).
Output: per-method skill exceedance evidence and ensemble improvement evidence.

(b) Skill and ensemble evaluation.

Figure 4: Overview of the feature-based model selection pipeline and the statistical evaluation used to assess forecasting skill for each method and for GLM-averaged ensembles.

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 1.4.5, 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 Avoiding Disempowerment

One additional risk to an AI forecasting system is the gradual disempowerment of humans. Many in the futurist community point to the real and currently destructive “laziness” engendered by a system that does all of the work for the human. For example, an aid grant maker may choose to simply accept all of the suggestions and ultimate forecast of an evaluation system, rather than critically examining the context and making the decisions independently. Even if an AI system could be more effective at producing a final verdict on the success of a system, human moral judgements should always be involved in the morally fraught work of international aid decision making. For example, the question of justice is an intrinsically human decision - if a country is undergoing a war, this may reduce the likelihood of success of a program, yet we must not simply aim for project success where the benefits to the people in the partner country would be enormous, if successful.

For this reason, the final probability of success was not provided to the user in the final system. Instead, all of the most relevant resources and statistics were exposed, and both relevant literature sources, academic literature, and references to wikipedia are exposed to the user, as well as various questions posed to the system and its method of breaking down the forecasting problem.

4.10 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).

Ashton, Helen Louise et al. (Dec. 2021). "A Puzzle with Missing Pieces : Explaining the Effectiveness of World Bank Development Projects". In: *Policy Research Working Paper Series* 9884. (Visited on 09/03/2025).

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).

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.

Bulman, David, Walter Kolkma, and Aart Kraay (Sept. 2017). "Good Countries or Good Projects? Comparing Macro and Micro Correlates of World Bank and Asian Development Bank Project Performance". In: *The Review of International Organizations* 12.3, pp. 335–363. ISSN: 1559-744X. DOI: 10.1007/s11558-016-9256-x. (Visited on 09/02/2025).

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).

Cherian, John J., Isaac Gibbs, and Emmanuel J. Candès (Dec. 2024). "Large Language Model Validity via Enhanced Conformal Prediction Methods". In: *Advances in Neural Information Processing Systems* 37, pp. 114812–114842. (Visited on 08/31/2025).

"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).

Datenlabor-Bmz/Awesome-Development-Cooperation-Data (June 2025). Datenlabor BMZ. (Visited on 08/22/2025).

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).

Eilers, Yota et al. (2025). "Volume, Risk, Complexity: What Makes Development Finance Projects Succeed or Fail?" In: *The World Bank Economic Review* (). DOI: 10.1093/wber/lhaf001. (Visited on 09/02/2025).

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).

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).

"Good Countries or Good Projects?" (Nov. 2013). "Good Countries or Good Projects? Macro and Micro Correlates of World Bank Project Performance". In: *Journal of Development Economics* 105, pp. 288–302. ISSN: 0304-3878. DOI: 10.1016/j.jdeveco.2013.06.003. (Visited on 09/02/2025).

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: ().

How Can You Use the Social Science Prediction Platform for Development Papers? (2025). https://blogs.worldbank.org/en/impactevaluations/how-can-you-use-the-social-science-prediction-platform-for-devel. (Visited on 08/22/2025).

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).

IATI Dashboard – IATI Activities (2025). https://dashboard.iatistandard.org/exploring-data/activities/. (Visited on 10/18/2025).

"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).

Kaiser, Micha et al. (Dec. 2025). "Leveraging LLMs for Predictive Insights in Food Policy and Behavioral Interventions". In: *Discover Food* 5.1, pp. 1–25. ISSN: 2731-4286. DOI: 10.1007/s44187-025-00552-x. (Visited on 10/17/2025).

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).

KfW Development Bank (2025). *IDEaL*. (Visited on 08/22/2025).

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).

Net ODA | OECD (2025). https://www.oecd.org/en/data/indicators/net-oda.html?oecdcontrol-03506f24e9-chartId=bb70a1f537&oecdcontrol-f42fb73652-var3=2023. (Visited on 09/02/2025).

Olken, Benjamin A. (July 2020). "Banerjee, Duflo, Kremer, and the Rise of Modern Development Economics*". In: *The Scandinavian Journal of Economics* 122.3, pp. 853–878. ISSN: 1467-9442. DOI: 10.1111/sjoe.12418. (Visited on 09/02/2025).

"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).

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.

Sustainability (IDOS), German Institute of Development and (2025). *Learning from KfW's ex-post evaluations? How conflicting objectives can limit their usefulness.* https://www.idos-research.de/discussion-paper/article/learning-from-kfws-ex-post-evaluations-how-conflicting-objectives-can-limit-their-usefulness-1/. (Visited on 09/02/2025).

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).

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems.* Vol. 30. Curran Associates, Inc. (Visited on 10/18/2025).

Vivalt, Eva (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, 16 December 2025