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, 14 January 2026

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 validation set, 300 later-starting activities in a dataset of 1,300 environmental and sustainability improving activities for 4 reporting organizations, the full forecasting system improves the prediction from 70.0% to 73.0% accuracy compared to a "pick the most common rating" baseline. For overall success ratings on a scale from approximately 1 to 6, the system can explain 19.4% of the variance in ratings compared to 10.2% for the baseline. I also produce an aggregate success metric averaging the z-score for ratings and commonly quantified activity outcomes relating to activity cost-effectiveness, and find the same model explains 22% of the variance in this metric. 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), protein folding (Jumper et al., 2021), biodiversity protection (Silvestro et al., 2022), and weather forecasting (Lam et al., 2023).

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) 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). As of 2022 pollution remains responsible for approximately 9 million deaths per year, corresponding to one in six deaths worldwide (Fuller et al., 2022).

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). 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* 2025). 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), 2025). 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 applicable to context-rich 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). 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 from various international aid organizations 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., n.d.).

This method differs in two key ways from prior literature. The first difference is that while many works in the literature have attempted to assess correlations between quantitative features and aid evaluation ratings, they have not focused on what knowledge would be available near the start of the activity, and do not assess out-of-time generalization of these correlations. In this work, I assess out-of-time generalization of feature importance. This is critical, because in order to improve aid decision making, one must assess the ability of models to forecast the outcomes of future interventions. The second difference is that in addition to standard statistical models, I implement judgemental AI forecasting, as a supplement and even a competitor to standard statistical 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.

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 a large 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), 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). 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.

One attempt which focused on specific outcomes of activities found their model using "embeddings" of LLMs could explain 70% of the variance of these outcomes, and assessed the performance of nonlinear models including the random forest model used in this work. However, they include features that could not be known at the beginning of the activity (e.g. actual duration), and do not assess out-of-time generalization, instead splitting randomly within the dataset, nor do they explicitly assess prediction performance for ratings

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 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

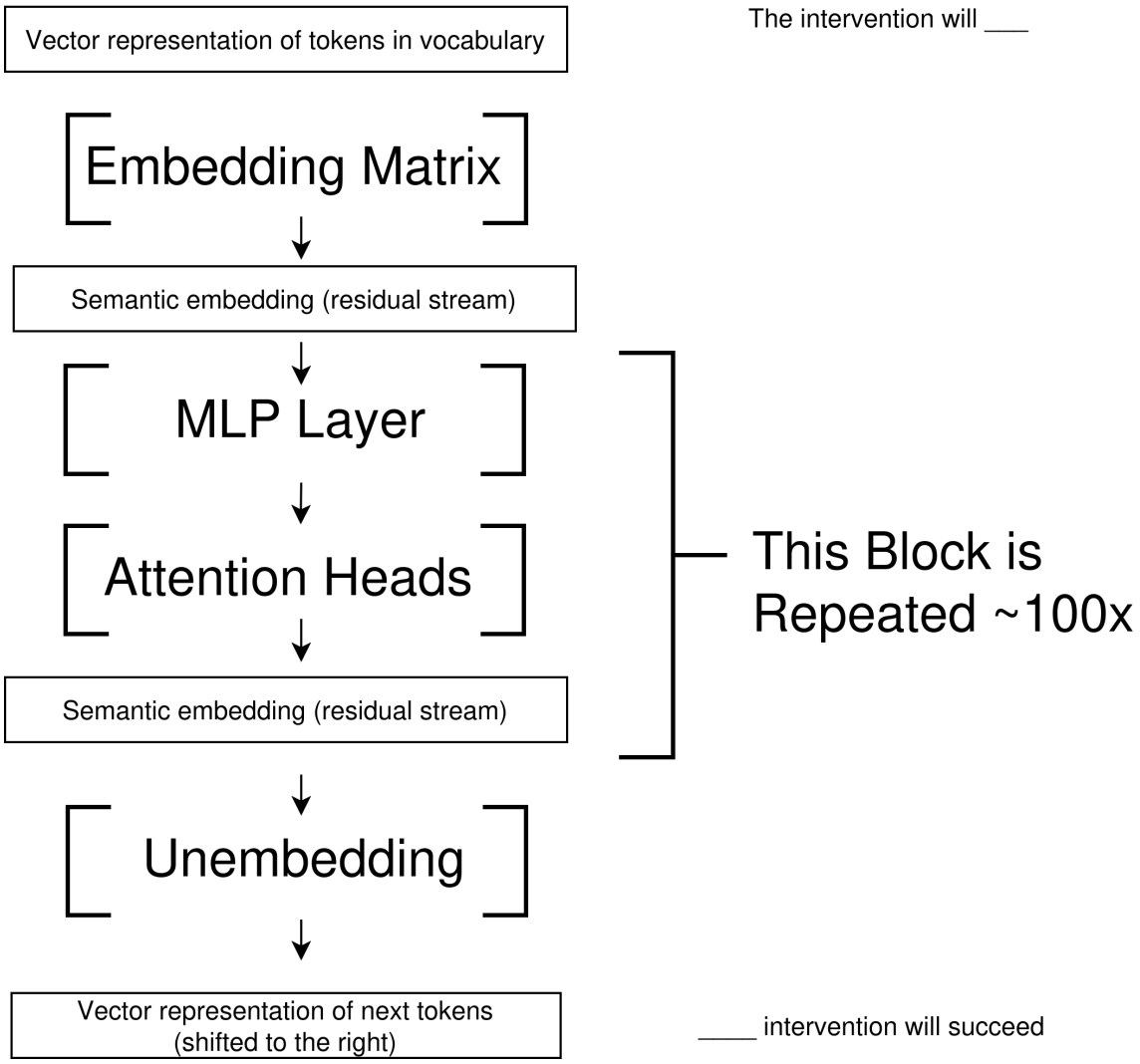

Figure 1: A simplified decoder-only transformer architecture.

were trialed on the validation set. Finally, the most promising version was used to predict evaluation outcomes on hundreds of evaluations.

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 the construction of a solar power plant 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. Recently, the number of grant evaluators at BMZ has been reduced. This work may also assist the remaining grant evaluation to make the most of the time they do have available.

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) (Schoenegger and Park, 2023). 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), 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). (*Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy | Science Advances* 2025) (Abolghasemi, Ganbold, and Rotaru, 2025) (Yan, Q., Seraj, R., He, J., Meng, L., and Sylvain, T., 2024).

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), 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).

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). 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))

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 Lu, 2025).

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., n.d.). 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). 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).

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). 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). 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) 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).

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)(Schoenegger, Jones, et al., 2025). 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.

1.4 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) (Schoenegger and Park, 2023). 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., n.d.). 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). 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). 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). 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). 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). This finding mirrors that of the neuroscience study (X et al., 2025) 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).

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) (See Table **??** and section 2.7). 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) (Bang et al., 2024).

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).

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) (Karger et al., 2024): 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).

In order to guard against leakage of information from the training, I selected deepseek as my forecasting model, due to its strong performance comparable to other models which have similar training cutoff dates (2023 for Deepseek V3.2).

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 on a six point scale within linked evaluation pdfs. 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 always 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, leading to a further exclusion of environmental or sustainability related activities.

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. An exception was project appraisal documents from the World Bank, or Project Information Documents, which were found to reliably not leak future information, and this was judged to be more trustworthy than extracting the creation date embedded in the activity document. Activities not meeting these requirements were also excluded, leaving 3,225 activities.

After passing all these filters, I finally attempted to extract the activity rating from the evaluation document using two separate methods. Because rating tendencies are systematically different for different reporting orgs, I needed sufficient data for training, validation, and testing for all organizations. I also restricted activities to those that were marked as "completed" in order to ensure comparability between rating scales, as the only activities not marked as “completed” were relatively recent and would dominate the held-out test set. I determined there were sufficient data for four reporting organizations:

The World Bank (957 activities), BMZ/KFW/GIZ (240 activities), the Asian Development Bank (ADB) (156 activities) and the UK Foreign Commonwealth and Development Office (FCDO) (127 activities).

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

- 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
- 23610: Heat plants
- 23630: Electric power transmission and distribution (centralised grids)
- 23631: Electric power transmission and distribution (isolated mini-grids)
- 23642: Electric mobility infrastructures
- 31130: Agricultural land resources
- 31210: Forestry policy and administrative management
- 31220: Forestry development
- 31281: Forestry education/training
- 31282: Forestry research
- 31291: Forestry services
- 32174: Clean cooking appliances manufacturing

- 41010: Environmental policy and administrative management
- 41020: Biosphere protection
- 41030: Biodiversity
- 41081: Environmental education/training
- 41082: Environmental research

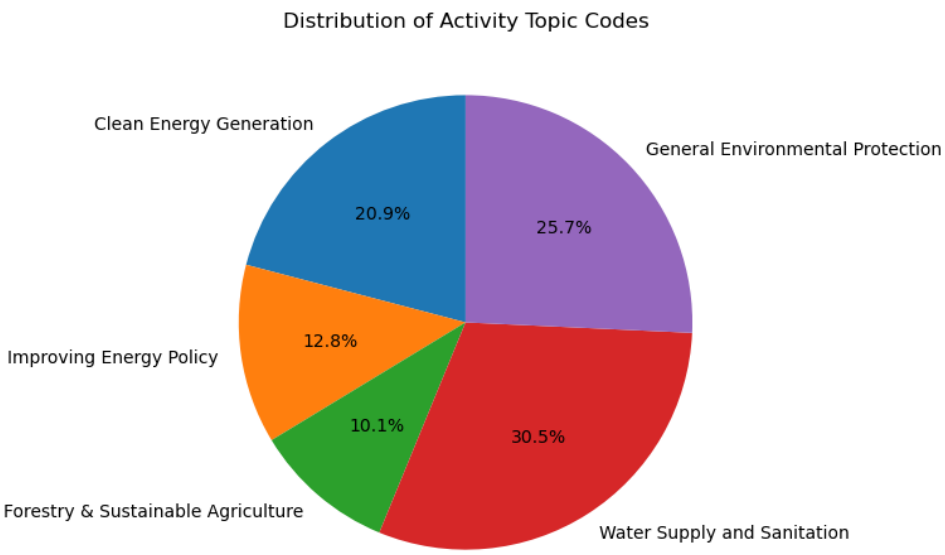

Figure 2: The split of topics analyzed from the dataset.

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. 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 the following 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”, “Mem-

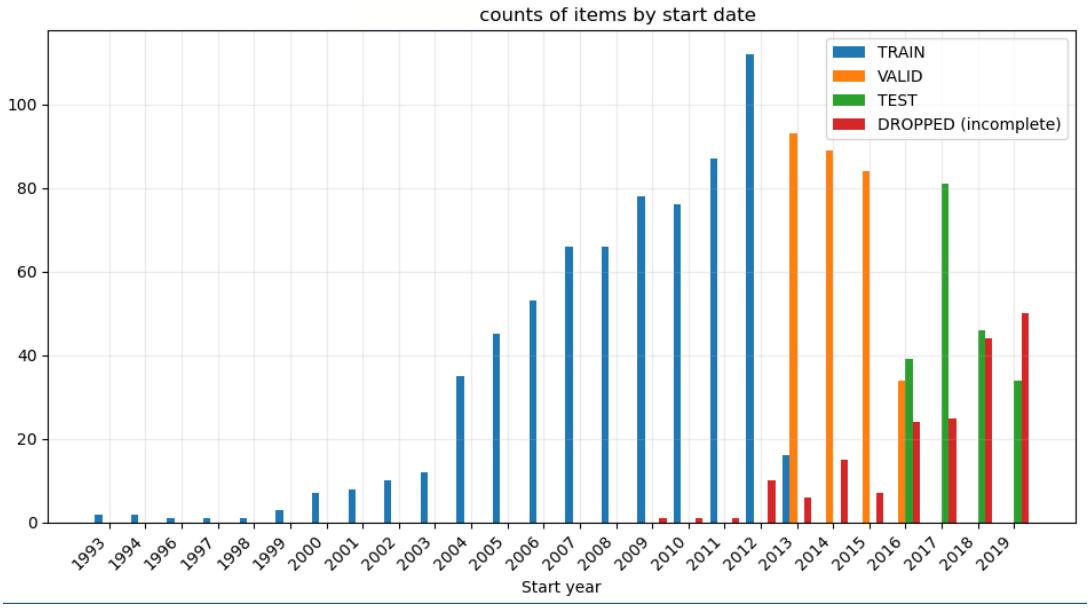

Figure 3: The activities included for predicting ratings, with the splits by count year. Incomplete activities, shown in red, were not used for prediction. Activity ids used for outcome prediction were given the same temporal boundaries.

orandum of understanding (If agreed by all parties)", "Tender", or "Contract".

I mark documents with "Objectives / Purpose of activity" or "Summary information about contract", 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 provide sufficient information to forecast with and sufficient information to evaluate that forecast, I require at least one "baseline" document and at least one "outcome" document per activity, with the baseline document at least one year prior to the 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 which were "Conditions", "Budget", "Tender", "Contract" with no other tags were excluded, as these were typically purely legal context, often containing very little evaluation or useful additional activity information.

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

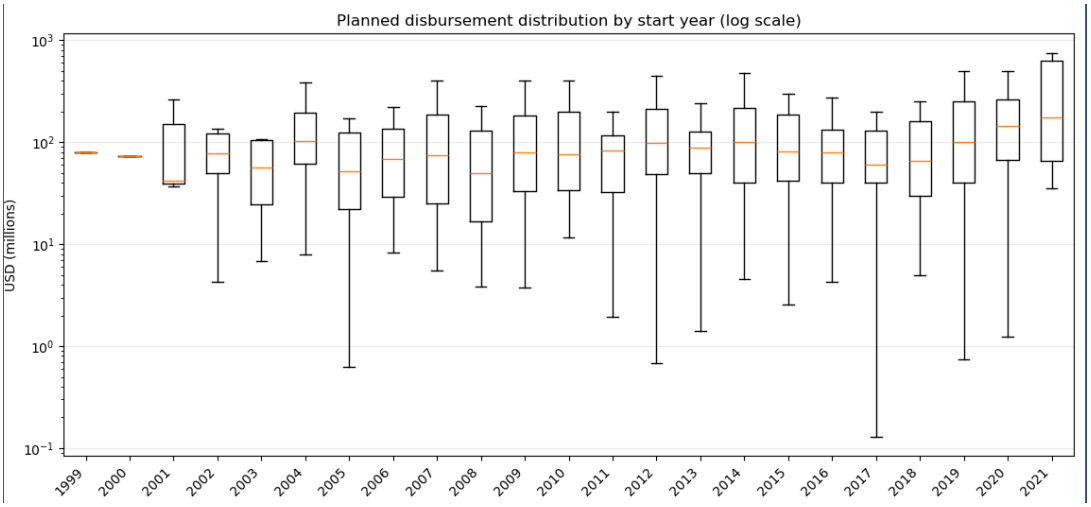

Figure 4: Total disbursements for IATI activities used for ratings, by year. There is no clear trend over time for activity size in the database.

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*.

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

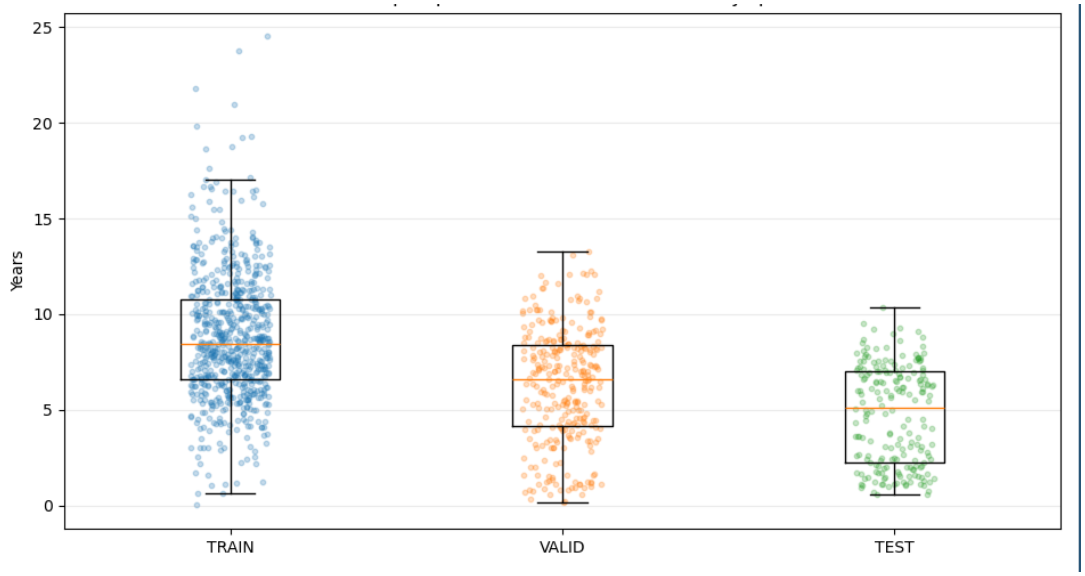

Figure 5: The durations of activities per split. More recently starting activities tend to be shorter, as they have not yet had time to complete and be evaluated. The out-of-distribution nature of validation and test sets increase the challenge of generalizing patterns from the training data.

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. After ranking, 2,312 documents had sufficiently informative activity information and activity evaluation documents.

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 activity. 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 contents", "references", or "activity description" were included and *gemini-2.5-flash* was queried to extract the overall rating from the documents.

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.

4. Interpreting Ratings Ratings were reported both with the rating itself, as well as a maximum and minimum possible rating. The World Bank rating scale from 1 ("Highly Unsatisfactory") to 6 ("Highly Satisfactory") was used as the template rating, and other ratings were attempted to match against this scale. Notably, BMZ/GIZ/KFW ratings were inverted to reach this scale. A "Satisfactory" score was considered equivalent to scores such as "successful", "On Track", or "met expectations". Scores listed as percentages or fractions were re-scaled to the 1 to 6 scale as well. In order to ensure ratings were fairly compared, only the top four most common organizations with ratings were included for training and validation of the forecasting system.

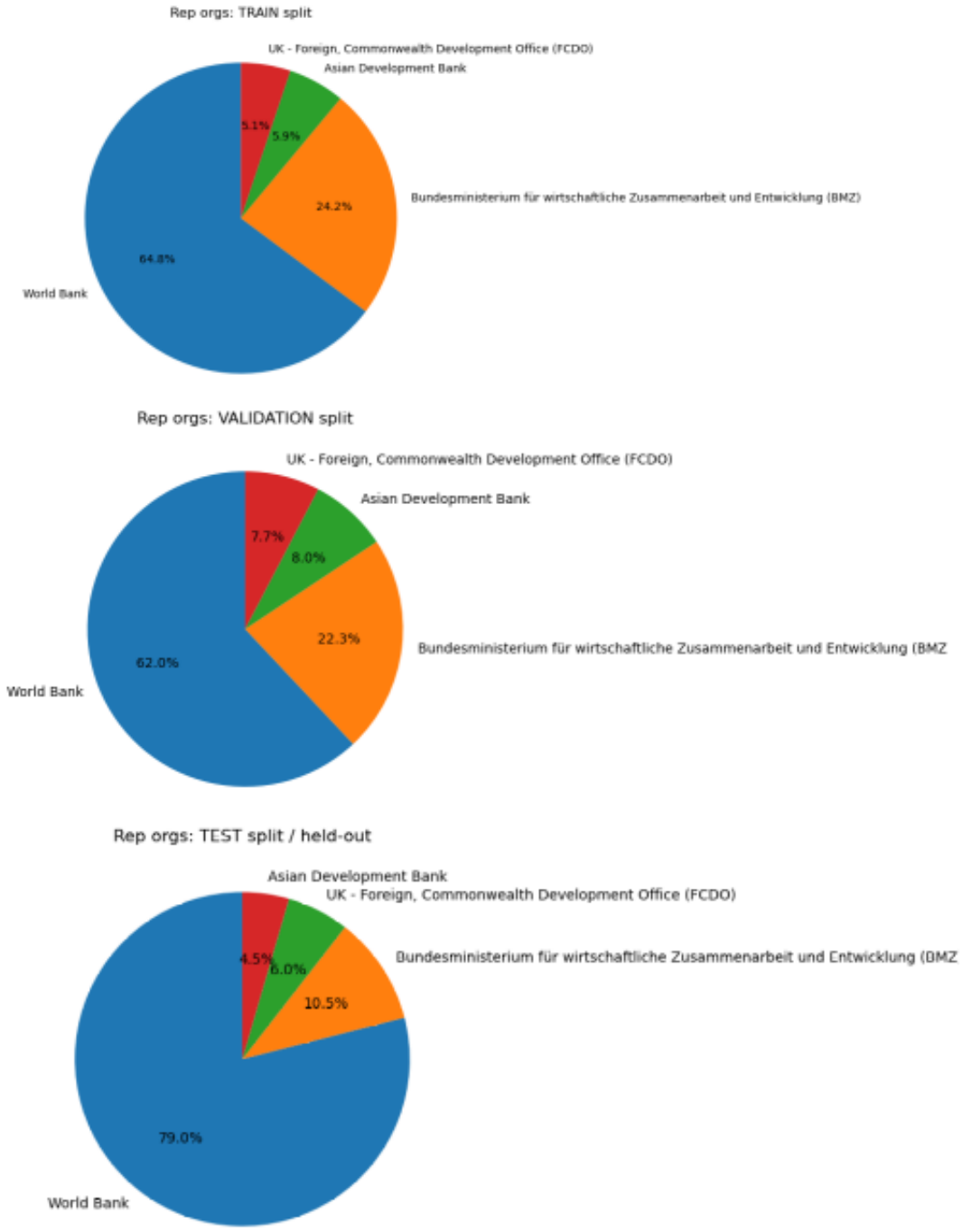

Figure 6: The breakdown of reporting orgs in the dataset which were used for training, validation and testing. The validation and test set are in this way significantly out-of-distribution.

‘’

2.4.1 Outcome Extraction

In addition to extracting ratings, a similar approach was used with *gemini-2.5-flash* to extract quantitative outcomes. I extracted quantitative outcomes from all pages which were categorized as outcomes, and marked as containing quantitative information (see Figure 7). Unlike with ratings, I did not limit to the top 4 reporting organizations, as reporting ratings is more susceptible to between-reporting-organization variation and gaming than the reportable quantitative outcomes of projects.

Once these PDF pages were extracted, I employed a combination of manual examination of the extracted outcomes and a report of common bigrams to identify outcome variables that could be compared between projects. For each common outcome category, I came up with a list of words and phrases that would commonly match reports of these outcome variables in the description, as well as appropriate units.

2.4.1.1 Keyword- and unit-based outcome parsers. For each outcome category, I implemented a dedicated parser that scans the extracted quantitative description, baseline, target, outcome, and units. Each parser follows the same general pipeline: (i) *filtering* by description keywords and unit constraints, (ii) *sanitization* of numeric values (e.g., dropping negative sentinels or implausible magnitudes), (iii) *normalization* to a canonical unit (e.g., hectares, tonnes, people), and (iv) *aggregation* to a single activity-level outcome.

2.4.1.2 Comparable outcome categories. Using manual inspection and frequency statistics over common bigrams, I defined a set of outcome categories that recur across evaluations and are interpretable across projects. The final set included:

- **Cost–benefit ratios (B/C):** benefit-cost ratio outcomes.
- **Rates of return:** economic rate of return (ERR/EIRR) and financial rate of return (FRR/FIRR), in percent.
- **Beneficiaries:** counts of beneficiaries (people).
- **Emissions reductions:** CO_2 or CO_2e reductions (total or per-year) in tonnes.
- **Water and sanitation connections:** counts of service connections, either new or repaired.
- **Pollution load removed:** wastewater pollutant load reductions (e.g., BOD/COD/nutrients), in tonnes and categorized by time basis (total, per-year, per-day).
- **Forest indicators:** trees/seedlings planted (counts) and area-based forest outcomes (reforested, under management, protected) in hectares.

```
SYSTEM:
You are extracting structured quantitative information from evaluation PDFs.
Return only valid JSON that matches the provided schema.  Do not include any
other text.

USER:
You are extracting information from the attached single page of an activity
evaluation PDF.
Extract quantitative outcomes that are due to the activity.


CONTEXT:
- ACTIVITY TITLE: {activity_title}
- ACTIVITY DESCRIPTION: {activity_description_truncated}
- PAGE NUMBER: {page_number}
- ACTIVITY TIMEFRAME: {start_date} to {end_date}


INSTRUCTIONS:
- For each indicator/metric found, add an entry to quantitative_outcomes.
- Each entry must include a short description of the metric and the
outcome_value.
- Where explicitly available, also include:
baseline_value (before the activity) and target_value (intended for the
activity).
- Use units only when the page clearly specifies them; keep units short (e.g.,
%, count, people, USD).
- Only report quantities that reflect the conclusion of the activity or the full
activity period.
- Leave out entries where it is unclear how to correctly enter them.


OUTPUT JSON SHAPE (schema-enforced):
"quantitative_outcomes":  [
"description":  "...",
"baseline_value":  <number optional>,
"target_value":  <number optional>,
"outcome_value":  <number required>,
"units":  "...  optional"
]
```

Figure 7: Quantitative outcome extraction prompt used with structured JSON output. The model receives a single-page PDF slice and returns a list of outcome metrics with optional baseline, target, and units where explicitly stated.

- **Irrigation outcomes:** increases in irrigated area (hectares), computed as a positive increase relative to baseline where available.
- **Energy outcomes:** installed generation capacity, in MW (or occasionally GWh where the source reported capacity in energy units).
- **Air quality (PM2.5):** PM2.5 reductions reported as concentration, emissions, or percent, kept as separate distributions.
- **Clean cooking stoves:** counts of stoves distributed/installed.
- **Agricultural yields:** yield increases expressed either as level changes (normalized to tonnes per hectare when possible) or as percent increases.

2.4.1.3 Implementation details by category. Each category required a custom python script, primarily because evaluation reports often contain multiple related indicators (e.g., component-level versus project-level rates of return) and frequently use heterogeneous units or phrasing.

Benefit–cost ratios (B/C). B/C ratio records were identified via a description keyword list (e.g., "benefit/cost ratio", "B/C") and a strict unit constraint: retained records must have empty units (since B/C is dimensionless). Non-numeric entries and implausible magnitudes were discarded. When multiple B/C indicators were present for an activity, regex heuristics were used to ranked records that explicitly referred to whole-project or overall ratios above component-specific ratios. The final activity-level B/C ratio was computed as the mean outcome across all indicators at the highest scope tier, as multiple extracted B/C quantities sometimes report the same project-wide metric in slightly different wording.

ERR/EIRR and FRR/FIRR. Rates of return were detected using keyword lists for economic and financial variants, with an explicit unit requirement of percent. Similar to B/C, I assigned a scope score to prioritize project-level rates over subcomponent rates. Where both before-tax and after-tax versions appeared, I preferentially retained before-tax values when both were present. In some evaluations, both "with" and "without" environmental co-benefits variants are reported; to improve comparability, I preferentially retained "without/excluding" versions when available (and otherwise dropped "with" versions only if doing so did not eliminate all candidates). Final activity-level rates were computed by averaging outcomes within the best scope tier.

Beneficiaries. Beneficiary outcomes were identified by requiring beneficiary language in the description and count-like units (e.g., people/persons/beneficiaries, including

"thousands" or "millions") while excluding monetary and percentage units. Values were normalized to people by applying multipliers implied by the unit string. For each activity, I selected the maximum realized beneficiary count among candidate records, reflecting the common structure of reporting multiple beneficiary subgroups where the overall total is typically the largest.

CO_2/CO_2e emission reductions. Emissions reductions were detected by combining (i) CO_2/GHG cues in either units or the description and (ii) reduction/avoidance language (e.g., "reduced", "avoided", "abated"). Unit parsing normalized values into tonnes using textual multipliers (thousand/million), and standard abbreviations (kt, Mt, MMT), plus kg-to-tonnes conversion when needed. Each record was classified as either a total reduction or a per-year reduction based on explicit annual language in the description or unit denominators. For each activity, I selected the largest reduction value within the preferred type tier (favoring totals when available, otherwise annual rates), yielding a single activity-level reduction with a canonical unit label.

Water and sanitation connections. Connections are prone to false positives (e.g., electricity grid connections, gas hookups, staffing-per-connection ratios). To reduce such errors, I required: (i) a connection token match (including multilingual variants), (ii) explicit water/wastewater/sanitation context in the description, and (iii) count-like units (or empty units when the description clearly conveyed the meaning). Candidate records were scored to prioritize household/domestic service connections and newly provided/functional connections. The chosen activity-level value was the highest-scoring record, breaking ties by larger realized counts.

Pollution load removed (wastewater). Pollution load outcomes were identified using reduction language (removed/reduced/abated/avoided) combined with "load"/"pollution"/"discharge" terms and mass-like units (tonnes/kg and variants). Because "load reduction" appears in unrelated contexts, I imposed a wastewater context filter (e.g., sewage, WWTP, effluent) and/or explicit pollutant tags (BOD, COD, nitrogen/phosphorus/TSS). I classified the time basis (per-year, per-day, or total) using unit and description patterns (e.g., "t/yr", "per annum"). Numeric values were normalized to tonnes. Records were grouped by (activity, pollutant tag) and selected by a priority rule emphasizing wastewater context and explicit per-year (or per-day) reporting; within the best tier, I selected the maximum normalized outcome.

Forest outcomes. Forest-related indicators were split into conceptually distinct distributions: (i) trees/seedlings planted (counts), (ii) area re/afforested or restored (hectares), (iii) forest area under management (hectares), and (iv) area protected (hectares). Area

records were retained only for pure area units (excluding per-hectare denominators and monetary units) and normalized using thousand/million modifiers. Tree-count records were retained only when units and description aligned with tree/seedling language and did not resemble beneficiary or other count metrics. For each activity and forest subcategory, I selected the maximum realized outcome (and retained associated baseline/target when available), reflecting the reporting convention that the most comprehensive figure is typically the largest.

Irrigated area increases. Irrigation outcomes were detected by irrigation/drainage keywords plus area units. Values were normalized to hectares using unit-specific conversions (e.g., acres, feddan, mu, and thousand/million hectare abbreviations). To express the outcome as an increase attributable to the activity, I computed `outcome_ha - baseline_ha` when baseline existed; otherwise, I treated the outcome as the increase (common in "area provided with improved irrigation" indicators). I retained only positive increases and then selected the maximum increase per activity.

Generation capacity. Energy generation capacity indicators were identified through "capacity" language combined with power-sector context (electricity/power/renewable/plant/turbine) and explicit power or energy units. I excluded common false positives such as "capacity building" and transport/wastewater contexts. Values were normalized to MW (or, in a small number of cases, to GWh when the reports used energy units to describe capacity), and per-activity selection favored total (stock) capacity over per-year variants where both appeared.

PM2.5 reductions and clean cooking stoves. PM2.5 indicators were detected using explicit "PM2.5" mentions and then split into three separate distributions based on units: concentration reductions (μg/m^3), emissions reductions (tons), and percent reductions (%). When baseline and outcome were available and the description did not already imply a reduction amount, I computed reductions as `baseline - outcome` where outcome improved relative to baseline. Clean cookstove outcomes were detected by stove keywords and count-like units, normalized using thousand/million modifiers, and aggregated as the maximum stove count per activity.

Agricultural yields. Yield-related outcomes were split into (i) level changes that could be normalized to tonnes per hectare, and (ii) percent yield increases. For level yields, I required crop/agriculture context to avoid financial "yield" false positives and normalized common yield units (e.g., tons/ha, tons/feddan, kg/mu) into t/ha. Yield increases were computed as `outcome - baseline` and retained only when positive. For each activity, I

aggregated by taking the mean increase across multiple yield indicators of the same type, yielding one per-activity value for t/ha increases and one for percent increases.

2.4.1.4 Aggregation and comparability strategy. Finally, I used the total disbursement for the activity reported by IATI and determined a rough estimate of the USD per unit outcome, except for Benefit-Cost Ratios, Rates of Return, and agricultural yield outcomes. I found on investigation that most outcomes were approximately log-normally distributed. I took the log10 of all categories except Benefit-Cost ratios, rates of return, and agricultural yields.

Splitting disbursements Unfortunately, I did not have access to outcome-level funding splits from the IATI database. In order to roughly represent the fact that dollar-per-unit spending can be allocated across several outcomes, I wrote a custom algorithm to evenly allocate total activity expenditures to what are usually distinctly funded outcomes. My procedure assigns each activity's total expenditure across the outcome components it reports, so later cost-per-unit calculations do not implicitly treat multi-outcome activities as having multiple full budgets. Components fall into three behaviors: outcomes relating to "beneficiaries reached" are always assigned the full budget, benefit/cost ratios and economic and financial rate of return are excluded from monetary allocation , and the rest of the outcomes are eligible for splitting. To avoid double-counting when two indicators are simply alternative measurements of the same underlying result, closely related indicators are first grouped into shared conceptual buckets, such as pairing protected area with area under management, pairing different yield-increase measures, and pairing tree planting with reforested area.

Once the components are bucketed, the algorithm gives each bucket an equal share of the activity's allocatable funding. Every component inside a bucket inherits that same share, meaning components that are "alternative measures of the same thing" share one slice rather than each taking a slice.

Carbon dioxide reductions are handled as a special case because they can act as a summary metric that overlaps with other mitigation outputs. If CO_2 reductions are reported without any closely linked mitigation outputs (such as improved stoves, added generation capacity, or forest-related actions like planting or management), then CO_2 reductions receive an equal share like any other bucket. If CO_2 reductions are reported alongside any of those linked outputs, it inherits the combined allocation already assigned to the linked outputs present for that activity. This prevents CO_2 from inflating allocated spending when it is essentially a derivative or co-reported consequence of the underlying mitigation actions.

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 for activity success forecasting.

Prediction baseline: always predict the most common rating for the reporting organization 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

- planned activity duration
- planned total disbursement
- whether the activity is primarily loan or grant-based
- the one-hot encoded reporting organization
- the Country Policy and Institutional Assessment (CPIA) score from the World Bank for that country
- the scope of the activity on a scale from 1-7, ranging from local to global
- the log(GDP/capita) of the countries where the activity takes place weighted by the percentage of the activity performed in each country.
- *gemini-2.5-flash*-generated evaluation on a score from 0 to 100 of:
  - how well financed the activity is
  - the activity integratedness within the broader activity ecosystem
  - the expected implementer performance
  - the ease of targeted outcomes
  - the degree of contextual challenge
  - the overall risk level
  - the activity's overall technical complexity.

The activity start date was not used, as there was no clear linear pattern with regards to overall activity success over time in the training data.

2.6 Experimental methods

2.6.1 Non-Parametric Bootstrap

The non-parametric bootstrap is a method used to diversify the training data, increasing the diversity in models that are trained many times. It can be used both for ensuring methods robustly improve performance on a diversity of different training setups, and in the case of training the random forest, increases independence between trees. This works by randomly sampling the same number of samples as exist in the training set, with replacement (the same training point may repeat more than once, at random).

2.6.2 GLM using IATI Features and Grades

Similar to the baseline prediction, the GLM is trained with ridge regression to reduce overfitting on noise and improve generalization.

2.6.3 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 *gemini-2.5-flash*-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, and 20 nearest neighbors using 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. Although the nearest neighbor method was used to collect examples for the LLM prompt, it was found that simply taking the weighted mean of ratings underperformed the "most common rating" method.

2.6.4 Random Forest

The Random Forest method is a statistical algorithm which constructs an ensemble of decision trees which would produce the correct output on the training data, and averages those decision trees. The averaging nature of the random forest algorithm reduces overfitting on the training data. The algorithm is inherently "regularized", penalizing an overly complex decision tree. The decision trees split based on value ranges of the features. By reducing the depth of the trees (the number of decision points where the decision tree splits), we can reduce the memorization of the training data from the trees, and

improve generalization of the model. Each decision also only considers a random fraction of the features, encouraging each tree to be more independent of each other and improving generalization further. The bootstrap method is also used to train trees, encouraging tree independences.

2.6.5 LLM Forecasting Method

The LLM forecasting method was decided upon by iteratively inspecting both the quality of the response, and the overall accuracy of the predictions made by the LLM. To generate the LLM forecasting methods, *gemini-2.5-flash* 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 pages marked as "deviations from plans", "delays or early completion", or "over or under spending" with a minimum forecasting relevance score of 3/10, and otherwise returning the pages with the highest forecasting relevance score.

Figure 8 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 ended approximately at or before the start of the activity being forecasted was then retrieved (see Section 2.6.3. 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, as well as a *gemini-2.5-flash*-generated summary.

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

2.6.6 Further Details on LLM Forecasting Method

The full prompt template for the LLM Forecast is shown in Figure 9.

Few-Shot Block In both methods, I use a k-nearest-neighbors (KNN) few-shot block of semantically similar activities in the training data (see Section 2.6.3 for how semantic similarity was determined). I selected a range of nearest neighbors. I asked the language model to extrapolate lessons about rating scales for the most similar "Highly Unsatisfactory", "Unsatisfactory" or "Moderately Unsatisfactory", the most similar "Moderately Satisfactory", and the most similar "Satisfactory" or "Highly Satisfactory" rated examples in the training data. A selection of $k = 3$ summarized mock forecasts was found to perform better $k = 1$, 5, or 7.

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 analysis, and (iv) the final evaluation outcome label.

Additional Prompts Two additional prompts were given, and inserted into the final forecast: (1) reasons the activity may have been evaluated as "Moderately Satisfactory" or worse, (2) reasons the activity may have been evaluated as "Moderately Satisfactory" or better.

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

```
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 8: 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.

```
SYSTEM:
You are an experienced international aid decision maker with a quantitative
mindset.  Respond with a comprehensive, thorough forecast of what the overall
evaluation rating of the activity will be, from the options of {options_text}.

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

### Lessons from similar activities ###
{knn_summary_text}
### End lessons ###

### Additional specific information about the activity that you summarized ###
{rag_synthesis_additional_info}
### End of additional information you summarized ###

ACTIVITY ID: {activity_id}
ACTIVITY TITLE: {activity_title}
ORIGINAL PLANNED START DATE: {planned_start}
ORIGINAL PLANNED END DATE: {planned_end}
ACTIVITY SCOPE: {activity_scope}
PLANNED TOTAL DISBURSEMENT (USD): {planned_total_disbursement_usd}
ACTIVITY LOCATION(S): {locations}
LOCATION GDP PER CAPITA, USD: {gdp_percap}
PARTICIPATING ORGANIZATIONS: {reporting_orgs}
IMPLEMENTING ORGANIZATION CATEGORY: {either "Government" or "NGO", otherwise
line not inserted}

ACTIVITY DESCRIPTION: {chatgpt_description}
ACTIVITY TARGETS: {targets_summary}
ACTIVITY CONTEXT: {activity_context}
ACTIVITY COMPLEXITY: {complexity_details}
ACTIVITY INTEGRATEDNESS: {how_integrated_description}
FINANCING DETAILS: {finance_summary}
IMPLEMENTER PERFORMANCE CONTEXT: {implementer_performance_text}
ACTIVITY RISKS:
{risks_summary}
ACTIVITY POSSIBILITIES: {possibilities_summary}

{training-set rating distribution text}
Here are a few reasons that you said the answer might be "Moderately
Satisfactory" or worse:
{insert_stage_s1_answer_here}
Here are a few reasons that you said the answer might be "Satisfactory" or
better:
{insert_stage_s2_answer_here}

YOUR TASK:
Aggregate your considerations above.  Think like a superforecaster (e.g. Nate
Silver).  On the very last line of your response, write 'FORECAST: ' followed
by exactly one option from this rating scale with no extra words:
{options_text}

Respond only in English.
```

Figure 9: Single-method multi-stage forecasting prompt. Stages s1 and s2 are run as separate calls, and their outputs are inserted into the final (s3) prompt via {insert_stage_s1_answer_here} and {insert_stage_s2_answer_here}.

```
SYSTEM:
You are an experienced international aid decision maker with a quantitative
mindset.  Your job is to identify missing-but-important information from activity
information documents and produce search phrases to find them in the documents.
Only target facts that would be knowable at the start of the activity; ignore later
implementation results.  You format the final lines of your response with exactly 5
phrases, with one line each per phrase:
PHRASE 1:  <query>
PHRASE 2:  <query>
PHRASE 3:  <query>
PHRASE 4:  <query>
PHRASE 5:  <query>

USER:
First, consider what information is available, and what is generally unavailable
but would be useful to know, in order to forecast the activity outcome on the
following scale:  {options_text}.
Second, generate five short search phrases to look up in the activity's documents,
customized to fill key informational gaps.

EXAMPLE QUERY PHRASES:
[Injected:  randomized list of example phrases]

ACTIVITY ID: {activity_id}
ACTIVITY TITLE: {activity_title}
ORIGINAL PLANNED START DATE: {planned_start}
ORIGINAL PLANNED END DATE: {planned_end}
ACTIVITY SCOPE: {activity_scope}
PLANNED TOTAL DISBURSEMENT (USD): {planned_total_disbursement_usd}
ACTIVITY LOCATION(S): {locations}
LOCATION GDP PER CAPITA, USD: {gdp_percap}
PARTICIPATING ORGANIZATIONS: {reporting_orgs}
IMPLEMENTING ORGANIZATION CATEGORY: {implementing_org_type}
ACTIVITY DESCRIPTION: {chatgpt_description}
ACTIVITY TARGETS: {targets_summary}
ACTIVITY CONTEXT: {activity_context}
ACTIVITY COMPLEXITY: {complexity_details}
ACTIVITY INTEGRATEDNESS: {how_integrated_description}
FINANCING DETAILS: {finance_summary}
IMPLEMENTER PERFORMANCE CONTEXT: {implementer_performance_text}
ACTIVITY RISKS:
{risks_summary}
ACTIVITY POSSIBILITIES: {possibilities_summary}

Provide the following format for your response:
COMPREHENSIVE REASONING ABOUT GOOD PHRASES: <extensive reasoning>
PHRASE 1:  ...
PHRASE 2:  ...
PHRASE 3:  ...
PHRASE 4:  ...
PHRASE 5:  ...
Respond only in English.
```

Figure 10: RAG phrase generation prompt (produces 5 document search phrases).

```
SYSTEM:
You are an experienced international aid decision maker with a quantitative
mindset.  Provide information related to forecasting the activity outcomes based on
the query phrases below using only the evidence excerpts provided.  Do not exclude
any relevant information.  Only include facts that would be knowable at the start
of the activity; ignore later progress or results.

USER:
ACTIVITY ID: {activity_id}

PHRASES:
1.  {phrase_1}
2.  {phrase_2}
3.  {phrase_3}
4.  {phrase_4}
5.  {phrase_5}

EVIDENCE EXCERPTS:
PHRASE: {phrase_1}
- [{source_doc_a} | {doc_type}] {retrieved_snippet_a}
- [{source_doc_b} | {doc_type}] {retrieved_snippet_b}
PHRASE: {phrase_2}
- [{source_doc_c} | {doc_type}] {retrieved_snippet_c}
...

Respond with relevant information about the activity from the excerpts that would
be useful to forecasting the eventual success of the activity, without losing any
relevant information or context.  Focus on providing information relevant to the 5
phrases above.

Respond only in English.
```

Figure 11: RAG synthesis prompt (summarizes retrieved evidence into forecast-relevant activity facts).

```
SYSTEM:
You are an experienced international aid decision maker with a quantitative
mindset.  Provide a balanced and thoughtful assessment of how similar
activities were rated.  Do not attempt to forecast the current activity outcome.
Only include information that would be knowable at the start of the activity;
ignore later implementation results.  Under no circumstances reveal actual
ex-post information.

USER:
You are extracting and analyzing information from the examples below as applies
to the current activity.

### EXAMPLE ACTIVITIES ###
[Injected few-shot block; for each neighbor:  title; risks; example forecast;
rating scale; final evaluation outcome]
### END EXAMPLE ACTIVITIES ###

ACTIVITY ID: {activity_id}
ACTIVITY TITLE: {activity_title}
ORIGINAL PLANNED START DATE: {planned_start}
ORIGINAL PLANNED END DATE: {planned_end}
ACTIVITY SCOPE: {activity_scope}
PLANNED TOTAL DISBURSEMENT (USD): {planned_total_disbursement_usd}
ACTIVITY LOCATION(S): {locations}
LOCATION GDP PER CAPITA, USD: {gdp_percap}
PARTICIPATING ORGANIZATIONS: {reporting_orgs}
IMPLEMENTING ORGANIZATION CATEGORY: {implementing_org_type}

Please respond as follows:
List all the separate reasons that the example forecasts described above went
well or badly, given their evaluations, if such considerations could apply to
this activity.  Rate the applicability of each consideration.  What are the
relevant lessons that can be learned as could apply to forecasting the outcome
of this activity?  Describe the key reasons each example was given the rating
they were.  Ensure the only information given is that which could be known or
reasonably forecasted at the start of the activity.

Respond only in English.
```

Figure 12: Prompt template used to summarize KNN few-shot examples into a compact "lessons from similar activities" block (used downstream in the final forecast prompts).

Ensembling Ultimately, I found that ensembling was too expensive of a method to use for forecasting.

2.7 Scoring Metrics

Accuracy The percent of the time the correct rating is forecasted. Non-integers are rounded to integers.

Side Accuracy The percent of correctly predicted "Satisfactory" or above vs "Moderately Satisfactory" or below (above or below 3.5). Approximately 50% of the training dataset sits above and approximately 50% sits below this boundary.

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.

Adjusted R^2 Adjusted R^2 is a version of R^2 that accounts for the number of regressors in the model. Unlike plain R^2, it penalizes adding predictors that do not meaningfully improve fit, making it more appropriate when comparing models with different numbers of features. It can decrease when irrelevant regressors are included, and it can be negative. Higher is better. While it penalizes extra parameters that may lead to overfitting, adjusted R^2 within a training set does not reflect model skill as accurately as out-of-distribution R^2.

2.8 Conformal Prediction

The outputs of statistical model predictions can be very helpful when a given minimum confidence interval (CI) is required, such as in developmental aid and cooperation interventions affecting the environment. "Conformal prediction" is a framework that provides distribution-free coverage guarantees: given a calibration set, predictions will contain the ground truth within a specified probability for future data generated from the same process as the training distribution, regardless of the underlying predictive model

(Cherian, Gibbs, and Candès, 2024). This is helpful so that 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 theoretical coverage guarantees with statistical backing.

2.8.1 Model For Fixed Width Conformal Prediction

To produce the fixed width conformal prediction, I reserved 10% of the training set ratings for the calibration, and trained a separate RF model on the remaining 90% of the train date. The result was the same error bar size for every prediction, theoretically guaranteed to cover 90% of outcome ratings assuming process exchangeability.

Consider a sequence of observations $\{(x_t, y_t)\}_{t=1}^{T}$ and a predictive model producing point forecasts $\hat{y}_t$. Define the residuals

$$e_t = y_t - \hat{y}_t.$$

In its simplest form, conformal prediction constructs prediction intervals using a fixed-width correction. Let

$$q_{1-\alpha} = \text{Quantile}_{1-\alpha}\left(|e_1|, \ldots, |e_T|\right).$$

where $\text{Quantile}_{1-\alpha}(\cdot)$ returns the smallest real number q such that at least a fraction $1-\alpha$ of the inputs are less than or equal to q. Because the process which generates q hasn't changed, for a new input x_{T+1}, the conformal prediction interval is given by

$$\hat{y}_{T+1} \pm q_{1-\alpha}.$$

This interval achieves marginal coverage at level $1-\alpha$, but its width is constant across time and does not adapt to changing uncertainty in the data.

2.8.2 Variance-Adaptive Conformal Prediction

I initially tried to produce such an error prediction by implementing a Bayesian Additive Regression Trees (BART) model (**quirogaBayesianAdditiveRegression2023**), which can be considered a bayesian version of the random forest model, where confidence intervals are added for each prediction. However, BART is relatively high-compute to train, and requires more tuning to reach the performance of a Random Forest model.

2.8.3 Variance Adaptive Conformal Prediction with a Ridge Regression Error Model

Because BART is computationally expensive to train and may miss out on essential features for confidence prediction, I also attempted my own CI model.

Therefore, I also tested a more customized approach to the problem domain. I theorized that most of the variance can be captured by the following features already available within the data:

- The count of missing features for a record which had to be median-imputed (*n_missing*)
- The standard deviation in predictions of the best predictor model (the random forest), with 30 randomly sampled bootstrapped data points from the earliest 70% of the training dataset (*bag_std*)
- The standard deviation of tree predictions within the RF model (*tree_std*)
- The magnitude of difference between the RF and the ridge regression model predictions (*abs_rf_minus_ridge*)
- The prediction of the random forest itself (*yhat_rf*)

I then calibrated this to produce a conformal prediction CI, where 90% of predictions fall within that interval. In order to generate the variable width, Ridge Regression predictions, The latest 30% of the training data was set aside. The first 2/3 of this set-aside data (123 points) was used to train the Ridge Regression model. The remaining 1/3 (68 points) was used to calibrate the error prediction model. As with the BART model, rather than a fixed 90% CI, I report a per-row confidence interval.

In order to train the error model, it was necessary to re-train an RF and Ridge regression model on the first 70% of the data in train, and use the next 20% to train the ridge CI prediction. Assuming that the process to generate the validation set is interchangeable with the process to generate the test set, conformal prediction theory tells us that we can say rigorously that the error bars I have calibrated against the last 10% of the training set will cover at least 90% of the true predictions. In other words, it is always <10% chance that the true rating will fall outside of the 90% confidence interval I generate. [NOTE: when i train on the held-out set, I will be able to put this theory to an even better test].

2.8.4 Model For Variance Adaptive Conformal Prediction

To account for heteroskedasticity, conformal prediction can be combined with model-implied uncertainty estimates. Suppose the predictive model provides an estimate of the

absolute error $\hat{\sigma}_t$ for an activity that started at time t. I define the normalized residuals as:

$$z_t = \frac{e_t}{\hat{\sigma}_t}.$$

I define the fixed halfwidth coefficient $q_{1-\alpha}$ as

$$q_{1-\alpha} = \text{Quantile}_{1-\alpha}\left(|z_1|, \ldots, |z_T|\right).$$

Because of the process exchangeability condition between activities starting at time T in the training dataset and at time $T+1$ outside of the dataset, the behavior of $\hat{\sigma}_T$ will be equivalent to $\hat{\sigma}_{T+1}$. The resulting prediction interval for y_{T+1} becomes

$$\hat{y}_{T+1} \pm q_{1-\alpha}\, \hat{\sigma}_{T+1}.$$

where $\hat{\sigma}_t$ denotes the model-implied estimate of the absolute error of the rating at time t. This construction preserves the distribution-free coverage guarantees of conformal prediction while allowing interval widths to adapt dynamically based on the model's own uncertainty estimates.

3 Results & Discussion

I provide a summary of the data made available from this work, provide an analysis of the success of the various methods at forecasting evaluation ratings, and provide preliminary results on the ability for the model to forecast specific outcomes.

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`. [NOTE: I will release once held-out set is analyzed.]

3.2 Predicting Overall Ratings

Overall, the forecasting system I produce is capable of predicting outcomes significantly above chance on out-of-distribution activities. Compared to prior work (**ashtonPuzzleMissingPieces2(**
I report a value consistent with an adjusted R^2 with training set (I report an adjusted R^2 of 0.450 for within-training set correlations on primarily world bank ratings, while others report at maximum an adjusted R^2 of 0.3 (**goldembergMindingGapAid2025**). I was not able to identify comparable out-of-distribution, time-ordered split analysis in the literature. As expected for prediction under distribution shift, my results on the out-of-time validation set were considerably weaker, with an R^2 of 0.17 for the random forest model, and an R^2 of 0.19 when incorporating the language model forecasting results correction + recency model.

3.2.1 Overfitting Corrections

R^2 was chosen as the "Adjusted R^2" has been used in similar work to evaluate model performance in the development aid literature and penalize overfitting by reducing the reported R^2 as a function of the number of input parameters. Mirroring similar reported methods in the literature, I calculated adjusted R^2 on the training points. While this is sensitive to overfitting, it is a common practice in the development aid literature. However, I find adjusted R^2 within the training set is highly sensitive to the specific parameters of the RF model and the subsequent degree of overfitting, such that adjusted R^2 increases to above 0.6 with default random forest parameters, while performance on the validation

set drops (See Table 1. I conclude that adjusted R^2 should not be used as a measure of forecast skill.

Relative to the default `RandomForestRegressor` configuration (e.g., `n_estimators`=100, `max_depth`=None, `min_samples_split`=2, `min_samples_leaf`=1, `max_features`=1.0, `bootstrap`=True, `ccp_alpha`=0.0), the specification used here deliberately constrains model capacity in ways that typically reduce overfitting. Trees are explicitly depth-limited (`max_depth`=14 rather than unbounded) and splits are only permitted when nodes contain substantially more data (`min_samples_split`=20 and `min_samples_leaf`=10), which smooths predictions by limiting fine-grained partitioning of the feature space. In addition, using a smaller feature subset at each split (`max_features`=0.488) increases tree diversity and reduces variance relative to the default that considers all features. The model also uses row subsampling (`max_samples`=0.86), further reducing variance by injecting additional randomness into each tree's training set. Increasing the number of trees (`n_estimators`=638) primarily improves stability via variance reduction rather than increasing effective complexity, while the pruning parameter (`ccp_alpha`=1.26×10^{-6}) remains close to the default of no pruning. Overall, compared to defaults, these choices trade some bias for a meaningful reduction in variance, making the fitted ensemble less susceptible to inflated in-sample fit and the attendant drop in validation performance.

3.2.2 Language Model Features

The language model derived features modestly aided prediction accuracy, in aggregate providing an improvement of about 4% additional explanation of the variance of outcomes out of the 9% discoverable by the RF model (See Table 1). The finance, integratedness, implementer_performance, targets, context, risks, and complexity features were directly inserted as grades from the model.

3.2.3 Embedding features

One LLM feature I add is the embedding of the LLM-generated "targets" field as a semantic representation of what each activity is trying to accomplish. This follows (**goldembergMindingGapAid2025**) in extracting the World Bank Project Development Objectives (PDO), as I find they are very similar to the LLM-generated outputs. I also attempted PDO extraction using regex methods, but found the results were noisy on matching, especially on non-world-bank projects, and did not improve prediction performance as much as embeddings on the llm-generated targets. I first normalize the LLM-extracted targets text into a stable canonical form (removing formatting artifacts, unescaping, splitting on separators, dropping "NO RESPONSE" tails, and deduplicating near-identical chunks). I then embed the cleaned targets text for each activity with the

gemini-embedding-001 model, yielding a single high-dimensional vector per activity that summarizes activity objectives (targets) in a continuous latent space.

I then replicate (**goldembergMindingGapAid2025**) and compress target embeddings using a two-stage dimensionality reduction pipeline. First, I apply PCA to reduce the embedding vectors to 50 dimensions and fit UMAP on the PCA outputs to produce 2D and 3D coordinate maps. I find the 3D embeddings (umap_x, umap_y, umap_z) perform better on prediction than 2D. This preserves enough local topology that activities with similar targets remain near each other in the compressed space. I also find qualitatively, that sectors with similar 2D vectors cluster around the activity environmental category, as in (**goldembergMindingGapAid2025**). While (**goldembergMindingGapAid2025**) find significant signal in deviations from average embeddings for countries or sectors, I do not find these theorized degree of "contextualization" features aid forecasting skill when I add them to my model.

3.2.3.1 Recency and LLM Adjustment Ridge Regression

I wanted to both correct the random forest model for temporal distribution shift (e.g., changing reporting practices, evaluation standards, portfolio composition, and macro conditions), and incorporate any usable information from the direct LLM forecasts. Even when the input features are stable, the conditional relationship $p(y \mid x)$ can drift, so a model trained on older activities can become mis-calibrated on newer ones. Furthermore, I found the LLM forecast predictions were significantly correlated with prediction error in the validation set.

3.2.3.2 Residual-correction formulation. Let $\hat{y}_i^{\mathrm{RF}}$ be the random forest prediction for activity i, and let $\hat{y}_i^{\mathrm{LF}}$ denote the LLM Forecast. I define the random-forest residual on the i'th activity as:

$$r_i := y_i - \hat{y}_i^{\mathrm{RF}}.$$

I then fit a ridge regression model to predict residuals from a small feature vector consisting of the RF prediction and (optionally) the LLM Forecast:

$$\hat{r}_i := \beta_0 + \beta_1 \hat{y}_i^{\mathrm{RF}} + \beta_2 \hat{y}_i^{\mathrm{LF}},$$

with an ℓ_2 penalty on (β_1, β_2) controlled by `alpha` (ridge strength). The corrected prediction is:

$$\hat{y}_i^{\mathrm{corr}} := \mathrm{clip}_{[0,5]}\left(\hat{y}_i^{\mathrm{RF}} + \lambda \hat{r}_i\right),$$

where λ is a scaling factor (set to 1.0 in my experiments) and clipping enforces the valid rating range between 0 and 5.

Table 1: Validation performance in predicting ratings across forecasting methods. Rows are sorted by ascending R^2 (higher is better). RMSE and Brier are lower-is-better; Side Accuracy is higher-is-better. Bold indicates the best value in each metric column. Variation between side-accuracy methods ("Moderately Satisfactory" or lower vs "Satisfactory" or lower) were not statistically significant. Side metrics use a binary split at rating ≥ 3.5 vs < 3.5 . The "recency correction" variants combine models using the 150 latest-starting activities in the training set for calibration/combination. The "no LLM features" Random Forest excludes the following features: finance, integratedness, implementer_performance, targets, context, risks, complexity, umap3_x, umap3_y, umap3_z.

Method	$R^2 \uparrow$	RMSE $\downarrow$	MAE $\downarrow$	Side Acc. $\uparrow$	Acc. $\uparrow$
Mode of reporting-org score baseline	0.102	0.897	**0.579**	0.700	0.530
Random Forest only	0.173	0.861	0.651	0.703	0.513
Random Forest, no LLM features	0.143	0.876	0.679	0.677	0.490
Ridge GLM	0.148	0.873	0.654	0.683	0.513
Random Forest (default params)	0.157	0.869	0.653	0.693	0.497
Ridge GLM + Random Forest (mean)	0.173	0.861	0.644	0.710	**0.537**
RF + recency correction	0.183	0.855	0.629	**0.733**	0.523
RF + LLM Forecast + recency correction	**0.193**	**0.850**	0.623	0.730	0.527

This is a simple stacked model: the RF provides the base signal, and ridge learns a low-capacity adjustment to remove systematic residual structure that appears in the recent/LLM-covered slice.

I tested two separate methods:

1. Recency correction (RF re-calibration on recent activities). In this variant I remove the LLM forecast entirely fixing $\beta_2 = 0$, but still calculate an offset β_0 and scaling β_1 on the 150 latest training examples.

2. LLM-informed correction (recency + LLM Forecast). In this variant, the ridge model uses both the RF prediction and the LLM Forecast as covariates on the activities where $\hat{y}_i^{\mathrm{LF}}$ is available. This allows the correction layer to learn a mapping from $(\hat{y}_i^{\mathrm{RF}}, \hat{y}_i^{\mathrm{LF}})$ to the residual r_i, effectively learning when the LLM Forecast contains signal about systematic RF error on the recent slice. The correction is applied only to activities where $\hat{y}_i^{\mathrm{LF}}$ exists; otherwise, predictions fall back to the uncorrected RF output.

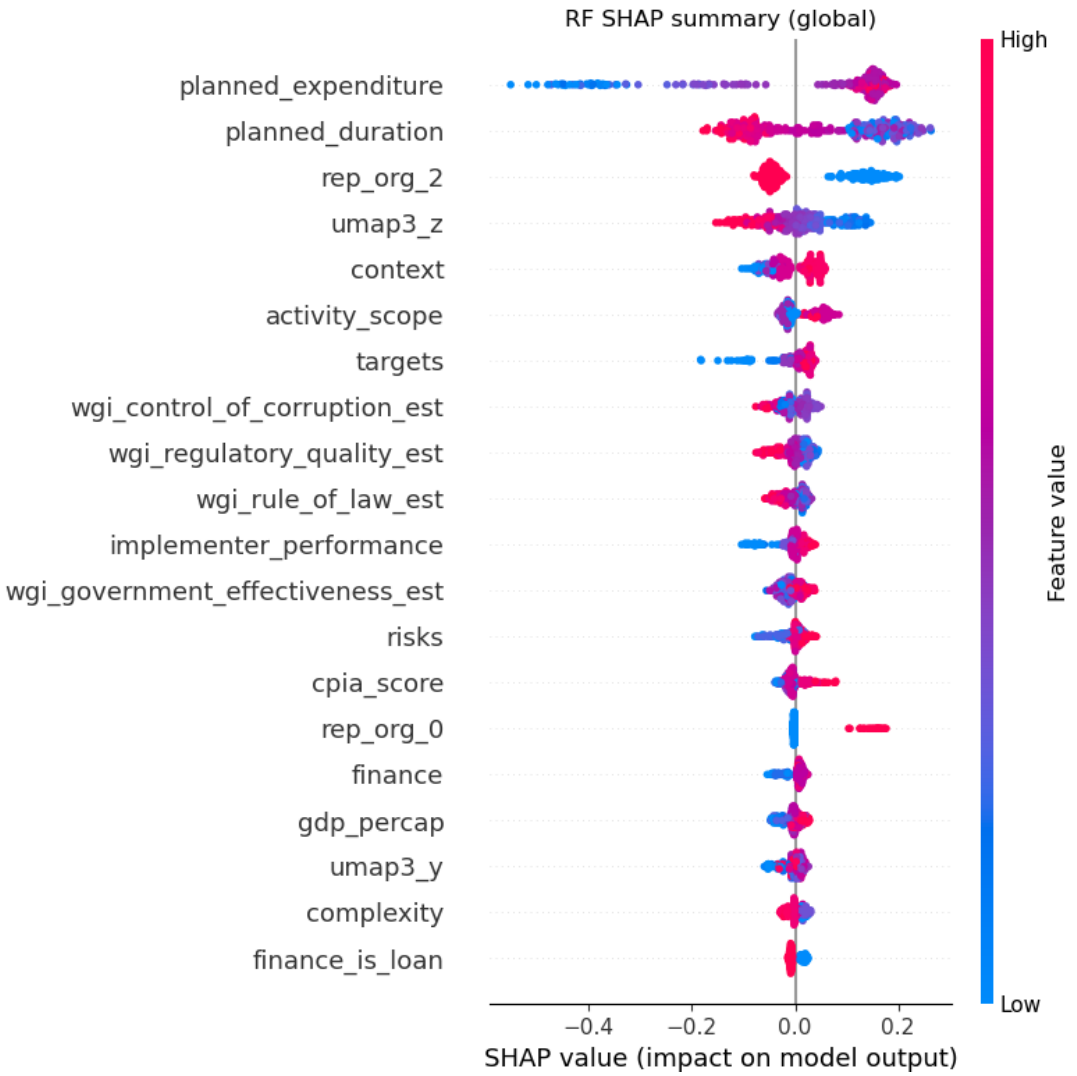

Figure 13: A SHAP analysis of ratings on the validation set from the RF. Red indicates an increase in the value of the feature, while blue indicates a below-average value. Points to the right of zero shift ratings up, points to the left of zero shift ratings down.

Table 2: Random-forest drop-one feature importance, sorted by impact. Each row reports the decrease in validation R^2 when the feature is removed (ΔR^{2*}), along with the share of training rows that were median-imputed for that feature (% missing**) and the training-set mean/SD (in the model feature space). The largest impacts come from planned budget and planned duration; governance indicators and semantic/LLM-derived summaries add moderate incremental signal; `rep_org_*` are one-hot reporting-organization indicators.

Feature	ΔR^{2*}	% missing**	Mean (train)	SD (train)
planned_expenditure	0.0940	38.18	18.026	1.273
planned_duration	0.0840	0.15	8.806	3.371
finance_is_loan	0.0320	4.26	0.635	0.482
rep_org_2	0.0310	0.00	0.648	0.478
wgi_regulatory_quality_est	0.0290	6.90	-0.401	0.488
finance	0.0290	2.50	81.634	9.268
targets	0.0280	0.15	66.625	13.442
rep_org_0	0.0270	0.00	0.051	0.221
umap3_y	0.0260	9.84	-1.615	1.023
wgi_government_effectiveness_est	0.0230	6.90	-0.386	0.473
wgi_rule_of_law_est	0.0230	6.90	-0.521	0.492
rep_org_1	0.0230	0.00	0.059	0.235
gdp_percap	0.0230	7.34	8.265	0.996
umap3_x	0.0220	9.84	0.460	2.776
wgi_control_of_corruption_est	0.0220	6.90	-0.550	0.467
wgi_political_stability_est	0.0210	6.90	-0.570	0.789
activity_scope	0.0200	0.00	3.185	1.397
risks	0.0190	0.59	64.439	19.752
context	0.0190	0.15	74.029	16.487
cpia_score	0.0170	48.75	3.499	0.382
complexity	0.0160	0.15	61.912	17.406
umap3_z	0.0160	9.84	1.599	1.869
implementer_performance	0.0150	0.29	82.044	11.749
integratedness	0.0090	1.32	83.266	5.568

* ΔR^2 is computed as $(R^2_{\text{full}} - R^2_{\text{dropped}})$ on the same validation split; larger values indicate greater marginal contribution under this ablation test.

** Percent of training rows with missing values for the feature, filled via median imputation (as used in the model pipeline).

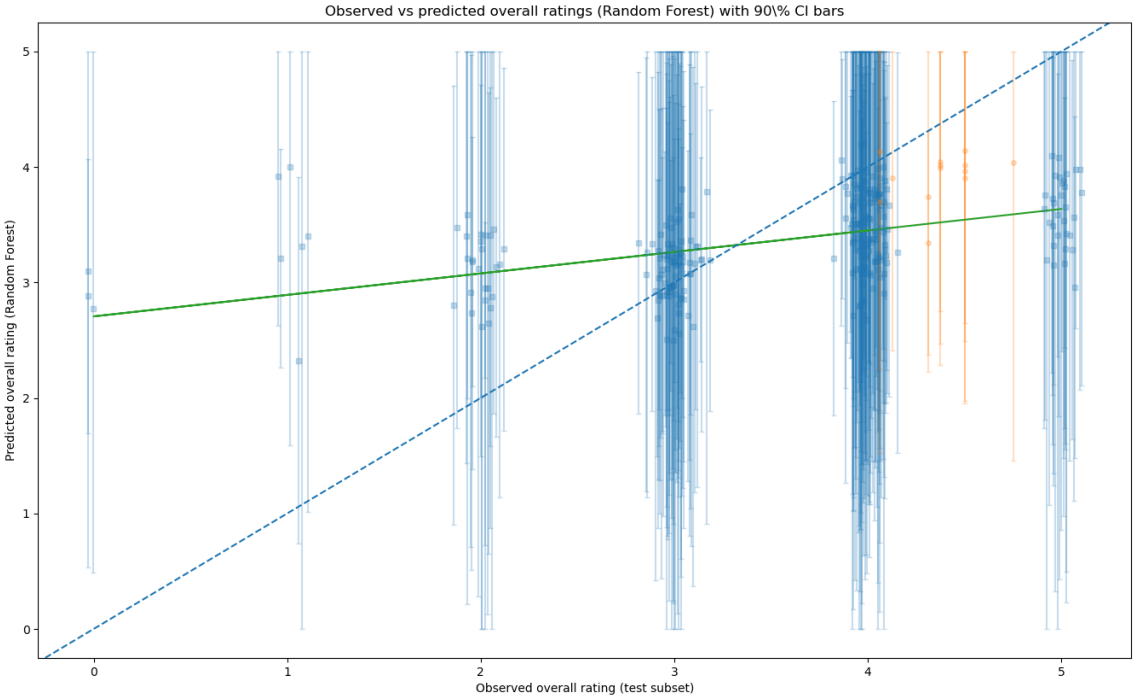

Figure 14: Observed versus predicted overall ratings (Random Forest) with calibrated 90% prediction interval (PI) bars (R^2 0.17, RMSE 0.862). The dashed diagonal indicates perfect agreement between observed and predicted ratings, and the green line shows the fitted trend. Random jitter on this plot's X axis was introduced to blue points to improve readability for confidence intervals. The orange points represent the few ratings which did not land on an integer between 0 and 5. See Section 3.4 for how the calibrated error bars were computed (Ridge Regression Error Model).

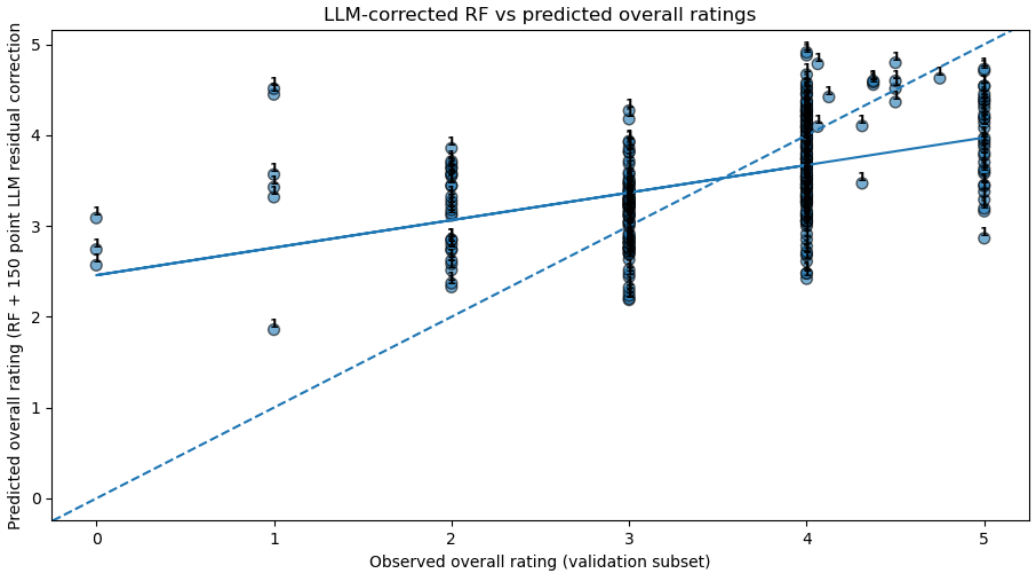

Figure 15: The scatter of observed vs predicted points on the held-out set for the LLM-corrected RF prediction. R^2 of 0.19, RMSE of 0.853.

3.3 Predicting Outcomes

In general, outcome predictions were weaker than ratings. Initially, I thought outcomes would be more predictable than ratings. A similar work found an adjusted $R^2 > 0.7$ for outcome ratings including in predicting beneficiaries reached (**goldembergMindingGapAid2025**), but this was including actual rather than planned durations, actual rather than planned financial disbursements, and several features including breakdowns of per-sector funding for activities and manager performance ratings from AidData that I did not include in my dataset. Furthermore the paper did not predict on cost-effectiveness, making outcome prediction a much easier task when given overall program spending, and lastly I find adjusted R^2 for within distribution is highly sensitive to model overfitting. As mentioned in previous sections, AidData was laboriously double-coded data entry, which introduces less error than the LLM and regex data extraction techniques I used. Furthermore, the outcomes with high detected correlation measured a lagged 5-year country-level indicator, while my data were extracted directly from the outcomes.

Several factors contributed to the difficulty of predicting specific outcomes from extracted IATI data:

- Unclear apportioning of funding towards each outcome
- Inconsistent measurement styles and definitions of terms like Benefit-Cost Ratio, which effects would be included in Economic Rate of Returns.
- Incorrect aggregation of multiple ratings within the documents

3.3.0.1 Outcome model training and evaluation. For each outcome distribution in Table 3, I trained a random-forest regression model that is identical (same features, preprocessing, and hyperparameters) to the model used for predicting ratings, but with the rating target replaced by the relevant activity-level outcome. A dummy variable for which outcome was also included. The ratings were included as a feature to aid learning about activity success. Models were trained using activity IDs in the training split with non-missing outcomes and evaluated on the validation activities. The counts reported in Table 3 correspond to the number of activities available in the validation split for each outcome. Outcome distributions with fewer than 10 activities in either the training or validation split were excluded.

In addition, a single aggregate Z-score was calculated, which subtracts the mean value of each outcome (including ratings) and divides by the standard deviation in the training set. For each activity, the mean value of the z-scores was taken for all dependent variables, including the rating. Due to its high predictability, I theorize that the z-score is a stronger indicator of activity success than activity rating alone, due to the prevalence of "gaming" activity ratings (**goldembergMindingGapAid2025**).

Table 3: Updated predictive performance of random-forest outcome models on the validation set. For cost-effectiveness outcomes, the target is USD per unit outcome and is modeled in log space; for the remaining outcomes the targets are modeled on their original scale. Outcomes are sorted by R^2 (descending).

Outcome	Units	Target transform	R^2	RMSE	MAE	$N_{\text{val}}/N_{\text{train}}/N_{\text{heldout}}$
All selected outcomes (z-aggregate activity mean)	z	none (z-score)	0.22	0.72	0.53	326 / 737 / 248
Rating	rating	none	0.15	0.88	0.67	324 / 732 / 244
Water and sanitation connections (cost-effectiveness)	connections	log(USD/connection)	0.11	0.69	0.49	30 / 84 / 10
Beneficiaries (cost per beneficiary)	people	log(USD/person)	0.05	0.97	0.67	102 / 220 / 59
Economic rate of return	percent	none	0.0084	22.0	15.0	128 / 379 / 65
Agricultural yield increase (percent)	percent	none	-0.0053	33.0	26.0	10 / 33 / 9
Benefit–cost ratio	ratio	none	-0.045	0.38	0.27	18 / 49 / 16
Generation capacity (cost-effectiveness)	MW	log(USD/MW)	-0.074	1.0	0.83	30 / 76 / 16
CO_2/CO_2e emissions reductions (cost-effectiveness)	tonnes CO_2e	log(USD/tCO_2e)	-0.14	1.5	1.2	27 / 47 / 13
Financial rate of return	percent	none	-0.28	17.0	14.0	49 / 170 / 21

In keeping with the results of (**goldembergMindingGapAid2025**), I do not find a strong correlation between activity ratings and z-scored outcomes.

I do find that failing to divide by the total disbursement as marked in iati increases predictability - the z-score aggregate correlation moves from 0.27 to 0.23 when we move from predicting the raw outcome to the cost-effectiveness (dollars per unit outcome). When dividing by each activity's disbursement for those that are marked as dollar-per-unit in Table 3, and looking at all outcomes except ratings, the correlation on z-scores drops to -0.01. However, including the ratings in the training (but not evaluating performance on predicting ratings) leads to an R^2 of 0.05 (n_{val}=163), which indicates training on ratings modestly improves performance in predicting outcomes, despite the lack of a clear linear relationship between z-scored ratings and z-scored outcomes.

Overall, little can be concluded from individual outcome correlations. For more detailed work, expert coding is likely required for robust extraction of outcomes, and funding breakdowns for projects should be used to more accurately evaluate cost-effectiveness, rather than the course assumption that all funding for a project goes to all outcomes. For example, for "beneficiaries reached", such results could be systematically different for different definitions of "Beneficiaries" - compare for example, indirect vs direct beneficiaries which were not disambiguated. New water or sanitation piping connections is a more clearly comparable outcome, although there may be systematic differences in the costs of

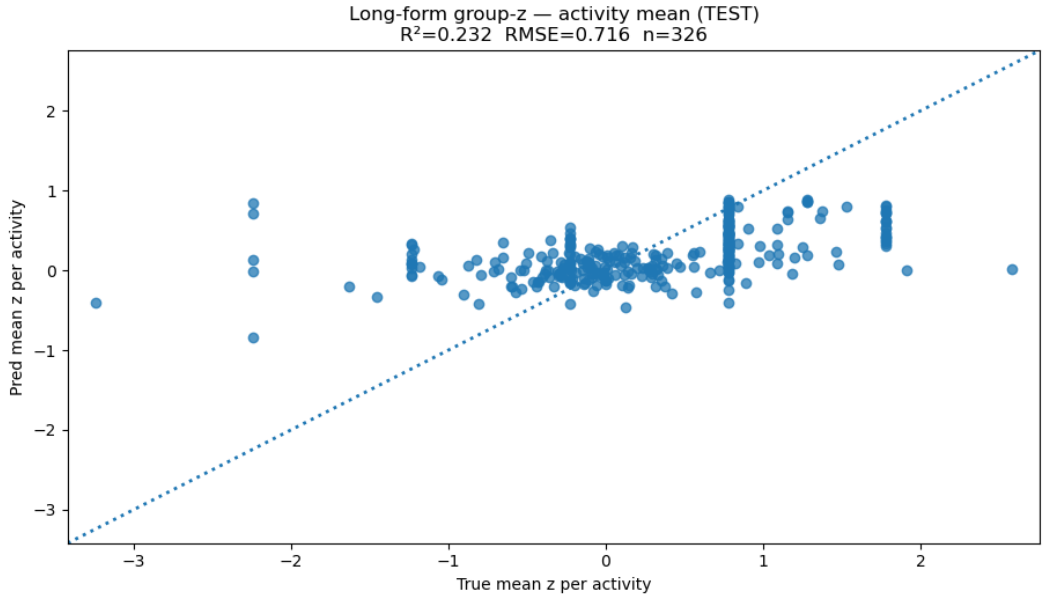

Figure 16: Validation: predicted versus observed absolute error. Each point corresponds to the mean of the random forest z-score prediction for ratings and all other outcomes listed in table 3, for activities in the validation set where at least one rating or outcome were predicted. The dashed line indicates perfect calibration ($y = x$).

sanitation connections and water supply that are not disambiguated by the model.

Despite these limitations, it appears that even a coarse coding of directly comparable activity outcomes is likely to provide a better sense of overall activity sense than evaluator ratings alone. I theorize the key reason is that while outcomes are difficult to code and reliably extract compared to ratings, ratings are more subjective, and prone to systematic differences between evaluators and reporting organizations. Given that these two errors are relatively uncorrelated, the meaned z-scored ratings+outcome aggregate indicator may be a better measure of activity success than any other available or derived metric from the IATI database.

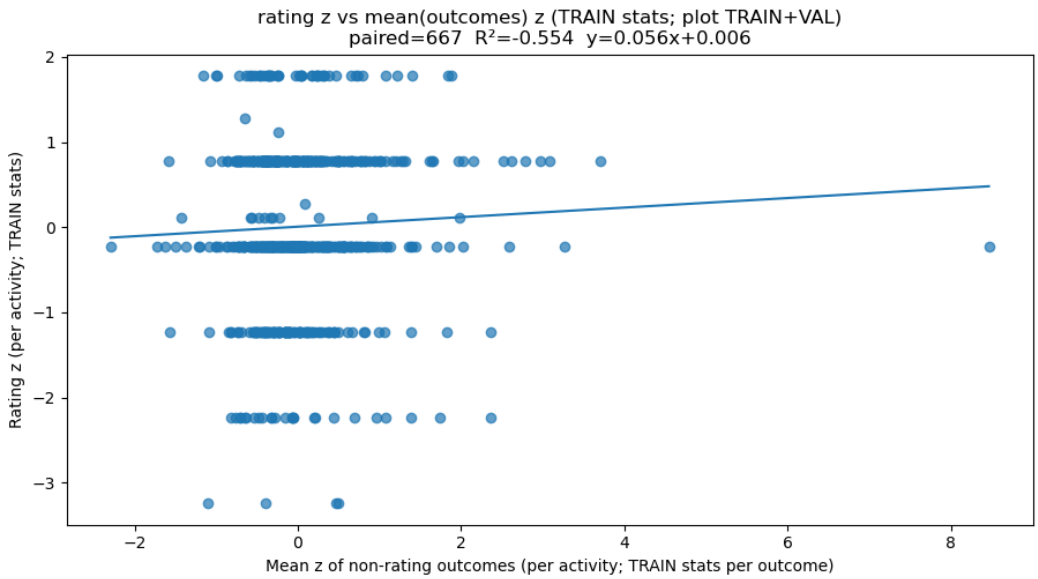

Figure 17: All activity's z-scored ratings plotted vs z-scored outcomes. There is a very low positive correlation (Pearson correlation of 0.044). The outliers near "8" averages a 433 percent Economic rate of return and a benefit-cost ratio of 1.6 for the "Jakarta Urgent Flood Mitigation Project", an unusually economically effective project, that should regardless not be removed simply for being an outlier. The project's efficiency was substantial thanks to significant loan savings. However, due to "the lack of clarity in the reported data, the limited application of FMIS findings, and the unfulfilled activities added at restructuring", the Overall Outcome rating for this outlier project to "Moderately Satisfactory". See `https://datastore.iatistandard.org/activity/44000-P111034`.

3.4 Conformal Prediction Results

3.4.1 Bayesian Additive Regression Trees

I found that the BART model was not only significantly worse at explaining the variance, and with a higher RMSE than the RF model, its confidence predictions were badly miscalibrated. A single model like BART appears to unable to properly capture the uncertainty of the training data under distribution shift. Furthermore, the BART model was found to be two orders of magnitude slower to train than the significantly higher performing RF model (as measured by R^2 and RMSE performance training on the full train set and checking against validation).

3.4.2 Ridge Regression Error Model Parameter Influence

I find relatively even influence of the different error terms in my CI prediction model (see Table 4). This explains why the BART was doing so badly at confidence interval prediction: it only had access to *tree_std*, and was therefore was missing 75% of the information about the source of confidence.

Table 4: Absolute influence on 90% CI per 1σ shift in the input variable.

Input variable	Absolute influence
tree_std	1.4
abs_rf_minus_ridge	1.1
bag_std	1.1
yhat_rf	1.0
n_missing	1.0

3.4.3 Ridge Regression Error Model

The Ridge Regression model does show statistically significant predictive power on the validation set (Spearman $\rho = 0.189$, p = 0.001) and can distinguish uncertainty at a coarse level: high-uncertainty tercile predictions have 35% higher mean absolute errors than low-uncertainty predictions. The distance between its prediction and the error of a prediction on validation is 0.406 on average, compared to "always predict the mean error", which has a distance of 0.432 from the true error, representing a small 6.2% improvement. This suggests the approach has potential but requires more training data to compete with simpler baselines. One explanation is the 70% of the train data was is insufficient to train a model better than picking the most common per-org rating, and thus the models used for the variable width CI error model were insufficiently similar to the RF model to reproduce its error modes [NOTE: we will see how this changes when running on the held-out set!]

Figure 18 shows how the Ridge Regression confidence prediction model's expected absolute error aligns with absolute error on the validation set, while Figure 19 aggregates this relationship into quantile bins (with bin counts annotated).

3.4.4 Fixed Width

Overall, the simple 90% fixed width performs better than either the BART or the Ridge Regression error model. It achieves closer to 90% coverage of the ratings in the validation set (92% fixed width vs 95% from the variable width model) on the validation set while using significantly less width (2.3 rating points fixed vs a mean of 3.7 for the variable width), indicating the variable-width prediction ability is too weak to enable superior CI intervals. Notably, setting the fixed width to the same mean 3.7 width as the variable model also achieves better coverage (97%) than the variable width model.

Figure 14 shows the resulting fixed width 90% prediction intervals (PI) around the random forest point predictions on the validation set.

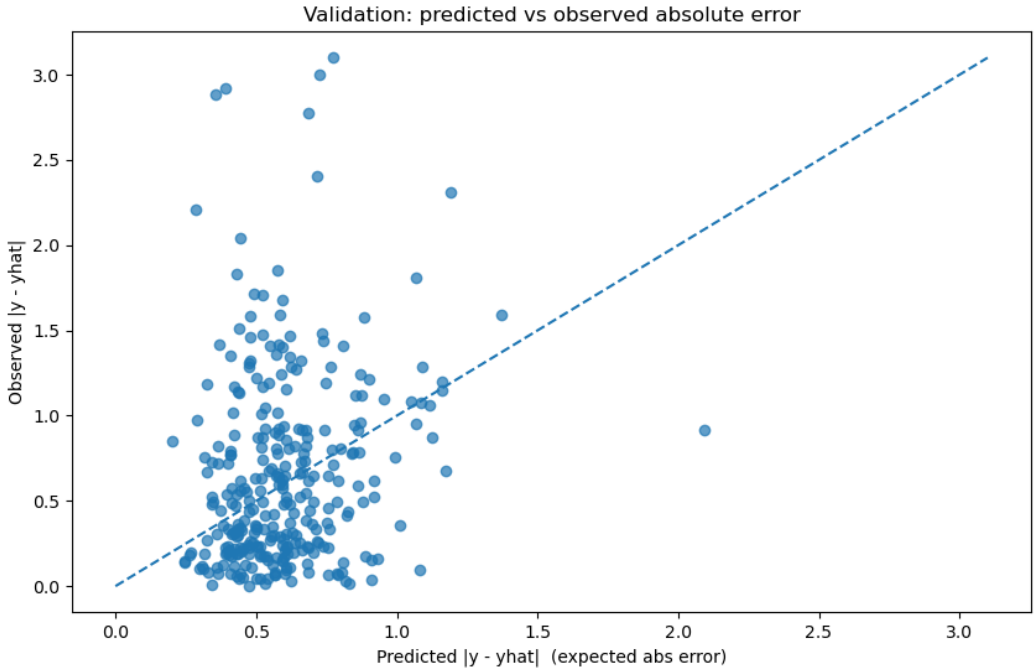

Figure 18: Validation: predicted versus observed absolute error. Each point corresponds to a record; the x-axis is the meta-model prediction of $|y - \hat{y}_{\text{rf}}|$ (expected absolute error) and the y-axis is the realized $|y - \hat{y}_{\text{rf}}|$. The dashed line indicates perfect calibration ($y = x$).

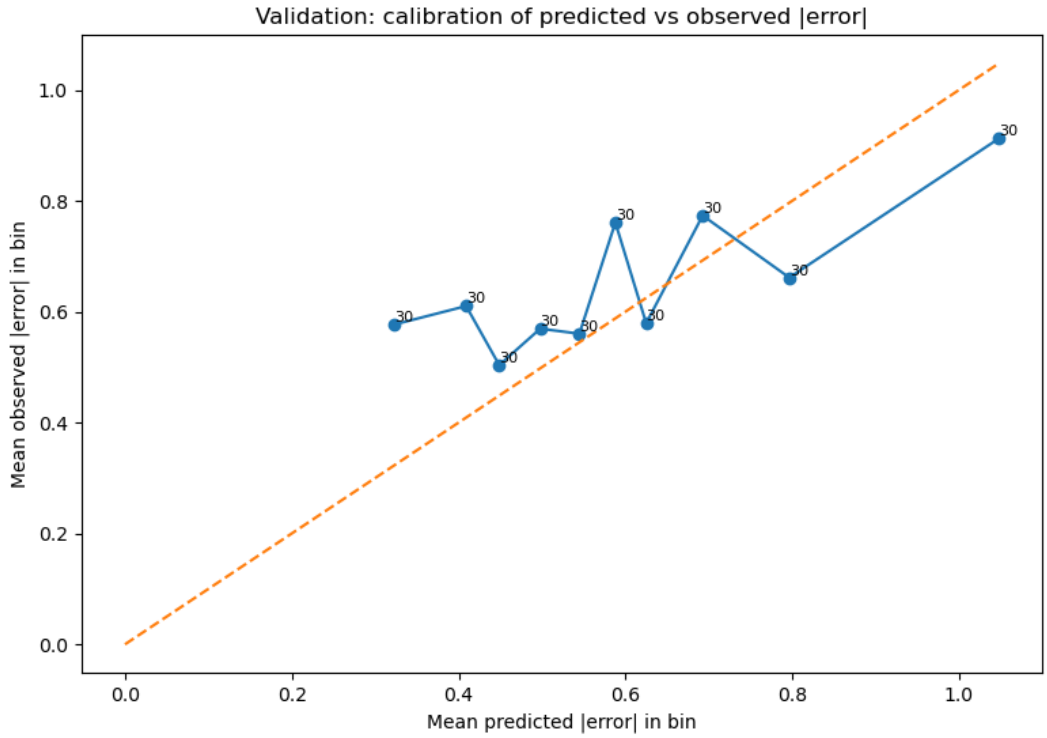

Figure 19: Validation: binned calibration of predicted versus observed absolute error. Points are bin means (with bin counts annotated). The dashed line indicates perfect calibration ($y = x$).

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).

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.

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).

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).

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).

Hewitt, Luke et al. (n.d.). "Predicting Results of Social Science Experiments Using Large Language Models". In: ().

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

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

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

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).

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).

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).

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).

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).

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, 14 January 2026