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, 23 February 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 create a profound difference in the rate of development in growing economies, improve the lives of the world's poorest, and often safeguard the environment and materially promote sustainability. However, international aid often fails to achieve its objectives. There have been few attempts in the literature to create models to explicitly predict the success of aid activities, and none focus solely 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 apply novel techniques to improve prediction accuracy, including using the reasoning abilities and information gathering ability of large language models (LLMs) to improve forecasts, introducing LLM summaries of various dimensions of activity documents, using recency adjustments, and defining a novel narrative similarity grade benchmark. I first assess overall success ratings on a scale from 1 to 6, then measure narrative forecast accuracy, and finally assess cost-effectiveness forecasting. The full forecasting system outperforms what could be extrapolated from the stated risks in activity documents alone for overall evaluations in tests against a validation set of 300 later-starting activities in a dataset of 1,300 environmental and sustainability improving activities restricted to 4 reporting organizations. Compared to the baseline of 65 % Pairwise accuracy for the risks extrapolation, the full system reached to 77 % [95% CI: 73 %, 81 %]. The forecasting system also explains 26 % [95% CI: 14 %, 37 %] of the variance in true ratings ($R^2 = 0.26$) in the validation set, compared to 10 % for a "pick the most common rating from reporting org" baseline, and 0% for the risks extrapolation. Finally, my forecasting system predicts which activity will have a higher cost-effectiveness among random pairs of activities in the IATI database with a success probability of 60 % [95% CI: 55 %, 66 %]. 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.

Table of Contents

1 Introduction

Note: for this draft, I focus on the methods and results, mostly leaving out the introduction, discussion, and conclusion. I am also waiting for feedback before running the full analysis on the held-out test set.

I set out to analyze the applicability and composability of machine learning techniques in forecasting the success of international aid activities using what could be known about the activities at the approval stage. The standard approach in the literature for assessing aid impact and success in environmental domains has historically involved exclusively "statistical forecasting": econometric techniques, linear regression models, and occasionally modern nonlinear models such as Random Forest (RF) or XGBoost (Goldemberg, Jordan, and Kenyon, 2025). However, a thorough assessment of methods to forecast activity success in this domain has yet to be conducted. In addition to statistical forecasting, I propose the use of Large Language Model (LLM) implemented 'judgemental forecasting", which allows forecasters to use tools including Fermi estimates, intuition, and information gathering to make a calibrated prediction on the likelihood of a given outcome (Halawi et al., 2024 | "Approaching Human-Level Forecasting with Language Models"), appropriate for forecasting techniques that can be improved by explicit reasoning processes beyond purely statistical correlations.

The modelling system I develop predicts how successful interventions will be in the context of international aid activities affecting the environment, utilizing statistical forecasting, judgemental forecasting, and a mixture of the two. I obtain metadata and PDF files from thousands of activities, and separate each record into information about the activity available at approval and evaluation information about how successful the activity was. I use this data to answer the following research questions in this work:

1. How do judgemental forecasting methods compare to statistical models in forecasting international aid overall success ratings and quantitative outcomes?
2. How do differing methods of combining judgemental and statistical forecasting compare in this domain?
3. What methods improve the accuracy of narrative (qualitative) forecasts in this domain?
4. What aspects of the activity available in my dataset at the beginning of the activity lead to higher or lower ratings?
5. How does forecasting aggregate cost-effectiveness compare to forecasting ratings?

This method differs in two key ways from prior literature, which has analyzed international aid evaluations. 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 at approval for the activity, and do not assess out-of-time generalization of these correlations. In this work, I assess out-of-time generalization ability of models. 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, not simply statically analyze a corpus of past activities. The second difference is that I implement judgemental AI forecasting, as a supplement to standard statistical models.

1.1 Prior Work

Within the domain of LLM use, there have been early attempts at using them to improve decision making for the environment. A recent tool called "climsight" summarizes and aggregates information about climate adaptation and mitigation (Koldunov and Jung, 2024), but stops short of making forecasts 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 (Callaghan et al., 2025).

Limited work has also been done using LLMs such as GPT-4 to serve as data sources for policy deliberation and multi-criteria assessment of climate and sustainability interventions, finding that 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 the unexplained residual from control variables on relevant country-level sector outcomes from the World Bank World Development Indicators. They also 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. Furthermore, in replicating and extending their method, I found their model worse than random chance at out-of-time prediction of outcomes, indicating severe issues with out-of-time generalization, consistent with validation and test set data leakage into their training set.

2 Methods

This thesis implements an LLM-based forecasting method for ratings, a statistical model for ratings free of any LLM features, a system combining the best LLM method and the best statistical model, with the advantages of both, and finally a modified system designed to predict the cost-effectiveness of activities from quantitative results extracted from activity evaluation documents. The system is built to forecast what the evaluation results will be for thousands of IATI records containing both a pre-intervention description of the activity, and an ex-post 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, and had quantitative and qualitative descriptions and results transcribed into a unified format.

2.1 Data Sources

After considering several data sources for prediction, including the OpenAlex publication repository of peer-reviewed evaluation documents and abstracts (Priem, Piwowar, and Orr, 2022), the IDEAL database of ex-post evaluations (KfW Development Bank, 2025), and the 3ie development database (*3ie Development Evidence Portal | 3ie* 2025), I decided to use the IATI database (*IATI Dashboard – IATI Activities* 2025), due to its substantial quantity of information available in textual format and extractable from the database records, its large quantity of activities, and the fact that new records are continuously added, making rerunning experimental methods in the future possible with little modification. While ex-post evaluations may provide sufficient information to describe the activity, relying on language models to completely remove information about the eventual outcome could introduce data leakage from past the start date. The IDEAL database and 3ie each had fewer than 300 hundreds of records for environmental topics each, far lower than IATI. Furthermore, although many millions of evaluations are available on OpenAlex, it proved difficult to reliably identify and de-duplicate published evaluations of environmental interventions.

The IATI database has several additional advantages. Records contain reliable "actual start" and "actual end" dates, and commonly also reporting when the activity was planned to start and end, and most records contain an overall evaluation rating on a six point scale within the linked evaluation PDFs. Activity records also reliably mark the reporting organization. The records used for this thesis have several activity information documents 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 also always

reported in the database, including if it is in the planning or completion/finalization stages, which is helpful information for forecasting.

A disadvantage of the IATI database is that it is sometimes inconsistent between reporting organizations as to how the data are filled in, and the format of documents was not always PDF, requiring me to write conversion scripts from other formats to PDF. Also, many download links were not functioning or required me to write custom web-scraping scripts to properly extract project documents in PDF format from the original websites where project documents were hosted. Planned start or end dates were often missing, leading to frequent exclusion of otherwise informative 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 that otherwise could have been included.

In addition to IATI, I also analyzed the AidData database (Tierney et al., 2011), which has double-coded data entry. While double-coded data entry introduces less error than the LLM and regex data extraction techniques I use, I found that the delay in hand-coding the ratings and results leads to relatively few recent environmental activities, and the lack of document links in AidData meant each activity record had far less extractable information. I found the overlap between AidData and my IATI training set was less than 30% of my IATI training set records. As AidData contained insufficient data for reliable forecasting on its own, and there was low overlap, I did not attempt to merge IATI with AidData or enrich IATI with AidData.

2.2 Data Filtering

Note: I have summarized a longer section here, contact me if you would like the original

The IATI dataset was progressively filtered from roughly 800,000 activities to 3,225 completed environment- and sustainability-related projects with both baseline and evaluation documents suitable for forecasting outcomes without future information leakage. Activities were first selected by sector codes and manual topic review, then constrained to those with downloadable documents, valid temporal ordering (baseline pdf uploaded at latest the end of first quarter of implementation period, evaluation at earliest last quarter), and sufficient metadata reliability. Additional restrictions ensured comparable rating scales and adequate sample sizes across four reporting organizations (World Bank, BMZ/KfW/GIZ, ADB, and UK FCDO). Documents were converted to PDF and dated primarily using PDF metadata, which was validated against document content, yielding a median 22-day discrepancy. To prevent leakage from early progress reporting or misdated files, documents were screened by manual inspection, LLM-assisted checks, and automated keyword searches, resulting in the exclusion of activities showing substantive outcome progress. Further filtering

removed purely legal or contractual documents, ultimately producing a curated dataset designed to reflect information realistically available near project start while preserving sufficient coverage for model training, validation, and testing.

In order to accurately extract 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:

2.3 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. Letter grades were selected in accordance with standard grading examples from Google Gemini documentation. 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 prioritized. 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 then had each of their pages categorized with what information was present on them. I determined by manual inspection this was more effective than semantic embedding search for long documents. To process the pages, I split pdfs 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 were 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. The two most relevant category choices among these were assigned for each page in the documents.

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 ratings. I found by looking at randomly selected pages that ratings often showed up on pages with a high rating for relevance to evaluation. Therefore, the first method sends each individual outcome page rated 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/KfW/GIZ 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. Cursory manual inspection did not reveal any leakage, and BMZ/KfW/GIZ overall ratings were less predictable than World Bank ratings, despite this data leakage possibility.

4. Interpreting Ratings Ratings were reported both with the rating itself, and 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 rescaled to match the 1 to 6 rating scale. Notably, BMZ/KfW/GIZ 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.

2.3.1 Outcome Cost-Effectiveness Extraction

In addition to extracting ratings, a similar approach was used with *gemini-2.5-flash* to extract quantitative cost-effectiveness-related outcomes. I extracted quantitative outcomes from all pages which were categorized as containing outcomes, and marked as containing quantitative information. Unlike with ratings, I did not limit these 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 (Goldemberg, Jordan, and Kenyon, 2025).

Once these PDF pages were extracted, I employed a combination of manual examination of the extracted outcomes and a frequency analysis of 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. Each script ensures implausible or non-numeric values are dropped. Domain-specific filters also reduced false positives (e.g., wastewater context for pollution loads, water context for connections, agriculture context for yields).

Comparable Outcome Categories

Using manual inspection and frequency statistics of common bigrams, I defined a set of outcome categories that recur across evaluations and are interpretable across projects. I report only those outcomes that remained after ensuring they contained values for at least 10 activities after filtering to the dates of the evaluation set (2013-02-06 until 2016-06-06). The final set included:

- **Benefit/cost 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.
- **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.

- **Energy outcomes:** installed generation capacity, in MW (or occasionally GWh where the source reported capacity in energy units).

Although I parse pollution load removed, forest indicators, trees planted, air quality (PM2.5), clean cooking stove distribution, and agricultural yield improvement, none of these contributed sufficient counts to be useful for further outcome forecasting.

I aggregate multiple extracted values per category by compiling all extracted outcomes in each outcome category for each activity and include them in an aggregation prompt for *gemini-2.5-flash* for aggregation into a final baseline, target, and outcome value. Where multiple extracted values conflicted, overlapped, or reflected different levels of aggregation (e.g., component-level versus project-level figures), the model was instructed to adjudicate between them and select or combine values to reflect a single coherent project-level outcome. The model was instructed to distinguish duplicates from additive components, prioritize project-level totals over subcomponent figures, aggregate only non-overlapping quantities, and return a null response when no coherent representative value could be determined.

Finally, I used the total disbursement for the activity reported by IATI and determined an approximate estimate of the cost in US Dollars per unit outcome (see "Splitting Disbursements" below), with the exception of benefit-cost ratios, rates of return, and agricultural yield outcomes. I found by empirical inspection that most outcomes were approximately log-normally distributed. I took the log_{10} of all categories except Benefit-Cost ratios and rates of return.

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 separately allocate total activity expenditures to non-overlapping funded sub-activities. 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. Benefit/cost ratios and economic and financial rate of return are excluded from monetary allocation, and other 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 portion

of the allocations 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 trees planted), 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 assigned to the CO_2 mitigating expenditures already present for that activity. This prevents CO_2 from inflating allocated spending when it is a co-reported consequence of other outcomes.

Outcome model training and evaluation

For each outcome distribution in Table 3, I trained a random-forest regression model containing similar features and hyperparameters to the model used for forecasting ratings, but with the rating target replaced by the relevant activity-level outcome. A one-hot encoded dummy variable for which cost-effectiveness outcomes were being averaged was also included. 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 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.

2.4 Baseline Methods

Always predict the most common rating for the reporting organization

This baseline technique provides a sanity check that more sophisticated methods are worthwhile. If this baseline outperforms more sophisticated models, it indicates that much of my methodology is overfitting on noise in the data.

Ridge regression with reporting organization and risk score

In order to assess how well calibrated overall risks are in activity documents, a ridge regression model was trained given both the LLM grade for the degree of risk assessed at the beginning of the project, as well as the dummy variable representing which organization was doing the rating. While by construction the full forecasting system also includes these features and thus with sufficient regularization must perform better than this baseline, it does provide some insight into how well we can infer the eventual rating from the risks alone in activity documents at approval. This score was only calculated for activities

where the LLM was able to successfully extract the grade for how risky the project is from the baseline documents. This method suffers from potential inconsistency in LLMs extracting the overall risk from baseline documents.

***gemini-2.5-flash* summary of risks**

A separate baseline used only for assessing the forecast grades of models, is submitting a summary of the risks identified from activity documents to the *gemini-2.5-flash-lite* grading LLM, and comparing these to proper LLM forecasts. This is a separate measure of the informativeness of risks identified in activity documents, and provides a sanity check that the LLM is not simply summarizing what the authors of activity of appraisal documents already highlight as risks. However, the lack of assertive claims about what will happen in a positive sense hinders this baseline metric relative to graded forecasts.

***deepseek-V3.2* provided only the activity summary and statistics**

I introduce several LLM methods closely informed by prior state-of-the-art methods in judgemental forecasting literature (Halawi et al., 2024) (Lee et al., 2025). In order to justify these methods, I compare to a simple baseline where all I provide the model is the activity ID, the activity title, an activity summary from *gemini-2.5-flash* of the 10 most important pages from the activity documents for forecasting the outcome, and the statistical prevalence of each outcome rating, and ask it directly to forecast the evaluation rating.

Ridge regression trained with non-LLM categories

In order to justify the addition of non-LLM categories, I use the baseline statistical categories used in prior literature and train a Ridge Regression Model on the outputs. All features included improve performance on the validation set for the final forecasting system, reducing performance when removed. I cite the sources which have found meaningful predictive effects or correlations between activity success or activity outcomes and the features listed.

The non-LLM features used were:

- planned activity duration (Goldemberg, Jordan, and Kenyon, 2025)
- planned total disbursement (consultation with BMZ officers), (Bulman, Kolkma, and Kraay, 2017) (Goldemberg, Jordan, and Kenyon, 2025) (H. L. Ashton et al., 2021) (Eilers et al., 2025)
- derived features: log(planned total disbursement), planned total disbursement per year
- whether the activity is primarily loan or grant-based (Bulman, Kolkma, and Kraay, 2017)
- the one-hot encoded reporting organization (Goldemberg, Jordan, and Kenyon, 2025)

- the Country Policy and Institutional Assessment (CPIA) score from the World Bank for that country (Bulman, Kolkma, and Kraay, 2017)
- the World Governance Indicators scores for control of corruption, government effectiveness, political stability, regulatory quality, and rule of law, as well as the mean of these values when none are median imputed (consultation with BMZ officers), (Goldemberg, Jordan, and Kenyon, 2025) (Ndikumana and Pickbourn, 2017)
- the scope of the activity on a scale from 1-7, ranging from local to global (Vivalt, 2020)
- the *log*(GDP/capita) of the countries where the activity takes place weighted by the percentage of the activity performed in each country. (Goldemberg, Jordan, and Kenyon, 2025) (Bulman, Kolkma, and Kraay, 2017)
- A dummy variable for each region as defined by the World Bank (AFE, Eastern and Southern Africa; AFW, Western and Central Africa; EAP, East Asia and Pacific; ECA, Europe and Central Asia; LAC, Latin America and the Caribbean; MENA, Middle East and North Africa; SAS, South Asia) (Goldemberg, Jordan, and Kenyon, 2025) (Vivalt, 2020)
- missingness counts for important features with missingness
- interaction terms for influential variables: gdp per capita times duration

The activity start date was not used as a feature, as there was no linear pattern with regards to overall activity success over time in the training data, and adding it tended to increase overfitting and reduce out-of-time generalization. As opposed to some results in prior literature (Vivalt, 2020) (H. L. Ashton et al., 2021), whether the implementer was a government, NGO, or another category was also not found to have any improvement in model performance and was excluded.

2.5 Experimental methods

2.5.1 Statistical Forecasting Methods

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, so I used this similarity method instead. I first weighted 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

number of neighbors using this method. I 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 15 similar ratings performed worse than the "most common rating" method and thus was not used as an experimental method to directly forecast ratings.

Random Forest

The Random Forest (RF) 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 RF 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, allowing each tree to be more independent of each other and improving generalization further. The bootstrap method is also used to train trees, encouraging tree independence.

XGBoost

The XGBoost (Extreme Gradient Boosting) method is a statistical algorithm which constructs an ensemble of decision trees sequentially, where each subsequent tree attempts to correct the errors made by the previous trees. Unlike the RF which trains trees independently and averages their predictions, XGBoost builds trees iteratively, with each new tree focusing on the residual errors of the ensemble.

The algorithm incorporates both L1 and L2 regularization terms to penalize model complexity and prevent overfitting on the training data. The decision trees split based on value ranges of the features, using a splitting criterion that accounts for the gradient and hessian of the loss function.

LLM-generated scores

Seven additional features were extracted using *gemini-2.5-flash*-generated evaluation on a score from 0 to 100 in order to enrich the activity with information from the project document PDFs from the start of the activity. These seven features were based on important aspects revealed by my literature review and consultation. I added one feature not introduced by others (the ease of targeted outcomes). Unlike technical complexity, which captures only one aspect of how difficult implementation is, or overall risk, which aggregates multiple threat sources including financial, legal, and regulatory, ease of targeted outcomes specifically reflects whether the activity's intended outputs are inherently measurable and achievable given typical implementation constraints which I

theorize to independently contribute to evaluation success.

The features added for this method included the seven features and two interaction terms

- *gemini-2.5-flash*-generated evaluation on a score from 0 to 100 of:
  - how well financed the activity is (consultation with BMZ officers), (Bulman, Kolkma, and Kraay, 2017) (Goldemberg, Jordan, and Kenyon, 2025) (Eilers et al., 2025)
  - the activity integratedness within the broader activity ecosystem (consultation with BMZ officers), (Vivalt, 2020)
  - the expected implementer performance including ownership (consultation with BMZ officers), (Bulman, Kolkma, and Kraay, 2017) (Goldemberg, Jordan, and Kenyon, 2025) (H. L. Ashton et al., 2021)
  - the degree of contextual challenge (consultation with BMZ officers)
  - the overall risk level (consultation with BMZ officers), (Eilers et al., 2025)
  - the activity's overall technical complexity (Eilers et al., 2025)
  - the ease of targeted outcomes
- Additional interaction terms added
  - governance times complexity
  - expenditure times complexity

Language Model Embeddings

Statistical models are unable to explicitly capture textual information. Accordingly, I insert this information via usage of embeddings, using the same Gemini embeddings model *gemini-embedding-001* of LLM-generated "targets" field (which summarize the objectives of what the activity aims to accomplish) as a semantic representation activity targets. This follows (Goldemberg, Jordan, and Kenyon, 2025) 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 projects not from the World Bank, 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 semantically encodes activity objectives.

I then replicate (Goldemberg, Jordan, and Kenyon, 2025) and compress target embeddings using a two-stage dimensionality reduction pipeline. This works via Principal Component

Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) (Healy and McInnes, 2024). PCA projects embeddings into a few directions that preserve most of their variation by finding orthogonal directions along which the data varies most, then projecting the data onto those directions. UMAP builds a graph of nearest neighbors in high-dimensional space, then optimizes a low-dimensional embedding to preserve those relationships. I first use PCA to reduce the embedding vectors to 50 dimensions and then fit UMAP on the PCA outputs to produce 2D and 3D coordinate maps. I find the 3D embeddings (umap_x, umap_y, umap_z) as features in the RF model perform better on prediction on the validation set for forecasting activity ratings than 2D or 4D, and thus chose 3 dimensions. This preserves enough local topology that activities with similar targets remain near each other in the compressed space. I also find from visual inspection, that sectors with similar 2D vectors cluster around the activity environmental category, replicating the finding in (Goldemberg, Jordan, and Kenyon, 2025).

By counting the word occurrences between features higher or lower on the UMAP x,y, and z axes, it is possible to investigate what aspects are reported by the embeddings, and the common sectors which they were categorized in. Broadly, low x values correspond to forestry, agriculture, water, and management related terms while high x correspond to energy sector and financing related terms. Low y correspond to similar terms as high x, both mostly in the energy sector. High y terms related to biodiversity, wastewater, and conservation. The z axis seems to relate more to urban vs rural: Low z corresponds to wildlife and undp related terms as well as "rural", while high z corresponds to sanitation and wastewater, as well as "urban" and "city". The z axis may also correlate slightly with contextualization, where the more "cookie cutter" objectives with regards to the sector in question are typically correlated with low z values. Lastly, the z axis also has a weak negative correlations with the number of countries: low z values tend to occur in fewer countries.

This method adds 4 additional features to the model:

- umap3_x: First UMAP coordinate capturing a spectrum from forestry, agriculture, and water management objectives (low values) to energy-sector and financing-oriented objectives (high values).
- umap3_y: Second UMAP coordinate separating energy-focused activities (low values) from biodiversity, wastewater, and conservation-oriented activities (high values).
- umap3_z: Third UMAP coordinate primarily reflecting an urban–rural axis, with lower values associated with rural and wildlife objectives and the UNDP, and higher values associated with urban sanitation and wastewater.
- a binary indicator for whether the umap embedding is missing (missing features are

always imputed using the median values in the training dataset)

Budget allocation for sector clustering

A key component of the efficacy of projects is how they allocate their funding. To do so, I query *gemini-2.5-flash* to identify the breakdown of outcome-related funding, where available in the baseline documents. Funding breakdowns were found to be important in prior literature (Goldemberg, Jordan, and Kenyon, 2025). However, unlike in (Goldemberg, Jordan, and Kenyon, 2025), I find that the measure of how much the funding allocations are split between different domains (the Herfindahl–Hirschman index (HHI)) does not improve performance on the validation set and was not added as a feature. The allocations of funding are required to approximately sum to the total expenditure of the project, with approximately 68% of projects having their budgets identified. If no high-scoring finance or budget pages were identified in the categorization step, I fallback to the first 10 pages of the highest rank baseline document.

In order to make the resulting dataset of budgets usable by the statistical model, I use the embeddings model to cluster the descriptions of the subsectors into 15 sector clusters (having tried 10, 15, and 20 clusters, I find 15 provides the optimal performance on the validation set). I sum the allocations within each cluster and report as a fraction of the total as a feature for the statistical model.

In summary, this adds the following 16 features to the statistical model:

- 15 features, one for each sector cluster expenditure percentage (Goldemberg, Jordan, and Kenyon, 2025)
- a binary missing indicator if there is no sector cluster information in the pdf. As for other features, the result is median imputed, meaning all features are set to 0 as that is the median for all sectors in the training data.

2.5.2 LLM Forecasting Methods

NOTE: I have a comprehensive overview of past literature involving LLMs in the introduction, but I've left it out of this draft.

Multiple studies have measured zero-shot LLM forecasting capability, and found that 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-V3.2* as my forecasting model, due to its strong performance and low cost. I also chose *gemini-3-pro* due to its strong performance. Lastly *gemini-2.5-flash* was selected only to assess the possibility of fine-tuning, as other more powerful models were unavailable for fine-tuning.

The LLM forecasting method was decided upon by iteratively inspecting both the quality of the response, and the overall accuracy of the forecasts 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.

To generate each mock forecast, I 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. I 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 were then retrieved (see Section 2.5.1. 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.

2.5.2.1 LLM Prompting Strategies

The full prompt template for the LLM Forecast is shown in Figure 1.

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 2.5.1 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 than $k = 1$, 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.

Ensembling Ensembling averages many individual forecasts of the same model, prompted in slightly different ways. I found ensembling was a relatively weak but robust method while validating, so due to its high API cost, I reserve the ensembling method for the final held-out set, which I will present once this thesis has reached completion.

```
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 1: 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}.

Fine Tuning In past work in a similar domain, fine-tuning significantly improved forecasting performance (Wen et al., 2025). I attempted to fine-tune *gemini-2.5-flash* using Vertex AI. To do so, I used Direct Preference Optimization (DPO), which requires one example of a good prompt, and one example of a bad prompt. Out of a random sample of 100 activity IDs in the training data, I forecasted the final forecast stage 5 separate times using *deepseek-V3.2*. I used 50 pairs where the model forecasted one rating increment closer than another rating, as the pairs of forecasts.

It was often the case that there were multiple options for the best or worse forecast due to the limited 6-point scale. In order to choose among forecasts that were equally good or equally bad, I also took the embeddings of the forecasts and found the cosine similarity to the embeddings of the mock forecast, and to embeddings of the outcome document and averaged these similarity scores. The most similar among equally good ratings were chosen as the good example for the DPO training pair, and the least similar among equally bad ratings as the bad example.

Once the 50 pairs were identified, I ran the fine-tuning using default settings over 20 epochs from Vertex AI (Learning rate multiplier of 1, Adapter size of 4, Beta of 0.1).

2.5.3 Recency and LLM Adjustment Ridge Regression

I wanted to both correct the RF 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 forecasts were significantly correlated with prediction error in the validation set.

Residual-correction formulation

Let $\hat{y}_i^{\mathrm{RF}}$ be the RF 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.

This is a simple stacked model: the RF provides the base signal, and ridge regression learns an adjustment to remove systematic residual error 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 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, forecasts fall back to the uncorrected RF output.

2.6 Scoring Metrics

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 RF, 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). I additionally estimate 95% confidence intervals for evaluation metrics using a percentile bootstrap, repeatedly resampling prediction-outcome pairs with replacement and computing the metric on each resample, taking the 2.5th and 97.5th percentiles of the resulting distribution.

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.

MAE (Mean Absolute Error) Measure of "average" distance, by taking the mean of the absolute value of the residuals. Lower is better. On a scale from 0 to 5, therefore worst possible value is 5, best possible value is zero. This method does not heavily penalize 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-time R^2.

Pairwise Probability Pairwise probability is evaluated as the proportion of pairs of individual predictions in the validation or test set that were correctly ordered from lower to higher rating. This method is insensitive to global shifts in ratings, which may occur due to events like the COVID 19 pandemic. It also is insensitive to calibration of the spread of possible outcomes. Given the significant noise related to globally relevant challenges, it can represent a more achievable and informative metric than R^2 or MAE for model forecasting skill. However, because pairwise probability does not reward ties, integer ratings artificially suppress performance compared to continuous predictions.

2.6.1 Choosing Rating and Cost Effectiveness Evaluation Metrics

After consideration of several metrics, I decided the Pairwise probability and R^2 were the most appropriate for assessing model skill in the domain of ratings and activity quantitative outcomes, while other metrics still provide useful insight. R^2 is sensitive to bias and outliers, which are important for assessing absolute predictive accuracy. However, a common use-case in aid funding decision making is comparing a pair of activities, or even a suite of many activities. In such a use-case, the Pairwise probability is more representative of what is needed. Furthermore, a weakness in R^2 is that global shifts in

the ratings and outcomes due to external factors can be unpredictable, and a model may by-chance capture these shifts without any real improvement in forecasting skill. Both are clearly interpretable: always forecasting the mean value is 0, equivalent to 0% of variance explained for R^2, while 50% is equivalent to the pairwise forecast of random chance.

2.6.2 Grading Free Form Forecasts

In addition to extracting the LLM forecasted rating, I also use *gemini-2.5-flash-lite* to grade narrative textual format forecasts. In order to do so, I had *gemini-2.5-flash* first summarize pages categorized as highly relevant to outcome evaluations, obtaining 10 pages with a score of at least 3/10 marked as "deviation_from_plans", "delays_or_early_completion", or "over_or_under_spending", or failing that, that were graded as highly relevant to activity evaluation. If no such pages were categorized, the first 10 pages of the highest ranked activity outcome evaluation were used. Next, for all ensembles of LLM forecasts, a grade was given for how similar the forecast was to the activity outcome using *gemini-2.5-flash-lite*. Grading was performed based on two key criteria: accurately identification of likely drivers of activity success or failure, and identification of likely outcomes being forecasted.

In accordance with the typical US grading scheme definitions, I define an "F" grade as 55 or lower and an A+ grade as approximately 97, with other grades defined in even intervals. I provide *gemini-2.5-flash-lite* the following rubric:

Grading scale

- A+/A/A-: Excellent forecast, highly accurate, attention on key drivers, multiple major events forecasted accurately.
- B+/B/B-: Good forecast, mostly accurate. A mix of correct and incorrect, but at least one major outcome was forecasted. At least one key driver identified.
- C+/C/C-: Adequate forecast, partially accurate or reasonable. Focus was adequate. Major outcomes incorrect, but some smaller aspects were correct.
- D+/D/D-: Poor forecast, although perhaps one or two small correct things. Mostly inaccurate. Wrong focus on drivers.
- F: Failed forecast, completely wrong or unsupported.

3 Results

I will discuss the activity overall rating and similarity grade forecasts, and aggregate Z-score forecast results below. First, I discuss my key findings with regards to the best method for evaluating international aid forecasting methods.

I have highlighted three key methods for measuring activity outcomes: evaluating forecast accuracy of overall evaluation ratings, evaluating forecast accuracy of overall aggregate cost-effectiveness Z-score, and scoring the narrative similarity between LLM forecast and the final outcome. In my experimentation, I discovered several reasons to be wary of activity ratings: they are often gamed as they are influential in future funding (Goldemberg, Jordan, and Kenyon, 2025), I find that they do not correlate strongly with activity outcomes (Pearson score of only 0.07), and I find that models with high forecasting accuracy sometimes have low similarity to the final forecast (see Figure 2). The cost-effectiveness aggregate Z-score suffers from inconsistency in its component parts between activities, its difficulty of interpretation, and the relatively low numbers of activity outcomes. I find that the method of grading narrative similarity is often a superior metric than the other two. I conclude that evaluating whether narrative forecasts are similar to the activity document outcomes using *gemini-2.5-flash-lite* is the best overall evaluation method. It is low-cost, costing only a few cents for 300 grades, and only takes around 2 minutes for those 300 grades. Furthermore, prior literature has found that the semantic embedding of activity results are correlated with quantitative country-level outcomes (Goldemberg, Jordan, and Kenyon, 2025), indicating that these are less likely to be "gamed".

I also set out to assess specifically how statistical models and LLM forecasting techniques compare. Both strategies were found to have both strengths and weaknesses:

Strengths

LLM Forecasting

- Can explicitly reason and identify missing information
- General reasoning skills may transfer across domains
- Can produce text-based forecasts of specific events

Statistical Models

- Cheap and fast to iterate
- Can use large numbers of features without context limits
- Can incorporate LLM forecasts and grades as features
- Efficiently generalize over large datasets
- Easier to prevent future leakage
- Mature methodology

Weaknesses

LLM Forecasting

- Expensive and slow to iterate
- Difficult to calibrate
- Fine-tuning is expensive, and unavailable for best models
- Limited interpretability of model reasoning
- Best models are closed-source
- Risk of training-data leakage
- Limited context window constrains attention and calibration

Statistical Models

- Cannot perform explicit reasoning
- Cannot directly use rich textual or world knowledge
- More prone to overfitting dataset-specific quirks

I will proceed to describe the results from each method in detail in the remaining sections.

3.1 Forecasting Ratings and Freeform Outcomes with LLMs

One research question of this thesis is how do differing methods of LLM forecasting apply in the context of forecasting international aid success ratings. I find that more generally capable models are better forecasters. I also find that the forecasting ability of the LLM to predict the rating does not clearly correlate with to the contents of the LLM forecast, and in fact find that the best rating produced from an LLM had the least accurate reasons for the forecasts. The strategy of incorporating additional information relevant to the forecasting question as well as references to similar activities does not clearly improve forecasting skill. I find no evidence of data leakage contaminating the ratings, with similar patterns of temporal correlation between statistical model average rating skill and LLM average rating skill as a function of start date. I find little evidence that fine-tuning can improve forecasting accuracy in this domain, although a thorough assessment of the potential for fine-tuning *gemini-2.5-flash* was cost prohibitive. I also find language models on their own consistently significantly underperformed statistical models in forecasting skill. However, using statistical models as the drivers of LLM forecasts, I find that statistical models significantly improve the quality of LLM narrative forecasts.

In addition to extracting the LLM forecasted rating, I also assess LLM grades. I find that the assessed similarity between forecasts and summaries of outcome evaluations is usually correlated with improved ability to predict ratings, with the exception that very low

context prompts are able to forecast accurately but for the wrong reasons (the "Baseline" method) (see Figure 2). Furthermore, I find that when the predicted rating using the best statistical model is injected into the prompt, LLM forecasts of activities become much more similar to the summaries of their outcome evaluations. This provides evidence that while language models can accurately forecast the reasons activities would succeed or fail if given the evaluation ratings, they generally lack the ability to correctly weigh the various factors to produce a calibrated forecast. We should be wary about using rating accuracy prediction as a sole benchmark for LLMs, given that the LLM prompt with the worst similarity to the actual rating also had the highest forecasting success. This is likely due to increased model ability when the context window is not as full - the model is better at forecasting when not overwhelmed with information in its context window, but lacking that key information worsens its ability to pick out the best reasons for the forecasts.

While ratings are not clearly correlated with better similarity grades, RAG, KNN retrieval, and extra reasoning of the models (stages 1 and 2) consistently improve narrative textual similarity to the outcome, albeit by a small margin. Simply presenting the grading model with the *gemini-2.5-flash* summary of risks of the project from the project documents lead to a forecasting score of 80.2 (between a "C+" and a "B-", corresponding to an "Adequate" forecast, with only "Adequate", rather than a good, focus on the drivers). However, the risks did not typically include any affirmative forecasts of outcomes, which may in part explain the low ratings. The best forecasting method in terms of similarity to outcomes was the forced ratings with a large amount of context (RAG+KNN) and included stage 1 and 2, achieving a grade of 88.5, which corresponds to a "B+" grade given by *gemini-2.5-flash-lite* (which sits midway between "Good" and "Excellent", meaning the forecasts were on average between "mostly accurate" and "highly accurate", with at least one major outcome forecasted, and at least one key driver of the outcome identified on average). Without any additional methods, *deepseek-v-3.2* can typically achieve a "B" grade on average, while use of the statistical model or of *gemini-3-pro* can lift scores to the "B+" range. The improvement in ability to forecast overall ranking improves significantly as well.

Fine Tuning I found the training loss for *gemini-2.5-flash* steadily reduced over the course of the fine-tuning, reducing by 95%. I tested this on an early subset of the validation data which excluded BMZ activities. The performance results were mixed. While the R^2 performance improved modestly, it remained worse than simply guessing the mean rating in the validation set, and remained far below the performance of the more powerful *gemini-3-pro*. Furthermore, the DPO objective is specifically intended to improve pairwise performance between two pairs of forecasts. However, the Pairwise probability score was slightly worsened by fine-tuning. The cost for the fine-tuning of 20 epochs for 50 pairs was close to 140 euros. Therefore, scaling up larger training points for fine-tuning does not seem to be a promising strategy, either to reach the performance of *gemini-3-pro*, or

to reach the performance of the more powerful combined statistical models.

RAG and KNN I find small differences in forecasting skill for the addition of RAG and KNN on forecasting ratings. However, the narrative similarity scores consistently improve on the validation set when RAG and KNN context is included. This indicates that while the LLM is able to identify more important information when reasoning about the final forecast, it is insufficiently calibrated to use the additional information it has gathered to improve the forecast itself.

LLM Model Selection I find no consistent difference between *gemini-2.5-flash* and *deepseek-V3.2* in forecasting skill. However, *gemini-3-pro* appears to be significantly better at forecasting than both, in both sets of activities achieving higher similarity grades, higher R^2, and higher Pairwise probability than other models. The only exception to this is the models which have been instructed to report an overall forecast which matches the forecast from statistical models and the single instance where *deepseek-v3.2* using only the minimal prompt predicts a modestly better R^2 and Pairwise probability for the ratings, while scoring a much worse overall forecast similarity score.

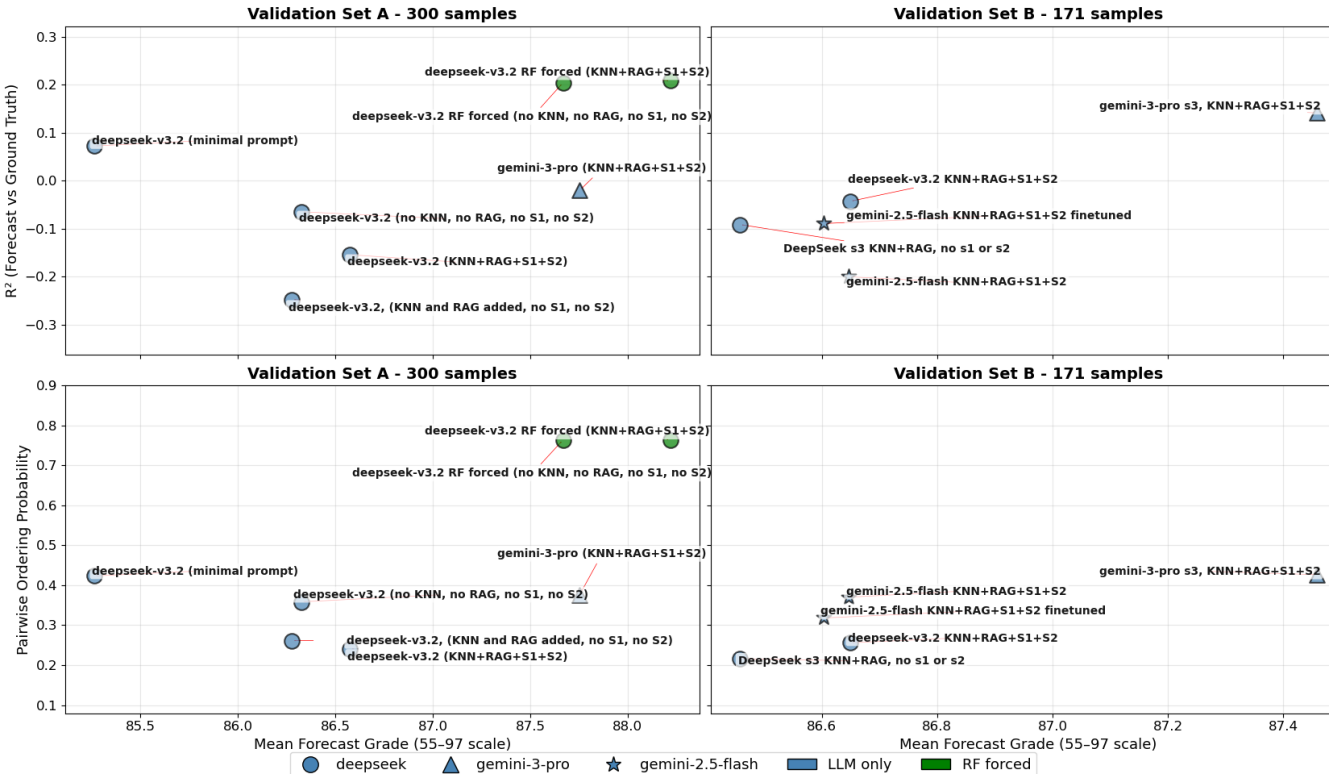

Figure 2: A comparison of various LLM forecasting methods. Validation set A contains 300 activities starting after the cutoff date of February 6th, 2013 and starting before the test set cutoff of June 6th, 2016. Validation set B contains an earlier set of 171 activities that included some in the test set and some in the training set from Validation set A, and excludes BMZ activities. "KNN" refers to the retrieval of 3 similar examples into the prompt, "RAG" refers to the model additioanlly retrieving information it decided it was missing from the available documents and adding to the prompt, "S1" refers to additional reasoning injected into the context considering why the outcome may be overall unsatisfactory, "S2" refers to additional reasoning in the context considering why the outcome may be overall satisfactory. Notably, *gemini-3-pro* outperforms other models when considering similarity grades. While information injection appears to help the forecasts come to their conclusions for the right reasons, it also appears to harm forecasting skill. When only the Activity ID, title, and a *gemini-2.5-flash* summary of the most important pages for activity forecasting were injected ("minimal prompt"), the LLM performed better on absolute ratings. Significantly better narrative grades are achievable when the model is forced to forecast the same rating as the RF model, but only if additional information such as RAG and KNN is given to the model. Because LLM forecasts are integers on a 0-5 point scale, Pairwise ordering is artificially suppressed and should be interpreted directionally.

3.2 Forecasting Ratings with Statistical Models

Overall, the forecasting system I produce is capable of forecasting evaluation ratings significantly above chance on out-of-time activities. Compared to prior work (L. Ashton et al., 2023), I report a value consistent with their adjusted R^2 on the training set. I report an R^2 of 0.34 and an adjusted R^2 of 0.29 for within-training set correlations on primarily world bank ratings, while others report at maximum an adjusted R^2 of 0.3 (Goldemberg, Jordan, and Kenyon, 2025). I was not able to identify comparable out-of-time, time-ordered split analysis in the literature. As expected for forecasting under data distribution shift, my results on the out-of-time validation set were somewhat weaker, with an R^2 of 0.23 for the RF model, and an R^2 of 0.26 when incorporating the language model forecasting results correction + recency model. I find recency provides better results than averaging the two models, due to their widely differing forecasting skill.

3.2.1 Overfitting Corrections

"Adjusted R^2" has been used in similar work to evaluate model performance in the development aid literature and penalizes 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 RF 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`=20), which smooths forecasts 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. Overall, compared to defaults, these choices trade some bias for a meaningful reduction in variance, making the fitted ensemble less susceptible to overfitting.

3.2.2 Embedding Targets

The language model derived features modestly aided forecast accuracy, in aggregate providing an improvement of about 7% additional explanation of the variance of outcomes out of the 26% 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.

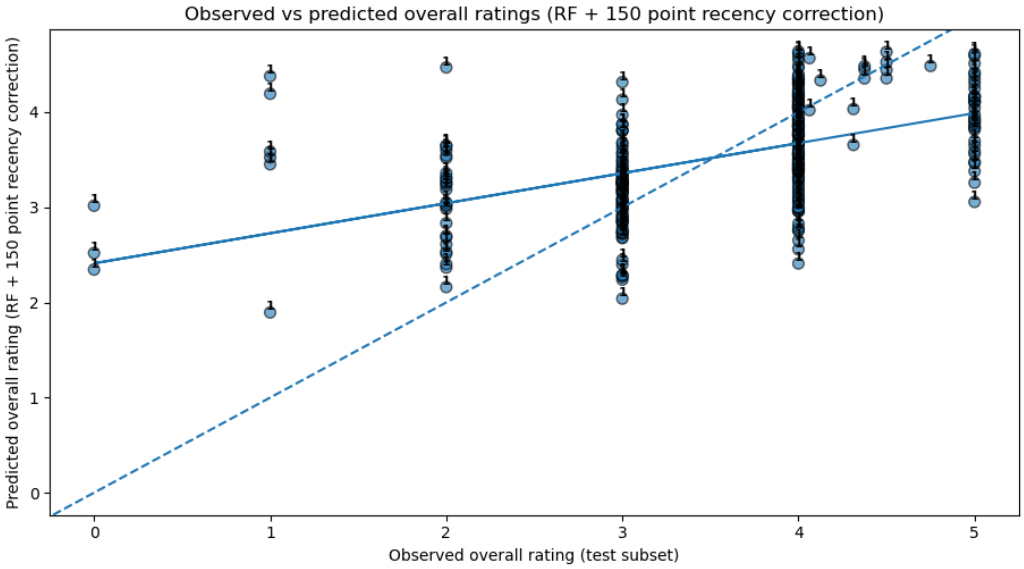

Figure 3: The scatter of observed vs predicted points on the validation set for the LLM-corrected RF prediction. $R^2 = 0.26$, Pairwise probability of 77 %.

SHAP analysis

I use SHAP (SHapley Additive exPlanations) to interpret model predictions by attributing changes in the output to individual features using Shapley values. For each observation, SHAP decomposes the prediction into a baseline (the mean model output) plus additive feature contributions. Each feature's SHAP value represents its average marginal contribution across all possible feature subsets.

An analysis of the split decisions using the SHAP model reveals that increased planned expenditure and decreased duration were found to strongly correlate with activity success, and the boolean dummy variable identifying reporting organization, in this case the World Bank which has lower ratings than the mean in this dataset, were the most important features for the RF model. Next, the urban/rural indicator from the embeddings tended to decrease ratings, also weakly indicating that less well contextualized ratings performed poorly. Finally, contextual challenge weakly shifted ratings downwards, while the ease of targets shifted model predictions upwards. Projects across many countries tended to perform slightly worse.

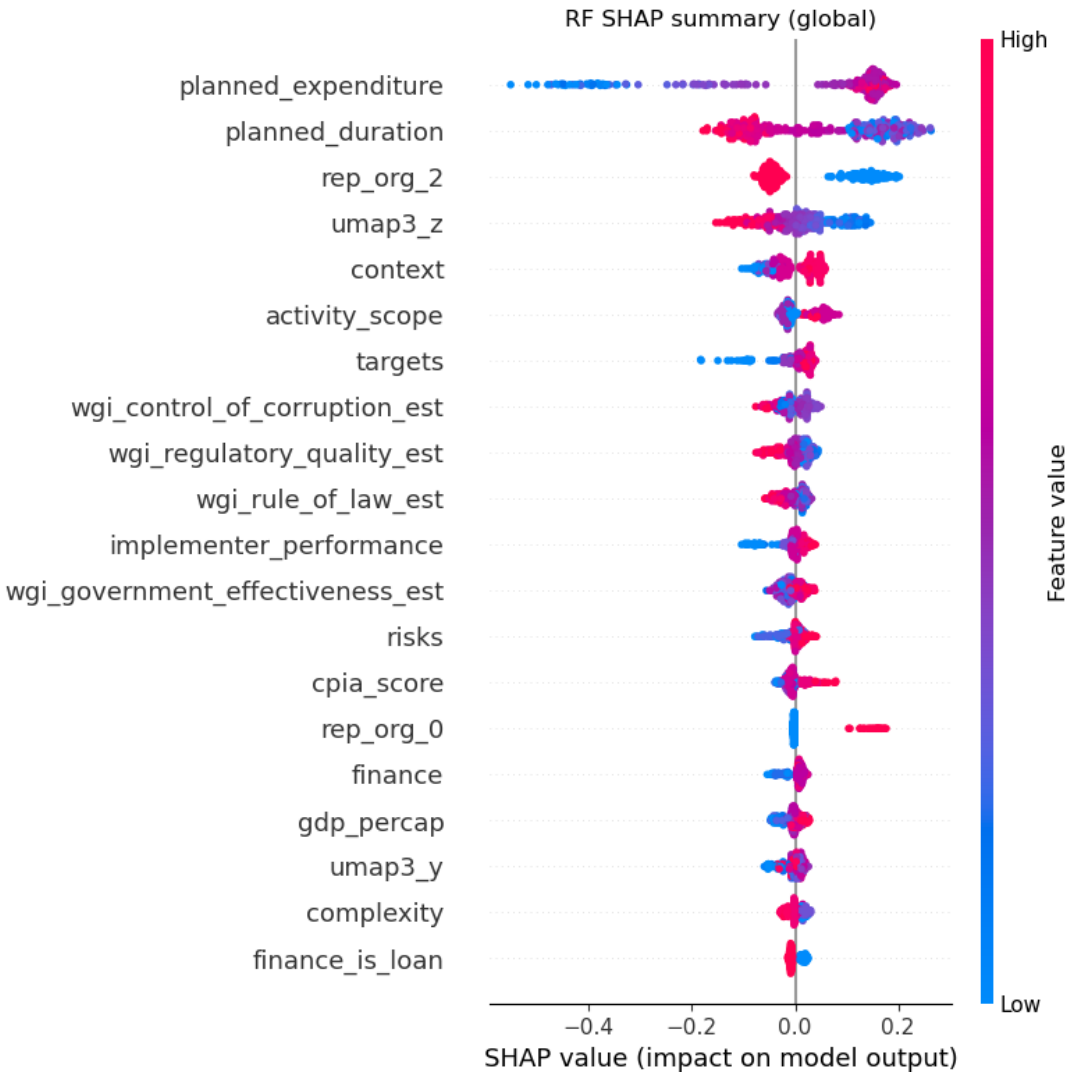

Figure 4: A SHAP analysis of ratings on the validation set from the RF. Red indicates an increase in the value of the feature, while blue indicates a below-average value. Points to the right of zero shift ratings up, points to the left of zero shift ratings down.

Table 1: Validation performance in forecasting ratings across forecasting methods. Rows are sorted by ascending R^2 (higher is better). RMSE and MAE are lower-is-better, others are better if higher. Bold indicates the best value in each metric column. I noticed that MAE was better when the metric was rounded, so I tried rounding my best model and found it beat the baseline metric on MAE. This is because the outcome ratings are integers, but the statistical models produce continuous prediction. The "recency correction" variants combine models using the 150 latest-starting activities in the training set for calibration/combination. XGBoost was found to underperform the RF when used as the LLM Forecast and recency combination base. The Pairwise metric is artificially suppressed for the "Mode of reporting-org score baseline" and the "RF + LLM Forecast + recency (rounded)" as these were always integers with frequent ties. The "no LLM features" RF excludes the following features: finance, integratedness, implementer_performance, targets, context, risks, complexity, umap3_x, umap3_y, umap3_z.

Method	R^2 ↑	RMSE ↓	MAE ↓	Side Acc. ↑	Acc. ↑	Pairwise ↑
RF + LLM Forecast + recency	**0.258**	**0.815**	0.590	**0.760**	**0.547**	**0.767**
RF + recency	0.254	0.817	0.593	0.750	**0.547**	0.767
XGBoost only	0.233	0.829	0.618	0.713	0.487	0.759
RF (default params)	0.214	0.839	0.622	0.703	0.507	0.757
RF + LLM Forecast + recency (rounded)	0.203	0.845	**0.535**	**0.760**	**0.547**	0.537
RF only	0.203	0.845	0.641	0.710	0.520	0.767
Ridge GLM + RF (mean)	0.198	0.848	0.644	0.697	0.523	0.756
XGBoost, no LLM features	0.189	0.852	0.661	0.700	0.493	0.729
RF, no LLM features	0.186	0.854	0.658	0.710	0.517	0.749
Ridge GLM	0.156	0.869	0.668	0.683	0.507	0.722
Mode of reporting-org score baseline	0.102	0.897	0.579	0.700	0.530	0.426
Ridge Baseline (risks + org only)	0.017	0.938	0.709	0.570	0.433	0.651

Table 2: Random-forest drop-one feature importance, sorted by absolue change in the prediction from the feature being increased by 1 standard deviation (Δ `pred_1sd`). 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 (`Miss` %}). The largest impacts come from planned budget and planned duration.

Feature	Miss %	$\Delta R^2 \uparrow$	ΔPairwise ↑	Δpred_1sd
planned_duration	0.0	0.0590	0.0305	-0.1028
expenditure_per_year_log	13.3	0.0051	0.0082	0.0978
planned_expenditure	8.7	0.0247	0.0210	0.0899
log_planned_expenditure	8.7	0.0326	0.0262	0.0642
activity_scope	0.0	0.0226	0.0249	0.0461
umap3_z	9.3	0.0249	0.0175	-0.0419
context	0.7	0.0316	0.0266	0.0403
cpia_score	44.7	0.0060	0.0149	0.0400
targets	0.7	0.0302	0.0243	0.0392
gdp_percap	10.0	0.0231	0.0201	0.0311
country_distance	9.3	0.0219	0.0193	0.0252
umap3_y	9.3	0.0260	0.0230	-0.0212
implementer_performance	1.0	0.0262	0.0226	0.0192
umap3_x	9.3	0.0341	0.0244	0.0183
sector_cluster_Capacity_Build	29.0	0.0442	0.0255	-0.0165
sector_distance	9.3	0.0368	0.0264	0.0154
complexity	0.3	0.0211	0.0201	-0.0106
finance	1.3	0.0155	0.0209	0.0076
governance_missing_count	0.0	0.0287	0.0227	-0.0070
risks	0.3	0.0206	0.0204	-0.0042
cpia_missing	0.0	0.0297	0.0236	-0.0033
sector_cluster_drinking_water	29.0	0.0222	0.0235	-0.0030
sector_cluster_managed_land	29.0	0.0262	0.0211	-0.0027
governance_composite	9.7	0.0061	0.0140	0.0023
region_EAP	0.0	0.0300	0.0252	0.0022
sector_cluster_Improved_transport	29.0	0.0204	0.0208	-0.0021
sector_cluster_Project_Management	29.0	0.0073	0.0193	-0.0019
sector_cluster_Institutional_cap	29.0	0.0224	0.0219	-0.0017
sector_cluster_reduced_CO2	29.0	0.0316	0.0248	0.0008
sector_cluster_Road_safety	29.0	0.0377	0.0257	-0.0007

3.3 Forecasting Cost-Effectiveness

In general, cost-effectiveness forecasts were weaker than ratings. A similar work found an adjusted $R^2 > 0.7$ for outcome ratings (Goldemberg, Jordan, and Kenyon, 2025), 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. My attempt to replicate their result revealed that they likely had training data leakage, but I have not had a response to my inquiries about whether the final code used to produce the tables in their paper was indeed the code containing the bug. Even if they correctly implemented their method, there are some reasons to think my outcomes would be harder to predict. Their paper did not predict on cost-effectiveness, making outcome prediction a much easier task when given overall program spending. Also, the outcomes with high detected correlation measured a lagged 5-year country-level indicator, which is less susceptible to reporting or extraction error, while my data were extracted directly from the outcomes using *gemini-2.5-flash*. Lastly, I found when replicating their methods that a model using their feature set only achieved an R^2 of approximately 0.1 for ratings, not the 0.3 which is implied in their paper.

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

- Rarity of quantitative outcomes.
- Unclear apportioning of funding towards each outcome, which is challenging to extract.
- Inconsistent measurement styles and definitions of terms like Benefit-Cost Ratio, or Economic Rate of Returns.
- Incorrect aggregation of multiple ratings within the documents. I find inaccurate aggregation was initially (artificially) increasing my prediction accuracy, as the errors were more predictable than the outcomes themselves.

I find outcome prediction to be challenging. My best model has an R^2 of 0.05, which indicates in absolute accuracy it likely beats a simple baseline of predicting the mean of the aggregate cost-effectiveness Z-score for the validation set with 95% confidence (see Figure 3). However, I cannot distinguish this absolute prediction accuracy from the baseline with 95% confidence. More encouragingly, I find my model is able to distinguish cost-effectiveness between pairs of progress, at a rate of 60% accuracy, compared to random chance of 50% with 95% confidence. The top two features of the model are UMAP x axis (higher x-axis values reduce cost-effectiveness), and the planned duration (longer durations increase cost-effectiveness). This stands in contrast with findings in the literature that better ratings are correlated with shorter durations. Although my model is not sufficiently strong to lend much confidence in this finding, I do find from the model interactions that

Table 3: Predictive performance of cost-effectiveness outcome RF model on the validation set. R^2, Pairwise ordering probability, and Spearman correlation are shown with 95% bootstrap confidence intervals. In general, only the aggregate had sufficient data to produce a meaningfully predictive model, although the model is only able to distinguish more cost-effective activities 60% of the time. Outcomes are sorted by R^2 (descending). The Benefit-Cost Ratios had identical forecasts for all examples, hence a pairwise and Spearman could not be properly calculated. All other outcomes had fewer than 10 items in training or validation. 95% confidence intervals are determined via bootstrap over the prediction set, which was only computed once.

Outcome	R^2	Pairwise (%)	Spearman	MAE	$N_{\text{train}}/N_{\text{val}}/N_{\text{test}}$
All selected outcomes (z-aggregate activity mean)	0.05 [-0.03, 0.13]	60 [54, 67]	0.31 [0.13, 0.46]	0.56	682 / 299 / 299
Water Connections Connections ($log_{10}(connections)$)	-0.01 [-0.39, 0.13]	66 [54, 77]	0.43 [0.07, 0.69]	0.70	71 / 25 / 25
Economic Rate Of Return Percent (percent)	-0.01 [-0.19, 0.03]	53 [45, 60]	0.09 [-0.14, 0.28]	15.39	293 / 103 / 103
CO_2 Emission Reductions Tonnes CO2E ($log_{10}(tonnes\ CO_2e)$)	-0.03 [-0.32, 0.00]	41 [27, 58]	-0.28 [-0.62, 0.18]	1.46	38 / 23 / 23
Benefit Cost Ratios Ratio (ratio)	-0.05 [-0.42, -0.00]	N/A	N/A	0.33	44 / 18 / 18
Financial Rate Of Return Percent (percent)	-0.15 [-1.19, 0.03]	51 [39, 64]	0.01 [-0.32, 0.36]	15.20	147 / 43 / 43
Generation Capacity ($log_{10}(Mw)$)	-0.32 [-1.13, 0.02]	62 [47, 75]	0.39 [-0.06, 0.68]	0.73	44 / 20 / 20

longer planned activities are more cost-effective.

In keeping with the results of (Goldemberg, Jordan, and Kenyon, 2025), I do not find a significant correlation between activity ratings and Z-scored outcomes. I found an overall Pearson correlation of only 0.07 between cost-effectiveness and ratings over the entire dataset for the Z-aggregate score (N=566). Overall activity ratings should therefore be used with caution when evaluating programs.

In conclusion, I find that outcome cost-effectiveness needs more work to extract sufficient data for reliable outcome prediction. Furthermore I find that obtaining cost-effectiveness metrics is more challenging than obtaining ratings, in contrast to the claims from prior literature (Goldemberg, Jordan, and Kenyon, 2025).

I do find that failing to divide by the total disbursement as marked in IATI increases predictability. This is in part because there is some noise inherent in the total expenditure, and also because when not dividing by expenditure, the model learns to predict linearly higher outcomes correlating with the expenditure. 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 0.05, (95% CI: -0.03, 0.13) and pairwise probability of 60% (95% CI: 54%, 67%) ($N_{\text{val}} = 299$), compared to an R^2 of 0.10 (95% CI -0.0982, 0.2625) and a pairwise probability of 68% (95% CI: 0.62, 0.74) (n_val=84) when forecasting the Z-score for CO_2 emission reductions, generation capacity, and water connections. While pairwise ordering ability for outcomes is statistically significant, the overall ability is weak and appears to be largely a function of expenditure and sector clusters.

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 coarse assumption that all funding for a project goes to all outcomes. New water or sanitation piping connections is a more clearly comparable outcome, although there may be systematic differences in the costs of 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 more robust ordering of overall activity cost-effectiveness than evaluator ratings alone, although outcome predictions are not statistically significantly better than forecasting the mean value for the outcome in a given sector in directly forecasting a single activity's cost-effectiveness.

4 Conclusion

I conclude by summarizing my findings regarding each research question posed in the introduction.

How do judgemental forecasting methods compare to statistical models in forecasting international aid overall success ratings and quantitative outcomes? LLM forecasts themselves consistently underperform statistical models in this domain, while having the disadvantage of being costlier and more difficult to iterate with. However, they provide meaningfully accurate narrative forecasts when combined with statistical models.

How do differing methods of combining judgemental and statistical forecasting compare in this domain? Embeddings of language models inserted into statistical models significantly improve forecasting ability. Furthermore, using embeddings as a means of selecting nearest neighbors and as a component of RAG retrieval gathered information that reliably increased the similarity grades between LLM forecasts and the eventual outcome over all tested LLM forecast configurations. Embeddings were also effective as a method for clustering activity disbursements as measured by improved

performance on the validation set.

Language models themselves were helpful in extracting grades for various aspects of the forecast, which as a group improved the statistical model forecast. While direct averaging did not demonstrate improved performance, I found that training a simple model to use the residual between the LLM and statistical model modestly improved forecasting skill across most metrics. There is some evidence that training a secondary model to incorporate the final LLM prediction with a RF prediction can also improve the overall forecast skill.

What methods improve the accuracy of narrative (qualitative) forecasts in this domain? It appears that the only reliable way of improving ratings over a range of different configurations was 1. to switch the model from the somewhat smaller *deepseek-V3.2* or *gemini-2.5-flash* models to the more expensive but more generally capable *gemini-3-pro* model, 2. to incorporate additional retrieval of information via RAG and inserting outcomes from similar past activities using the KNN technique and 3. to directly prompt the model to come to the same conclusion as the RF model.

Additional prompts to elicit reasons the outcome may go well or badly (stages 1 and 2 in my methods) sometimes deteriorated scores in certain model configurations, although the best performing models used these stages in its context window when forecasting. Similarly, fine-tuning sometimes helped, and sometimes deteriorated scores.

What aspects of the activity available in my dataset at the beginning of the activity lead to higher or lower ratings? As others in the literature have reported, decreased duration and increased planned expenditures tend to correlate with higher activity ratings (Vivalt, 2020) (L. Ashton et al., 2023) (Eilers et al., 2025). However, there is moderate evidence from the cost-effectiveness RF model that cost-effectiveness of quantitative outcomes does not clearly correlate with increased spending, and that increased durations also increase cost-effectiveness. Rating organizations tend to differ systematically, and by-sector differences in ratings appear to be more significant in general than regional differences. However, this was not the core focus of this work and needs further research.

How does forecasting aggregate cost-effectiveness activity outcomes compare to forecasting ratings? The current dataset appears insufficient to forecast activity outcomes, both in numbers of quantitative outcomes and noisiness inherent in their extraction. There is some evidence in the validation set that an aggregate measure of cost-effectiveness can be used to rank promising activities, with a 60% chance of a correct ordering compared to 50% which would be arrived at by random chance.

Works Cited

References

3ie Development Evidence Portal | 3ie (2025). https://developmentevidence.3ieimpact.org/. (Visited on 08/18/2025).

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

Ashton, Louise et al. (Feb. 2023). "A Puzzle with Missing Pieces: Explaining the Effectiveness of World Bank Development Projects". In: *The World Bank Research Observer* 38.1, pp. 115–146. ISSN: 0257-3032. DOI: 10.1093/wbro/lkac005. (Visited on 01/10/2026).

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).

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

Callaghan, Max et al. (Feb. 2025). "Machine Learning Map of Climate Policy Literature Reveals Disparities between Scientific Attention, Policy Density, and Emissions". In: *npj Climate Action* 4.1, p. 7. ISSN: 2731-9814. DOI: 10.1038/s44168-024-00196-0. (Visited on 02/04/2026).

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

Goldemberg, Diana, Luke Jordan, and Thomas Kenyon (Feb. 2025). "Minding the Gap: Aid Effectiveness, Project Ratings and Contextualization". In: *The World Bank Economic Review*, lhaf005. ISSN: 0258-6770. DOI: 10.1093/wber/lhaf005. (Visited on 01/04/2026).

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).

Healy, John and Leland McInnes (Nov. 2024). "Uniform Manifold Approximation and Projection". In: *Nature Reviews Methods Primers* 4.1, p. 82. ISSN: 2662-8449. DOI: 10.1038/s43586-024-00363-x. (Visited on 02/12/2026).

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

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

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

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).

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

Ndikumana, Léonce and Lynda Pickbourn (Feb. 2017). "The Impact of Foreign Aid Allocation on Access to Social Services in Sub-Saharan Africa: The Case of Water and Sanitation". In: *World Development* 90, pp. 104–114. ISSN: 0305-750X. DOI: 10.1016/j.worlddev.2016.09.001. (Visited on 01/19/2026).

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

Tierney, Michael J. et al. (2011). "More Dollars than Sense: Refining Our Knowledge of Development Finance Using AidData". In: *World Development* 39.11, pp. 1891–1906. ISSN: 0305-750X. DOI: 10.1016/j.worlddev.2011.07.029. (Visited on 02/04/2026).

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).

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).

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.*

Declaration of AI Use

ChatGPT 5 and Claude Sonnet were used occasionally to write individual sentences in the Methods section summarizing code and to format equations in LaTeX. All LLM authored text was carefully checked for validity. ChatGPT 5.2 and Claude Sonnet were also used to provide feedback for spelling, writing style and grammar.

Potsdam, 23 February 2026