

Methodology Note

[1. Overview](#)

[2. Data Source](#)

[3. Building the activity panel](#)

[4. LLM classification method](#)

[5. Correction methodology](#)

[5.1 Estimating the error rates and confidence intervals](#)

[5.2 Correcting activities with evidence as a component](#)

[5.3 Applying the corrections](#)

[6. Limitations](#)

[Appendix A: LLM prompt for classifying development activities](#)

[Appendix B: Top 10 donors by gross disbursement from OECD CRS](#)

1. Overview

This methodology note describes the process followed to estimate how much development agencies spend on evidence generation. Evidence generation is defined here as research activities that estimate the causal impact of a policy/intervention/program. This classification is not readily available, so we used the International Aid Transparency Initiative (IATI) data as a repository of development project information to investigate how many projects contain evidence generation and how much is spent on it. We limit the universe to the top nine development agencies by disbursement size, as captured by IATI reporting, between 2010 and 2025. The data and sample construction process are outlined in Sections 2 and 3.

After building a panel of project spending by year and donor, we perform textual analysis on the project descriptions to determine whether each project contains evidence generation. In practice, we use Claude Sonnet 4.6 and ask the Large Language Model (LLM) to classify each project activity into evidence generation vs. not evidence generation. We also ask for a confidence level on that decision (low, medium, or high) and a reasoning. If the project is classified as evidence, we ask it to classify further whether the project is primarily evidence generation or only contains an evidence component. In total, we classified 79,736 development projects, resulting in 631 positive cases. We include more details on the LLM classification method and process in Section 4.

In Section 5, we perform two corrections on the raw classifications to obtain disbursements on evidence generation:

1. We correct for potential errors in the raw classifications into evidence-generation by estimating classification error rates. To do this, we stratify the sample into six groups by their evidence classification (evidence vs. not evidence) and classification confidence (high, medium, low). We estimate average error rates per stratum by manually checking a random sample of 152 projects. We then apply the Clopper-Pearson exact method to obtain 95% confidence intervals around the error rates.
2. We correct for evidence spending for evidence-as-component projects, wherein only a portion of the project activity involves evidence generation. We cap the disbursement for evidence-as-component projects at \$1.4 million, the median cost of impact evaluations, as estimated using verified spending on independent impact evaluations from the Millennium Challenge Corporation (MCC).

Applying these two corrections, we end up with 745 projects classified as evidence, with around \$1.6 billion spent on evidence between 2010 and 2025.

Our analysis has several limitations, which we outline in Section 6. Among the most important are: reporting on IATI is voluntary, so the results are driven by the reporting practices of the organisations; the LLM classifies projects based on their descriptions, not on the project documentation or what actually happened to each project; and the corrections are based on a small sample and on a single agency used to price evidence-as-a-component activities. In summary, this is our best guess of how much nine development agencies have spent on evidence generation since 2010.

2. Data Source

Our main source of information is the International Aid Transparency Initiative (IATI). This is an open data standard, launched at the Third High Level Forum on Aid Effectiveness in Accra, Ghana, in 2008, through which aid organizations publish details of their activities. Currently, IATI has a Secretariat hosted by the United Nations Development Programme (UNDP), and the service is provided through the United Nations Office for Project Services (UNOPS). Reporting to IATI is voluntary, which means that development organizations control what data is released and the level of detail.

In February 2026, we downloaded the full IATI dataset as a single SQLite database from IATI Tables, the IATI Secretariat's relational version of the published XML, using the build dated 13 January 2026. Donors report their information as XML files, and the IATI Secretariat then transforms the information into relational databases that the public can access in various formats. This means that our data comes from a snapshot built in January 2026. Any changes made to the data by publishers since then will alter the relational database that the public can access today. Finally, we end up with a dataset containing 203 tables, roughly one per reporting field, around 909,000 activities, and 1,775 reporting organizations. In our analysis, we used the information from the tables for activity, transactions, related activities, descriptions, sector, and money values.

As a complementary exercise, we tried using information from the OECD Creditor Reporting System (CRS)¹. Appendix B shows the top 10 development organizations by gross disbursements. We observe some similarities but also clear discrepancies, as well-known aid donors like Japan and France are missing from our analysis. The challenge with the CRS data is that the project descriptions are generally uninformative, with description lengths ranging from 47 to 144 characters in total. We ultimately use IATI for analysis due to the richness of the descriptions of its development projects.

3. Building the activity panel

We start by defining the sample of organizations we want to study. We are interested in knowing how the top 10 development organizations by disbursement spend on evidence generation between 2010 and 2025. According to IATI, disbursements are “outgoing funds that are placed at the disposal of a recipient government or organisation, or funds transferred between two separately reported activities.” This means we are looking at the actual money spent rather than money committed. In Table 1, we show the ranking of reporting organizations by disbursement size and the final inclusion decision for each of them.

Table 1. Top aid organizations by disbursements, 2010 onwards

Rank	Publisher	Disbursements , USD bn	Activities with disbursements	Share of activities with a disbursement	Decision
1	World Bank	518.6	4,051	96%	Selected
2	USAID	316.8	39,116	76%	Selected
3	US Department of State	167.7	21,979	not computed	Set aside: non-development US agency
4	Asian Development Bank	165.3	2,705	96%	Selected
5	Inter-American Development Bank	135.6	1,390	90%	Selected
6	FCDO (UK)	120.5	3,173	99%	Selected
7	US Department of Defense	82.7	11,411	not computed	Set aside: non-development US agency
8	Sweden (Sida)	74.2	17,664	90%	Selected
9	Islamic Development Bank	64.7	877	48%	Selected
10	African Development Bank	61.6	2,183	78%	Selected

¹ OECD Creditor Reporting System, full bulk parquet file, accessed 5 March 2026 via the oda-reader Python package (version 1.4.2), covering activities reported through 2024

11	Germany (BMZ)	55.9	3,343	12%	Selected, then dropped (reporting convention)
12	Global Fund to Fight AIDS, Tuberculosis and Malaria	52.0	1,940	not computed	Set aside: vertical fund
13	Netherlands (Ministry of Foreign Affairs)	51.6	8,924	98%	Selected

First, we removed the US State Department and the US Department of Defense because they are not development agencies (in the case of the State Department, this was true at the time of this analysis). Similarly, we removed the Global Fund for being a vertical fund. We end up with 10 development organizations: the World Bank, the United States Agency for International Development (USAID), the Asian Development Bank (ADB), the Inter-American Development Bank (IDB), the Foreign, Commonwealth, and Development Office (FCDO), the Swedish International Development Cooperation Agency (SIDA), the Islamic Development Bank (IsDB), the African Development Bank (AfDB), the Federal Ministry for Economic Cooperation and Development (BMZ), and the Netherlands Ministry of Foreign Affairs.

Ultimately, we decided to also remove BMZ, given that they only report disbursement data on 12% (N=3,343) of their project activities. Comparatively, when looking at the size of commitments, BMZ reports closer to \$168.6 billion across almost 33,000 activities. This means that even though BMZ reports all of its commitments, its disbursements do not follow the same reporting practice. Notably, the Agence Française de Développement (AFD) is also missing from our analysis, due to data limitations. Despite being a large donor, AFD only started reporting their activities on IATI from 2018, midway through our period of analysis. Additionally, we exclude the World Bank Trust Fund, a reporting organization that appears separately in the IATI dataset. This is to avoid double counting, as 92% of these funds are funded by the same governments and development agencies we are including in the analysis. Table 2 shows the final organizations used in our analysis.

Table 2. Top 9 donors included in the analysis

Donor	Activities	Disbursements, USD billion	Share of disbursements
World Bank	4,051	518.6	34.3%
USAID	39,098	317.6	21.0%
Asian Development Bank	2,704	165.6	11.0%
Inter-American Development Bank	1,390	135.6	9.0%
FCDO (UK)	3,173	120.5	8.0%
Sweden (Sida)	17,338	74.3	4.9%

Islamic Development Bank	877	64.7	4.3%
African Development Bank	2,183	61.6	4.1%
Netherlands MoFA	8,922	51.6	3.4%
Total	79,736	1,510.2	100%

To define the activities that belong to the period of analysis we decided to filter activities whose first commitment was done in 2010 or onwards. According to IATI, a commitment is defined as “A firm, written obligation from a donor or provider to provide a specific amount of funds, under particular terms and conditions, for specific purposes, for the benefit of the recipient”. In this way, we are able to identify projects that we know started at least in 2010. Transactions from these activities without a date or monetary value are excluded from the analysis. By the time we accessed the data we know that the last full year of reporting available is 2024, so values from 2025 should be considered as incomplete.

On the other hand, activities in IATI can have sub-activities, and each of these levels of reporting can have its own transactions. According to IATI's documentation, an activity is called a “parent” and a sub-activity is called a “child.” We match a child's transactions to its parent only when the child has exactly one parent and the parent is published by the same reporting organisation. Child activities with no resolvable parent, or with more than one, stay as their own activity. The activity metadata (identifier, donor, title, description) are those of the rolled-up activity, i.e., the parent where a rollup happened. In the case of our 9 donors, out of 105,656 child activities, only 63 have more than one parent. However, none of these 63 activities carries a transaction date or a monetary value, so they do not enter the analysis.

Transactions are then summed to a monetary value at the activity-year level; hence, an activity's total is the sum over the years. Activities whose net total disbursement is zero or negative are kept in the counts but excluded from every money figure. We end up with 79,736 activities, of which around 3% have a net zero or negative disbursement. We include a clean description of every activity, which is what the LLM will read to classify whether the project generates evidence or not. We also include one primary DAC 5-digit purpose code per activity to identify the sector of the activity.

Finally, to get the final panel for analysis, we made two adjustments. First, Sweden reports one project once per recipient country as separate activities that share a base identifier with a numeric suffix, each repeating the full project amount. We identified 500 duplicated activities, verified by matching the money value to the cent, the description, and the shared base identifier. After deduplicating, we end up with 199 unique activities. Second, 46 activities could not be classified because the API returned an empty or unusable response, even when the request was sent again. These represent 0.006% of the total disbursement money and were distributed as follows: USAID with 18, Sweden with 25, the Netherlands with 2, and the Asian Development Bank with 1. After removing Sweden's duplicate projects and the descriptions the LLM could not

use, we end up with a frame of 79,736 activities, one row each, and 225,036 activity-year rows, totalling \$1.51 trillion in disbursements.

4. LLM classification method

We define evidence generation as an activity that is explicitly aimed at establishing causal impact. This includes randomized controlled trials (RCTs), quasi-experimental designs (difference-in-differences, regression discontinuity, instrumental variables, etc.). We exclude monitoring and evaluation (M&E) systems and M&E capacity building, feasibility and preparation studies, non-causal research (such as literature reviews, policy analysis, and market studies). We used an LLM model (Claude Sonnet 4.6) through the API, for text analysis on the descriptions of the activities². We used a prompt that included the definition, the rules, the output schema, and nine examples. The prompt asked the LLM model to make a decision on whether the activity generates evidence (true or false). If the activity was classified as evidence, we asked whether it is primarily an evidence activity or a larger project with an evidence component. Additionally, we asked the model to include its level of confidence in the evidence classification (high, medium, or low). Finally, we asked the model to include the reasoning behind its decisions for inspection by a reviewer. The prompt used to instruct the LLM model is included in Appendix A.

5. Correction methodology

We are interested in correcting the raw LLM classification along two dimensions. First, we want to include uncertainty in the money and count of activities classified as evidence, given that the LLM classifies activities based on their descriptions, and the descriptions are themselves a function of each aid organization's reporting practices. To address this, we estimate error rates based on a sample of manually reviewed projects. Second, we want to correct for projects where only a component of the full project generates evidence. This gives a more realistic figure for the spending that goes into evidence when a project is large and only a portion of its money goes into a rigorous evaluation. To address this, we look for a benchmark of the cost of impact evaluations and apply the median value across all activities classified as evidence-component projects.

5.1 Estimating the error rates and confidence intervals

We defined six strata in the population, given by the joint categories of evidence and confidence level. The six strata are: (1) evidence, high confidence; (2) evidence, medium confidence; (3)

² The API calls to classify the activities were run in May of 2026. We used the default temperature of 1 to make the calls, so the raw classification itself might not be reproducible. Currently, Claude AI models don't offer changing the temperature parameter so any similar work can face the same issue.

evidence, low confidence; (4) not evidence, high confidence; (5) not evidence, medium confidence; (6) not evidence, low confidence.

The population sizes of each stratum are shown in Table 3. From each stratum we draw a random sample of 27 projects, except for the “low-confidence, evidence” stratum, where only 17 projects reside. This stratum is checked in full given its size. With this sample of 152, we manually review each project by looking at the documentation and information available. This means that the correction is done with information beyond what the LLM had access to in the description.

Table 3. LLM raw classifications by evidence and confidence level

LLM label	Confidence	Activities
Evidence	High	256
Evidence	Medium	358
Evidence	Low	17
Not evidence	High	75,953
Not evidence	Medium	882
Not evidence	Low	2,270

We estimate average error rates per stratum from the manual review of projects. This is shown in Table 4. Illustratively, in the “high-confidence, evidence” stratum, 2 of the 27 projects reviewed should not have been classified as evidence-generating. Thus, the average error rate in this stratum is estimated at 7.4%.

As we can see from Table 4, high-confidence classifications are reliable in both directions. However, medium-confidence classifications err both ways, and low-confidence, evidence classifications are mostly wrong. One reassuring fact is that classification accuracy seems to move in sync with the level of confidence, although there are errors in almost all strata, as expected.

Following this exercise, we use the Clopper-Pearson exact method to estimate a confidence interval around the error rates. The logic behind this method, applied to this exercise, is to ask: if we change the true error rate in the population, what is the probability of seeing the error rate we got in this particular sample? The upper and lower bounds of the interval correspond to the error rates for which there is still at least a 2.5% probability of seeing an error rate like the one we estimate from the sample.³

³ Illustratively, consider the “high-confidence, no evidence” stratum. There were zero errors found within the manually verified sample of 27 projects in this stratum. However, a true error rate of 12.8% would still allow for greater than 2.5% probability of finding 0 errors within a sample of 27. As we see in the final results, this stratum drives the high level of uncertainty in the upper limits of our confidence intervals – as 12.8% of 75,953 activities is large.

Let s be a stratum, N_s be the number of activities in stratum s , n_s the number of activities sampled in stratum s , y_s the number of classification errors found from the sample. Our estimate of e_s , the error rate in stratum s , is given by:

$$\hat{e}_s = \frac{y_s}{n_s}$$

We then derive the Clopper-Pearson interval from the Binomial distribution:

$$P(Y_s = y_s) = \sum_{y_s=0}^{n_s} \binom{n_s}{y_s} e_s^{y_s} (1 - e_s)^{n_s - y_s}$$

To estimate an interval we fix the sample size n_s and number of errors y_s , changing only the true error rate e_s . The lower bound estimate of the classification error is the error rate where the probability $P(Y_s \geq 2.5\%)$ and the upper bound is the error rate where the probability $P(Y_s \leq 2.5\%)$.

Table 4. Error rates and intervals

Stratum s	N_s	n_s	y_s	$\hat{e}_s$	95% confidence interval
Evidence, high	256	27	2	7.4%	0.9% to 24.3%
Evidence, medium	358	27	15	55.6%	35.3% to 74.5%
Evidence, low (census)	17	17	14	82.4%	exact
Not evidence, high	75,953	27	0	0.0%	0% to 12.8%
Not evidence, medium	882	27	8	29.6%	13.8% to 50.2%
Not evidence, low	2,270	27	1	3.7%	0.1% to 19.0%

5.2 Correcting activities with evidence as a component

Of the raw evidence classification, 87% of the money sits in large projects (infrastructure, health, and social protection) that contain only one evaluation. For example, for donors like the World Bank, all evidence-generating activities are classified as evidence-as-component projects, totalling a raw figure of \$9.5 billion. Counting the full programme treats all of that money as evidence spending, which would vastly overestimate the value of evidence spending. We therefore correct the disbursement value for evidence-as-component projects.

We used the Millennium Challenge Corporation (MCC) as a benchmark to estimate an average cost of conducting an impact evaluation. The MCC publishes a list of all its independent evaluations, tagged by type (impact, performance, or multiple), and its evaluator contracts are federal contracts visible in [USAspending.gov](https://www.usaspending.gov), the open data portal for US spending, so each evaluation can be linked to what the evaluator was paid.

The process to price an impact evaluation at MCC is as follows:

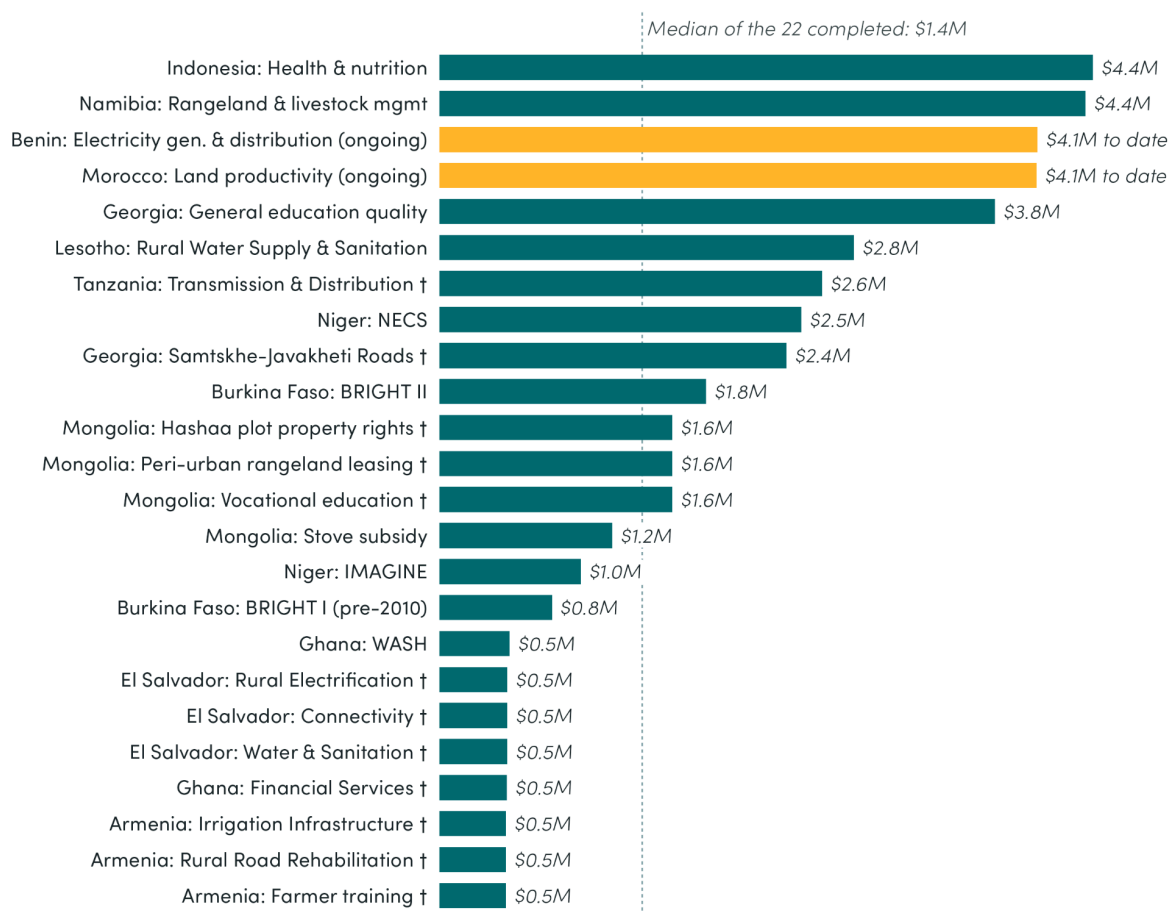
- We begin by scraping the mcc.gov list of independent evaluations. Of 176 entries, we found 32 projects with the tag “impact evaluation,” across 20 programs.
- We then pull every MCC contract award with activities between 2010 and 2024 from the USAspending platform. This totals 2,485 awards. We filter these awards by their descriptions using keywords like “evaluation,” “survey,” “baseline,” and “impact.” We end up with 891 candidate awards.
- Then we look up each of the independent evaluation entries in MCC’s Evidence Platform Catalog, which has detailed information about the evaluator, country of performance, and dates of implementation.
- Finally, we match each of the awards from the USAspending platform with the 32 independent evaluations based on country of performance, contract recipient (evaluator), and dates of implementation.

Table 5. Data sources used to price an impact evaluation by MCC

Source	Where	Taken	What it gave
MCC independent-evaluations list	mcc.gov, inline 'var rows' array with type tag and catalog link	2026-10-09	176 evaluations; 32 tagged Impact in 20 programmes
MCC contract awards	USAspending API, spending_by_award, awarding agency MCC, award types A to D, activity 2010 to 2024	2026-10-09	2,485 awards
Award details	USAspending API, awards/<id>	2026-10-09	891 candidate evaluator orders
Evidence Platform catalog pages	mcc.icpsr.umich.edu (saved by hand from a browser)	2026-10-12	56 pages: evaluator, type, method, contract start

For the 32 evaluations, we were able to find awards for 24 of them. As 2 of those 24 are still ongoing evaluations, we focus on the remaining 22. Figure 1 shows the results of matching evaluations to spending. The median spending on the 22 completed evaluations is around \$1.4 million. We therefore cap all activities classified as evidence-as-component projects at \$1.4 million to get a more accurate sense of the true value of evidence spending.

Figure 1. What MCC paid its independent evaluators per impact evaluation



What MCC paid its independent evaluators per impact evaluation: USAspending obligations to the evaluator. 24 of 32 impact evaluations have a linked contract. Gold = ongoing evaluations, shown at cost to date. † Contract also paid for other evaluations; its value is split equally among them. Source: USAspending contract awards (pulled 9 Oct 2026), linked to the mcc.gov list of independent evaluations and the MCC Evidence Platform.

5.3 Applying the corrections

We use the estimated error rates and the estimated disbursement value of evidence-as-component projects to reach a corrected estimate of total disbursements on evidence generation. Concretely, we make two key corrections:

1. We multiply the raw evidence classification by a probability that the activity comprises evidence generation. The probability comes from the error rate estimated from the manual review of the activity's stratum (as outlined in section 5.1);
2. We replace the activity's full disbursement with the value its evidence would be worth if it is evidence generation. This is equal to the full disbursement value for primary evidence projects, and at most \$1.4million for evidence-as-component projects (as outlined in section 5.2).

Suppose that activities are indexed by i . Each activity belongs to one of six strata s , defined by its evidence label from the LLM classification (evidence/ not evidence), and a confidence on that label (High/ Medium/Low). Namely, the six strata are: (1) evidence, high confidence; (2) evidence, medium confidence; (3) evidence, low confidence; (4) not evidence, high confidence; (5) not evidence, medium confidence; (6) not evidence, low confidence.

For each activity, we define:

- L_i , an indicator variable equal to 1 if the activity was labeled as evidence-generating; and 0 otherwise
- C_i , an indicator variable equal to 1 if the evidence-generating activity is a component of a larger project (evidence-as-component project); and 0 otherwise, i.e. evidence-generation is the primary activity
- T_i , the total net disbursement (USD) for the activity from 2010 to 2025

Using the results from Table 4, we estimate the error rate of each stratum, e_s , as the share of reviewed activities whose label was wrong:

$$\hat{e}_s = \frac{y_s}{n_s}$$

where n_s is the number of projects reviewed in the stratum, and y_s is the number of incorrect classifications. The estimated error rate for each stratum, $\hat{e}_s$, is applied uniformly to every activity in stratum s .

We use the estimated error rate to estimate the probability a project is truly evidence-generating, p_i . The estimate of p_i is given by:

$$\hat{p}_i = \begin{cases} 1 - \hat{e}_{s(i)} & \text{if } s = \text{evidence with High, Medium, or Low confidence} \\ \hat{e}_{s(i)} & \text{if } s = \text{not evidence with High, Medium, or Low confidence} \end{cases}$$

We therefore assume that the probability that a raw classification is truly evidence-generating is constant for each activity within the same stratum s .

We then use the estimated cost of an impact-evaluation project to adjust for the disbursement value of each activity as follows:

$$v_i = \begin{cases} T_i & \text{if labelled evidence and primary } (L_i = 1, C_i = 0) \\ \min(T_i, K) & \text{if labelled evidence as a component } (L_i = 1, C_i = 1) \\ \min(T_i, K) & \text{if labelled not evidence } (L_i = 0) \end{cases}$$

where v_i is the disbursement value of activity i , and K is the cost of an impact-evaluation project. As described in section 5.2, we assume $K = 1,400,000$. Note also that if a project is labeled as not evidence-generating ($L_i = 0$), we also cap its disbursement value at \$1.4million. This is equivalent to assuming that all evidence generating activities that are incorrectly classified as not evidence are evidence-as-component activities (rather than primary ones).⁴

Given the above, the corrected number of evidence activities, N^{corr} , and the corrected disbursement value, M^{corr} are given by:

$$N^{corr} = \sum_i \hat{p}_i, \quad M^{corr} = \sum_i \hat{p}_i v_i$$

Worked example

In Table 6 we present a basic exercise with hypothetical single activities, and illustrative corrections applied to the evidence count and disbursement values.

Table 6. Hypothetical activities and corrections

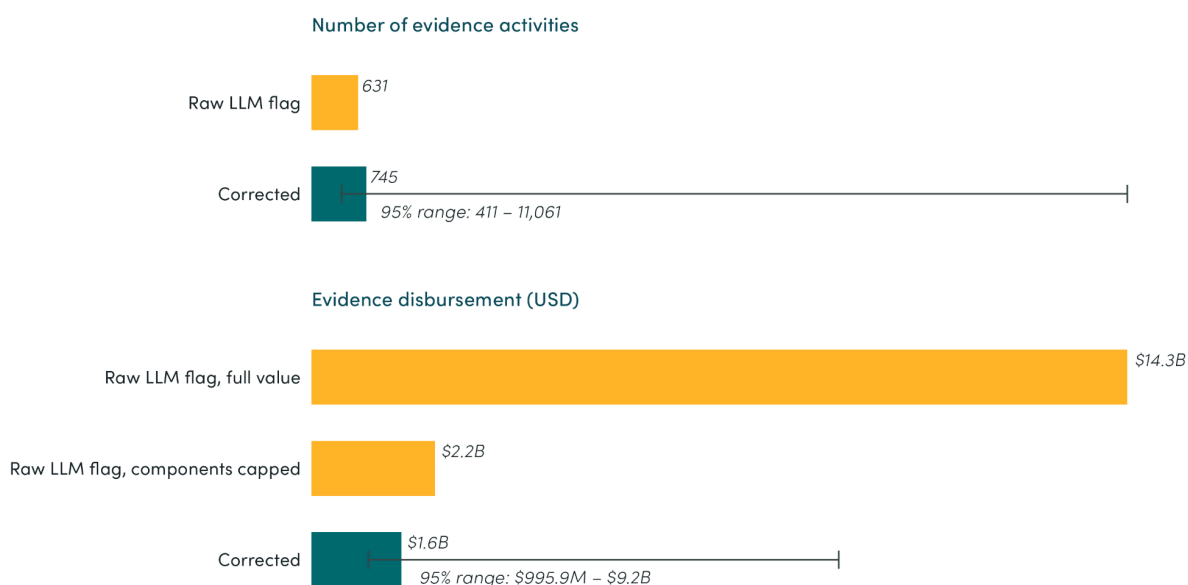
Strata, i.e. model label and confidence	Total T_i	Probability $\hat{p}_i$	Value v_i	Added value to the aggregate evidence-count	Added value to total disbursement
Evidence, high, primary	5.0M	$1 - 0.074 = 0.926$	5.0M	0.926	4.63M
Evidence, high, component	5.0M	0.926	$\min(5.0M, 1.4M) = 1.4M$	0.926	1.30M
Evidence, medium, component	0.8M	$1 - 0.556 = 0.444$	$\min(0.8M, 1.4M) = 0.8M$	0.444	0.36M
Not evidence, medium	5.0M	0.296	1.4M	0.296	0.41M
Not evidence, low	5.0M	0.037	1.4M	0.037	0.05M

Estimating the confidence intervals on disbursement

Finally, the 95% confidence intervals on disbursements are applied in the same way. Instead of the average error rate, we apply the estimated lower and upper limits of the 95% range, as indicated by table 4. The aggregate results of each of the corrections, with confidence intervals, can be seen in Figure 2.

Figure 2. Aggregate raw and corrected counts and spending on evidence

⁴ As primary evidence projects tend to cost more than \$1.4 million, this means that our estimates of total disbursement are likely an underestimate of total spending on evidence generation.



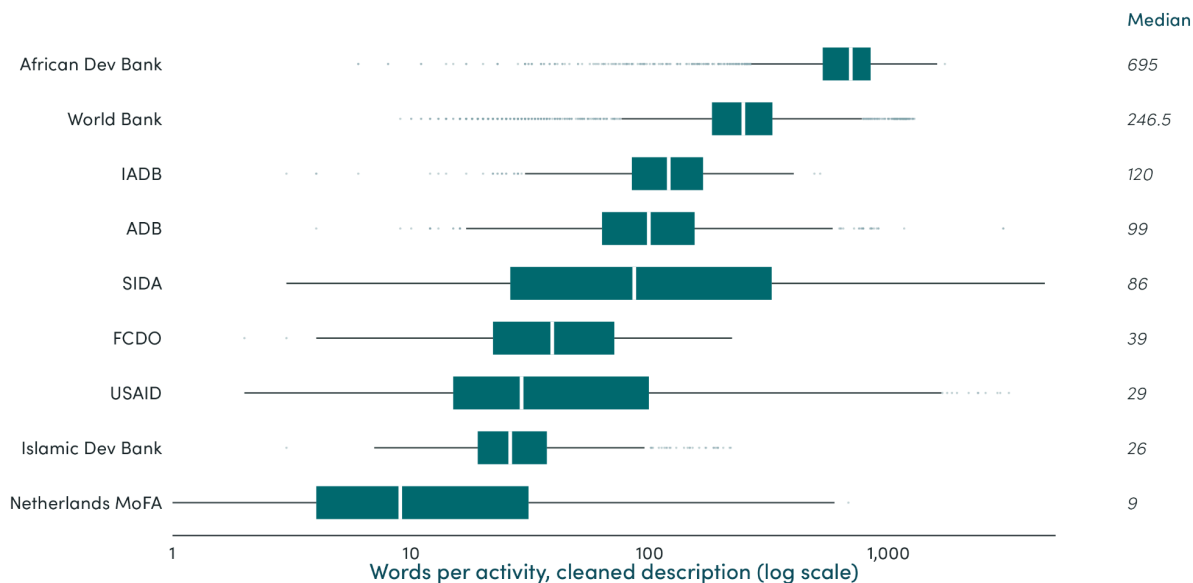
Bars = point estimate; whiskers = 95% range from the stratum error rates.
 Corrected with stratum error rates from a manual review of 152 activities (Clopper-Pearson 95% ranges); rates applied uniformly within each stratum.
 Evidence components valued at min(project total, USD 1.4M); projects the LLM missed are all valued as components.
 Activities with net total disbursement <= 0 excluded from money.

6. Limitations

This exercise is our best approximation given the constraints that exist. Some of the most significant limitations of our analysis starts with the data source itself. We are using IATI, which is a self-reporting open data repository of development projects. This means that the analysis – and identification of evidence generating activities – is limited by the reporting practices of the individual organisations.

For example, we show in Figure 3 how the length of project descriptions vary by donor. Whereas the African Development Bank uses an average of 672 words, donors like the Netherlands Ministry of Foreign Affairs and the Islamic Development Bank use 34 and 33 respectively. This directly affects the likely accuracy of project classifications, as the large language model conducts the classifications solely based on project descriptions.

Figure 3. Description length by donor



Boxes = 25th to 75th percentile, white line = median, whiskers = 1.5 x interquartile range, dots = activities beyond it. Empty descriptions excluded. Deduplicated to activity level. Source: IATI.

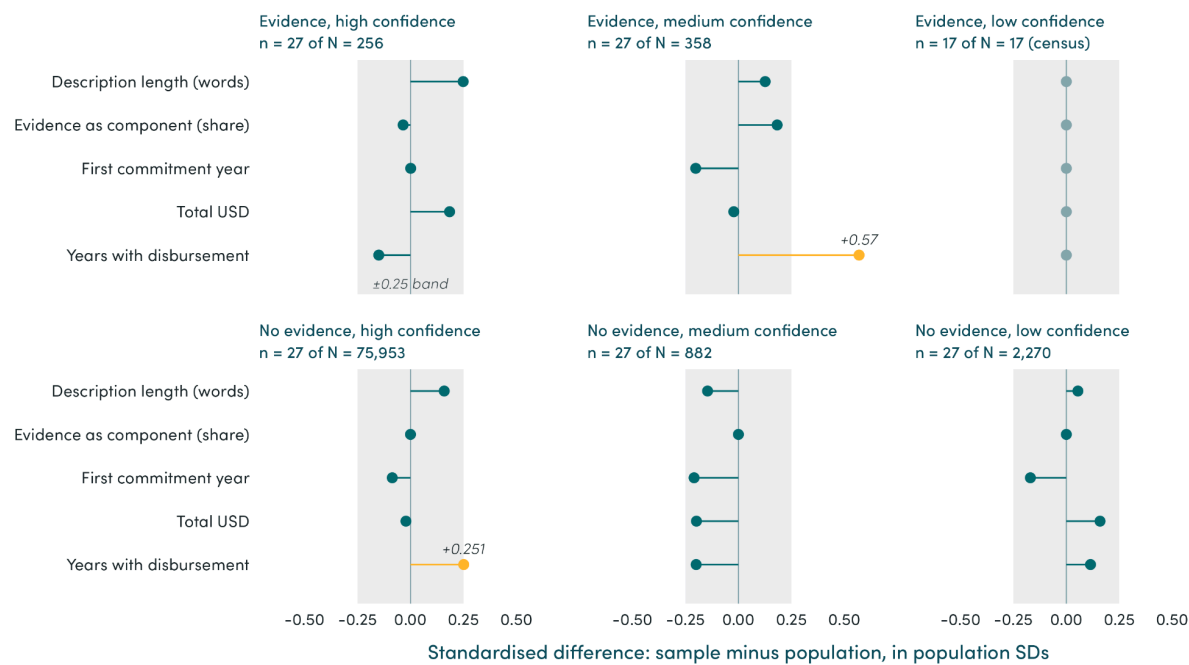
Furthermore, IATI is reported at the level of the organisation and development project. This means we are likely double counting aggregate spending on evidence generation where there is substantial cross-organisational funding of evidence activities. For instance, organisations like USAID and FCDO fund evidence generation activities that are then conducted by the World Bank. This would likely show up as two separate activities in our data (e.g. an FCDO evidence generation project and then components in World Bank projects). At the same time, the restriction to development projects means we could also be undercounting evidence spending activities, as not all evidence related expenses are registered as development activities. This includes, for instance, personnel or operational costs of research staff that sit within organisations whose salaries are not paid by projects reported in IATI.

There are also several limitations in the results driven by the assumptions and estimation procedure. We have wide confidence intervals on our estimates. This is driven by the number of projects we were able to manually review per stratum (N=27), and importantly, the large number of projects within the “no evidence, high confidence” stratum (N=75,953).⁵ In Figure 4, we illustrate how the manually verified samples differ when compared with the population across across five dimensions of interest: description length in words, the share of evidence as a component, first commitment year of the project, the total monetary value of the projects and the number of years with a disbursement. Overall, we find no meaningful differences except for project length (i.e. number of years with non-zero disbursements) in medium-confidence evidence and high confidence not evidence groups.

⁵ We found zero errors in our manual checks of the “no evidence, high confidence” stratum, but estimate 95% confidence intervals on the error rates of 0-12.8%. The upper bound gives us high uncertainty on evidence classifications in this stratum.

Lastly, we make several broad assumptions when applying the corrections. First, we assume that error rates are uniform within each stratum, whereas it is possible that errors differ along other dimensions. Second, we also assume a constant valuation of “evidence-as-a-component” projects. In practice, the cost of an impact evaluation in component projects is likely to differ by the donor, sector, or size of the project.

Figure 4. Population and sample comparisons



23 of 25 comparisons are within ± 0.25 SD, the usual balance threshold (shaded); gold points fall outside it.
Simple random sample of 27 activities per stratum; the 17 low-confidence evidence activities were all reviewed (census, grey points).

Appendix A: LLM prompt for classifying development activities

You are a research assistant helping classify development projects based on whether they include an evidence generation component.

DEFINITION:

Evidence generation means any activity explicitly aimed at establishing causal impact — i.e., determining whether a project or intervention caused specific changes in outcomes. We focus on evidence generation with counterfactual analysis. This includes:

- Randomized controlled trials (RCTs)
- Quasi-experimental studies (difference-in-differences, regression discontinuity, instrumental variables, propensity score matching, event studies, etc.)
- Observational causal inference studies
- Impact evaluations explicitly aimed at attributing changes in outcomes to an intervention
- Baseline surveys conducted with the explicit intention of a future causal impact evaluation
- Funding or commissioning others to conduct any of the above

This does NOT count as evidence generation:

- Monitoring and evaluation (M&E) systems or capacity building, unless they explicitly include a causal impact evaluation component
- Feasibility studies
- Preparation studies
- Non-causal research (e.g. literature reviews, policy analysis, market studies)
- Mentions of "impact evaluation" that refer to M&E system strengthening rather than actual causal studies
- Projects that mention impact evaluation only as a vague future goal without concrete plans

IMPORTANT RULES:

- A large project should be classified as YES if AT LEAST ONE component explicitly involves causal evidence generation, even if the rest of the project does not
- The word "impact evaluation" alone is not sufficient — look for whether the project is actually conducting or funding causal research or just strengthening M&E systems
- Mentions of future impact evaluation without concrete plans in the current project do NOT count

OUTPUT:

Respond only in JSON format with exactly these fields:

```
{  
  "evidence_generation": true or false,
```

```

"evidence_as_component": true or false,
"confidence": "high", "medium", or "low",
"reasoning": "one or two sentences explaining your classification"
}

```

DEFINITIONS FOR OUTPUT FIELDS:

- "evidence_generation": true if the project includes any evidence generation activity as defined above
- "evidence_as_component": true ONLY if the project is primarily about something else (e.g. infrastructure, service delivery, capacity building) but contains at least one component that involves causal evidence generation. Set to false if the entire project is about evidence generation, or if evidence_generation is false.

EXAMPLES:

Description: "Project Development Objective (from Project Appraisal Document) The Project Development Objective is to improve forest management; and increase access to markets and basic services by small forest producers including indigenous and campesinos in selected Northern Provinces of Argentina The project development objective of the Forests and Community Project for Argentina is to improve forest management; and increase access to markets and basic services by small forest producers (including indigenous people and campesinos) in selected Northern Provinces. This project is comprised of five components. (1) The sustainable livelihoods component will improve the livelihood of small forest producers and promote sustainable forest landscape management, with special emphasis on the Chaco Ecoregion, through: (a) the preparation of Integrated Community Plans (PICs); (b) the implementation of PICs; and (c) the design and implementation of subprojects. (2) The natural forest management and civil society component will support carrying out of a national forest and climate awareness campaign to expand and maintain support for the implementation of the Forest Law through the establishment and/or expansion of existing community radio to distribute information on sustainable natural resources management, climate resilience and indigenous rights. (3) The skills development and technology transfer component will finance strengthening the technical, managerial skills of: (i) indigenous and campesino community members; (ii) small forest producers; and (iii) technical service providers and extension agents working in the Selected Northern Provinces, through the provision of short training courses including both classroom and community level instruction and financing of scholarship grants. (4) The forest monitoring component will support the expanding Secretariat of Environment and Sustainable Development (SAyDS) technical and institutional capacity to manage forest information and monitor changes in forest cover. (5) The project management and impact monitoring component will finance: (i) provision of support for the technical and administrative coordination of the project, (ii) development of an impact evaluation system to identify changes in outcomes directly attributable to Project activities and (iii) development and implementation of a grievance redress mechanism to identify and resolve Project related grievances."

Output: {"evidence_generation": true, "evidence_as_component": true, "confidence": "high", "reasoning": "Component 5 explicitly finances an impact evaluation system to identify changes in outcomes directly

attributable to project activities. The project is primarily a forest management project with evidence generation as one component."}

Description: "EN: (Nigeria): With FY 2022 funds this activity will:..Provide cross-sectoral, Mission-wide support on strategy, project, and activity-level monitoring, evaluation, and learning; and..Provide third-party, field-based monitoring, and data gathering to augment the Mission's evidence base for accountability and decision making..Due to increasing insecurity across the country, strict Embassy security restrictions for chief of mission staff, and a generalised non-permissive implementation environment traditionally monitoring, evaluation, and learning has become difficult. The goal of the Monitoring, Evaluation and Learning Support Activity (MELSA) activity is to provide third party MEL tools to staff and partners and help transform USAID/Nigeria into a premier practitioner of program monitoring and evaluation rooted in organizational learning and adaptation. MELSA will provide continuous, on-the-ground, on-demand and systematic support to all USAID/Nigeria technical teams and implementing partners on performance monitoring, data verification, performance and impact evaluations, project- and activity-level monitoring, evaluation and learning plans, organizational learning, capacity building, short term technical assistance, and knowledge management. This activity cross-cuts the USAID Nigeria CDCS supporting all objectives as well as being closely aligned with Goals 1- 4 of the 2022-2026 Joint Strategic Plan..."

Output: {"evidence_generation": false, "evidence_as_component": false, "confidence": "high", "reasoning": "Although the project mentions impact evaluations, it is an M&E system strengthening activity focused on monitoring tools and capacity building, not on generating causal evidence."}

Description: "To strengthen the impact of global aid by learning lessons and generating evidence for design of new programmes. The Strategic Impact Evaluation Fund (SIEF), over 2011/12-2021/22, will be channelled through a World Bank executed Trust Fund. It is expected that SIEF will deliver a minimum of 54 high quality impact evaluations of key development interventions focusing on human development outcomes. In addition, SIEF will deliver 20 capacity building workshops and events to improve understanding of impact evaluation techniques."

Output: {"evidence_generation": true, "evidence_as_component": false, "confidence": "high", "reasoning": "The entire project is dedicated to funding high quality impact evaluations of development interventions. Evidence generation is the primary purpose, not a component."}

Description: "This feasibility study for the development of agro-industrial parks falls within the framework of implementation of the recommendations of the high-level conference organized by the African Development Bank in Dakar on agricultural transformation in Africa. It also aligns with Democratic Republic of Congo (DRC)'s priorities of transforming agriculture from a subsistence sector into a real driving force for the development of the entire Congolese economy. This study will conduct technical, financial, environmental and social feasibility studies, as well as prepare the business model for the establishment of three new agro-industrial parks in the country's centre region. Its total cost is estimated at UA 1 524 582 (USD 2 149 661), net of taxes and customs duties. The study will, for each of the parks, comprise two components, especially: (i) diagnosis phase and pre-feasibility studies that will lead to the preparation of a development master plan, including a strategic environmental and social assessment (SESA) for each agro-industrial park; (ii) detailed

technical, economic and financial study phase, including environmental and social impact assessments of core infrastructure and the development of a research programme."

Output: {"evidence_generation": false, "evidence_as_component": false, "confidence": "high", "reasoning": "This is a feasibility study assessing technical, financial and environmental aspects of agro-industrial parks. No causal impact evaluation is involved."}

Description: "The proposed AGRIZIDERP preparation study will address various aspects of the value chain starting from production to consumption including technical, institutional, social, economic, and environmental aspects. It is expected that after the conclusion of the feasibility study, the Bank in coordination with Government of Mozambique (GoM), Development Partners and Private Sector will finance the program under Feed Africa Initiative. The PPF advance requested will be used to finance the feasibility study of a The Agro-processing Zone and Resilience Program of Limpopo Special Economic Zone (AGRIZIDERP), as described in this proposal The concerned program aims to catalyze a process that results in a market system that is competitive, inclusive, and resilient. To ensure good program quality at entry, implementation, monitoring of system changes and at impact evaluation and to also fast track the preparation of this important program, comprehensive feasibility studies including marketing analysis, detailed engineering designs with Bill of Quantities (BoQ) and detailed designs need to be completed before program preparation and appraisal stages."

Output: {"evidence_generation": false, "evidence_as_component": false, "confidence": "high", "reasoning": "Although impact evaluation is mentioned, it refers to a future goal after program preparation. The current project is a feasibility study with no causal evidence generation component."}

Description: "The intervention will support parts of CGD's ongoing research programmes with focus on three key research themes; 1. Aid effectiveness and aid channels and partners 2. Development finance 3. Technology and new paths to development"

Output: {"evidence_generation": false, "evidence_as_component": false, "confidence": "high", "reasoning": "The project supports research on aid effectiveness and development finance but does not involve causal impact evaluation methodology."}

Description: "To produce a body of research evidence and analysis indicating practical ways of changing institutions so as to increase economic growth in poor countries. This will enable a better understanding of how institutions improve and change economic performances, and by 2020 developing countries will be able to use this understanding to improve their economies through developing better institutions."

Output: {"evidence_generation": false, "evidence_as_component": false, "confidence": "medium", "reasoning": "The project aims to produce research evidence on institutions and economic growth but there is no explicit mention of causal research methods or impact evaluation."}

Description: "SV: This proposed project study the impact of agricultural extension services on farmers in Kenya. Using unique data we will compare a treatment group with a control group of farmers that have been exposed to a donor-financed agricultural sector program in Kenya. In addition to the micro-analysis the project will also explore the broader impact using an economy-wide multi-sectoral model. In both the micro and the sectoral model approach distributional, gender and poverty effects will be discussed., This proposed project

study the impact of agricultural extension services on farmers in Kenya. Using unique data we will compare a treatment group with a control group of farmers that have been exposed to a donor-financed agricultural sector program in Kenya. In addition to the micro-analysis the project will also explore the broader impact using an economy-wide multi-sectoral model. In both the micro and the sectoral model approach distributional, gender and poverty effects will be discussed."

Output: {"evidence_generation": true, "evidence_as_component": false, "confidence": "medium", "reasoning": "The project compares a treatment and control group to study the causal impact of agricultural extension services, indicating a causal study. Confidence is medium because it is unclear how treatment and control groups were assigned."}

Description: "EN: Mali Food Security Policy Research Program"

Output: {"evidence_generation": false, "evidence_as_component": false, "confidence": "high", "reasoning": "The description contains no mention of causal research methods, impact evaluation, or any evidence generation activity."}

Now classify the following project description:

Description: "{description}"

Appendix B: Top 10 donors by gross disbursement from OECD CRS

Table B1. Top 10 donors by gross disbursement (OECD CRS)

Rank	Donor	Gross disbursement (USD millions)
1	United States	560,217
2	EU Institutions	359,570
3	Germany	328,424
4	International Bank for Reconstruction and Development	313,263
5	International Development Association	262,611
6	Japan	239,314
7	Asian Development Bank	206,789
8	United Kingdom	182,018
9	France	169,492
10	Inter-American Development Bank	148,000

Source: OECD Creditor Reporting System, full bulk parquet file, accessed 5 March 2026 via the oda-reader Python package (version 1.4.2), covering activities reported through 2024