Introducing the OPERA database

SEI brief
September 2026

Kira Kappe
Adis Dzebo

1. Introduction

The emergence of artificial intelligence (AI) and large language models (LLMs), in combination with natural language processing (NLP), has created new opportunities for large-scale extraction and classification of evidence from unstructured policy documents. This can help researchers and policymakers identify patterns in how governance and political contexts shape policy success and explain why similar interventions succeed in some settings but not others.

This brief introduces the last stage in the process of using AI and LLMs for large-n information extraction, described in Kappe and Dzebo (2025), and using NLP methods combined with policy evaluation frameworks to classify extracted summaries and measure policy success (Dzebo and Kappe, 2026).
In this brief, we present the Outcomes and Process Evidence from Reviews and Audits (OPERA) database (Dzebo et al. 2026), which comprises 111 436 summaries, extracted from 2839 national policy evaluations and independent policy audit documents.

A core objective of the database is to find evidence for the relationship between process and outcome and use both to evaluate policy success. This follows from the theoretical premise that the quality of policy processes determines policy outcomes (Dzebo, Shawoo, and Browne 2025; Nilsson et al. 2024). As this brief demonstrates, however, the relationship between evaluation evidence and policy success is far more complex, and OPERA's most interesting findings are embedded within this complexity.

2. Evaluating policy success

The analytical framework underlying the data extraction is composed of:

- outcome variables – i.e. effectiveness, efficiency, outcomes and impact, and attribution – which provide an overview over what was achieved, and
- process variables – i.e. agenda-setting, policy formulation, content coherence, policy implementation, and engagement and inclusion – which explain how it was achieved.

Building on an established composite indicator methodology (Greco et al. 2019; OECD et al., 2008), the database aggregates summary classifications into per-variable performance scores, based on which an overall document success score (0–100) is assigned. The scoring method consists of several components that ensure (a) inclusion of all nine variables, if available; (b) award of quality bonuses where multiple top-tier classifications co-occur; and (c) use of confidence values and standardized effective sample size to avoid scoring inflation from large documents. Final scores are grouped into five success tiers, allowing for cross-document comparison and the identification of success drivers across variables (Dzebo and Kappe, 2026).

BRIEF SUMMARY OF THE PROCESS

The SEI AI Reader extracts 81 758 analytical summaries from 2164 documents.

- NLP classifiers assign each summary a categorical value by variable. Fine-tuning improved six of nine variables, achieving mean test accuracy of 78.6% to 90.1% across the nine variables.
  Five benchmark variables (i.e. policy objective, success criteria, theory of change, evaluation method and evaluation findings) are extracted and classified to anchor what each evaluation set out to assess and how rigorously it did so.
- A composite scoring methodology aggregates summary classifications into outcome (variables 1–4) and process (variables 6–10) basis scores, applying confidence weighting, quality bonuses and cross-variable consistency gates.
- Document scores are mapped to a 0–100 scale and assigned to one of five success tiers: elite, gold, good, moderate and poor.

3. What does the database cover?

The entire dataset includes evaluation documents from 21 policy sectors and 116 countries and the European Union, covering almost all relevant thematic policy areas per geographical region. Climate and energy policy evaluations dominate among the sectors, with 22.4%, and, together with finance, public sector and health, compose 50% of total evaluation document data (Figure 1).

Geographically, European countries dominate with 45%. Second is Asia with 21%, followed by North America, South America and Africa, each with 7% to 9%. By income classification, 55% of evaluation documents originate from high-income countries, while lower-middle and low-income countries together account for 15%.

Figure 1: Evaluation document coverage across regions and sectors

Documents by sector and region

All 2,164 documents in the OPERA database, 1,533 audits and 631 evaluations. Climate and energy, the sector of this paper, are the top two rows.

Documents: 1 to 4 | 5 to 14 | 15 to 34 | 35 to 74 | 75 and above | — none

	Europe 975	Asia 463	North America 204	South America 187	Africa 155	Oceania 105	Central America 62	Regional 13	Total
Climate	113	57	54	16	7	15	4	3	269
Energy	119	37	23	16	4	10	5	2	216
Finance	146	32	9	7	9	8	4	1	216
Public Sector	99	43	9	24	13	6	5	—	199
Health	68	44	20	17	26	8	5	2	190
Biodiversity	45	21	23	17	9	14	12	—	141
Water	42	41	28	8	4	12	5	1	141
Agriculture	32	32	6	10	20	10	2	—	112
Education	56	14	3	8	19	2	1	—	103
Employment	66	7	2	5	5	4	3	1	93
Urban	22	30	1	7	5	1	1	—	67
Forestry	10	14	3	19	8	1	4	—	59
Industry	17	23	1	5	2	1	2	1	52
Infrastructure	27	13	3	6	1	1	1	—	52
Mobility	38	5	4	4	1	—	—	—	52
Inequality	19	13	3	5	5	—	3	—	48
Oceans	15	4	8	—	3	8	2	1	41
Rural	13	16	1	5	2	1	—	—	38
Land Use	11	7	2	3	4	3	—	—	30
Security	13	5	—	4	4	—	1	—	27
Gender	4	5	1	1	4	—	2	1	18
Total	975	463	204	187	155	105	62	13	2,164

Source: OPERA database

4. What are the main findings?

The database includes 2164 evaluation documents analysed in terms of geography, sector, variable, sub-variable and many other factors that can't be covered in the scope of this brief. Therefore, we present three overarching findings and their patterns: the distribution of success scores, the relationship between policy process and outcome, and the contextual differences in success across sectors and geographical regions. Each is discussed in turn below.

4.1 Aggregated success scores and tier distribution

The success score distribution is strongly skewed toward failure. Almost 76% of evaluation documents fall within the lower two tiers (moderate and poor), and only 8.4% reach gold or elite tiers, indicating that the evaluation document found proof of success across both process and outcome and multiple variables (Table 1). Reaching the upper tiers requires that the evaluation document reports evidence of success across both outcome and process variables, which few documents do comprehensively.

A defining feature of the database is the stark contrast in scoring between audits and evaluations. Audits – which compose 71% of the total evaluation documents, with 74% of the audits coming from high-income countries – achieve a mean score of 23.2. Policy

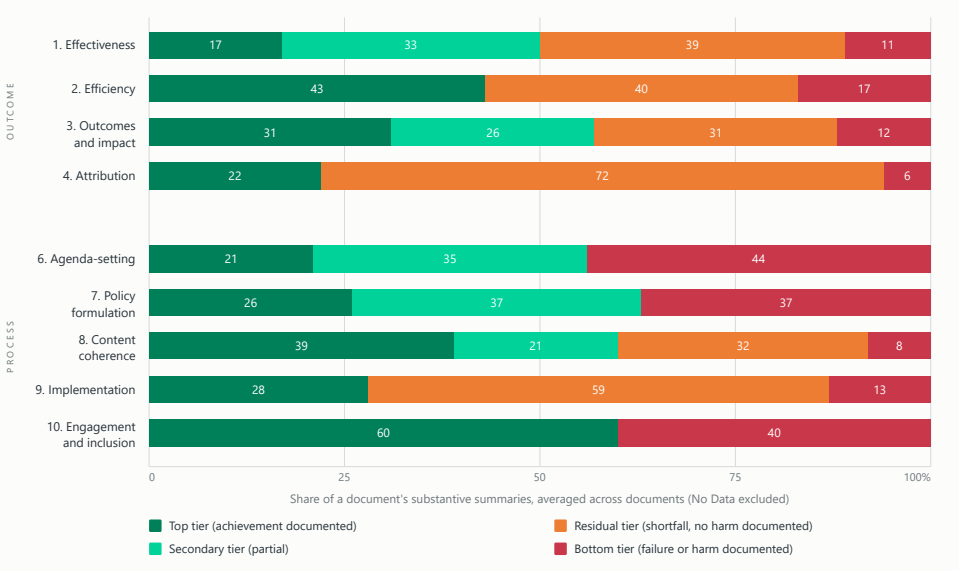
evaluations, in contrast, represent primarily middle-income countries (80%) and achieve a mean score of 56.3. Despite having a similar number of variables and being spread fairly evenly between process and outcome, audits almost exclusively apply less rigorous methodologies and report failure at two to five times the rate of evaluations across all nine variables. They are more prone to assessing compliance and adherence to regulatory standards. Evaluations, which apply more rigorous methods, assess impact against stated objectives to a much greater extent and are structured to identify and attribute achievements (Peters and Pierre 2023). This structural distinction between document types, rather than differences in policy performance itself, is the primary driver of the score distribution and must be accounted for when interpreting success patterns in the database.

Table 1. Success tier distribution (N = 2839 documents)

Tier	Range	Count	Percentage
Elite	≥ 90	67	3.1%
Gold	75–89	115	5.3%
Good	50–74	340	15.7%
Moderate	25–49	621	28.7%
Poor	< 25	1021	47.2%

Based on the tier distribution, it follows that most of the evaluation documents, particularly audits, failed to document policy success. Figure 2 shows the per-variable scoring distribution, where only 17.3% of the policies were reported to have been effective, while 49.3% were not effective. Similar distributions are documented in efficiency, outcomes and impact, agenda-setting and content coherence. Attribution, formulation and implementation record top-tier evidence in 21.8%, 25.6% and 27.6% of cases respectively. The only variable with a large positive achievement is meaningful engagement and inclusion (59.9%), albeit with a coverage of only 31% of the documents scored.

Figure 2: Classification per variable divided into positive, partial or no effect, and negative performance. (N = 2164).



Source: OPERA database

4.2 Process-outcome relationship: how strong is the connection?

A central question for policy evaluation is whether well-designed and well-implemented policies deliver better results. The answer is a qualified yes. Process quality and outcome achievement are moderately and uniformly coupled, with a correlation of about 0.44 in evaluations and 0.35 in audits at score level, and 0.55 in both on a length-free ordinal measure. Three checks suggest this link is real rather than a statistical artefact. It is not created by mixing the two document traditions, since the association inside evaluations alone is as strong as the association across the whole corpus. It does not fade under rigorous designs, holding between 0.37 and 0.46 from the weakest evaluation methods up to randomised trials, with no reliable difference between them. And it is moderate rather than decisive: documented process quality tracks roughly a fifth to a third of the variation in documented outcomes, and the rest moves independently of it.

Beyond the strength of the connection, our analysis of the relationship between process and outcome has another useful finding, concerning the direction. For policy evaluations, process evidence says much more about failure than about success. The correlation rises from 0.37 to 0.70 where failure is documented rather than success. In audits, however, the pattern reverses. Positive process categories are three to four times over-represented among effective outcomes. Good process appears necessary for success but not sufficient. A policy can do everything right and still fail because of external factors, political dynamics or system complexity. But when process breaks down, whether through weak coordination, poor implementation or exclusion of stakeholders, failure becomes significantly more likely.

This asymmetry has practical implications, and they invert between the two traditions. For auditors, the productive question is “which process features produce success?”

whilst evaluators should rather consider “which process failures most reliably prevent it?”. The OPERA database supports both lines of inquiry.

4.3 Geographical and sectoral differences in policy success

When scores are aggregated by region, a pronounced geographical difference emerges. African countries record the highest average score (47.9), followed by Asia (46.8) and South America (35.0), while North America (21.3) and Europe (26.6) score lowest. More broadly, low-income countries average 47.0 while high-income countries average 26.0.

An important explanatory factor lies in the above-mentioned audit-evaluation contrast. In high-income countries, 95% of documents are audits. In low-income countries, 79% are evaluations. Within evaluations the income gradient disappears altogether, at 54.5, 56.0, 57.6 and 52.8 across low, lower-middle, upper-middle and high income. The gap therefore reflects who gets audited and who gets evaluated, not who governs better. Europe is the clearest illustration of this effect. Its low score is connected to the data, where 94% are audits. Extrapolating Europe’s 62 evaluations gives an average score of 50.4.

Sectoral patterns tell a different story. Among sectors with at least 20 documents, rural development (43.0) and industry (40.3) score highest, while biodiversity (23.4) and oceans (26.0) score lowest, a spread of about 20 points. These results suggest that the conditions for demonstrating policy success vary systematically by policy domain. This points to another productive use of the database. Beyond drawing conclusions from aggregate scores, the OPERA database is designed to also support focused analysis within sectors and country groups, where structural comparisons are more meaningful.

5. Pattern analysis through qualitative case study analysis

By examining how gold and elite evaluations within a specific sector perform, or how cross-tier policies with common objectives or theories of change perform in their own terms, we can reveal which policy instruments and institutional arrangements have worked under comparable conditions. This was done for forestry policies. Dzebo and Kappe (2026b) demonstrate this approach for forestry, examining nine policies across six countries under varying institutional capacities and baseline conditions.

The forestry case illustrates why within-sector analysis surfaces patterns that aggregate correlation cannot. The process-outcome relationship described above takes on concrete meaning when examined in context. The forestry analysis identified three conditions whose joint presence distinguishes successful from unsuccessful implementation: clear authority, adequate resources, and functioning monitoring (Dzebo and Kappe, 2026b). Where all three align, interventions deliver. Where one lags, results weaken. Where several fail simultaneously, outcomes turn negative. Instrument design, the coherence between policy instrument and goal, emerges as the strongest predictor of success, but the same instrument logic produces divergent results across contexts, confirming that the instrument-context-fit matters as much as instrument choice. Moreover, where policies pursue dual objectives, such as forestry goals

compounded with socio-economic targets, environmental achievement typically outpaces socio-economic gains unless complementarity is structurally embedded in the design in the form of active synergies and coherence. Thus, a sectoral analysis enables patterns that aggregate correlation cannot, empirically showing that implementation breakdowns reliably predict failure, while successful implementation is necessary but not sufficient for positive outcomes.

These are exactly the kinds of actionable, context-specific patterns, visible in process documentation but invisible in aggregate statistics, that the OPERA database can help researchers to identify across all 21 sectors.

6. Conclusion: how can the OPERA database support research on policy success patterns?

The OPERA database shows that AI can enable systematic document analysis and evidence synthesis at scale. It supports research on policy success patterns in three concrete ways:

by enabling selective country profiles that identify which variables score lowest and which co-occur most consistently with successful outcomes
by allowing sector-specific analysis of high-scoring evaluations to surface which policy instruments and theories of change have worked under varying institutional and resource conditions, and
by exposing document type, methodological rigour and reporting completeness as filterable dimensions, which enables researchers to disentangle genuine policy performance from differences in evaluation practice.

Together, these capabilities create a working tool for replicating policy successes across scale, rather than a static reference.

The OPERA database in its current form is not intended to be a representative policy evaluation sample in the statistical sense. It comprises a subset of policy evaluation documents from a multitude of databases, including 3iE Development Evidence Portal, European Environment Agency (EEA) and the International Organization of Supreme Audit Institutions (INTOSAI) among others. It consists primarily of English-language documents (85%) and needs to be complemented in the future for better representativeness. Moreover, audits and policy evaluations are merely an intermediary variable for analysing policy success. They comprise the accumulated output of policy evaluation practice as produced by governments, multilateral organizations and research institutions, and therefore reflect the priorities and capacities of evaluating organizations.

Published by

Stockholm Environment Institute
Visiting address: Textilgatan 43
Post and deliveries: Virkesvägen 1A
120 30 Stockholm, Sweden
Tel: +46 8 30 80 44
www.sei.org

Copyright © September 2026 by Stockholm Environment Institute. This work is licensed under Creative Commons Attribution 4.0 International.

To view a copy of the licence, visit creativecommons.org/licenses/by/4.0

DOI:

https://doi.org/10.51414/sei2026.038

Author contact

kira.kappe@sei.org
adis.dzebo@sei.org

Media contact

ulrika.lamberth@sei.org

Visit us: sei.org

Stockholm Environment Institute is an international non-profit research institute that tackles climate, environment and sustainable development challenges.

We empower partners to meet these challenges through cutting-edge research, knowledge, tools and capacity building. Through SEI's HQ and seven centres around the world, we engage with policy, practice and development action for a sustainable, prosperous future for all.

References

Dzebo, A., Kappe, K. (2026). Classifying and benchmarking LLM-extracted summaries for policy evaluation. (SEI Brief). https://doi.org/10.51414/sei2026.033

Dzebo, A., Kappe, K., Babis, W., & Cabre, M. M. (2026). Outcomes and Process Evidence from Reviews and Audits (OPERA) Database (Version 1.0) [Dataset]. https://polevaldata.kolle44server.org/

Dzebo, A., Shawoo, Z., & Browne, K. (2025). Does Policy Coherence Make National Implementation of Global Sustainability Agendas More Successful? Annual Review of Environment and Resources, EG50. https://doi.org/10.1146/annurev-environ-111523-102337

Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2019). On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. Social Indicators Research, 141(1), 61–94. https://doi.org/10.1007/s11205-017-1832-9

Kappe, K., & Dzebo, A. (2025). Can AI produce reliable and consistent data analysis? (SEI Brief). https://doi.org/10.51414/sei2025.040

Nilsson, M., Hackmann, H., Skoba, Y., Guilanpour, K., Oni, T., Dzebo, A., Reyers, B., Zusman, E., Olsen, S. H., & Onoda, S. (2024). Seeking synergy solutions: Policies that support both climate and SDG action [Expert Group on Climate and SDG Synergy]. United Nations Department for Social and Economic Affairs and the United Nations Framework Convention on Climate Change. https://sdgs.un.org/basic-page/seeking-synergy-solutions-four-thematic-reports-55697

OECD, European Union, & European Commission – Joint Research Centre. (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD. https://doi.org/10.1787/9789264043466-en

Peters, B. G., & Pierre, J. (2023). Same, same but different? The expansion of auditing and its consequences for policy evaluation. In Handbook of Public Policy Evaluation (pp. 93–103). Edward Elgar Publishing. https://www.elgaronline.com/edcollchap/book/9781800884892/book-part-9781800884892-13.xml

SEI Stockholm Environment Institute