SEI Stockholm Environment Institute

Classifying and benchmarking LLM-extracted summaries for policy evaluation

SEI brief
August 2026

Adis Dzebo
Kira Kappe

1. Introduction

This is the second of three briefs on methods for AI-enabled systematic literature review and evidence synthesis. In the first, we discuss how large language models (LLMs), as well as SEI's AI Reader, could support information extraction from unstructured documents at scale and the steps that need to be taken to ensure reliable and accurate results (Kappe & Dzebo, 2025).

The SEI AI Reader demonstrated that LLMs can extract analytically usable summaries from unstructured evaluation documents at approximately 85% agreement with human coding. However, extraction alone does not yield comparable data.

This brief explains how the resulting 81 758 summaries, extracted from 2164 policy evaluations and independent audits, are converted into a comparable evidence base for detailed analysis of the processes and conditions that were responsible for policy success and failure in 116 countries and across 21 sectors. This data and its classification are the major component of the Outcomes and Process Evidence from Reviews and Audits (OPERA) database (Dzebo et al., 2026).

2. Devising a structure for analytical processing

The SEI AI Reader (Babis et al., 2024) is designed to enable a structured analysis where dependent and independent variables are defined, and additional context provides flexibility to iteratively fine-tune through variable-specific input and guidelines.

The policy evaluation framework applied in the first brief was continuously used throughout the AI-processing pipeline. It is structured around three layers. Benchmark variables (Part 0) capture policy objectives, success criteria, theory of change, evaluation methods, and main findings, anchoring *what* each evaluation set out to assess and *how rigorously* it did so. Outcome variables (Part 1) cover effectiveness, efficiency, outcomes and impact, and attribution, capturing *what was achieved*. Process variables (Part 2) cover agenda-setting, formulation, content coherence, implementation, and engagement and inclusion, capturing *how it was achieved*. This separation enables us to study process-outcome relationships in policy evaluation.

A pre-processing step provided a decisive improvement in data quality. It included text cleaning (e.g. removing headers and footers and margin text) and processing all

documents from PDF to AI-friendly Markdown format using a tailored optical character recognition (OCR) process that extracted tables and figures separately from the main document, using three separate models,[1] and merging them together in a post-processing step.

This step increased mean summaries per document from 21 to 35 and provided a balanced distribution between outcome and process summaries. The extracted data is summarized in Table 1.

Table 1. Database overview

Metric	Database value
Total documents	2164 (1533 audits, 631 evaluations)
Total summaries	81 758
Mean summaries per document	37.8 (median 33, range 9 to 173)
Countries	116 and the European Union
Sectors	21
Benchmark summaries	15 093
Outcome summaries	37 368
Process summaries	32 750

3. Variable classification with natural language processing

Using LLMs for repeated classification of text would incur significant financial costs as well as contribute to unnecessary environmental impact due to their energy consumption. Results would also be non-deterministic because the same input would produce different outputs at different runs (Atil et al., 2025). Instead, a lighter sentence transformer encodes each summary as a dense vector that is passed to a logistic regression classifier trained separately for each of the nine variables (Bui et al., 2016). Production runs three encoders, assigned per variable: an adapted gte-modernbert-base at 768 dimensions, an adapted bge-large-en-v1.5 at 1024, and the unadapted gte-modernbert-base.

The two adapted encoders were fine-tuned on 8852 contrastive text pairs drawn from manually labelled summaries across 1281 documents, held document-disjoint from the locked test set. Domain fine-tuning adapts the model's vector space to evaluative language, improving its ability to distinguish phrases such as "achieved 40% of the target" from "achieved 90% of the target", which generic models treat as near-identical ((Zhang et al., 2025). Its weakness is that contrastive training treats every pair as simply positive or negative, which fits ordinal categories poorly (Hoffmann et al., 2022).

The OPERA database at a glance: four steps from text to evaluation score

1. The SEI AI Reader extracts 81 758 analytical summaries from 2164 documents.
2. Five benchmark variables, the policy objective, success criteria, theory of change, evaluation method and findings, anchor what each document set out to assess and how rigorously it did so.
3. Nine fine-tuned natural language processing (NLP) classifiers assign all summaries a classification based on its variable belonging and manual validation, achieving production-grade accuracy.
4. Confidence-weighted basis scores and quality bonuses for outcome and process variable summaries aggregate to a document score on a 0 to 100 scale, mapped to one of five success tiers.

4. Model training and inference

Systematic A/B testing of both researcher-validated summaries and random database examples compared the base model against the fine-tuned. This resulted in a hybrid combination where six fine-tuned variable models performed better than the base model. Three variables – attribution, outcomes and impact, and content coherence – which depend on auxiliary information rather than evaluative signals, performed worse after fine-tuning. These variables depend on contextual and definitional reasoning, for example recognizing what constitutes a strong causal claim, or whether a policy aligns with international framework mechanisms.

A parallel ablation process showed that the justification text, an LLM-generated motivation summary for each extraction was shown to improve accuracy between 6 and 11 percentage points. In contrast, similar tests with benchmark summaries gave mixed results, mainly due to crowding of the 512-token context window of the model.

This combination produced mean production test accuracy of 90% across the nine variables, ranging from 81.1% (Effectiveness) to 100% (Agenda-setting and content coherence). Table 2 summarizes the performance for all nine variables.

Table 2. Classification of model performance

Variable	Model*	Classes	Test acc.	95% CI	CV acc.	Macro F1	Test n
1. Effectiveness	gte adapted	5	78.6%	68.7 to 86.0	75.6%	0.80	84
2. Efficiency	gte adapted	4	88.1%	78.2 to 93.8	86.7%	0.88	67
3. Outcomes and impact	bge-large, adapted	5	90.1%	81.7 to 94.9	88.7%	0.88	81
4. Attribution	gte-base	4	80.4%	71.6 to 86.9	74.9%	0.80	102
5. Agenda-setting	gte-base	4	79.7%	68.8 to 87.5	76.5%	0.79	69
6. Formulation	gte-adapted	4	86.6%	77.6 to 92.3	81.1%	0.87	82
7. Content coherence	bge-large, adapted	5	89.2%	80.7 to 94.2	84.3%	0.90	83
8. Implementation	bge-large, adapted	4	82.6%	73.2 to 89.1	76.9%	0.84	86
9. Engagement	bge-large, adapted	3	87.7%	78.7 to 93.2	83.8%	0.88	81

Note 1: Test accuracy on a locked, document-disjoint held-out set.
Note 2: All nine heads are flat multi-class and include No Data as a predicted class, so the figures measure the full task. The intervals are wide because the held-out sets hold 67 to 102 rows.

5. Refining and quality control

Two additional post-training and post-inference steps were applied for increased reliability. First, each prediction carries a probability that is assigned one of four confidence levels, Strong, Moderate, Weak or Unclear, calibrated per variable against the manually classified summaries and matched to the model's accuracy, so that an accurate variable is allowed a relaxed threshold for its top label whereas a less accurate one requires a higher probability before its prediction is trusted. Calibration is achieved during training rather than afterwards, by selecting each head's regularisation strength on out-of-fold expected calibration error, so no post-hoc temperature is applied to any of the nine variables (Guo et al., 2017). The confidence level in turn weights each summary's contribution to the aggregated score, ensuring that low-quality predictions are attenuated rather than discarded.

Table 3. Variable categories with summary scores

Variable	Top-tier score	Mid-tier score	Bottom score	Negative score
1. Effectiveness	Effective (90)	Partially Effective (60)	Not Effective (10)	Ineffective (5)
2. Efficiency	Efficient (90)		Not efficient (20)	Inefficient (5)
3. Outcome and impact	Outcome and impact (95)	—	No outcome (20)	Negative outcome or impact (5)
4. Attribution	Strong attribution (95)	—	Weak or no attribution (20)	Negative attribution (5)
5. Agenda setting	Strong agenda-setting (90)	Mixed (60)	—	Conflictive Agenda-Setting (5)
6. Formulation	Well-drafted (90)	Adequately Drafted (60)	—	Poorly Drafted (5)
7. Content coherence	Coherent (90)	Partially coherent (60)	Not coherent (20)	Incoherent (5)
8. Implementation	Effective (90)	—	Not effective (10)	Ineffective (5)
9. Engagement	Meaningful (90)	—	—	Insufficient (5)

Two cleaning steps run before a summary reaches an encoder. The first is a quality screen: each summary is checked for defects such as truncation, empty page references, table or inline HTML markup and/or heavy LaTeX formatting. Markup artefacts are stripped where readable text survives and the remainder are excluded, leaving 81 630 of the 81 758 summaries clean. The second is text normalisation. Evaluation summaries open with formulaic scaffolding, “The audit report notes that”, “The coordination mechanisms described include”, “Stakeholder participation was assessed through”, which carries no signal about the class while consuming the encoder’s limited attention. If cleaning would leave fewer than five words or 25 characters, the original is kept. Page citations are deliberately not stripped. Removing them was tested and degraded every variable measured, because the encoders were fine-tuned on text that included them.

Finally, two within-variable adjustments were implemented:

- A confidence downgrade, which sends a variable’s top class to No Data when its probability falls below that variable’s Moderate boundary. It is score-neutral on all nine variables, because the boundary coincides with the Weak to Moderate confidence cut.
- A global length cap, which holds any summary shorter than 20 words to at most Moderate confidence, so that a thin fragment cannot carry full weight.

The categories of variables and summary scores are presented in Table 3.

6. Aggregating summaries for document-level evaluation

A simple average of summary classifications is inadequate for producing comparable document scores, because of the varying number of summaries per evaluation and the asymmetry between four outcome variables and five process variables. Instead, a scoring methodology was developed, following established composite indicator methodology (Greco et al., 2019; OECD et al., 2008). The methodology consists of four scoring components.

The first two components include individual basis scores for outcome and process variables. The two sides are aggregated differently because their evidence behaves differently. The four outcome variables are independent dimensions of success and are not equally important, so each is judged in its own right (Munda and Nardo, 2009; Naidoo, 2020). The five process variables are stages of a single implementation pathway. Process documentation coverage is more spares, with only a quarter of documents covering all five stages. Here, five equal slots with the strongest available process evidence, one per stage, rewards breadth where it exists (Kara et al., 2022). In both cases, the maximum basis score was 50 points.

Second, two quality-bonus methods complemented the outcome and process basis scores. A bonus was awarded only when at least two top-tier summaries co-occur across multiple variables. Bonuses increased with the number of top-tier summaries per variable and could add 10 points in total for both outcome and process basis scores.

A dampening effect was applied based on the mean confidence, as well as for long documents with a larger amount of summaries, using pivoted length normalisation (Singhal et al., 1996; Kish, 1965). Finally, despite the theoretical maximum score being 109.25, scores were capped at 100 and a mid-tier ceiling was put on those evaluations that did not have any top-tier classified summaries.

The resulting score (0–100) was categorized in five success tiers: Elite (≥ 90), Gold (75–89), Good (50–74), Moderate (25–49), and Poor (< 25). The final tier distribution of the 2839 evaluations documents and other major findings from the OPERA database are documented in Brief 3.

7. Conclusion

This brief introduced the text classification methodology, combining policy evaluation frameworks, policy implementation theory and natural language processing (NLP), used in the OPERA database (Dzebo et al., 2026). The methodology was applied to more than 81 758 summaries extracted from 2164 policy evaluations and independent audits. The scoring methodology is deliberately optimized for identifying well-evidenced success. Below the top two tiers, low scores may reflect genuinely unsuccessful policies, incomplete reporting, or both. To address the complementary question of why policies fail, a separate failure-mechanism classification was developed and will be discussed in subsequent publications.

References

Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., & Baldwin, B. (2025). *Non-Determinism of "Deterministic" LLM Settings* (arXiv:2408.04667). arXiv. https://doi.org/10.48550/arXiv.2408.04667

Babis, W., Munoz Cabre, M., Dzebo, A., Martelo, C., Salzano, C., Torres Morales, E., & Arsadita, F. (2024). *SEI AI Policy Reader* [Dataset]. Stockholm Environment Institute. https://gptbatchpolicyprocesstool.streamlit.app/

Bui, D. D. A., Del Fiol, G., & Jonnalagadda, S. (2016). PDF text classification to leverage information extraction from publication reports. *Journal of Biomedical Informatics*, *61*, 141–148. https://doi.org/10.1016/j.jbi.2016.03.026

Clark, C., & Divvala, S. (2016). *PDFFigures 2.0: Mining Figures from Research Papers*. JCDL. http://pdffigures2.allenai.org/

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/

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

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning – Volume 70, ICML'17*, 1321–1330. https://dl.acm.org/doi/10.5555/3305381.3305518

Hoffmann, D. T., Behrmann, N., Gall, J., Brox, T., & Noroozi, M. (2022). Ranking Info Noise Contrastive Estimation: Boosting Contrastive Learning via Ranked Positives. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(1), 897–905. https://doi.org/10.1609/aaai.v36i1.19972

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

Kish, L. (1965). Sampling Organizations and Groups of Unequal Sizes. *American Sociological Review*, *30*(4), 564–572. https://doi.org/10.2307/2091346

Li, Z., Liu, Y., Liu, Q., Ma, Z., Zhang, Z., Zhang, S., Guo, Z., Zhang, J., Wang, X., & Bai, X. (2025). *MonkeyOCR: Document Parsing with a Structure-Recognition-Relation Triplet Paradigm* (arXiv:2506.05218). arXiv. https://doi.org/10.48550/arXiv.2506.05218

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

Smock, B., Pesala, R., & Abraham, R. (2021). *PubTables-1M: Towards comprehensive table extraction from unstructured documents* (arXiv:2110.00061). arXiv. https://doi.org/10.48550/arXiv.2110.00061

Zhang, Y., Zhu, Y., Zhu, Z., Liu, P., Xie, P., & Wu, C. (2025). A Domain-Finetuned Semantic Matching Framework Based on Dynamic Masking and Contrastive Learning for Specialized Text Retrieval. *Electronics*, *14*(24). https://doi.org/10.3390/electronics14244882

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

DOI:

https://doi.org/10.51414/sei2026.033

Author contact

authorcontact@sei.org
authorcontact@sei.org

Media contact

mediacontact@sei.org

Visit us: sei.org

Notes

1 The three models used were Microsoft TATR (Table Transformer) for table detection and structure recognition (Smock et al., 2021), and pdffigures2 for figure and caption extraction (Clark & Divvala, 2016). The main text was processed with MonkeyOCR (Li et al., 2025).