SEI Stockholm Environment Institute

Can AI produce reliable and consistent data analysis?

SEI brief, August 2025

Kira Kappe

Adis Dzebo

With recent advances in artificial intelligence (AI) tools for research, tasks that would have previously overwhelmed even the most dedicated research teams, such as the systematic analysis of thousands of policy documents, are increasingly within reach. These tools now provide advanced document analysis capabilities even to researchers lacking technical expertise or advanced coding skills.

However, one question still lingers: how can we ensure that AI-generated data meets the quality standards demanded by academic scrutiny?

Here, we explore the potential of AI tools for systematic policy analysis, while also examining the challenges and pitfalls that may prevent AI from fully delivering on its promise.

Evidence-based policy evaluation

In a pilot project conducted by SEI researchers, we aim to assess the advantages, risks and limitations of using AI tools in academic research. Our focus is on policy evaluation analysis. We are currently conducting a large-scale review of outcome and impact evaluations of policy implementation, as well as independent audit reports with the help of SEI's AI Reader.

Our objective is to uncover the drivers of successful policy implementation in different countries and to extract insights into what enables effective outcomes across diverse national or thematic contexts. Focusing on climate policy, this work aims to address the persistent challenge of linking policies to successful outcomes and to identify patterns of effective implementation within specific national, socio-economic and governance contexts, thereby supporting more informed policymaking.

SEI AI READER

The SEI AI Reader (beta) is a document analysis tool developed in 2024 by SEI researchers (Babis et al., 2024). It utilizes large language models (LLM), such as ChatGPT, to assist with literature reviews and extract policy-relevant data across various national and thematic contexts. It features a user-friendly interface with customizable input options:

- A main query panel for targeted prompts.
- A structured table where users can define (i) key independent variables, (ii) instruct what data needs to be extracted for each variable and, (iii) provide examples to illustrate what should (and should not) be extracted.

The tool's simple interface is designed to ensure accessibility even for non-technical users.

Creating an analytical framework

To guide our analysis, we conducted a human-led literature review of established policy evaluation frameworks such as those by the OECD and the European Environmental Agency, to identify suitable variables for extracting the key elements that determine policy success. Our analytical framework is structured around two broad themes, focusing on their correlation and causality.

1. **Criteria for establishing successful policy implementation**, including effectiveness, efficiency, outcomes and impacts, attribution and spillover effects.
2. **Principles of effective policy implementation processes**, such as agenda-setting, policy formulation, content, implementation and stakeholder engagement.

Recognizing that policy processes vary across governance systems, our analytical approach is structured around three "universal" characteristics of effective policy processes: coordination, coherence and integration. Our analytical framework ultimately comprises 12 independent variables and a checklist of 46 questions.

To conduct the analysis, we utilized a newly compiled global database of impact and outcome evaluations, along with extensive repositories of independent audits of national policy implementation. Both databases include metadata and direct access to thousands of documents.

Operationalization and pilot review

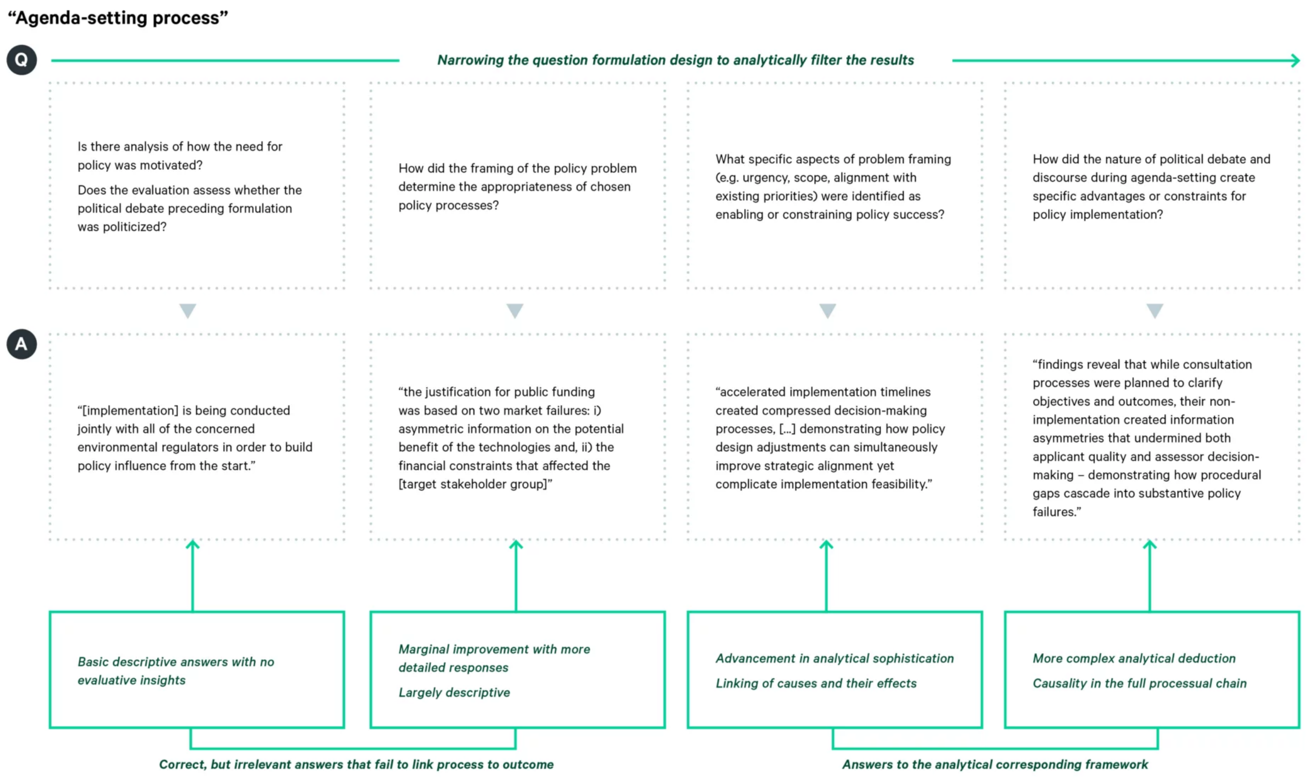
With our analytical framework and policy document dataset ready, we initiated a pilot study to test the tool's capabilities. We began by analysing four documents – both manually and using the AI Reader. We conducted approximately 10 iterative runs on each document using prompts of varying specificity and detail.

The purpose of this iterative process was to calibrate the tool by refining the input query, question formulation and context specification to match the accuracy of human analysis.

Prompt design and tool calibration

We quickly discovered that our initial analytical framework, designed with broad exploratory questions, was not effective in extracting the relevant information. The tool often returned generic, repetitive answers with no evaluative insights or omitted responses altogether.

Through iteration where each question was redesigned and reformulated, we managed to arrive at a solution where answers became more sophistically advanced and analytically complex. This process is reflected in Figure 1, showing how we eventually arrived at answers that corresponded to our analytical framework.



"Agenda-setting process"

Q — *Narrowing the question formulation design to analytically filter the results*

Is there analysis of how the need for policy was motivated? Does the evaluation assess whether the political debate preceding formulation was politicized?	How did the framing of the policy problem determine the appropriateness of chosen policy processes?	What specific aspects of problem framing (e.g. urgency, scope, alignment with existing priorities) were identified as enabling or constraining policy success?	How did the nature of political debate and discourse during agenda-setting create specific advantages or constraints for policy implementation?

A

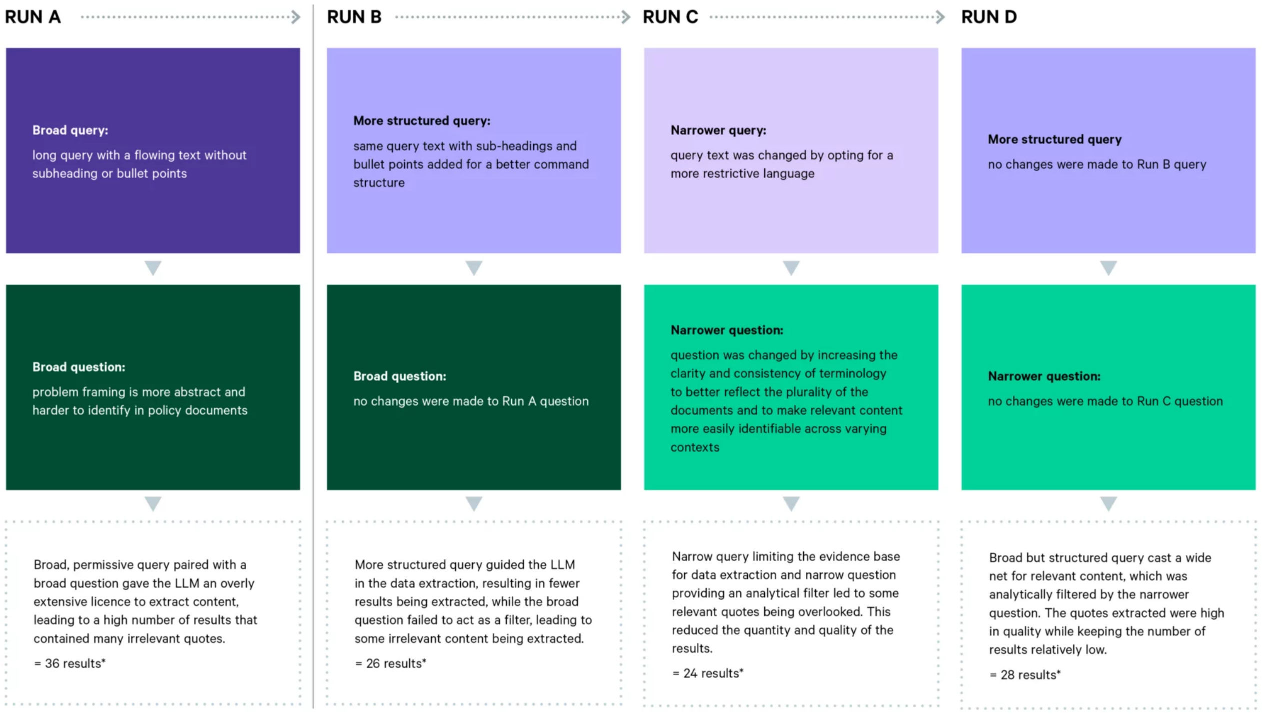
"[implementation] is being conducted jointly with all of the concerned environmental regulators in order to build policy influence from the start."	"the justification for public funding was based on two market failures: i) asymmetric information on the potential benefit of the technologies and, ii) the financial constraints that affected the [target stakeholder group]"	"accelerated implementation timelines created compressed decision-making processes, [...] demonstrating how policy design adjustments can simultaneously improve strategic alignment yet complicate implementation feasibility."	"findings reveal that while consultation processes were planned to clarify objectives and outcomes, their non-implementation created information asymmetries that undermined both applicant quality and assessor decision-making – demonstrating how procedural gaps cascade into substantive policy failures."
Basic descriptive answers with no evaluative insights	*Marginal improvement with more detailed responses* *Largely descriptive*	*Advancement in analytical sophistication* *Linking of causes and their effects*	*More complex analytical deduction* *Causality in the full processual chain*

Correct, but irrelevant answers that fail to link process to outcome

Answers to the analytical corresponding framework

Figure 1: Formulating the right question.
Graphic: Mia Shu / SEI.

Similarly, iterative refinement of queries and questions revealed how query-question (mis-)matches significantly influence both the quality and quantity of responses from the AI Reader. As shown in Figure 2, different combinations applied in individual runs (A-D) yielded distinct outcomes. The findings suggest that a broad but structured query combined with a specific, analytical question provides the most effective balance. This combination enables the AI reader to cast a wide net for all relevant content while applying a narrow filter that recognizes multiple types of evidence as relevant.



RUN A

Broad query: long query with a flowing text without subheading or bullet points

Broad question: problem framing is more abstract and harder to identify in policy documents

Broad, permissive query paired with a broad question gave the LLM an overly extensive licence to extract content, leading to a high number of results that contained many irrelevant quotes.

= 36 results*

RUN B

More structured query: same query text with sub-headings and bullet points added for a better command structure

Broad question: no changes were made to Run A question

More structured query guided the LLM in the data extraction, resulting in fewer results being extracted, while the broad question failed to act as a filter, leading to some irrelevant content being extracted.

= 26 results*

RUN C

Narrower query: query text was changed by opting for a more restrictive language

Narrower question: question was changed by increasing the clarity and consistency of terminology to better reflect the plurality of the documents and to make relevant content more easily identifiable across varying contexts

Narrow query limiting the evidence base for data extraction and narrow question providing an analytical filter led to some relevant quotes being overlooked. This reduced the quantity and quality of the results.

= 24 results*

RUN D

More structured query no changes were made to Run B query

Narrower question: no changes were made to Run C question

Broad but structured query cast a wide net for relevant content, which was analytically filtered by the narrower question. The quotes extracted were high in quality while keeping the number of results relatively low.

= 28 results*

**Results per sample question across all tested documents*

Figure 2: Choosing the right query – question combination.
Graphic: Mia Shu / SEI.

A third parameter supporting the gradual improvements made through iteration was the “context variable”, which allowed us to define terms, provide contextual guidelines to a specific independent variable and incorporate “do’s and don’ts” from previous run results to correct inconsistencies.

Other refinements added at later stages of the prompting included instructing the AI to provide original answers to each question, along with separate justifications. These justifications offered additional insights into why specific information was extracted and contextualized the answer in light of the question posed. This structure is expected to support later analysis, helping to more systematically assess and rank the relevance of extracted answers.

Accuracy and reliability

We assessed the tool's accuracy by comparing its responses to our manual assessment, using a simple quantitative scoring system. On average, the AI achieved approximately 85% accuracy across all four documents compared to our own analysis.

A perfect score was not possible due to the subjective nature of the topic – even our team did not always agree on the correct answer. Sometimes, the AI Reader provided insights missed by human analysis; in other cases, it failed to extract relevant information.

Importantly, there was significant correlation between human and AI analysis on which questions lacked available data, suggesting the AI could resist the urge to "please" by fabricating answers. Notably, we did not observe hallucination of facts and wrongful answers, likely due to safeguards in our prompt design (e.g. requiring page references, direct quotes and explicit instructions not to hallucinate).

Consistency

We also cross-analysed the answers from multiple runs to assess the consistency of the tool's outputs and determine how often answers were the same or similar, comparable or significantly different. The findings are cautiously promising: consistency across the four documents ranged from 69% to 90%.

One recurring issue was that while 3 or 4 runs often produced similar results, one run would occasionally differ. The reason for this discrepancy remains unclear.

As a potential solution for the final analysis of the full dataset, we suggest running each document twice to help identify and compensate for potential outliers or inaccuracies. However, some inconsistency is likely unavoidable when working with AI tools, due to the inherent "black box" nature of LLMs such as ChatGPT.

Additional considerations when conducting multiple runs include the environmental impacts of repeated processing and the added workload of comparing and consolidating the results.

Implications and potential for scalability

Returning to our central question – "how can we ensure that AI-generated data meets the quality standards demanded by academic scrutiny?" – our pilot review suggests that, despite some limitations, the consistency, accuracy and reliability of AI-generated data are sufficiently high.

This makes advanced tools like the SEI AI Reader a promising solution for overcoming the methodological challenges and time constraints involved in systematically processing and synthesizing the vast and growing body of climate policy evaluations, particularly grey literature. These tools can help derive actionable insights from past policymaking experiences.

Our next step is to expand the analysis to a larger set of documents to validate the current calibration and assess the scalability of our approach before proceeding to the main analysis.

Even at this early stage, our findings suggest that AI tools hold considerable potential for analysing large volumes of policy-related documents, supporting the identification of common patterns in successful policy implementation across varied national and thematic contexts.

This research contributes to closing persistent empirical gaps by:

1. enhancing understanding of how coherent policymaking contributes to effective implementation outcomes, and
2. identifying the specific, localized conditions that explain “what works, where, how and why?” (Browne et al., 2023; Dzebo et al., 2025).

Published by

Stockholm Environment Institute
Linnégatan 87D 115 23 Stockholm,
Sweden
Tel: +46 8 30 80 44
www.sei.org

DOI:

http://doi.org/10.51414/sei2025.040

Author contact

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

Media contact

ulrika.lamberth@sei.org

Cover photo

buraratn / Getty Images.

Visit us: sei.org
Twitter: @SEIresearch
@SEIclimate

References

Babis, W., Muñoz Cabré, M., Martelo Llerena, C., Salzano, C., Torres-Morales, E., & Arsadita, F. (Forthcoming, 2025). *SEI AI Reader* [Dataset]. Stockholm Environment Institute.

Browne, K., Dzebo, A., Iacobuta, G., Faus Onbargi, A., Shawoo, Z., Dombrowsky, I., Fridahl, M., Gottenhuber, S., & Persson, Å. (2023). How does policy coherence shape effectiveness and inequality? Implications for sustainable development and the 2030 Agenda. *Sustainable Development, 31*(5):3161-3174. https://doi.org/10.1002/sd.2598

Dzebo, A., Shawoo, Z., & Browne, K. (2025). Does policy coherence make national implementation of global sustainability goals more successful. *Annual Review of Environment and Resources*, EG50. https://doi.org/10.1146/annurev-environ-111523-102337.

Nilsson, M., Hackmann, H., Sokona, Y., Guilanpour, K., Oni, T., Dzebo, A., & Onoda, S. (2024). *Seeking synergy solutions: policies that support both climate and SDG action.* Expert Group on Climate and SDG Synergy. UN Department of Economic and Social Affairs. https://sdgs.un.org/sites/default/files/2024-06/Thematic%20Report%20on%20Climate%20and%20SDGs%20Action-060824.pdf