United Nations
Sustainable Development Group
System-Wide Evaluation Office

Utilizing United Nations evaluation evidence in support of the 2024 QCPR

SWEO Learning Paper #1

February 2025

In 2024, the UNSDG System-Wide Evaluation Office (SWEO) led a pilot cross-system initiative aiming to provide user-friendly mapping and summary products of United Nations evaluation evidence to support engagement in the 2024 Quadrennial Comprehensive Policy Review (QCPR) process,[1] and ultimately contribute to more effective UN development system support to the implementation of the 2030 Agenda. The initiative produced:

i. interactive evidence maps featuring United Nations evaluations, published between 2021 and 2024, mapped against priority areas of the 2020 QCPR and SDGs
ii. summaries of UN evaluation evidence on five priority topics:
   a. the Resident Coordinator system,
   b. UN development system regional architecture,
   c. funding quality,
   d. food systems; and
   e. humanitarian-development-peace linkages

This initiative provided a strong proof of concept for UN system-wide evaluation evidence mapping and summaries. It also generated significant learning to be taken forward, both in terms of its substantive findings and the methodology employed (which leveraged artificial intelligence (AI) through a large language model (LLM) for evidence classification). This paper provides an overview of the key takeaways from the initiative for the UN's evaluation, knowledge management and evidence-based policy making communities.

1. Context

The initiative was conducted in the context of a rich but fragmented body of UN evaluations with limited utilisation of this evidence base to contribute to high level UN system decision making and to inform intergovernmental bodies and processes. Key elements of this context included:

Evidence generation side	Evidence use side
- The UN system produces around 1000 evaluation reports per calendar year[2] - UNEG database is incomplete and contains inconsistent tagging of	- Increased Member State demand for reporting on UN contributions to development results and the implementation of mandates,

[1] The QCPR is the mechanism through which Member States review and guide UN operational activities for development through the work of the United Nations Economic and Social Council (ECOSOC) and the General Assembly.

[2] Internal SWEO estimate based on extraction of reports from the UNEG evaluation database, identification of gaps in that dataset and gathering of reports from evaluation offices.

reports (by evaluation type, theme etc.) - There is no UN system-wide interactive evaluation evidence mapping platform	including possible gaps and challenges (A/Res/78/166) - Limited use of evaluation evidence in UN development system-wide reporting (e.g. SG reports on QCPR and the QCPR monitoring framework – primarily informed by monitoring surveys)
Gap to be bridged between evaluation evidence availability and use at UN system/strategic/intergovernmental levels	

2. Objectives

The primary objective of the initiative was to make UN evaluation evidence more accessible to users, to support evidence-informed decision-making by Member States and UN development system entities in the context of the 2024 QCPR process.

The secondary objectives, which are the main focus of this paper, were to:

(i) Provide a proof of concept for UN system-wide evaluation evidence mapping, led by the newly established UNSDG System-Wide Evaluation Office

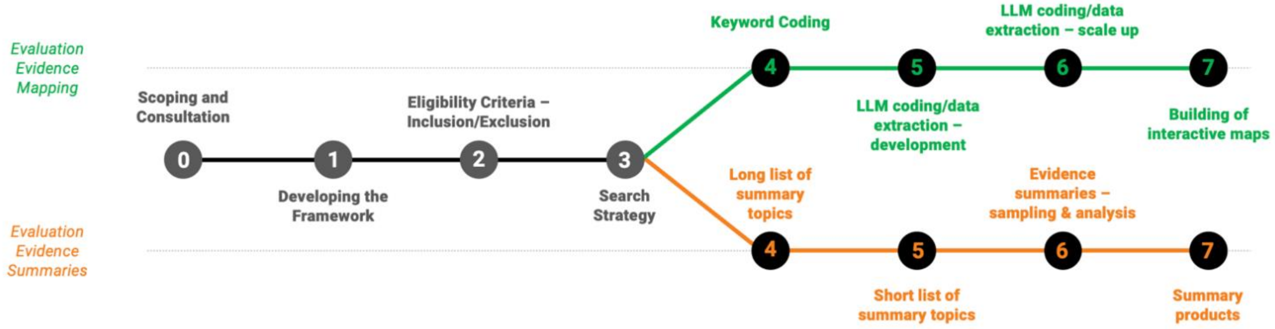
(ii) Initiate discussions around trends in evaluation practice and opportunities for improvement in evaluation knowledge management across the UN system

(iii) Generate learning for the UN evaluation and evidence synthesis community in the use of artificial intelligence in evidence mapping

3. Methodology

The interactive evaluation evidence maps, and the evidence summaries were produced between April and October 2024 in the following phases (Figure 1):

Figure 1: Summary of Overall Methodology


Evaluation Evidence Mapping
Evaluation Evidence Summaries
Scoping and Consultation
0
1
Developing the Framework
Eligibility Criteria – Inclusion/Exclusion
2
3
Search Strategy
Keyword Coding
4
5
LLM coding/data extraction – development
LLM coding/data extraction – scale up
6
7
Building of interactive maps
Long list of summary topics
4
5
Short list of summary topics
Evidence summaries – sampling & analysis
6
7
Summary products

A detailed breakdown of the phases is presented below:

0 **Scoping and consultation:**

Initial consideration of the context and opportunities. A Concept Note was drafted and presented to the UN Evaluation Group (UNEG) for consultation and feedback. An inter-agency Management Group (UNDP, UNFPA, WFP) comprising Senior Evaluation Officers and synthesis specialists was formed. Resource mobilisation was conducted.

1 **Developing the framework:**

The existing and agreed QCPR monitoring framework[3] was confirmed as the most appropriate starting point for the mapping of evaluation evidence against QCPR priorities.

As the typical evaluation scope within the UN system differs somewhat from the categories in the framework, it was necessary to adapt the framework to allow for the most useful categorisation and presentation of evaluation evidence. Certain topics were broken down into subcategories, e.g. "capacity development" into data/statistics and science-technology-innovation, and "cross cutting issues" into gender equality, youth, human rights and disability. Other topics were merged together, e.g. different measures on UNDS funding quality and development system reform implementation (see Annex 2 for details). The SDGs were selected as a secondary mapping framework.

Given the knowledge management/accessibility aims of the initiative, an interventions vs. outcomes framework was not deemed appropriate. Instead, the axes for mapping were QCRR theme/priority (or SDGs) and type of evaluation.

2 **Eligibility criteria – inclusion/exclusion:**

SWEO's long term ambition is to map all evaluation evidence produced by the UN system in real time through "living maps",. Considering the then-active intergovernmental discussions on the QCPR in 2024, the primary eligibility/inclusion criteria for the pilot initiative was a pragmatic consideration of the types of evaluation that were most likely to include evidence of UN system-wide significance and/or UN system contribution to the SDGs. This initiative sought to map evaluations meeting the below definitions, published by UN entities in the 2020-2024 QCPR cycle:

- ***Country-level*** *– including evaluations of country portfolios, country strategies, country programme documents, country-wide crisis responses, multi-year partnerships with a national government*

- ***Regional*** *– including evaluations across multiple countries, covering portfolios of projects/programmes, regional strategic plans/policies, typically managed from a regional office etc.*

- ***Thematic*** *– including global/corporate evaluations of a particular thematic area of the organisation's work, covering multiple interventions in multiple countries (this may be referred to as a "global programme" in some entities).*

[3] A Monitoring Framework to track the implementation of QCPR resolutions has been maintained by UN DESA since 2012 QCPR resolution. The 2020-2024 framework was developed through inter-agency consultation in 2021 and includes 5 axes and 24 topic areas. It draws on a number of system-wide data sources, most notably annual surveys of UN entity HQs, UNCT members, RCs and programme country governments, administered by DESA. - https://ecosoc.un.org/sites/default/files/2024-05/QCPR-MF_SGR2024-AdvancedVersion-13May2024.xlsx

	Evaluation Evidence Mapping	Evaluation Evidence Summaries
	- ***Strategic/Policy*** – *including global/corporate evaluations of strategies and policies (organizational or sectoral), including evaluations of organizational strategic plans and/or policy quality/implementation.* - ***Joint/pooled funding*** – *including evaluations of joint programmes delivered by 2+ entities and portfolios of projects funded by multi-partner pooled funds* - ***Evaluation syntheses*** – *including studies that synthesize or summarise findings from a body of existing evaluation work, across an organization/thematic area/geographic area etc*. These categories covered around 25% of the estimated 4000 evaluation reports produced in the UN system from 2021 - 2024. This sample excluded **project evaluations** covering a single project, implemented by a single UN entity, in a single country (which are estimated to account for up to 75% of UN evaluations). While potentially relevant at system level, such evaluations were assumed, for the purposes of this pilot initiative, to be of more limited system-wide/strategic relevance.	
3	**Search strategy:** An iterative search strategy was utilised to aggregate all UN system evaluations that met the aforementioned inclusion criteria. 1. Extraction of evaluation meta-data from the UNEG database – this provided useful starting point for the search and an indication of the overall volume of evaluations produced by the UN system each year (~1000) in the period in question. However, review of the data suggested that the database had major gaps in coverage, and inconsistency in tagging across different entities made the confident identification of the above types of evaluation very difficult. 2. UN entities (UNEG and UNSDG members) were requested to review lists of evaluations exported from the UNEG database, to confirm that listed evaluations met the criteria and to add evaluations that were missing[4]. 3. UN entity responses were screened by SWEO for relevance, to remove duplicates, and tagged against the evaluation types above. Extensive data cleaning was required to standardise fields, correct titles/dates and confirm valid URLs. This resulted in a final sample of 950 evaluation reports, representing approximately 25% of evaluations published by the UN system in the period.	
4	**Keyword coding:** A rapid, preliminary mapping of the reports was conducted through Excel formula true/false keyword searching of titles and descriptions (where they existed) and manual tagging of certain categories of report (e.g. by type or by entity) to certain QCPR themes/priorities. For example, manually coding of all OHCHR reports to	**Long list of summary topics:** This basic preliminary mapping enabled the identification of a long list of topics where there would likely be a sufficient number of evaluation reports and sufficient depth and range of evidence to analyse and identify key insights in the summary products.

[4] 32 entities sent validated/updated lists of reports. 5 entities responded to confirm that they had published no evaluations meeting the criteria during the period. SWEO screened extracted UNEG database lists for 6 entities that did not respond. A further 16 entities did not respond and had no evidence of relevant evaluations (meeting inclusion criteria) on their website. 2 entities did not respond but did have evidence of relevant evaluations (meeting inclusion criteria) on their website which were reviewed and included by SWEO.

Evaluation Evidence Mapping		Evaluation Evidence Summaries
	"human rights" (regardless of keyword search outcomes).[5]	
5	**LLM coding/data extraction – development:** The sample of reports were classified by QCPR priority and SDG using the following techniques: - Manual classification of a purposively selected subset of 75 reports[6] by SWEO to serve as a "test set" - Prompt engineering for LLM classification (without model development/training) - Multiple applications of the prompt to the manually tagged set of reports with quality checking/manual review to identify perceived errors against the test set and opportunities for prompt refinement - Quality and accuracy improved gradually with each iteration (See Findings/Lessons Learned section below for further details on this process)	**Short list of summary topics:** The long list of summary topics was reduced following consultation on priorities with intended end users and giving consideration to the availability of existing (and forthcoming) inter-agency evaluation syntheses. Five priority topics were selected (see Final Products section).
6	**LLM coding/data extraction – scale up:** The final prompt, refined following multiple iterations, was used to classify all 950 reports and included: - Relevance scores for each QCPR priority/SDG and each report - Extraction of supporting text passages (with sections/page numbers) to facilitate review and build confidence - A generated summary/explanation of if/how each QCPR priority/SDG was included in the report and why the report has been tagged or not The same prompt was also used to extract report meta-data in a more consistent manner than was found in manually maintained databases – e.g. extraction of the countries covered by the evaluation and UN agencies involved in joint evaluations. To fill a key gap in the available data set, generative AI was also used to produce	**Evidence summaries – sampling & analysis:** The preliminary mapping of evaluations and familiarity with the sample of evaluations enabled the team to quickly develop bespoke sampling frameworks to identify the 30-50 most relevant evaluations for inclusion in each summary. Data extraction and analysis was conducted against a simple analytical framework developed for each topic using tools such as MaxQDA, Dedoose and NVivo. Experimentation with LLMs helped accelerate extraction processes and broaden samples. Sampling and analysis was conducted in 4-6 weeks following the same processes as evaluation synthesis but **without rigorous screening for quality of evaluations.**

[5] This was a pragmatic step used to identify a long list of possible summary topics *before* comprehensive mapping of the evidence base was complete. It was specific to the circumstances of the pilot and would not necessarily be repeated.

[6] 75 reports were selected to include (1) all evaluation types, (2) all UN entities within the sample, and, on the basis of the keyword tagging alone, (3) all 24 QCPR priorities.

Evaluation Evidence Mapping		Evaluation Evidence Summaries
	standardised report abstracts (summarising report scope, findings, conclusions and recommendations in 250 words).	
7	**Building of interactive maps** The enriched datasets generated by the LLM-assisted coding/data extraction were presented using EPPI Mapper (a free, open source software for the creation of evidence gap maps). SWEO / UNDP IEO developed an Excel template for the conversion of datasets to the JSON format required as the input for EPPI Mapper – thus avoiding an additional step of manual coding in EPPI Reviewer.	**Summary products** Summaries were drafted in a common, user-friendly/accessible format of between 10-15 pages. Shorter two-page briefs were also produced. The summaries included: - Background on the QCPR mandate/official reporting - 5-10 key insights from evaluations on specific topics for consideration by system-wide policy makers and intergovernmental bodies, e.g. emerging issues, themes or recurring findings, conclusions and recommendations across different contexts and entities. - Annotated and hyperlinked bibliography

4. Final products

Links to the final evidence summaries and interactive maps are provided below.

A visible shift - the independent resident coordinator I Brief

UNSDG SYSTEM-WIDE EVALUATION OFFICE

A visible shift – the independent resident coordinator

Summary of United Nations evaluation evidence

United Nations

United Nations development system reform at the regional level - slow progress I Brief

UNSDG SYSTEM-WIDE EVALUATION OFFICE

United Nations development system reform at the regional level – slow progress

Summary of United Nations evaluation evidence

United Nations

Unlocking quality funding I Brief

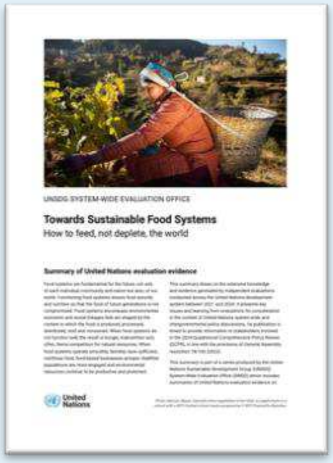
UNSDG SYSTEM-WIDE EVALUATION OFFICE

Unlocking quality funding

Summary of United Nations evaluation evidence

United Nations

Towards Sustainable Food Systems I Brief


UNSDG SYSTEM-WIDE EVALUATION OFFICE

Towards Sustainable Food Systems
How to feed, not deplete, the world

Summary of United Nations evaluation evidence

United Nations

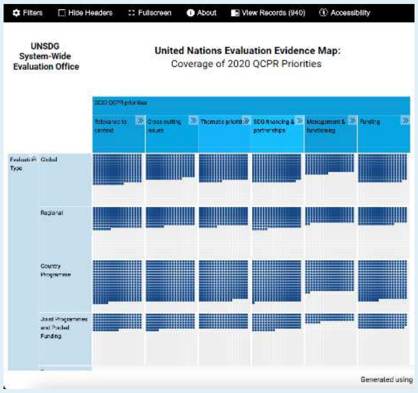
Building a whole of system response to complex settings I Brief

UNSDG SYSTEM-WIDE EVALUATION OFFICE

Building a whole of system response to complex settings

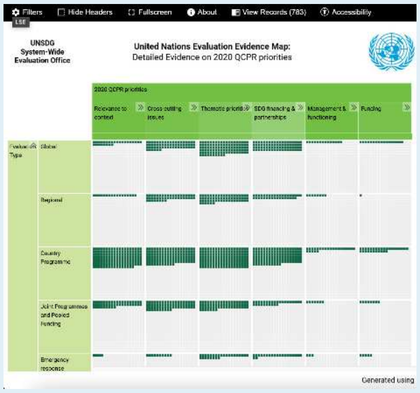
Summary of United Nations evaluation evidence

UN Evaluation Evidence Map: Coverage of 2020 QCPR Priorities



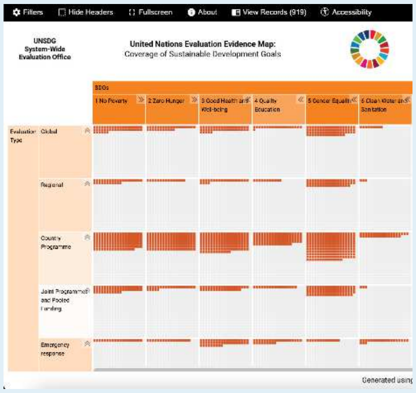
For the question: *How many evaluations cover topic x? Does y evaluation cover topic x?*

UN Evaluation Evidence Map: Detailed Evidence on 2020 QCPR Priorities



For: *Which evaluations contain the most / greatest depth of evidence on topic x?*

UN Evaluation Evidence Map: Coverage of Sustainable Development Goals



For: *How many evaluations cover SDG x? Does y evaluation cover SDG x?*

5. Findings/lessons learned

Substantive findings on evaluation in the UN system

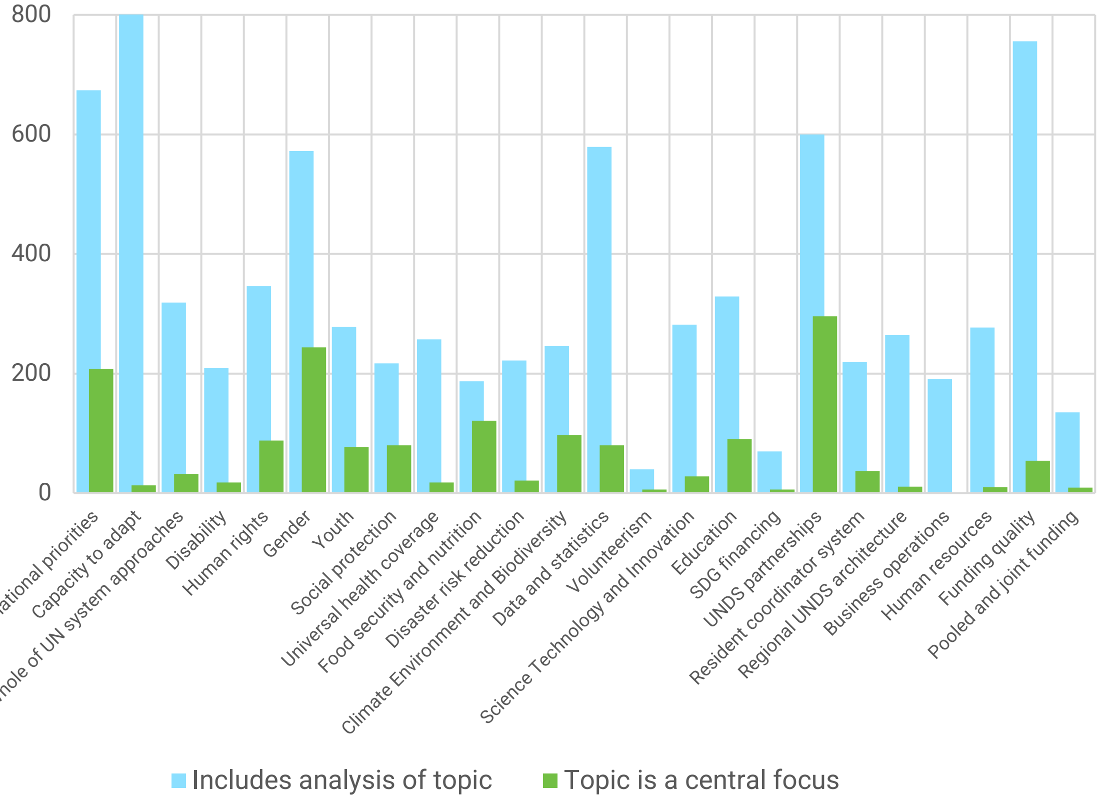
The mapping exercise provides some insights on the overall state of evaluation across the UN system. However, it should be taken into account that only approximately 25% of UN evaluations published between 2020 and 2024 were included in the analysis, and that these findings are not conclusive.

- Key areas of **evidence density identified against QCPR priorities** (Figure 2) include: relevance to national priorities, gender equality, core thematic priorities (e.g. social protection, food security, climate/environment, and education)
- Notable **evidence gaps against QCPR priorities** (Figure 2): Results-based management ('capacity to adapt'), disability inclusion; SDG financing; RC system/regional architecture; human resources/business operations; pooled/joint funding
- The use of relevance scales (0-3) in the LLM-assisted tagging methodology highlighted notable **variance in the extent to which analysis of QCPR priorities are "included" in evaluations or are a "central focus**". For example, topics such as results-based management and funding are covered by many reports with few reports having these topics as a central focus

These are both positive and negative findings. On one hand, this exercise demonstrates that UN evaluation reports often contain a rich depth of information within reports, which is not evident from report titles and compounded by the limited use of abstracts by most evaluation functions. . On the other hand, it illustrates gaps in evidence on important systemic, strategic and "enabling" factors for the overall effectiveness and functioning of the UN development system.

These findings were confirmed by the more detailed analysis of the sampled reports for the evaluation evidence summaries. This analysis also found a varying depth of analysis on systemic and cross-cutting issues such as funding quality/modalities (as opposed to funding availability), donorship, system-wide agendas, UN coherence/coordination, "business models", incentive structures etc.

Figure 2: Classification of evaluation reports by QCPR priority (2021-2024)



Lessons learned on AI-assisted evaluation evidence mapping

The evidence classification and mapping side of this initiative also generated considerable learning on the potential for acceleration and automation using AI / large language models in such tasks. While some issues and inaccuracies may still be present, the SWEO considers that the outputs, produced in a relatively short period of time, to be a strong proof of concept. Key learnings from the exercise include:

- **Maturity of commercially available LLMs**: the pilot used commercially available and easily accessible large language models[7]. Evidence classification testing and scale up to the full sample of reports was conducted through prompt engineering, without resource-intensive LLM training or model development. This approach met the requirements of the initiative and was very cost effective. It is important to note that the AI "use case" demonstrated by the pilots was focused on the *classification* of reports/evidence extraction and the generation of abstracts and meta data; it was not used to produce AI-assisted *synthesis* of evidence. SWEO expects technological advancements within AI to allow for use of increasingly accurate and mature LLMs in the near future.

[7] Google Gemini 1.5 Pro (with Anthropic Claude 3.5 Sonnet and OpenAI GPT-4 used as alternatives). Models were selected primarily for their large context window.

- **Testing/validating "percentage accuracy" against human coding is challenging:** the original intention of the methodology was for human experts to manually code a sample of evaluation reports against the mapping framework (QCPR priorities) and then work towards a target accuracy percentage with LLM processing, e.g. a 95% tagging accuracy of AI-tagged priorities against human expert tags. In reality, it was very difficult to test accuracy against the test set in a quantitative/statistical way. This was due to having multiple human experts and the natural human inconsistencies in tagging styles, even with prior agreement of a strict tagging protocol . Inconsistencies in the test set were sometimes greater than those tagged by the LLM.

- **Human/expert familiarity with the evidence to be classified by the LLM is nonetheless vital**: while the manual coding of a sample of reports did not result in a "gold standard" test set of reports to assess a percentage accuracy of LLM tagging against, the process was still vitally important. It ensured that the team involved in the testing of LLM tagging had a high level of familiarity with the reports being processed. This was important for effective and efficient review of test outputs and reengineering of the prompts. e.g. allowing quicker identification of "errors" or possible hallucinations.

- **Human-LLM collaboration:** the testing and development process for evidence extraction/coding was better characterised as an interaction or collaboration between human and artificial intelligence. The two compared and discussed their coding and arrive at a coding decision as two human researchers might do when selecting studies for inclusion in a systematic review ("double screening"). In this regard, classification prompts which ask the language model to *explain* decisions on a classification (or "show working") and extract relevant text passages were especially helpful. During the development and testing process, this helped to identify areas of weak comprehension by the model (or unclear prompting by the human) or possible hallucination and enable refinements in the next iteration. In the final product, the same "explainers" also provided transparency on classification decisions to the end user.

- **Implications for small teams/low budgets**: the approach described above can be implemented with a small team. Such teams will benefit from the inclusion of an AI/Machine Learning Specialist to lead on prompt engineering and data processing, and dedicated time from evaluation professionals to provide familiarity with a test set of reports and fast feedback on the quality of test outputs. These dynamics produced good "human-LLM collaboration" results (as explained above) in a short period of time. It should also be noted that the initiative benefitted from regular and meaningful engagement from its inter-agency management group (including specialists in evidence synthesis and mapping). This group provided invaluable advice and external validation of the approach. The main methodological trade off in this approach is between (i) the size of the manually tagged sample for testing and development and (ii) the number of iterations/revisions to prompts (re-engineering). In a small team, it may be most beneficial to prioritise multiple rapid iterations between prompt engineer and evaluation specialist.

6. Remaining challenges/limitations and next steps

The initiative highlighted some challenges and limitations facing AI-assisted evaluation evidence mapping, some of which can be addressed in further initiatives by SWEO and some which require adjustments to knowledge management practices across the UN system. These are set out below.

Issue		Implication / next steps
Data inputs	**No complete and consistently and comprehensively tagged UN evaluation database**. As set out in the Methodology section above, this initiative highlighted some limitations of the UNEG evaluation database.	This learning has been shared with ongoing efforts to revamp the UNEG evaluation database, following the recent launch of a new UNEG website. SWEO recommends the exploration of automated inclusion of UN evaluation reports in the database, pulled using APIs from UN entity databases, to remove human error and improve consistency in database edits and basic tagging.
	Evaluation report formats and structures vary significantly within and across UN evaluation offices. This can create challenges for machine readability and classification of reports. Inconsistencies include the availability and length of Executive Summaries and Abstracts, presence of certain evaluation sections (background, introduction, conclusion), overall writing style of qualitative sections and document file types and design.	UNEG should consider working on the establishment of more common standards for evaluation report presentation / structure / meta data. A possible first step may be to standardise inclusion of abstracts in evaluation reports. This could be assisted, for the existing repository and moving forward, by generative AI. This initiative used an LLM to generate standard 250 word abstracts for 950 reports – reviews suggest that these are of acceptable quality.
Data processing	**LLM comprehension (or quality of prompt engineering) varied across topics**, with lower overall accuracy suspected on certain topics. Examples included: - Challenges distinguishing between support to national data/statistics systems and internal UN data/statistics	To the extent possible, mapping frameworks should not include categories with vague parameters or overlapping categories of evidence. Challenges can be overcome to an extent through sustained collaboration between substantive experts and a skilled prompt engineer/AI specialist.

	- Challenges extracting and classifying relevant evidence on concepts such as Results-Based Management - Challenges distinguishing between SDG financing/UNDS funding sources/pooled and joint funding	Maturity of available LLMs will also continue to improve.
	LLM security settings / "prohibited content": In some instances, commercially available LLMs may refuse to process UN evaluation reports on grounds of harmful or prohibited content (e.g. female genital mutilation, child marriage, and other child protection issues).	Providers may be willing to remove security settings upon request from a UN entity. However, some guardrails on prohibited content are deeply embedded in the models and cannot be removed. Direct collaboration with LLM companies may be required in the future to explore potential solutions.
Next steps / opportunities	**Static nature of pilot maps:** the pilot outputs maps produced in 2024 are static and will not update in real time as new reports are published. This results in decreasing utility as time progresses.	The ambition of the next phases of this work should be to move towards living evidence maps, this would require either: - Embedding LLM evidence classification in existing databases to identify and classify evidence contained in evaluation reports periodically or in real time. - Protocols for the updating of existing evidence maps with newly published reports, which are detailed but also consistently adhered to
	Making evidence mapping and summaries more responsive to demand	The 2024 QCPR (and A/Res/78/166) provided an opportunity and entry point to pilot this work and demonstrated its potential value to decision makers. SWEO aims to make evidence mapping and summaries part of its ongoing multi-year work-plan. This would mean that topics for summaries/synthesis are consulted upon in similar way to

		topics for new system-wide evaluations and are planned ahead of key decision points (e.g. summits/HLPFs).

This initiative was a collaboration between SWEO and evaluation offices across the United Nations. It provided user-friendly mapping and summary products of United Nations evaluation evidence to support engagement in the 2024 QCPR. The initiative was coordinated by SWEO, with substantive contributions from the following entities:

Funding

IFAD Investing in rural people

UNHCR The UN Refugee Agency

unicef

WFP World Food Programme

Management group

UNDP IEO Independent Evaluation Office United Nations Development Programme

UNFPA IEO Independent Evaluation Office

WFP World Food Programme

The **United Nations Sustainable Development Group System-Wide Evaluation Office (SWEO)** is responsible for the provision of independent evaluation evidence to strengthen learning, transparency, and accountability, to incentivize joint work and collective learning, and to conduct system-wide evaluations and advance evidence on the United Nations development system's contribution towards achievement of the 2030 Agenda for Sustainable Development and the Sustainable Development Goals.

To ensure independence, impartiality, and credibility, the UNSDG SWEO is a standalone independent office within the United Nations Secretariat. The Office is led by the Executive Director, reporting directly to the Secretary-General, but exercising operational independence in the performance of the evaluation function.

https://www.un.org/system-wide-evaluation

un-systemwideevaluationoffice@un.org

United Nations
Sustainable Development Group
System-Wide Evaluation Office