International Labour Organization | Evaluation Office

i-eval THINK Piece, No. 29

Artificial intelligence is only as good as the knowledge behind it

Lessons from building the artificial intelligence features of *i*-eval Discovery

By Guy Thijs | May 2026

i-eval AI Assistant

At the time of writing, Guy Thijs was serving as Director of the ILO Evaluation Office, with his retirement scheduled for July 2026. Over nearly four decades with the ILO, he has consistently focused on system development and knowledge management. This interest was evident during his tenure as Director of IPEC and continued through his work on knowledge management platforms and communities of practice in the Asia and the Pacific. He further advanced these interests as Director of the Evaluation Office, with a strong emphasis on system improvement and knowledge management. This *i*-eval Think Piece has benefited from peer comments and inputs by Ja Eun Lee and Janette Murawski, both of whom played an important role in supporting the development of the AI Assistant and helped refine the article by providing valuable feedback. In addition, generative AI (ChatGPT, OpenAI, version 5.2) was used to support language editing, improve clarity, and enhance the overall flow of the text. The use of AI was limited to stylistic and editorial assistance; all substantive judgments, interpretations, and conclusions remain the responsibility of the author. This approach is consistent with emerging good practices on the transparent use of AI tools in research and evaluation, which recommend disclosure of AI support while ensuring full human oversight and accountability.

▶ Contents

1. From knowledge management to *i*-eval Discovery

The ILO Evaluation Office's (EVAL), flagship database, i-eval Discovery is not an isolated innovation, but the result of a long trajectory of institutional learning in evaluation and knowledge management within the ILO. Its origins can be traced back to the early 2000s, when a comprehensive system, known as i-track, was developed under the International Programme on the Elimination of Child Labour (IPEC), at the time one of the ILO's largest and most operationally complex programmes, with extensive requirements for monitoring, reporting, and evaluation. To manage this scale, IPEC invested in what was then a highly advanced knowledge management system, enabling the systematic capture and organization of evaluation findings – including lessons learned, good practices and recommendations - across countries, regions and interventions.

When IPEC discontinued its use of the system, EVAL recognized its strategic value and adopted it in 2010, transforming it in 2016 into what became i-eval Discovery. This transformation with support from the Information and Technology Management Department (INFOTEC) was significant. The platform evolved into a sophisticated, user-oriented repository, incorporating visual mapping features and intuitive navigation that allowed evaluation reports and related documentation to be explored by country, region, themes, and other areas of interest. It enabled users to identify good practices, extract lessons learned, and review management responses to evaluation recommendations (the latter being integrated from EVAL's Automated Management Response System (AMRS) that was developed in 2018. In doing so, i-eval Discovery moved beyond document storage towards structured and curated knowledge management.

The strength and quality of this repository proved decisive. It enabled the Evaluation Office—despite being a relatively small team—to operate with a level of knowledge management capacity that was, in many respects, ahead of its time. Without sustained investment over time in structured, curated, and quality-controlled evaluation evidence, the development of the i-eval AI Assistant would simply not have been possible.

At the same time, this experience highlights a broader institutional issue. As a knowledge-centred organization, the ILO produces a substantial volume of evidence, yet the systematic use of this knowledge has not always kept pace with its production. Lessons learned from evaluations as well as research findings represent a critical form of institutional capital, directly relevant to improving policy and programme design, but their potential is not always fully realized. The challenge has never been the absence of knowledge, but rather the ability to access, connect, and apply it in a timely and meaningful way.

> Lesson 1: Knowledge management is not a support function, but a strategic capability.

2. From repository to use

Up to 2025, despite these advances, a fundamental challenge remained: the existence of a well-structured repository does not automatically translate into effective use - evaluation knowledge requires syntheses. Its value lies not in individual reports, but in the ability to draw as much evaluative evidence and insight as possible across multiple evaluations, contexts, and timeframes.

In practice, this process remains resource intensive. Users must identify relevant reports, extract information, and construct their own synthesis, often under time pressure. As a result, a persistent gap remains between the production of evaluation knowledge and its effective use.

Even the most advanced repository remains limited if users cannot efficiently transform information into insight. It was precisely this gap that the ILO Evaluation Strategy 2023–2026, developed in 2022, sought to address by positioning the introduction of artificial intelligence (AI) as a strategic indicator of performance. The objective was to move beyond retrieval towards immediate synthesis, enabling users to interact with evaluation knowledge in a more direct and timely way that is tailored to their needs – thus meeting the evaluation strategy's end target of 2025.

This is where the i-eval AI Assistant introduces a qualitative shift. By enabling users to pose natural-language questions and receive structured, evidence-based syntheses, the system significantly lowers the barrier to using evaluation knowledge. Instead of navigating multiple reports and manually extracting findings, users can obtain a concise synthesis—typically in the form of a two-to-three-page analytical response—drawing on multiple source-cited evaluations across the repository.

This functionality directly addresses one of the most persistent constraints in evaluation use: time. Tasks that would traditionally require several days of reviewing and comparing reports can now be completed in a matter of minutes. At the same time, the requirement that all outputs be source-cited and accompanied by a bibliography ensures that this increase in efficiency does not come at the expense of credibility. Users are able not only to access synthesized insights, but also to verify and trace them back to the original evidence. We also used system prompting to ensure that if a question is not about evaluation reports, the bot politely declines to answer. These strict guardrails ensured the system (and cost burden of tokens) would not be exploited by asking non-evaluation queries.

> Lesson 2: Access is not use — functional ease, timing, and opportunity determine uptake

3. From repository to AI assistant

It is important to distinguish between i-eval Discovery and the i-eval AI Assistant which was launched in December 2025. i-eval Discovery is the repository and knowledge platform, providing structured access to more than 1700 evaluation reports and a broader body of evaluation-related material, including lessons learned, good practices, and management responses—amounting in total to more than 5000 documents in PDF format. The i-eval AI Assistant is the latest development built on top of this foundation and is embedded within i-eval Discovery as a dedicated tab on the dashboard.

The i-eval AI Assistant enables users to pose natural-language questions and receive structured syntheses of evaluation evidence, typically in the form of a concise two-page-three-page response. These syntheses draw across multiple reports simultaneously, allowing users to move from individual findings to patterns and lessons emerging across the evaluation portfolio. These outputs are systematically source-cited and accompanied by a bibliography, ensuring that all statements can be traced back to the underlying evaluation reports.

> Lesson 3: In evaluation, an answer without a source is not evidence.

4. Co-developing *i*-eval Discovery and the *i*-eval AI Assistant

The development of the i-eval AI Assistant within i-eval Discovery was undertaken in close collaboration with INFOTEC and a dedicated IT team responsible for building the system's infrastructure and was characterized by a highly iterative and engaged process. This was not a situation in which requirements were defined once and handed over for implementation. Instead, the EVAL played a continuous and central role in shaping the system.

> Lesson 4: AI development in evaluation must be anchored in user experience and measurable value, not IT expertise alone.

In the early stages, considerable effort was invested in articulating needs, translating evaluation practice into functional specifications, and developing a clear case for investment. This included preparing a detailed analysis of expected efficiency gains, scoping the application according to target users, and assessing the return on investment.

Once the project was approved, the work shifted into an intensive phase of co-development. The process was structured around frequent—often daily—15-minute huddles over a period of six months with the IT team, allowing for rapid feedback and continuous refinement. Through iterative testing, the Evaluation Office was able to identify where technically correct responses did not align with evaluation logic.

> Lesson 5: System quality cannot be assessed by technical metrics alone; it must be validated against how target groups interpret and use evidence.

5. When technical relevance is not enough – designing for diversity

Early outputs from the system were technically strong. The i-eval AI Assistant retrieved relevant information and generated coherent responses. However, from EVAL's perspective, these outputs revealed a critical limitation: they were often analytically narrow.

The system tended to draw heavily from a single report, retrieving multiple highly relevant excerpts from the same source. This revealed what can be described as the "dominant document" problem. While technically correct, the system was effectively amplifying one source rather than synthesizing across many. The outputs appeared robust, but were in fact narrow, reflecting repetition rather than synthesis and allowing a single report to disproportionately shape the entire answer. This directly contradicts evaluation practice, which relies on comparing and integrating evidence across multiple sources.

From a technical perspective, this behaviour is inherent to retrieval-augmented generation (RAG) systems. These systems retrieve semantically similar texts ("chunks") from a knowledge base and generate responses based on that retrieved content. Because similarity scoring favours closely matching language, the retrieval layer tends to select multiple chunks from the same document when it aligns closely with the query. From an IT perspective, this represents optimal performance. From an evaluation perspective, however, it results in a lack of diversity in evidence.

> Lesson 6: Technical relevance (similarity) does not equal evaluative quality (synthesis).

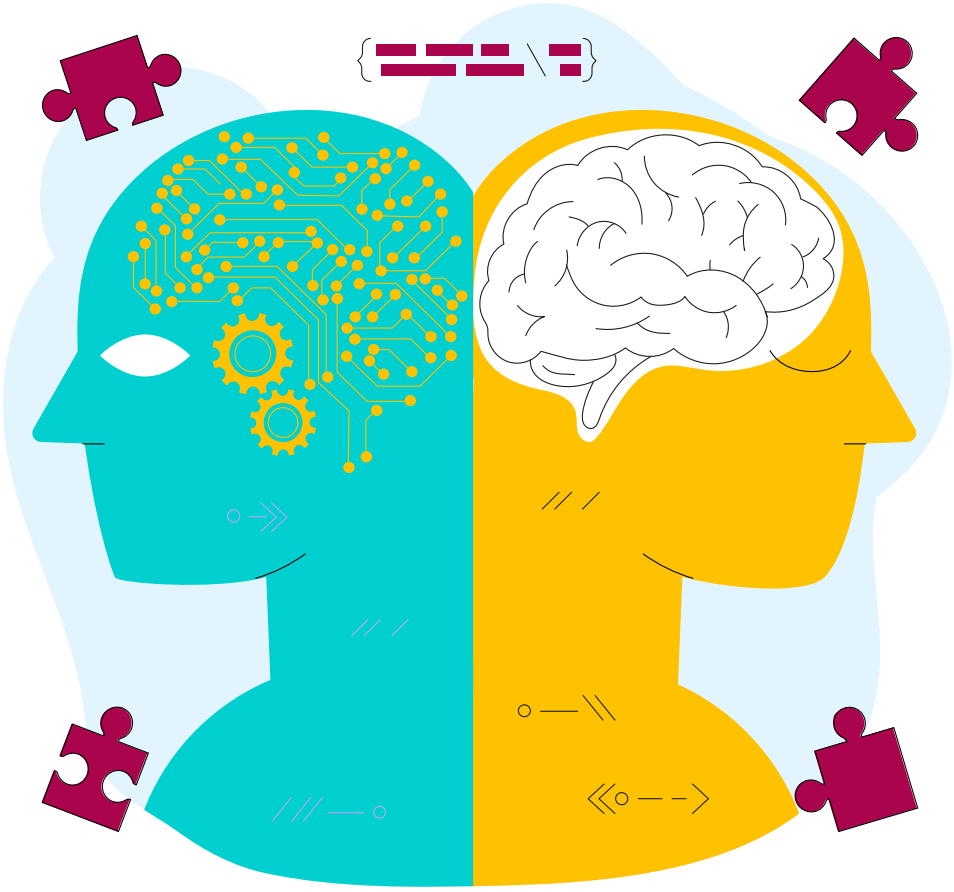
The lack of diversity in retrieved evidence therefore emerged as a critical challenge. When left to operate purely on semantic similarity, the system consistently returned clusters of highly similar chunks from a single report. While internally consistent, these outputs were not representative of the broader evidence base.

This is where the interaction between IT expertise and evaluation substance expertise proved decisive. The IT team provided the technical foundation, ensuring that the system performed efficiently and retrieved relevant information. EVAL, however, assessed outputs through a different lens—focusing on credibility, balance, and representativeness of evidence. What appeared technically optimal was not necessarily analytically sound.

> Lesson 7: Diversity in evidence does not emerge automatically in AI bots—it must be deliberately designed. In evaluation, diversity of evidence is more valuable than depth from a single source.

In response, the EVAL introduced a decisive request: a strict limit of a maximum of five content chunks per report. This was not a technical adjustment proposed by developers, but a requirement grounded in evaluation logic. By using a "5 chunk rule" limiting the contribution of any single report, the system was forced to engage with a broader set of sources across the repository.

The impact of this intervention was immediate and significant. Outputs became less repetitive, more balanced, and more reflective of the diversity of evaluation evidence. What had previously been extended summarization began to approximate synthesis.



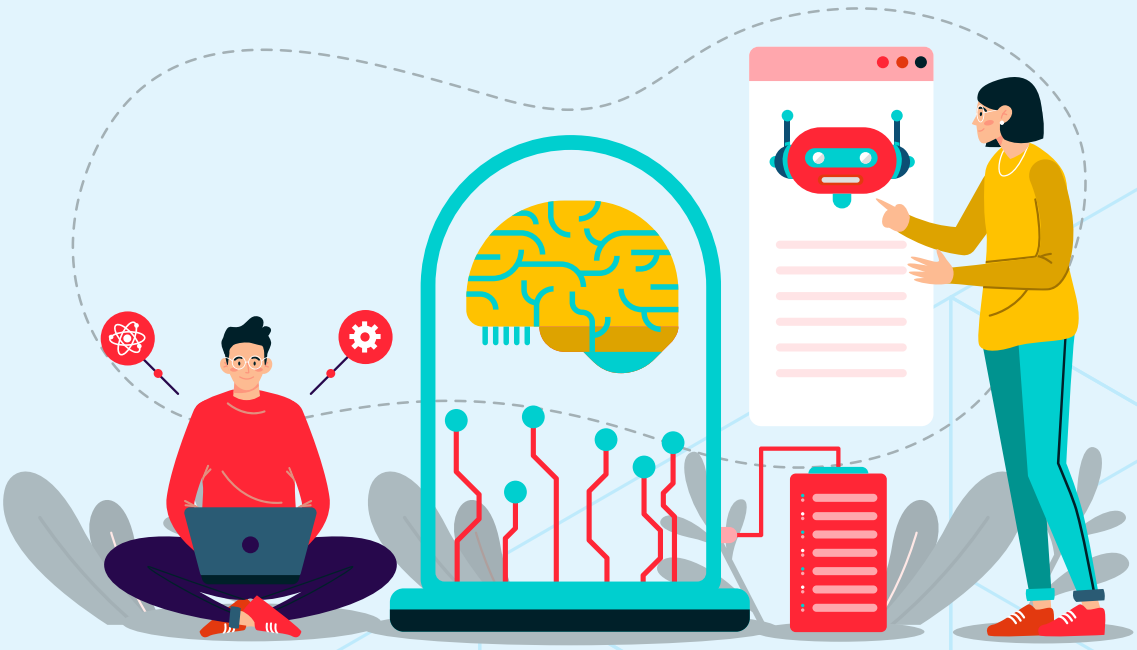
6. Managing expectations and use

As the i-eval AI Assistant moves into broader use, managing expectations becomes essential. AI systems often create an expectation of immediate, comprehensive, and definitive answers. In practice, however, the Assistant provides structured syntheses of available evidence—typically in the form of concise two-page analytical responses—rather than final judgments.

This distinction is critical. The system accelerates access to evaluation knowledge by reducing the time required to retrieve, compare, and synthesize evidence across multiple reports. Tasks that would traditionally take several days can now be completed in a matter of minutes. At the same time, the requirement that all outputs be grounded in the repository, fully source-cited, and accompanied by a bibliography ensures that this increase in efficiency does not come at the expense of credibility. Users are able not only to access synthesized insights, but also to verify and interrogate the underlying evidence.

The design of the AI Assistant—particularly the emphasis on grounding and diversity of sources—also shapes how users engage with it. By drawing on multiple reports rather than relying on a single source, the system reinforces a more balanced understanding of evaluation findings. However, this also requires users to interpret outputs appropriately. The Assistant does not replace evaluative reasoning; it supports it by providing structured entry points into the evidence base.

> Lesson 8: AI supports evaluation use—it does not replace it.



7. From repository to intelligence

i-eval Discovery represents the platform that makes this shift possible, while the i-eval AI Assistant represents the realization of that shift—from storing knowledge to actively mobilizing it. The transition from repository to intelligence is not simply a technological step, but the result of a long-term institutional process.

> Lesson 9: AI becomes effective only when embedded in strong, continuously evolving knowledge systems.

The experience demonstrates that the effectiveness of the AI Assistant is directly linked to the strength of the underlying knowledge system. The repository is not static as many sources from newly designed AI bots tend to be but is dynamic as it is continuously updated and curated over time, providing the structured and quality-controlled evidence base on which the system operates. Without this foundation, the ability to generate credible, up-to-date, grounded, and diverse syntheses would not exist.

At the same time, the development process highlights the importance of integration—between systems, between functions, and between expertise. The evolution from i-track to i-eval Discovery, and now with an embedded AI Assistant, reflects a sustained effort to connect knowledge management, evaluation practice, and technological innovation. This continuity has allowed the Evaluation Office to move ahead of its time, despite operating with limited financial and human resources.

The introduction of AI therefore does not represent a departure from previous efforts, but a continuation of them. It builds on years of investment in structuring evaluation knowledge and extends this work by enabling real-time synthesis and interaction with the evidence base. In doing so, it addresses the long-standing challenge of moving from access to use.

8. Concluding reflections: An ongoing journey

The development of the i-eval AI Assistant, embedded in i-eval Discovery, demonstrates how the ILO Evaluation Office has moved beyond managing evaluation knowledge towards actively mobilizing it. The experience has generated several important lessons. Knowledge management must be treated as a strategic capability; access alone does not guarantee use; source-cited synthesis is essential for credibility and trust; information technology expertise and evaluation expertise must be brought together; and diversity of evaluative evidence must be deliberately embedded into AI systems if outputs are to support meaningful evaluation use.

At the same time, this remains an ongoing journey. The i-eval AI Assistant is now available to more than 3,000 ILO staff members in headquarters and the field, as well as consultants with ILO guest accounts. Yet this is only the beginning of a broader ambition. The Assistant is at the centre of EVAL's knowledge management strategy to make evaluative evidence more accessible, usable and actionable across the ILO.

EVAL's broader vision is to transform i-eval Discovery into a comprehensive one-stop shop for evaluation knowledge, bringing together evaluation evidence, policies, guidance materials, tools, workflows, independent evaluation quality assessments, and project performance cards within a single platform. By making these resources AI-accessible through a unified interface, users would be able to seamlessly navigate between evidence, guidance, and performance information, significantly strengthening the use of evaluation for learning, accountability, decision-making, strategic planning and project design.

Looking ahead, a major milestone has already been achieved with senior management's approval in principle to finance the extension of the i-eval AI Assistant beyond the Office to ILO constituents and external users. This represents a potentially transformative step towards democratizing access to evaluative evidence and unlocking the full value of more than two decades of accumulated evaluation knowledge. The planned public rollout would substantially expand the reach, visibility and use of evaluation findings while positioning the ILO at the forefront of AI-enabled knowledge management and evidence-informed decision-making within the multilateral system.

The experience also has relevance beyond the ILO. Through exchanges with colleagues in the United Nations Evaluation Group (UNEG), it became clear that many organizations are grappling with similar challenges in moving from evaluation production to effective knowledge use. While the development of the i-eval AI Assistant was undertaken with relatively modest resources compared with larger initiatives, the Evaluation Office was agile enough to move early and deliver a functioning solution. This experience has both benefited from and contributed to broader system-wide reflections on how artificial intelligence can be aligned with evaluation principles, particularly credibility, traceability, transparency and diversity of evidence.

The potential utility of the AI Assistant extends well beyond supporting project design. By enabling rapid synthesis of evidence across themes, countries, regions and areas of ILO activity, the system can support speeches, briefings, policy discussions, strategic planning, programme development and decision-making. Its value lies in making accumulated evaluation knowledge available precisely when it can influence action.

The final lesson is therefore institutional as much as technical: artificial intelligence becomes useful when it strengthens the use of existing knowledge. For the ILO, the i-eval AI Assistant provides a practical mechanism for bringing evaluation evidence closer to decision-making while preserving the principles that make evaluation credible—traceability, balance, diversity of evidence and professional judgement. More broadly, this experience demonstrates that meaningful innovation in evaluation depends not only on the size or budget of an Evaluation Office, but on sustained commitment to knowledge, clarity of purpose, systems thinking, and the ability to translate evaluation principles into practical, accessible and usable tools.

ilo.org

International Labour Organization
Route des Morillons 4
1211 Geneva 22
Switzerland