AFRICAN DEVELOPMENT BANK GROUP

IDEV
Independent Development Evaluation
African Development Bank

gLOCAL EVALUATION WEEK 2026: IDEV SESSION

Leveraging AI Across the Evaluation Cycle: Insights from MDB Evaluations

KEY QUESTION: "How can MDB evaluation functions integrate AI into their workflows responsibly, and what methodological, ethical, and institutional safeguards make the difference?

Event overview

The Independent Development Evaluation (IDEV) function of the African Development Bank Group (AfDB or "the Bank Group") convened a session on 3 June 2026 as part of gLOCAL Evaluation Week 2026 to share practical lessons from integrating Artificial Intelligence (AI) into evaluation workflows and to examine methodological, ethical, and institutional safeguards required for responsible AI adoption. The session brought together evaluation professionals from independent evaluation functions of the AfDB, International Fund for Agricultural Development (IFAD) and the Islamic Development Bank (IsDB) to advance collective knowledge on how AI can accelerate evidence generation while preserving the rigour and credibility.

The session opened with remarks from Karen Rot-Münstermann, Evaluator General of the AfDB, who framed the significant opportunities as well as the concrete risks associated with the adoption of AI in evaluation. This was followed by a presentation from Daouda Drabo, Data Scientist at IDEV, who highlighted practical applications of AI in evaluation syntheses and thematic evaluations, drawing on IDEV's real-world experience. The panel discussion, moderated by Prof. Howard White, Professor of Evidence-Based Policy at the University of Lanzhou and Senior Research Associate at the Research and Evaluation Centre (REC), featured Andrew Angunko, Chief Quality and Methods Advisor at IDEV/AfDB, Hansdeep Khaira, Senior Evaluation Officer at IFAD-IOE and Oguz Ceylan, Lead Evaluation Specialist at IEvD/IsDB.

The discussion examined how AI tools, particularly generative AI, are being integrated into evaluation workflows across MDB evaluation functions, the efficiency gains they generate, and the safeguards required to prevent hallucinations, bias, and traceability gaps. Together, the presentations and panel discussion underscored that AI holds real potential to accelerate evidence generation, but only when deployed with structured human oversight, robust quality assurance, and transparent documentation with evaluator judgement always at the centre.

Key messages

1. **AI is a powerful tool for evaluation, but a complement, never a replacement.** AI can accelerate document screening, evidence extraction, thematic clustering, and synthesis, increasing efficiency and broadening scope. However, AI cannot make evaluative judgements, assign performance ratings, or frame conclusions and recommendations as humans can. These remain the exclusive responsibility of evaluators. Over-reliance on automated outputs erodes critical judgement and compromises the credibility of findings.
2. **Traceability and human validation are non-negotiable safeguards.** Every AI-assisted evidence extraction must include the full metadata document title, year, page, section, evaluation questions, and theme to enable verification and ensure accountability. Human

validation must operate at two levels: confirm that extracted evidence is present in the source document, and assess its relevance to the evaluation question. Omissions identified during validation must be supplemented through manual review.

3. **Data quality and prompt quality determine output quality.** The principle of "garbage in, garbage out" applies with full force to AI-assisted evaluation. If the input data are not of high quality, or the prompts are poorly designed, the outputs will be unreliable. Pre-testing prompts on a sample of documents before applying them to the full corpus is a critical quality assurance step. Clean, well-structured, and non-confidential source data is a prerequisite for responsible AI use.
4. **Institutional governance frameworks and cross-institutional learning are essential.** Data protection, confidentiality, transparency, and accountability requirements must be embedded in formal institutional AI strategies before AI use is scaled. MDB evaluation functions must promote cross-institutional learning. Networks such as the Evaluation Cooperation Group (ECG) should be used deliberately to share use cases, safeguards, and emerging standards across institutions.

Selected insights from the discussions

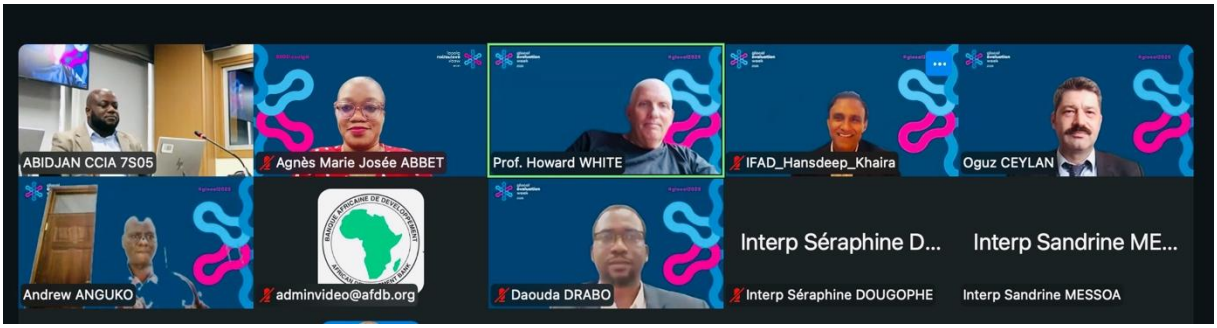
- **IDEV's six-stage AI workflow for evaluation syntheses and thematic evaluations.** IDEV has developed a structured workflow for AI-assisted evidence extraction comprising six stages: (1) anchoring all AI use in evaluation questions from the inception report; (2) parallel manual and AI-assisted document collection feeding a consolidated database; (3) human validation of the document base; (4) evidence extraction using detailed, multi-page prompts with metadata; (5) human validation of extracted evidence to verify relevance and identify omissions; and (6) AI-assisted evidence analysis to identify recurring themes, enabling and hindering factors, and cross-cutting patterns. Microsoft Copilot is the Bank-approved tool currently in use.
- **IFAD-IOE: two-year AI strategy and a complex evaluation as an incubator**. IFAD's Independent Office of Evaluation (IOE) developed a two-year AI strategy in 2024. The corporate-level evaluation of IFAD's replenishment commitments, spanning seven to eight years of evidence, served as an incubator for AI, supporting transcription analysis, thematic evidence extraction, document classification, and organisation of material around sub-evaluation questions.
- **IsDB-IEvD: from prompts to agents and a knowledge extraction tool**. The Islamic Development Bank's Independent Evaluation Department is making a strategic shift from individual prompts to standardized prompts, and increasingly to AI agents, standardised automated processes that improve consistency, reduce variability, and align outputs with institutional standards.
- **Lack of coordination between institutions**. A critical observation raised by the moderator and endorsed by all panellists is that MDB evaluation functions are conducting AI experiments largely in parallel, without sufficient structured cross-learning. Institutions are independently finding similar solutions to similar problems. The ECG represents an existing mechanism for cross-institutional coordination that should be used more actively to document shared standards, exchange use cases, and develop common governance frameworks.
- **The AI-generated completion report risk.** The use of AI to auto-generate project completion reports from project documents and supervision reports should serve as an evidence-mapping tool within this process, summarising supervision reports, identifying recurring issues, and organizing material, rather than as an automatic report writer. All claims must be verified against source documents, and an audit trail of documents used, prompts applied, outputs generated, and judgements made must be maintained.

Strategic implications for independent evaluation functions

- Invest in structured staff capacity development on AI tools, prompt engineering, human validation protocols, and ethical AI use, ensuring that evaluators have the skills to deploy AI responsibly and to interrogate its outputs critically.
- Activate a formal cross-institutional AI learning exchange through the ECG Working groups to share use cases, safeguards, and emerging governance standards moving from parallel experimentation to collective knowledge building.
- Establish mandatory traceability standards for all AI-assisted evaluation work, requiring full metadata documentation (document, year, page, section, evaluation question, theme) for every AI-extracted evidence item.

Conclusion

The session reinforced a defining principle of AI adoption in independent evaluation: AI is most valuable when used responsibly, with human judgement always at the center. The practical lessons shared by IDEV, IFAD-IOE, and IsDB-IEvD demonstrate that AI can genuinely accelerate evidence generation, reducing weeks of manual review, enabling larger evidence bases, and supporting more systematic synthesis, but only when deployed with structured workflows, rigorous human validation, full traceability, and clear institutional governance. As AI tools continue to evolve rapidly, MDB evaluation functions have both the opportunity and the responsibility to build collective knowledge, share standards, and ensure that the rigour and independence that evaluation depends on are strengthened, not compromised, by the age of AI.



From left to right:

First row: Damase Sossou (Principal Evaluation Capacity Development, IDEV/AfDB), Marie Josee Abbet (Evaluation Capacity Development Consultant, IDEV/AfDB), Prof. Howard White (Moderator, Professor of Evidence-Based Policy, University of Lanzhou), Hansdeep Khaira (Senior Evaluation Officer at IFAD-IOE), Oguz Ceylan (Lead Evaluation Specialist, IsDB-IEvD).

Second row: Andrew Anguko (Chief Quality and Methods Advisor, IDEV/AfDB and Daouda Drabo (Data Scientist, IDEV/AfDB).