Navigating AI & Digitalization in M&E

A Three-Lens Framework for M&E and Evaluation Professionals and Institutions

25 Feb 2026
Working Draft

Acknowledgements

This white paper was conceptualized and prepared by Douglas Glandon, with the support and encouragement of Jos Vaessen (Head of the Global Evaluation Initiative). Valuable input and feedback were provided by the following colleagues (in alphabetical order): Alice Macfarlan, Andres Ignacio Gonzalez Flores, Claudia Paola Olavarría Manríquez, Ketevan Nozadze, Patricia Rogers and Patrizia Cocca. Alice Macfarlan and Patricia Rogers also provided substantive input to the development and refinement of the GEI frameworks behind the 'evidence needs' and 'workflow' lenses. The author used a generative AI tool (Anthropic Claude) to assist with language refinement and editorial clarity. All ideas, analysis, and conclusions are solely those of the author, and any errors remain the author's own.

Disclaimer

1. Introduction

Artificial intelligence and digitalization are reshaping how evidence is generated, analyzed, and used. Across the monitoring and evaluation (M&E) profession, practitioners encounter a growing volume of AI-related tools, guidance, training, and organizational policies. Some of this is useful. But the sheer volume and diversity of the landscape – spanning technologies from satellite imagery analysis to large language models, and audiences from individual practitioners to those responsible for strengthening national M&E systems – has made it difficult for professionals and institutions to determine what is relevant, reliable, and appropriate for their specific circumstances.

This paper offers a framework to help M&E and evaluation professionals and institutions navigate this complexity. It proposes a way of thinking – a set of complementary perspectives that help practitioners and institutions ask better questions about if, where, how, and under what conditions AI and digital tools may have a role in their work. The framework is organized around three lenses – evidence needs, workflow, and capability – each capturing a distinct relationship between practitioners and technology. Applied together, these lenses support more informed, context-sensitive decisions about AI in M&E.

Throughout this paper, "AI" is used as shorthand for the broader nexus of artificial intelligence and digitalization. These are deeply interconnected: most AI applications depend on digital data infrastructure, and many digital innovations in M&E involve AI components. Similarly, "evaluation" is used broadly to encompass the full range of evaluative activities – from needs assessments and design studies through implementation monitoring, impact evaluation, and evidence synthesis. The framework applies across this spectrum, recognizing that the "M" and "E" of M&E draw on overlapping but distinct professional activities and evidence needs. In this paper, we use the term 'AI tool' to refer to a specific technology (e.g., a large language model, a computer vision model), while 'AI-enabled approach' refers to the use of such a tool for a particular evaluative purpose (e.g., using natural language processing for thematic analysis of stakeholder feedback).

2. Reframing the discourse

The field of M&E has responded to the rise of AI with energy and creativity – issuing policies, convening discussions, experimenting with new tools, and developing practitioner-oriented guidance. This paper seeks to build on that foundation by addressing a dimension that existing contributions have tended to treat separately: the integration of evidence, workflow, and capability considerations in a way that supports practical decision-making. The following subsections describe three features of the current discourse that this framework is designed to complement and extend.

2.1 AI treated as a monolithic technology

A natural starting point for many organizations has been to develop policies, convene discussions, and offer training on 'AI' as a category. This is a reasonable response to a rapidly evolving landscape. But it also tends to obscure critical differences. Computer vision applied to satellite imagery, a predictive machine learning model for program targeting, and a text classification using a large language model (LLM) for qualitative coding differ in the data they require, the skills needed, the ways they can go wrong, the ethical considerations they raise, and the evaluative questions they can address. Developing undifferentiated guidance on "the strengths and limitations of AI" is roughly as useful as guidance on "the strengths and limitations of research methods" without distinguishing between randomized controlled trials and participatory action research.

Furthermore, discussions about "AI" in the M&E field are often shaped by recent rise and use of LLMs, which are currently the most visible form of the technology. As a result, the term is sometimes used as shorthand for this particular — and in several respects distinctive — subset of AI applications. Similarly, organizational AI policies that apply a single set of rules across all applications can face a difficult balance: guidance may become either very general, limiting its practical usefulness, or closely tailored to prominent use cases while leaving other parts of the landscape less clearly addressed.

2.2 Tool-first rather than need-first orientation

Much AI-related content in the M&E space adopts a tool-first orientation: beginning with an AI technique and illustrating potential applications. This is useful for practitioners who already know they want to use a particular technology. However, it is often less well suited to more common situations: a practitioner facing a specific evidence need who wants to understand whether an AI-enabled approach might help address it; someone facing time, budget or bandwidth constraints in their evaluative work who wants to explore whether AI could help; or a professional or institution considering what competencies and capabilities are required. This is not a challenge unique to AI – good evaluation practice has always started from the question rather than the method – but the pace and visibility of AI adoption have made the tool-first pull particularly strong. Such a tool-first orientation can introduce a framing effect: by starting with the technology and working toward the application, it implicitly positions AI as a solution. A need-first orientation instead begins with the evaluative purpose and considers AI as one of several possible means. For example, a practitioner seeking to understand patterns in a large volume of stakeholder feedback might consider automated thematic clustering, using a natural language processing (NLP) model, but a need-first approach would also surface alternatives – such as stratified sampling with systematic human coding, or structured categorization against a pre-defined analytical framework – and the choice among them depends on the nature and quality of the data available, the type of insight required, the cost and time requirements of different approaches, and the analytical capacity of the team.

2.3 Fragmented contributions

Existing contributions to the literature and body of guidance materials often focus on particular dimensions of the relationship between AI and M&E or evaluation: applications mapped to evaluation phases (Branchini & Vallina Acha, 2025; Jacob, 2025) or to MERL activities and data types (Bruce, Gandhi & Vandelanotte, 2020), competency implications (Mason, 2023), evaluation industry dynamics (Nielsen, 2023), ethical dimensions (Reid, 2023), responsible practice guidelines (UK Evaluation Society, 2025), or others. Each provides valuable insight and advances our understanding in its respective area.

In practice, however, these dimensions tend to come together. A practitioner deciding whether to use NLP for qualitative analysis is at the same time considering evidence needs (what specifically can be learned and will the findings be credible?), workflow (how does this fit my process and timeline?), and capability (does my team have the skills to apply it appropriately?). These dimensions often interact, with choices in one area shaping possibilities in the others. A framework that helps practitioners consider them jointly can draw from existing contributions and support more integrated decision-making, rather than approaching each dimension in isolation.

3. Three Lenses: A Framework

3.1 The three lenses

The framework is organized around three lenses – evidence needs, workflow, and capability – each capturing a distinct relationship between people and technology. M&E and evaluation professionals and institutions typically approach AI questions from one of these three starting positions:

Evidence needs: *"What kinds of evidence do we need, and which types of AI-enabled approaches might help us generate each?"*

This lens is epistemic. The practitioner has identified one or more evidence needs and wants to know whether specific AI-enabled approaches might expand what is possible – addressing questions that were previously infeasible, or enabling richer, more timely, or more granular evidence than conventional approaches allow. For example, a practitioner working on an evaluation of a rural electrification program might be interested in predictive modeling (a subset of machine learning) using satellite-derived nighttime light data because of the potential to measure changes in economic activity at a geographic resolution and temporal frequency that household surveys alone cannot provide.

Workflow: *"What are the constraints in our evaluative work processes, and could AI-enabled approaches help address them?"*

This lens is operational. The practitioner faces constraints – in data access, processing capacity, analytical bandwidth, or communication reach – and wants to know whether a specific type of AI-enabled approach can address them. This includes efficiency gains but extends to accessing new types of data and expanding what can be done with evaluation products. For example, an evaluation team managing a large portfolio of

country evaluations might ask whether AI-assisted translation and summarization (using a large language model, for example) could enable them to produce tailored briefs for different national stakeholders from a single evidence base.

Capability: *"What do we need to know and have in place to work responsibly in an AI-influenced environment?"*

This lens is developmental. The practitioner or institution wants to know what skills, knowledge, ethical frameworks, and institutional capabilities are needed to remain competent, credible, and responsible. For example, a national M&E directorate might ask what governance policies and staff competencies are needed before any AI tools are introduced into their evaluation processes.

These three lenses capture genuinely distinct relationships between practitioners and technology. The evidence needs lens treats AI as potentially expanding what is epistemically possible. The workflow lens treats AI as a means of operational enhancement. The capability lens treats AI as a professional and institutional development challenge.

Each lens is organized using an existing GEI framework that provides its internal structure. The evidence needs lens draws on the GEI policy/program cycle, which organizes evidence needs by the stages at which they inform policy and program decisions. The workflow lens draws on the GEI task framework for evaluative activities, which organizes evaluative work into task clusters. The capability lens draws on the GEI Evaluation Competency Framework, which organizes professional competencies by domain. These frameworks provide the categories in the detailed mapping tables presented in Sections 4, 5, and 6.

Notably, this framework does not include a lens organized around a taxonomy of AI technologies (e.g., "uses of predictive machine learning models," "uses of NLP," "uses of computer vision"). This is intentional. A technology-taxonomy lens starts from the tool and works toward applications – precisely the tool-first orientation this framework is designed to move beyond. The three lenses are practitioner-centered: each starts from where practitioners are – what they need to know, what they need to do, what they need to become.

This framework also does not have a separate lens for ethical and responsible practice; rather it is a dimension that runs through all three, manifesting differently in each. Section 8 details how this works in practice.

3.2 How the lenses relate

The lenses are complementary perspectives that overlap in practice. A person seeking guidance on how to apply NLP to stakeholder feedback analysis may be primarily focused on gaining new analytical insight into patterns across a large dataset. But incorporating NLP is likely to add steps to their workflow and requires skills in prompt design, output validation, and qualitative methodology – drawing on the other two lenses. Similarly, a person seeking a training module on prompt engineering may be primarily interested in building their capabilities, but the most relevant prompting approaches will vary depending on what evidence they are trying to generate and how an AI tool fits into their specific workflow.

Most substantive decisions about AI in M&E draw on two or all three lenses. The framework's value lies in helping practitioners recognize which considerations are in play and attend to each deliberately. It follows that the same AI-enabled approach will frequently appear in discussions of more than one lens – and should. In this way, the lenses are not a sorting mechanism that assigns each AI tool to a single category, but rather a set of complementary questions to ask about any given use of AI.

A PRACTICAL NOTE ON USING THE FRAMEWORK EFFECTIVELY

The three-lens framework is designed to be accessible to all evaluation professionals. It is most powerful, however, when those working on it bring, or can draw on, a working knowledge of AI techniques, including their inner workings, limitations, and what good implementation looks like in practice.

To ensure that we arrive at the right conclusions when using the framework — correctly identifying whether AI techniques can add value to an evaluation and accurately assessing any implications for workflow or team competencies — we recommend that application of the framework draws on a mix of technical and evaluative expertise. This means ensuring the right combination of people is involved not only during implementation, but when making decisions of whether to use AI techniques or not.

4. Lens 1: Evidence Needs

4.1 The core idea

A practitioner's choice of approaches, methods, and tools – including AI – should be driven by the specific evidence needs of the task at hand. At its most basic, this means starting from the question: what do we need to know, and what are the most appropriate ways to find out? AI enters the picture only when a specific application offers a credible way to generate evidence that is more valid, more timely, more granular, or more feasible than what conventional approaches can provide.

This lens draws on the GEI policy/program cycle framework, which provides a structured way to organize the broad landscape of evaluative approaches according to the stages at which evidence informs policy and program decisions. The framework identifies stages in the policy and program cycle, the central questions that arise at each stage, and established approaches to answering those questions ("ways of answering" in the GEI framework). AI applications are situated within these ways of answering: not as replacements but as tools that may support, augment, or extend them.

4.2 The policy/program cycle

The policy/program cycle framework serves as a boundary object between the producers and users of evaluative evidence. That is, it provides shared language and structure that helps M&E professionals and decision-makers develop a common understanding of the fit between evidence needs and approaches to address them – even when they bring different expertise and perspectives. This is important because it grounds the conversation about AI in M&E within the broader conversation about evidence-informed decision-making, rather than treating AI as a standalone topic.

The framework also conveys the broad array of evaluative approaches that can strengthen evidence-informed decision-making – not only ex-post, summative evaluations, but also needs assessments, design studies, implementation monitoring, and evidence synthesis, among others. This breadth matters for AI: different types of AI applications are relevant to different types of evaluative work, and a framework that encompasses this range is more useful than one focused narrowly on formal evaluations.

The GEI policy/program cycle identifies five stages, each with central questions and established ways of answering. Table 1 provides a high-level summary with an illustrative AI application for each stage; a detailed mapping of all central questions, ways of answering, and illustrative AI applications is provided in Appendix A.

Table 1: Evidence needs across the policy/program cycle and illustrative AI applications

Stage	Central questions (illustrative)	Illustrative AI/digital applications
1. Understand the situation	What is happening? Why? What should be prioritized?	Machine learning (ML) for pattern detection in large administrative datasets as part of a situation analysis, enabling identification of geographic clusters to target for an intervention that manual review of the same data would be unlikely to surface.
2. Explore options	What are the options? What effects are anticipated? How do they compare?	ML-based document screening and relevance classification for accelerating an evidence synthesis process, reducing the volume of literature requiring human review while maintaining methodological rigor through human oversight of inclusion decisions and quality assessment.
3. Design the intervention	How is it expected to work? What is needed for implementation? How will we know if it works?	ML-based semantic retrieval (matching documents by conceptual relevance) across program documentation to inform intervention logic, identifying relevant precedents across a larger evidence base than manual searching would practically cover.

Stage	Central questions (illustrative)	Illustrative AI/digital applications
4. Support implementation	Is it being delivered as planned? What results are emerging? How should it adapt?	Automated anomaly detection using an ML model for real-time program monitoring data (e.g., health management information systems), flagging unusual service delivery patterns across districts weeks before they would surface in routine quarterly reporting.
5. Assess results and implications	What difference did it make? How and why did it work (or not)? What are the implications?	Causal ML methods (e.g., causal forests) applied within randomized controlled trials to identify how treatment effects vary across subpopulations – for example, revealing that a cash transfer program had substantially larger effects on school enrollment for girls in rural areas than for other groups, enabling more targeted program design.

4.3 Key considerations when applying this lens

The potential for AI to contribute to M&E is not evenly distributed across all types of evidence needs. Some questions and ways of answering are inherently more amenable to AI applications than others, and this distinction is partly epistemic in nature. Ways of answering that involve large-scale pattern recognition, data integration across heterogeneous sources, or systematic scanning of large evidence bases are well-suited to current AI capabilities. For example, predictive machine learning models can detect anomalies in program monitoring data that may be invisible to human analysts reviewing the same datasets, and NLP techniques (e.g. text classification models) can accelerate evidence synthesis by processing volumes of literature beyond what a review team could feasibly read.

In contrast, ways of answering that are fundamentally normative or deliberative – such as convening stakeholder judgment on priorities, weighing competing values, or interpreting findings in light of context-specific political and cultural considerations – involve a different kind of knowing. These are not simply more complex empirical tasks; they require human judgment about what matters and why, which cannot be derived from data alone regardless of how much data is available. AI tools may play supporting roles in these areas (for example, by compiling and organizing stakeholder input to support structured deliberation, or facilitating logistics of stakeholder engagement), but the core epistemic work remains human.

Between these two poles lies a substantial middle ground. Thematic analysis of qualitative interview data, for example, involves both pattern recognition and interpretive judgment. LLMs are increasingly capable of processing and categorizing large volumes of text, and – under certain conditions and guardrails – may synthesize information more consistently than a team of human coders working under time pressure with their own biases and limitations. But the validity of such analysis depends

on whether the categories and interpretations are meaningful in context – something that requires human evaluative judgment. Recognizing where a specific evidence need falls along this spectrum is one of the key practical contributions of this lens.

The same AI application can serve different evidence needs at different stages of the policy/program cycle – and may be appropriate for one purpose but not another. For example, NLP applied to open-ended stakeholder feedback during an initial situation analysis, where the goal is to surface broad themes and priorities from a large volume of input, serves a different evidence need than NLP applied to interview transcripts during an impact evaluation, where the goal is rigorous thematic analysis contributing to causal claims – with substantially different quality assurance demands.

5. Lens 2: Workflow

5.1 The core idea

The workflow lens asks how specific AI and digital tools could enhance evaluative work processes. This lens draws on the GEI task framework for evaluative activities[1], which organizes evaluative work into task clusters: from managing the overall process through framing, team engagement, design, data collection and analysis, to reporting and application of findings.

The framework applies to evaluations but also to other evaluative activities – such as needs assessments, monitoring reviews, evidence syntheses, real-time learning exercises, and so forth – that draw on evaluative thinking without constituting a formal evaluation. Not all task clusters will be equally relevant to every evaluative activity – routine program monitoring, for example, may draw heavily on data collection and analysis but involve minimal formal design, while an evidence synthesis may focus primarily on framing and reporting.

When implementing this work, practitioners face constraints in time, budget, data access, processing bandwidth, and communication reach. AI can address these in multiple ways: making existing tasks faster or cheaper, enabling access to data sources and processing approaches previously out of reach, and expanding how evaluation products can be tailored and disseminated.

Not all workflow enhancements are alike. It is useful to distinguish three types. First, *automation of routine tasks*: well-defined, repetitive tasks with minimal interpretive judgment – such as transcription, format conversion, and data cleaning – where AI frees practitioners to focus on higher-order work, with quality assurance focused on verifying accuracy. Second, *augmentation of human judgment*: tasks where human interpretation is essential, but AI improves consistency, coverage, or speed, such as qualitative coding with AI-suggested codes for human review, where the challenge is maintaining analytical rigor and validity. Third, *enabling previously infeasible processes*: tasks not practically possible without AI, such as continuous real-time monitoring of implementation at population scale using satellite imagery or other remote sensing data, or near-simultaneous production of audience-tailored evaluation products in

[1] The framework was developed based on review and adaptation of established guidance resources including the Manager's Guide to Evaluation and Rainbow Framework.

multiple languages from a single evidence base. This category matters because practitioners focused only on improving existing workflows may overlook the possibility that AI opens up fundamentally different ways of generating and using evaluative evidence. Where AI enables entirely new processes, the workflow lens may intersect with the evidence needs lens – for example, when satellite-based monitoring not only changes *how* implementation is tracked but also generates evidence of a type and granularity that previous field-based approaches could not provide.

Across these categories, effective practice depends on the intentional design of the interaction between human judgment and AI capability – sometimes described as a 'human-in-the-loop' approach. This means defining structured points in the workflow where human input shapes AI processing, where AI outputs are reviewed before being acted upon, or where iterative exchange between the two improves the quality of both contributions. Getting these interaction points right is not only a workflow design question – it also depends on the judgment, skills, and institutional conditions that the capability lens (Section 6) addresses.

Table 2 provides a summary of how AI applications and digital tools intersect with the task clusters of the evaluative activities framework; a more detailed mapping at the individual task level is provided in Appendix B. Some of the AI-enabled approaches listed here will be familiar from the evidence needs discussion in Section 4 – reflecting the fact that the same approach often warrants consideration through more than one lens.

Table 2: Workflow – evaluative activity task clusters and illustrative AI applications

Task clusters	Illustrative AI/digital applications
Manage the overall process	NLP-supported analysis of program documentation and stakeholder communications to map relationships, interests, and influence when planning a complex, multi-stakeholder evaluation – surfacing connections across a large volume of documents that manual review would be unlikely to identify.
Frame the evaluation (or evaluative activity)	ML-assisted literature scanning and clustering to rapidly identify relevant prior evaluations, research, and methodological approaches when developing the evaluation framework – reducing the time spent on manual literature review while broadening the range of relevant past examples the team can draw on.
Engage the team	Structured matching of team member competency profiles against technical requirements of the evaluation design, helping identify where additional expertise (e.g., in geospatial analysis, advanced statistical methods, etc.) is needed.
Design the evaluation (or evaluative activity)	Using a large language model as a sounding board for reasoning through evaluation design alternatives given the evaluation questions, context constraints, and available data. Purpose-built tools integrating AI into curated repositories of evaluation approaches (such as BetterEvaluation) are in development but not yet available.
Conduct data collection and analysis	AI-powered speech-to-text applications for transcribing interviews across multiple languages, combined with semi-automated coding to identify recurring themes – enabling an evaluation team to process a larger volume of qualitative data while maintaining human oversight of interpretation and meaning-making.

Task clusters	Illustrative AI/digital applications
Report and apply findings	Using large language models for producing tailored summaries of evaluation findings for different audiences (policymakers, program managers, community stakeholders) from a single evidence base – expanding the reach and usability of evaluation products beyond what the team could produce manually.

5.2 Key considerations when applying this lens

Several further considerations apply when using the workflow lens. Some evaluative tasks are time-consuming for good reasons: careful reading of qualitative data, stakeholder deliberation, and iterative refinement all benefit from sustained human attention. Automating these processes may save time while reducing analytical quality. The relevant question is not simply "can AI do this faster?" but "what is gained and what is lost?"

AI-enabled workflows can also introduce new dependencies – e.g., on connectivity, commercial platforms, data formats, and specialized maintenance – that need to be anticipated and managed.

Benefits are unevenly distributed. AI tools perform less reliably in languages that are poorly represented in their training data. Applications that depend on structured digital data only work where that data infrastructure already exists. And while AI can help under-resourced teams do more with less, those same teams may in some cases lack the expertise and/or institutional safeguards to catch errors in AI-generated outputs.

6. Lens 3: Capability

6.1 The core idea

The capability lens asks what M&E professionals and institutions need to know, be able to do, and have in place to engage with AI and digitalization responsibly. Where the first two lenses are instrumental (i.e., does a specific AI-enabled approach serve a specific evidence or workflow purpose?), this lens is about preparation and readiness: what must practitioners and institutions develop to navigate an AI-influenced professional environment? It encompasses technical skills, critical judgment, ethical reasoning, and institutional frameworks.

This lens draws on the GEI Evaluation Competency Framework, which identifies five competency domains – professional, technical, managerial, interpersonal, and contextual – applicable across various evaluation roles (evaluators, team leaders, managers, commissioners, experts, and users of evaluations). The emergence of AI does not require wholesale rewriting of this framework. Most of what makes a good M&E professional remains unchanged. Rather, AI requires additional, AI-specific competencies within the existing structure. Table 3 provides a summary of how illustrative AI-specific competencies map to each domain; a more detailed breakdown is provided in Appendix C.

6.2 Individual competencies

Table 3: Capability – AI-specific competencies within the GEI framework

Competency domain	Illustrative AI-specific competencies
Professional	Identifying ethical considerations specific to different types of AI use in evaluation (e.g., bias in training data, consent for AI processing of participant information); pursuing AI-related professional development through reputable sources and sharing emerging knowledge with peers.
Technical	Understanding which AI techniques are relevant to which evaluation use cases (e.g., distinguishing when NLP, computer vision, or predictive modeling might be appropriate); interpreting AI outputs critically, including recognizing common pitfalls such as overfitting, hallucination, and spurious pattern detection.
Managerial	Planning AI integration into evaluations, including scoping resource needs, timelines, and risks specific to different types of AI applications; building evaluation teams with appropriate AI-relevant skills and establishing clear role delineation between AI-related and conventional tasks.
Interpersonal	Communicating AI concepts, capabilities, and limitations to non-technical stakeholders in accessible terms; facilitating constructive discussion when stakeholders hold different views about the appropriateness of AI in a particular evaluation context.
Contextual	Identifying relevant local and institutional policies governing AI use in evaluation (e.g., data protection regulations, organizational AI policies); adapting AI approaches to local conditions, including language availability, digital infrastructure, and cultural norms around technology use.

An important area for future development is extending this to other professional roles in M&E systems – including policymakers, evaluation commissioners, and senior officials – who require foundational AI literacy competencies for oversight, accountability, and informed commissioning.

6.3 Institutional and organizational capabilities

Individual competencies are insufficient without institutional support. This section addresses the organizational and system-level conditions that enable or constrain effective AI use in M&E. Key areas include:

Data readiness: Availability, quality, interoperability, and governance of data – the foundation on which any AI application depends.

Governance and policy frameworks: Specific, context-appropriate policies guiding when and how AI may be used in evaluative work, including ethical review processes.

Infrastructure and tools: Computing resources, software, connectivity, and procurement processes that determine what is practically feasible.

Role clarity and team composition: How AI-related roles and responsibilities fit within evaluation teams and organizational structures.

Institutional learning processes: Mechanisms for capturing, sharing, and applying experiences with AI applications across evaluations and over time.

Diagnostic tools can help institutions and national M&E systems assess these capabilities systematically. GEI's M&E Systems Analysis (MESA) framework, for example, provides a structured approach to diagnosing M&E system strengths and weaknesses; the development of an AI/digitalization readiness module within MESA is under active exploration.

Note: This section outlines institutional capability dimensions without constituting a full institutional capability framework. Developing a comprehensive, validated framework for institutional AI readiness in M&E is a substantial undertaking flagged for future work.

6.4 Sequencing and readiness

Not every institution is ready to deploy a particular type of AI application. The capability lens supports a sequencing perspective: before investing in, say, machine learning for heterogeneous treatment effect analysis, are foundational capabilities in place – data infrastructure, governance mechanisms, basic digital literacy? This is particularly important in contexts where foundational M&E system strengthening is still underway. Premature introduction of AI applications in such settings risks wasting resources, producing poor-quality outputs, and eroding trust in both AI and evaluation. That said, carefully bounded pilot projects can serve as a valuable entry point – building institutional familiarity with AI tools while surfacing the specific capability gaps that more sustained investment would need to address.

6.5 Key considerations when applying this lens

In many cases, critical judgment may matter more than technical skill. As AI tools become more accessible, the professional challenge increasingly shifts from "can I use this tool?" to "should I use it, and can I critically evaluate its outputs?"

The relationship between individual capabilities and institutional conditions is complex and varies by type of AI application. For some applications – particularly accessible tools like large language models – individuals can learn skills they immediately apply to their own work, and benefits may spread organically as colleagues observe results. This bottom-up adoption is double-edged: organizations may benefit from employee initiative but also face risks when individuals use AI tools without appropriate governance (information security, data privacy, methodological integrity). Governance policies that manage this duality – enabling responsible individual use while maintaining organizational safeguards – are among the most pressing institutional needs. For other applications that require specialized infrastructure, data pipelines, or team-level coordination, institutional investment is a precondition for any meaningful use.

Access to AI tools, training, and infrastructure is unevenly distributed across organizations and countries. Without deliberate attention to equity, AI risks widening existing professional inequalities rather than narrowing them.

7. Applicability Across Levels

Each lens can be applied at multiple levels of the M&E system:

Individual or project level: "What evidence does this evaluation need?" / "What process constraints am I facing?" / "Do I have the skills?"

Organizational level: "What evidence does our portfolio need?" / "Where are systemic process constraints?" / "What institutional capabilities do we need to build?"

National M&E system level: "What policy questions could AI help the system answer?" / "Where are system-level bottlenecks?" / "What system-wide readiness conditions are needed, including whether foundational M&E capacities should be strengthened before AI applications are introduced?"

The questions change across levels, but the three-lens structure remains relevant. This multi-level applicability is particularly important for organizations supporting evaluation capacity development, which operate across individual capacity development, organizational strengthening, and national M&E system support.

8. Ethics and Responsible Practice

Ethical and responsible AI practice is among the most critical considerations in this space. This framework addresses ethics not as a standalone topic but as a dimension that runs through each lens, manifesting differently depending on which aspect of AI use is under consideration.

How ethics manifests across the three lenses

Through the evidence needs lens: ethical questions center on the validity and fairness of AI-generated evidence. E.g., could biases in training data lead to conclusions that misrepresent certain groups? Are the limitations of AI-generated evidence transparently communicated to users?

Through the workflow lens: ethical questions center on quality, consent, and the distribution of benefits. E.g., does automation maintain quality standards? Are data subjects informed about AI processing of their information? Do efficiency gains accrue primarily to evaluators while risks fall on the evaluated?

Through the capability lens: ethical questions center on professional responsibility and institutional accountability. E.g., are practitioners deploying tools they cannot critically evaluate? Do institutions have the governance mechanisms to ensure meaningful human oversight?

This differentiated treatment – showing how a principle like "ensure fairness" translates into different practical considerations depending on the lens – enables practitioners to identify the specific ethical questions most relevant to their situation, rather than working from generic checklists.

9. Why Integration Matters

Each lens individually provides a useful but partial view. The framework's central proposition is that the three lenses are interdependent in practice: any decision about AI in M&E simultaneously involves evidence, workflow, and capability considerations, whether or not the decision-maker recognizes this. Considering the lenses together surfaces trade-offs and interactions that a single-lens view would miss.

Two examples illustrate this.

Consider using an LLM for drafting evaluation reports. Through the workflow lens alone, this looks straightforward: it saves significant drafting time, the tools are widely accessible, and the capability requirements appear low. But adding the evidence needs lens introduces complexity. A well-prompted LLM synthesizing a large body of evaluation findings could, in principle, identify patterns or connections across the evidence base that a time-pressed human author might miss – a non-trivial potential contribution. However, LLMs can also flatten nuance, fabricate plausible-sounding content, and strip away the contextual judgment that makes evaluation findings meaningful. And adding the capability lens surfaces the most critical risk: evaluators need the judgment to distinguish when AI-generated text accurately reflects the evidence and when it subtly distorts it – a competency that is difficult to develop and easy to overestimate. A workflow-only view would adopt enthusiastically; the integrated view adopts cautiously, with substantial human oversight and clear protocols for verification.

Now consider satellite-based nighttime light imagery for evaluating rural electrification and economic development programs. Through the evidence needs lens, this is potentially transformative: satellite sensors capture nighttime luminosity data that correlates strongly with economic activity, electricity access, and urbanization – enabling measurement of program outcomes at a geographic resolution and temporal frequency that household surveys cannot match. The same data can serve multiple evidence needs: tracking electrification progress at the town or village level, measuring levels of nighttime light as a proxy for economic activity, and identifying geographic disparities in program reach. Nighttime light data has been used to evaluate rural electrification programs, monitor the economic impacts of COVID-19 lockdowns, and estimate subnational GDP where official statistics are unreliable. Through the workflow lens, it may replace or supplement expensive field-based data collection with scalable, consistent, and repeated observation – particularly valuable for monitoring large-scale infrastructure programs across wide geographic areas and extended time periods. But through the capability lens, critical constraints can emerge: interpreting nighttime light data requires understanding of what satellite sensors can and cannot detect (light saturation in bright areas, insensitivity to economic activity that does not produce light), skills in geospatial data processing, and – crucially – ground-truthing against local survey or administrative data to validate what the

imagery represents in a specific context. Without this ground-truthing, which requires both the technical skills to execute and access to ground-level data for comparison, impressive satellite imagery can create false precision. The integrated view suggests a strategic investment: the evidence and workflow value is high, but realizing it requires deliberate capability building in geospatial analysis, validation methods, and interdisciplinary team composition.

This kind of integrated analysis – assessing an AI application through all three lenses simultaneously – generates decision-relevant insight. An application can be transformative for evidence needs and demanding for capability simultaneously. It can have high workflow value and negligible evidence-needs value. It can appear attractive through one or two lenses while presenting risks visible only when all three are considered together. The framework supports this by making the trade-offs explicit and structured.

10. Applications of the Framework

The following are offered as propositions for how the three-lens framework can be applied in practice.

10.1 Clarity and transparency in AI-related guidance and communication

When recommending, documenting, or discussing AI use in M&E, professionals and organizations can use the three-lens structure to make their contributions more specific and more transparent. This applies across multiple channels:

Guidance and documentation. Rather than generic recommendations about "AI in evaluation," guidance notes and organizational policies can specify which lens or lenses an AI application primarily addresses. For example: "This application of satellite-based nighttime light analysis for evaluating rural electrification programs primarily addresses an evidence need related to the scale, granularity and temporality of outcome measurement, enabling village-level tracking of electricity access and economic activity proxies that household surveys alone cannot provide at comparable frequency or spatial resolution (Lens 1). It has high workflow impact during data collection and analysis stages, replacing or supplementing periodic field visits with continuous, scalable observation across the entire program area – particularly valuable for large-scale infrastructure programs spanning multiple districts or provinces (Lens 2). It requires substantial capability investment, including skills in remote sensing data processing and interpretation, understanding of what nighttime luminosity does and does not measure (e.g., saturation effects in bright areas, insensitivity to non-light-emitting economic activity), and – most critically – ground-truthing against local household survey or administrative data to validate what satellite observations mean in a specific program context (Lens 3)." This level of specificity helps practitioners assess relevance to their own situation.

Professional events. Conference panels and workshops on "AI in evaluation" can generate confusion when participants unknowingly operate from different lenses – one discussing evidence potential, another discussing workflow tools, another discussing

capability gaps. The three-lens structure can help organizers and participants specify which dimension is under discussion and can be used to deliberately structure sessions that address multiple lenses on the same topic.

Terms of reference and reporting. Commissioners can use the framework to specify AI-related expectations, and evaluators can use it to structure their reporting on AI use. For example, inception reports could describe the evidence need addressed, integration into the evaluation workflow and associated quality assurance measures, and relevant team capabilities and ethical safeguards.

10.2 Prioritization and investment decisions

Organizations and systems developing AI strategies can use integrated three-lens analysis to make differentiated investment decisions rather than adopting blanket directives. By assessing AI uses against evidence needs, workflow, and capability considerations together, decision-makers can identify which uses warrant strategic investment, which require prior capability building, which represent low-risk workflow improvements that can be adopted readily, and which may look attractive from one perspective but pose risks when viewed through the others.

Appendix D provides illustrative assessments of four AI applications demonstrating how this kind of integrated analysis supports more nuanced decision-making than single-lens assessments.

10.3 Knowledge curation and resource navigation

The three-lens structure can serve as an organizing architecture for curating AI-related resources in a knowledge repository. Resources in such a repository can be tagged according to one or more lenses: a training resource on using AI-powered translation tools for multilingual evaluations, for instance, might be tagged to a specific competency area within the capability lens (e.g., AI-augmented data collection and analysis) and to a specific task within the workflow lens (e.g., a data processing task in the "conducting data collection and analysis" task cluster), while it may not have a specific evidence needs tag since translation assistance is not about generating new evidence.

This structure supports practitioners who approach the repository with a specific problem in mind – an evidence gap, a workflow constraint, a capability need – and can equally serve those interested in a particular AI-enabled approach, by surfacing a range of lens-specific considerations associated with that approach. In either case, when a practitioner finds a resource through one lens – say, searching for workflow tools related to qualitative data processing – they can also see what other lens-related tags are associated with that same resource, surfacing relevant capability or evidence considerations they might not have otherwise encountered. The structure can also help identify gaps: for example, where available guidance on a tool addresses evidence and workflow considerations but not the capability requirements for responsible use. Of note, this tagging approach depends on the specific resources available in the repository and whether they contain sufficient detail to enable such tagging.

11. Conclusion

This framework is offered as a structured starting point – to be tested, enriched, and refined through application. The M&E profession is well-equipped to navigate this moment. The same capacities that define good evaluation – rigorous thinking, ethical commitment, contextual sensitivity, methodological pluralism – are precisely those needed to engage with AI responsibly. The challenge is not to become technologists, but to remain evaluators.

We invite the global M&E and evaluation community to apply, test, challenge, and refine this framework.

References

Branchini, B., & Vallina Acha, B. (2025). Use and application of artificial intelligence in public policy evaluation: A scoping review. Journal of Policy Evaluation, 1, 114–133. https://doi.org/10.4995/jpeval.2025.23837

Bruce, K., Gandhi, V.J., & Vandelanotte, J. (2020). Emerging Technologies and Approaches in Monitoring, Evaluation, Research, and Learning for International Development Programs. MERL Tech Report #4. https://merltech.org/wp-content/uploads/2020/07/4_MERL_Emerging-Tech_FINAL_7.19.2020.pdf

Jacob, S. (2025). Artificial Intelligence and the Future of Evaluation: From Augmented to Automated Evaluation. Digital Government: Research and Practice, 6(1), 10:1–10:10. https://doi.org/10.1145/3696009

Mason, S. (2023). Finding a safe zone in the highlands: Exploring evaluator competencies in the world of AI. New Directions for Evaluation, 2023(178–179), 11–22. https://doi.org/10.1002/ev.20561

Nielsen, S.B. (2023). Disrupting evaluation? Emerging technologies and their implications for the evaluation industry. New Directions for Evaluation, 2023(178–179), 47–57. https://doi.org/10.1002/ev.20558

Reid, A.M. (2023). Vision for an equitable AI world: The role of evaluation and evaluators to incite change. New Directions for Evaluation, 2023(178–179), 111–121. https://doi.org/10.1002/ev.20559

UK Evaluation Society. (2025). AI in evaluation: Good practice guidelines for practitioners. Version 2.0. https://evaluation.org.uk/new-ai-guidance-launched-for-evaluators/

Appendix A: Evidence Needs – Detailed Mapping

This appendix provides a detailed mapping of the GEI policy/program cycle stages, central questions, ways of answering, and illustrative AI and digital applications. Where no substantive AI application is identified, or where AI's role is limited to supporting rather than performing the activity, this is noted. *Note that the listing of an AI/digital application in this table does not indicate a recommendation or endorsement.*

Stages	Central questions	Ways of answering	Illustrative AI/digital applications
Stage 1: Understand the situation	*What is happening?*	Assessing the current situation	Data integration across heterogeneous sources; dashboarding and visualization
		Analyzing existing data	ML for pattern and anomaly detection in large administrative datasets
		Consulting stakeholders	NLP-assisted analysis of stakeholder feedback data (e.g., open-ended survey responses, complaint records, consultation transcripts)
	Why is it happening?	Analyzing potential causes and contributing factors	ML-assisted analysis of linked administrative datasets to identify correlates and potential drivers (e.g., spatial clustering of service delivery gaps alongside infrastructure and demographic variables)
		Synthesizing existing evidence	NLP-assisted evidence synthesis and literature review
		Mapping system dynamics	Systems dynamics and network analysis modeling (not AI-specific)
	What should be prioritized?	Ranking or scoring needs	Multi-criteria analysis tools *(note: most established tools are rule-based rather than AI-driven; AI-enhanced versions are emerging but not yet widely validated in M&E contexts)*
		Deliberating collectively on priorities	AI may support logistics and preparation for stakeholder deliberation processes (e.g., scheduling, compiling background materials, organizing input), whether in person or virtual, but the deliberative process itself is a fundamentally collaborative and human activity
		Mapping existing assets and resources	AI-assisted data integration to compile resource inventories across sources
Stage 2: Explore options	*What are the options?*	Scanning for existing interventions	NLP-assisted searching and classification of intervention descriptions across evidence repositories and program databases
		Synthesizing evidence about interventions	AI-assisted systematic review (ML-based screening and prioritization of search results, with human oversight of inclusion decisions and quality assessment)
		Generating ideas with stakeholders	AI may support preparation for collaborative ideation processes (compiling background materials, structuring options from prior evidence), whether in person or virtual

Stages	Central questions	Ways of answering	Illustrative AI/digital applications
	What are the anticipated effects in this context?	Synthesizing effectiveness evidence	NLP-assisted synthesis of effectiveness literature; AI-aided transferability assessment is an emerging area
		Assessing transferability to this context	Limited current applications; context-sensitive judgment remains primarily human
		Anticipating future conditions and effects	Predictive outcome modeling; scenario analysis; agent-based and systems dynamics modeling
	How do the options compare?	Weighing options against multiple criteria	Multi-criteria decision analysis tools (AI may assist with data compilation, but value weighting involves stakeholder judgment)
		Comparing costs and outcomes	Simulation-based cost-effectiveness modeling *(note: established tools are increasingly incorporating ML components for parameter estimation, though core modeling approaches remain largely conventional)*
		Convening expert or stakeholder judgment	AI may support structured elicitation processes (e.g., Delphi panels) through compilation and analysis of responses
Stage 3: Design the intervention	*How is the intervention expected to work?*	Articulating the intervention's logic	LLMs as a "thought partner" for articulating program logic
		Specifying causal mechanisms	Limited current applications; causal reasoning remains primarily human
		Learning from design evidence	AI-assisted retrieval of design evidence from comparable programs
	What is needed for successful implementation?	Assessing implementation requirements	AI-assisted scanning of implementation research for comparable contexts
		Learning from implementation research and past evaluations	AI-assisted scanning of implementation research and evaluation databases
	How can we manage the risk of negative effects?	Identifying, assessing, and managing risks	Risk scanning across program documentation and evidence bases
	How will we know if it works?	Defining success criteria and indicators	LLMs as a drafting aid for indicator development (e.g., based on program descriptions and comparable evaluations)
		Designing with evaluation and learning in mind	AI-assisted retrieval and comparison of evaluation designs from comparable programs (e.g., searching evaluation registries and reports to identify design precedents for similar interventions and contexts)
Stage 4: Support implementation	*Is the intervention (still) appropriate?*	Gathering stakeholder perspectives on fit	Text classification and sentiment detection (both NLP techniques) on feedback and complaints data; social media monitoring; real-time beneficiary satisfaction analysis
		Reviewing continued relevance	LLM-assisted scanning of changing context indicators

Stages	Central questions	Ways of answering	Illustrative AI/digital applications
	Is it being delivered as planned?	Tracking activities and outputs	Real-time analysis (using predictive and classification ML models) of program/MIS data; remote sensing and computer vision for monitoring observable changes (e.g., infrastructure construction, land use change, nighttime light patterns, etc.); automated compliance checking
		Assessing implementation fidelity	AI-assisted (using various ML modeling techniques) comparison of planned vs. actual implementation using program data
		Documenting implementation processes	Using an LLM for automated transcription and summarization of implementation meetings and field reports
	What results and insights are emerging?	Tracking early outcome signals	Adaptive learning and early warning systems using program monitoring data
		Gathering real-time feedback	AI-powered chatbots or SMS analysis for beneficiary feedback; sentiment analysis
		Identifying unexpected developments	Anomaly detection in program monitoring data
	How should it be improved or adapted?	Capturing and applying learning	LLM-assisted synthesis of monitoring data and implementation lessons
		Testing improvements iteratively	Platform-based A/B testing for digital interventions; rapid feedback loops using real-time program data *(note: AI-automated rapid cycle evaluation is an emerging concept but not yet an established practice)*
		Adapting to changing conditions	Scenario modeling to anticipate effects of contextual changes (not AI-specific)
Stage 5: Assess results and implications	*What happened?*	Measuring pre-defined outcomes	Automated data integration from multiple sources; geospatial visualization; timeline reconstruction
		Identifying emergent and unexpected outcomes	Anomaly detection; NLP analysis of open-ended responses and qualitative data for unanticipated themes
	What difference(s) did the intervention make?	Analyzing the intervention's contribution	NLP-assisted coding of evidence for qualitative comparative analysis and process tracing (note: these are emerging applications that require careful validation)
		Comparing to what would have happened otherwise	Causal ML for heterogeneous treatment effect estimation (e.g., causal forests integrated into RCTs to identify subpopulations for whom effects differ)
		Engaging stakeholders in assessing impact	AI (e.g., LLMs and other NLP techniques) may support compilation and organization of stakeholder perspectives to inform collaborative assessment, but evaluative judgment remains human
	How and why did it work (or not)?	Investigating causal mechanisms	NLP-assisted coding of qualitative evidence relevant to hypothesized causal pathways (e.g., systematically identifying and organizing passages across interviews, focus groups, and documents that speak to specific mechanisms); causal interpretation and mechanistic reasoning remain human tasks

Stages	Central questions	Ways of answering	Illustrative AI/digital applications
		Linking implementation to outcomes	AI-assisted integration of implementation and outcome data for process tracing
		Analyzing how context shaped results	Causal ML-assisted subgroup and interaction analysis (e.g., causal forests or similar methods to explore how effects varied across contexts, geographies, or population groups)
	How successful was it and what are the implications?	Synthesizing judgments about overall success	Inherently human evaluative judgment; AI may support by compiling and visualizing evidence summaries across evaluation questions to inform deliberative processes
		Drawing out implications for decisions	AI (e.g., LLMs) may assist in generating structured decision briefs or scenario presentations tailored to different audiences, but identifying what the findings mean for policy or practice requires human judgment

Appendix B: Workflow – Detailed Mapping

This appendix provides a detailed mapping of the GEI task framework for evaluative activities, specific tasks, and illustrative AI and digital applications. *Note that the listing of an AI/digital application in this table does not indicate a recommendation or endorsement.*

Task clusters	Tasks	Illustrative AI/digital applications
Manage the overall process	Establish decision-making and governance	AI-assisted (NLP-powered) stakeholder mapping and analysis
	Identify and engage stakeholders	Network analysis of program documentation; automated communication scheduling
	Secure and allocate resources	AI-assisted resource planning and budget estimation tools
	Establish quality review processes	AI-assisted ethics screening checklists
	Ensure ethical issues are addressed	Automated compliance monitoring against ethical protocols
Frame the evaluation (or evaluative activity)	Define what will be evaluated	AI-assisted (using LLMs or other NLP techniques) literature scanning and summarization of prior evaluations
	Develop/revise Theory of Change	AI (LLMs and other Generative AI models) as thought partner for drafting ToC and articulating assumptions and logic
	Clarify purpose, users, and uses	AI-assisted review of comparable evaluation frameworks and use cases
	Develop key evaluation questions	AI-assisted review of evaluation question banks and comparable designs
	Determine criteria and standards	AI-assisted scanning of relevant standards and benchmarks
	Draft and translate instruments	AI-assisted translation and drafting of data collection instruments
Engage the team	Identify necessary qualities and skills	Structured matching of competency profiles against evaluation design requirements
	Develop terms of reference	AI-assisted ToR drafting based on templates and comparable evaluations
	Compare options and select team	AI-assisted comparison of candidate profiles against requirements
Design the evaluation (or evaluative activity)	Develop evaluation design	AI-assisted navigation of evaluation design options
	Identify important design elements	AI-assisted review of methodological literature for design components
	Arrange review of design	Simulation-based design testing to anticipate data challenges
Conduct data collection and analysis	Conduct surveys	AI quality checks on survey data (consistency, completeness, outlier detection)
	Conduct interviews	AI-powered speech-to-text transcription and translation
	Conduct observation	Remote data collection via satellite imagery, mobile sensor data, digital trace data

Task clusters	Tasks	Illustrative AI/digital applications
	Conduct document review	AI-assisted (using NLP models) document analysis and information extraction
	Clean and prepare data	AI-assisted (using ML models) data cleaning, deduplication, and harmonization
	Code and classify data	Semi-automated coding and classification with human review
	Harmonize data across sources	AI-assisted data matching and integration across heterogeneous datasets
	Analyze patterns and themes	Pattern recognition in large or unstructured datasets; NLP-assisted theme extraction
	Conduct causal analysis	Causal ML for heterogeneous treatment effect estimation (e.g., causal forests within RCTs)
	Conduct geospatial analysis	Geospatial mapping and remote sensing analysis
Report and apply findings	Develop reporting products	Generative AI (e.g., LLMs) for drafting, summarizing, and reformatting evaluation reports
	Produce audience-tailored outputs	AI-generated tailored summaries for different stakeholder groups
	Create visualizations	AI-generated data visualizations and interactive dashboards
	Disseminate findings	AI-assisted dissemination planning and multi-channel distribution
	Support use of findings	Chatbot-mediated access to evaluation findings for non-technical users
	Follow up on use of findings	AI-assisted recommendation tracking and implementation monitoring

Appendix C: Capability – AI-Specific Competencies

This appendix provides a more detailed mapping of draft AI-specific competencies (to be further developed/refined) to the GEI Evaluation Competency Framework domains and competency areas.

Note: The numbering gaps in the second column are due to the fact that not all competency areas in the GEI framework have associated AI-specific competencies in the current draft.

Competency domain	Competency area	Illustrative AI-specific competencies
1. Professional	1.3 Upholds ethical standards	Describes ethical considerations in AI use; addresses AI-related ethical considerations in evaluation protocols; advises colleagues on privacy safeguards in AI-related data collection and processing
	1.5 Pursues professional development	Identifies reputable AI/data science information sources; explains fundamental AI concepts to peers; presents case studies of AI in evaluation
2. Technical	2.2 Applies appropriate methodologies	Describes major AI techniques and matches them to evaluation use cases; describes how (a particular type of) AI works and the implications for evaluation
	2.4 Uses appropriate data collection and analysis	Identifies key uses of AI-augmented data collection and explains associated risks; implements AI-augmented data collection and/or analysis
	2.5 Interprets data and draws valid conclusions	Identifies common AI pitfalls affecting evaluation quality, e.g., overfitting, hallucinations; describes key performance metrics of AI approaches
3. Managerial	3.1 Plans evaluations	Describes key considerations and risks for AI integration; describes resource needs for AI-augmented evaluation steps
	3.2 Builds a team with the right skills	Identifies team roles needed for AI approaches; coordinates multi-disciplinary teams with clear AI-related role delineation
	3.4 Ensures responsible research practices	Identifies key data governance frameworks applicable to AI in evaluations; implements relevant AI-related data governance measures
4. Interpersonal	4.2 Communicates clearly	Explains basic AI concepts in plain language to stakeholders; facilitates discussion around AI approaches, findings, and limitations
	4.3 Maintains intellectual openness and humility	Encourages input on AI approaches; recognizes own limitations in AI expertise; responds constructively to feedback about AI approach selection
5. Contextual	5.2 Understands institutional context	Identifies relevant context-specific policies relating to AI; adapts AI approaches to align with local policies
	5.3 Adapts to context and stakeholders	Recognizes need to adapt AI approaches to local context, e.g., language, infrastructure; implements context-appropriate AI solutions

Appendix D: Illustrative Three-Lens Assessments

This appendix illustrates how integrated three-lens analysis can support decision-making about AI in M&E. Each assessment examines an AI application through all three lenses, demonstrating how the interaction between lenses generates insights that no single lens produces alone.

Lens	Profile A: Nighttime light remote sensing as a proxy for economic development	Profile B: ML for heterogeneous treatment effect analysis in RCTs	Profile C: NLP for qualitative data analysis	Profile D: LLMs for drafting evaluation reports
Evidence value	High – enables measurement of electrification, economic activity, and urbanization at spatial and temporal resolutions that household surveys cannot match; same data can inform situation analysis, implementation monitoring, and outcome evaluation.	High – moves beyond average treatment effects to identify which subpopulations benefit most (or least) from an intervention, enabling more targeted and equitable program design. *Note that subgroup effect estimates should typically be treated as hypothesis-generating rather than definitive.*	Moderate – can improve consistency and coverage across large qualitative datasets, reducing the risk that themes present in the data are overlooked due to volume; validity depends heavily on human oversight of categories and interpretation.	Limited but potentially non-trivial – an LLM synthesizing a large evidence base could surface patterns a time-pressed human might miss, but also risks flattening nuance and fabricating content.
Workflow impact	High – may replace or supplement expensive field-based data collection with scalable, repeated satellite observation across wide geographic areas and extended time periods.	Moderate – adds a specialized analytical step to the impact evaluation workflow; most valuable when integrated into existing RCT analysis pipelines rather than conducted as a standalone exercise.	Uncertain – accelerates coding and initial theme identification but requires substantial human review of AI-generated codes and themes to ensure analytical validity, which may add time and effort.	Moderate – saves drafting time for experienced evaluators who have the judgment to critically review every output against the underlying evidence.
Capability requirements	High – requires geospatial data processing skills, understanding of what satellite sensors can and cannot detect (e.g., light saturation, non-light-emitting activity), and access to ground-level data for validation. Ground-truthing is a non-trivial	High – requires strong foundations in experimental design and causal inference, statistical programming skills, understanding of sample size requirements for reliable subgroup estimation, and the ability to interpret	Moderate to high – requires understanding of both qualitative methods and NLP limitations to use responsibly.	Deceptively low – tools are easy to use; the hard part is judging when AI-generated text accurately reflects the evidence and when it subtly distorts it.

Lens	Profile A: Nighttime light remote sensing as a proxy for economic development	Profile B: ML for heterogeneous treatment effect analysis in RCTs	Profile C: NLP for qualitative data analysis	Profile D: LLMs for drafting evaluation reports
	undertaking requiring both technical expertise and local survey or administrative data for comparison.	and communicate heterogeneous effects to non-technical audiences.		
Integrated implication	Consider strategic investment. Evidence and workflow value is high but realizing it requires deliberate capability building in geospatial analysis, validation methods, and interdisciplinary team composition. Without ground-truthing, impressive imagery can create false precision.	Consider strategic investment with technical partnerships. Transformative for evidence quality in well-designed RCTs, but the specialized statistical skills required mean most evaluation teams will need to partner with or hire quantitative specialists. Focus initial investment on capacity to interpret and communicate results.	Invest in capability with clear guardrails. Useful as augmentation; risky as replacement. Build team capability in both qualitative methods and NLP limitations. Establish clear human-in-the-loop quality protocols before deployment – including independent review of AI-generated coding against human-coded subsamples.	Exercise caution. Adoption often driven by time pressure rather than evaluative value. High risk of undermining credibility without robust human-in-the-loop review processes. Adopt only with strong verification protocols, evaluators capable of detecting subtle distortions, and transparent disclosure of AI's role in the drafting process

global
evaluation
initiative