CGIAR | Independent Advisory and Evaluation Service

Considerations and Practical Applications for Using Artificial Intelligence (AI) in Evaluations

Technical Note

D. Cekova, L. Corsetti, S. Ferretti and S.Vaca

June 2025

Correct citation: Cekova, D., Corsetti L., Ferretti, S. and Vaca, S. (2025). *Considerations and Practical Applications for Using Artificial Intelligence (AI) in Evaluations. Technical Note.* CGIAR Independent Advisory and Evaluation Service (IAES). Rome: IAES Evaluation Function. https://iaes.cgiar.org/evaluation

Cover image: Digitalization, AdobeStock.

Considerations and Practical Applications for Using Artificial Intelligence (AI) in Evaluations

Technical Note

Diana Cekova, Lea Corsetti, Silva Ferretti and Sara Vaca

Title	Considerations and Practical Applications for using AI in Evaluations. Technical Note
Purpose	To encourage and guide the integration of AI into CGIAR's evaluation practice by introducing and clarifying key concepts, legal considerations, and ethical standards. It aims to inform the revision of the CGIAR-wide Evaluation Framework and provide practical tools—such as software recommendations and prompt examples—to support responsible and effective use of AI in evaluation activities.
Audience	The primary audience for this document includes evaluators, evaluation managers, and commissioners involved in evaluation activities within IAES and broader CGIAR. The document may also be of value to a broader group of stakeholders within and beyond CGIAR, such as internal staff involved in designing evaluations, consultants developing evaluation frameworks or analyzing data, policymakers commissioning evaluations, and organizations involved in developing AI policies.
Framework and Policy Reference	This Technical Note supports operationalizing the CGIAR-wide Evaluation Framework and Evaluation Policy (2022)
Soliciting Input and Feedback	The IAES welcomes queries and feedback on this beta version of the Technical Note to inform the next version. The approach and methods will continue to evolve based on practical experiences and emerging industry standards for the ethical integration of AI into evaluation processes. Users in CGIAR and beyond are encouraged to contact the Evaluation Function within the Independent Advisory and Evaluation Service (IAES) of CGIAR.
Contact	IAES-Evaluations@cgiar.org

Acknowledgments

This Technical Note was prepared by a diverse team. Their work was overseen and guided by the Evaluation Function lead Svetlana Negroustoueva, under the overall direction of Allison Grove Smith, director of CGIAR Independent Advisory and Evaluation Service (IAES). The authors of this Technical Note wish to thank the peer reviewers and key contributors for their valuable insights: financial and digital inclusion advisor, Catherine Mclennan Highet, and researcher and consultant on technology, innovation, and entrepreneurship, John Kieti. We also express our sincere gratitude to senior evaluation manager, Ibtissem Jouini for her valued contribution.

The development team thanks the IAES staff in Rome who provided great support, ensuring the smooth conduct of technical and administrative processes related to developing this Technical Note.

Contents

List of Tables

List of Figures

List of Boxes

Table of Acronyms

AEA	American Evaluation Association
AGI	Artificial General Intelligence
AI	Artificial Intelligence
AR4D	agricultural research for development
CEN-CENELEC	European Committee for Electrotechnical Standardization
EES	European Evaluation Society
EF	Evaluation Function
EU	European Union
FAIR	Findable, Accessible, Interoperable, and Reusable (data principles)
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
IAES	Independent Advisory and Evaluation Service (part of CGIAR)
ICT4D	Information and Communication Technologies for Development
IEEE	Institute of Electrical and Electronics Engineers
ISO/IEC	International Organization for Standardization/International Electrotechnical Commission
KII	Key Informant Interview
LLM	Large Language Model
MERL	Monitoring, Evaluation, Research and Learning
ML	Machine Learning
NLP	Natural Language Processing
OECD	Organization for Economic Cooperation and Development
PII	personally identifiable information
RAG	Retrieval-Augmented Generation
SC	System Council
SLM	Small Language Models
ToR	Terms of Reference
UN	United Nations
UNDP	United Nations Development Program
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNFPA	United Nations Population Fund
VOPES	Voluntary Organizations for Professional Evaluation
XAI	Explainable AI

Executive Summary

The CGIAR 2030 Research and Innovation Strategy commits organizational change with seven ways of working, including "Making the digital revolution central to our way of working". In that context, Artificial Intelligence (AI), introduces both opportunities and risks to evaluation practice. Guided by the CGIAR-wide Evaluation Framework, integrating AI tools requires a governance approach to balance innovation with ethical responsibility, ensuring transparency, fairness, accountability, and inclusivity. This Technical Note encourages and guides CGIAR evaluators to ethically explore, negotiate, and experiment with AI tools:

Explore: Evaluators are invited to discover how AI, especially GenAI, can enhance evaluation efficiency, from scoping and data analysis to reporting. The Note provides practical guidance on AI applications and examples to support creative yet responsible exploration.

Negotiate: Integrating AI should be openly discussed with commissioners, stakeholders, and teams. The Note prioritizes jointly defining boundaries, expectations, and ethical parameters (transparency, accountability, and data sensitivity) at each phase of evaluation.

Use AI Responsibly: While AI tools are evolving, evaluators are encouraged to pilot and iterate their use. The document supports experimentation through practical tips, prompt examples, and tool selection criteria, all while emphasizing documentation and learning from each use case.

Effective AI governance is grounded in core **principles: Transparency** requires clear documentation of AI tool usage, data sources, model limitations, and decision-making processes. **Accountability** involves assigning responsibility for AI decisions and outputs and establishing oversight and redress mechanisms. **Fairness and inclusion** must proactively mitigate bias and discrimination, with particular attention to underrepresented groups and data gaps. **Data privacy** and security must align with applicable data protection regulations and ensure secure handling practices. **Human oversight** ensures that evaluators retain control over processes and can intervene as needed.

In operationalizing ethical AI governance in CGIAR evaluations, due diligence is required in **assessing AI tools** for ethical alignment before deployment: reviewing the transparency of vendors, the documentation of models, and their intended use cases. Where relevant, components involving AI—especially those engaging human subjects or sensitive data—should undergo ethics review. AI applications must be adapted to the local and cultural contexts in which evaluations are conducted, as what is suitable in one setting may be inappropriate in another. Additionally, participants should be informed about the use of AI systems and the implications of data collection or processing to ensure informed consent.

Ethical AI governance should be embedded in the entire **evaluation lifecycle**. During the *design phase*, evaluators should define AI tools to use, why they are selected, and assess risks. In *data collection*, AI tools should be used in ways that uphold data privacy and protection standards and avoid reinforcing harmful stereotypes or excluding groups. During the *analysis* phase, the role of AI in supporting interpretation should be documented, with an acknowledgment of limitations or biases. In *dissemination, documentation, and reporting*, AI's contribution, limitations, and human validation should be disclosed. By rapid adaptation of content across formats, languages, and complexity levels, AI opens possibilities for broader, more inclusive communication of findings. Finally, the *follow-up* phases should include a reflection on the ethical implications observed and how these lessons can improve future evaluations.

By embedding methodological flexibility into the evaluation processes, AI adoption would contribute to integrity, equity, and learning in an era of rapid technological advancement. This Technical Note is a conversation starter—as a **"Beta" version**, it will evolve based on responsible real-world experimentation and continuous reflection. Evaluators are encouraged to be responsive to stakeholder input throughout the evaluation processes, to ensure relevance, accuracy, and inclusivity.

1. What is the AI Technical Note

The CGIAR 2030 Research and Innovation Strategy, launched in 2022, sets the stage for doing business differently to ensure that research provides real solutions for development. CGIAR is committed to change the way it works, following seven new ways of working (WoWs). One of the WoWs is "Making the digital revolution central to our way of working".

1.1. How This Note Supports the CGIAR Evaluation Framework

The IAES Terms of Reference drive operations of the Evaluation Function (EF): to implement the System Council (SC) endorsed Multi-Year Evaluation Plan (2025-27 Workplan for CGIAR's IAES (SC/M21/DP5)). Operationalizing CGIAR-wide evaluation framework and policy involves development and revision of the evaluation guidelines and related materials. The scope of this Technical Note includes evaluations under the **CGIAR-wide Evaluation Framework** and Policy (2022).

This Technical Note has been developed to encourage and guide CGIAR evaluators to explore, negotiate, and experiment with AI tools while upholding ethical standards. **The specific aims** of this Note are threefold:

1. to inform the work of IAES EF, its staff and consultants on key aspects relating to the use of AI for their work at CGIAR and beyond, including relevant terms and concepts, legal and regulatory frameworks, to towards a responsible, ethical and effective use of AI.
2. to provide practical guidance: software recommendations, prompt creation examples, and curated resources to support effective and responsible AI integration in the evaluative activities under IAES.
3. to provide a framing and set a base for revision of Policy to reflect responsive and credible integration of AI tools for evaluation practice in CGIAR post launch of CGIAR's Digital Strategy. This Technical Note aligns to the following principles of the CGIAR-wide Evaluation Framework: relevance, use and utility; transparency; ethics and equity; and credibility and robustness, therefore ensuring AI applications support effective evaluation while maintaining integrity and ethical standards.

1.2 What To Expect in this Technical Note

This Note intents to address current and broader aspects of the CGIAR's forthcoming organization-wide AI strategy, i.e. open digital solutions, ethical AI frameworks, data governance, and innovation hubs across CGIAR centers. In fact, the objectives of this version of the Technical Note are to:

- Introduce AI-related concepts and terminology through a comprehensive glossary
- Provide context on responsible and ethical AI governance frameworks
- Examine critical ethical considerations for AI use in evaluation
- Build evaluator competencies for effective AI integration, including:
    - Developing AI literacy and practical skills
    - Managing data sensitivities and privacy concerns
    - Creating appropriate workflows and supervision mechanisms
    - Ensuring transparent documentation of AI use
- Outline **practical** AI applications across evaluation phases with detailed examples
- Offer **guidance** on effective AI conversations and prompt engineering strategies
- Present **curated** resources, tools, and example prompts for immediate application

This Technical Note (TN) comes at a unique inflection point where the full potential of generative AI in evaluation is still unfolding—its usefulness depends on the specific needs, capacities, and contexts of each user and organization. This Note promotes thoughtful experimentation, transparent decisions, and adaptive use in complex environments like CGIAR, where evaluation needs are diverse and evolving.

Setting this expectation upfront emphasizes that this TN takes an exploratory rather than prescriptive approach. Integrating AI into evaluation requires mindful experimentation more than rigid instructions. The Note is not a manual for quick application but a practical tool for building awareness, confidence, and discernment in navigating AI's evolving role in evaluation. Experimentation should be responsible, negotiated, and grounded in context, not blind. It is not a quick-start guide or a "how-to" manual. Instead of fixed workflows, it highlights areas where evaluators can develop and share competencies. These areas—focused on different evaluation phases—are meant to inspire exploration and support responsible use in dynamic settings like CGIAR, where needs change rapidly.

This TN does not provide step-by-step instructions for immediate use by evaluators with no AI background. Just as a short-term evaluator needs to have foundational training for focus groups, evaluators will also require a thorough understanding of AI's risks, trade-offs, and implications to use it effectively. However, the Note is accessible to evaluators on short-term assignments to help them reflect on what is feasible, responsible use, and necessary support. As one practitioner put it after "three sleepless nights of experimentation," the real breakthrough comes not from pre-packaged solutions, but from developing the judgment needed to align AI use with complex evaluation challenges.

Consider this Note as a conversation starter, to evolve over time based on responsible real-world experimentation and continuous reflection. As AI technologies and their governance frameworks rapidly change, as more solutions are being generated and shared, this Note will be periodically revised to incorporate emerging proven practices and regulatory developments in CGIAR and beyond, including in the evaluation industry. Considered **a Beta –version,** the timely revision of this TN will contend with related AI policy changes in CGIAR, digital and AI strategy and evolution in supporting softwares[1]. The aims of this TN, objectives, and intended users present an opportunity for CGIAR to position itself as a peer in AI-aided evaluations, documenting and sharing its AI experiences to inform best practices in evaluation industry.

2. What is AI

Generative AI (GenAI) has firmly established itself as a transformative technology that is here to stay. The proliferation of AI-driven services is increasingly evident across various sectors, with significant implications for global development. The Organization for Economic Cooperation and Development (OECD) released a press release warning that GenAI is set to exacerbate existing regional divides. This divide becomes even more pronounced when considering nations in the Global South, where historical inequities and limited digital infrastructure already create significant barriers to technological adoption. Initial evidence on AI's potential suggests that using GenAI tools in the workplace can significantly improve performance in specific tasks. However, these productivity gains are unevenly distributed.

[1] The evaluation function under IAES is responsible for maintaining and updating the TN, with contributions acknowledged. Revisions will follow IAES protocols, with revision history and contributors documented in each edition.

Figure 1. What is AI

As Krishna (2024) highlights, AI technologies developed primarily in the Global North often embed values and biases that may not align with diverse cultural contexts, potentially deepening rather than bridging existing digital divides. Current forms of AI should therefore be viewed as tools to enhance and streamline certain processes, not replace human labor or judgment. When thoughtfully deployed, AI can help navigate complexity, enhance efficiency, generate timely insights, and support data-informed decision-making—but these benefits must be accessible to all regions of the world to avoid technological colonialism.

As AI continues to advance and become more sophisticated, it is increasingly important to understand its applications—particularly in the emerging field of Generative AI (GenAI). Within the evaluation sphere, AI is already present in various tools and methods. For reference, definitions of AI and its key subtypes, including GenAI, are provided in the glossary at the end of this document.

AI is becoming ubiquitous in daily lives, even predating the recent rise of GenAI. From Zoom's AI-powered meeting assistants and note-takers to personalized recommendations on Netflix and YouTube, **AI has been working behind the scenes to enhance consumer digital experiences for at least a decade**. However, it is important to move beyond passive AI adoption and critically assess how AI systems shape knowledge production and decision-making.

2.1. AI in Evaluation: Concepts

GenAI is a tool to help support and streamline some processes in light of often heavy workloads and tight deadlines. However, it is not a tool to replace human efforts. For instance, MAXQDA[2] and other qualitative research tools increasingly incorporate sophisticated AI-assisted coding capabilities, but these still require human validation as they can lead to misinterpretations or miss contextual nuances. They serve as accelerators of human-led analysis rather than autonomous solutions. For dissemination and learning, while GenAI can support writing, editing and QA processes, the current capabilities of the software should not be over-estimated. It is crucial to carefully consider the ethical implications and potential risks associated with submitting a document generated by GenAI, as the evolving technology may have limitations or biases that could affect integrity and credibility of the work.

From a labor perspective, while concerns exist about over-reliance on GenAI-produced content that potentially undermines human expertise, there is also an opportunity to leverage AI for routine tasks, freeing evaluators to focus on higher-value work that requires human judgment and insight. **Successfully leveraging AI requires simultaneously optimizing the value derived from its ability to enhance human tasks, and the effectiveness of safeguards, validation, and quality control measures necessary to manage the inherent risks**. The key is finding the right balance where AI enhances, rather than merely

[2] MAXQDA is a comprehensive software program designed for qualitative and mixed methods data analysis, enabling evaluators to organize, analyze, and visualize diverse data types including text, audio, video, and survey data.

substitutes, human contribution, with appropriate safeguards, validation, and quality checks in place to mitigate risks.

2.2. How GenAI Works

GenAI is powered by Large Language Models (LLMs) that are trained on vast amounts of data, drawn from sources such as Common Crawl (web data), Project Gutenberg (books), GitHub (code), news archives, scientific papers, and Wikipedia. These models also commonly incorporate data from more controversial sources, including Reddit, Twitter/X, Facebook, Instagram, and other social media platforms, raising important ethical considerations about consent, bias, and representation in LLM training data. These models learn patterns in grammar, syntax, and semantics, enabling them to predict the next word in a sequence based on the contextual cues provided by the preceding words. GenAI is built on neural networks, a type of machine learning inspired by the way the human brain processes information. These models do not operate through explicit rules or logic but instead identify statistical patterns in vast datasets to generate new content.

Key components and concepts in GenAI include:

- **Statistical representation**: GenAI models are not knowledge platforms(?) themselves, but rather statistical representations of the knowledge contained in their training data. At the core of GenAI models such as LLMs is unsupervised learning—they are trained on massive amounts of text (or images, audio, and other data, depending on the model), but they do not analyze or assess information in the way humans do. Unlike a search engine, which retrieves relevant sources, or a researcher, who compares and evaluates information, GenAI simply predicts the next word, pixel, or note based on probability distributions. It does not weigh evidence, assess credibility, or understand the value of a source—it merely reconstructs language as a 'stochastic parrot'.
- **Temperature settings**: By adjusting model 'temperature', developers can influence how creative or predictable the AI's outputs will be. Higher temperatures result in more varied and human-like responses, while lower temperatures produce more conservative, robotic-sounding outputs.
- **Sequential word prediction**: GenAI constructs sentences one word at a time, choosing from probability distributions of potential words at each step. This process allows the AI to generate coherent, contextually relevant responses.

A fundamental issue with GenAI is copyright and fair use.[3] The models were trained on huge amounts of available text, images, and other content under the assumption that they are not copy-pasting, and thus not violating copyright. However, this creates blurry legal and ethical boundaries and many of these legal challenges are in the courtroom during writing of this TN. The way in which GenAI reuses knowledge, styles, and structures makes it nearly impossible to set a strict boundary between inspiration and infringement.

1. **Explainability**: GenAI operates through layers of hidden connections, where billions of parameters determine how text, images, or data are generated. Unlike traditional models with explicit rules, the underlying neural networks function as 'black boxes'-even developers cannot trace the specific reasoning behind each output. While Explainable AI (XAI) techniques offer partial insights, this

[3] As of the writing of this document, there are several ongoing lawsuits alleging copyright infringement still ongoing, particularly in the USA.

inherent lack of traceability and interpretability creates significant challenges for evaluation practice, particularly when accountability and transparency are ethical requirements.

2. **Bias and reinforcement of dominant narratives**: AI models learn from existing texts meaning they default to the most statistically prevalent patterns. This can lead to reinforcing stereotypes, over-representing mainstream perspectives, and overlooking marginalized voices. In evaluation, where context, diversity, and representation matter, this creates a risk of producing skewed insights that mirror existing power imbalances rather than challenging them. Research by Ashwin et al. (2023) confirms this concern, demonstrating that when analyzing interviews with displaced Rohingya people, LLMs introduced non-random biases that correlated with interviewee characteristics, performing worse than simpler supervised models trained on high-quality human annotations.
3. **Attribution and source awareness**: AI does not retrieve, reference, or compare sources. Rather, it generates text probabilistically, because it cannot situate information within a clear body of knowledge. Unlike a researcher who engages with historical context, academic discourse, or expert sources, AI responds without an inherent sense of where knowledge comes from or how it fits within a larger debate. This is a major challenge for evaluation, where the ability to trace insights back to a verifiable source is critical. Without attribution, the credibility, intent, or validity of AI-generated outputs cannot be assessed. The increased use of Retrieval-Augmented Generation (RAG[4]) permits LLMs to potentially cite sources, to mitigate for key challenges in verification and attribution.
4. **Misinformation and hallucinations:** Despite ongoing efforts by AI providers, controlling misinformation on a large scale remains difficult. AI models can still be manipulated, and new problematic content types are emerging, such as increasingly believable synthetic photos and videos. AI models generate fluent responses even when lacking true knowledge, resulting in ***hallucinations*** (confidently incorrect, even fictitious, outputs). In evaluation, this can result in fabricated references, misinterpretation of data, and plausible-sounding (but false) insights, posing a critical risk for evidence-based decision-making.
5. **Replicability:** Unlike traditional analytical methods, GenAI operates probabilistically, so it does not always produce the same output for the same input. Temperature settings influence this, controlling the level of randomness in responses. Even at low temperatures, responses can vary significantly due to the probabilistic nature, making true replicability impossible. Additionally, personalized GenAI outputs depend not only on the prompt but also on interaction histories, user configurations, and model fine-tuning. For example, in agricultural research for development (AR4D), GenAI tools increasingly adapt to individual users' styles and preferences—sometimes implicitly—based on prior interactions, input patterns, and system settings. This personalization introduces variability, challenging consistency and replicability in evaluation tasks.

3. How to use AI

3.1. GenAI Applications: Chatbots, Copilots, and Beyond

When incorporating AI into evaluation practice, **evaluators will encounter different modalities of engagement with these technologies.** Across all approaches, **effective implementation requires**

[4] RAG connects LLMs to external, often real-time data sources such as search engines or institutional/internal documents, retrieving relevant context to inform text generation. This can permit more trustworthy and factually grounded responses, rather than relying solely on the LLM's internal training data.

balancing human expertise with AI capabilities, recognizing that optimal results for evaluation emerge from complementary strengths rather than replacement. The **key is leveraging AI's computational power while applying human judgment, contextual knowledge, and ethical consideration**.

The integration of AI into evaluation workflows necessarily shifts some aspects of control from traditional manual approaches. This transition in agency is evident when directly interfacing with AI tools but becomes more nuanced as AI functionality becomes embedded within familiar software and automated processes. Increasingly, AI serves not merely as a tool but as a mediator of interactions with information, colleagues, and evaluation stakeholders. Maintaining critical awareness of this evolving relationship with work processes and data is essential as the boundaries between human and AI contributions become less distinct.

- AI tools can transform evaluation practices through three distinct modalities of use—ranging from direct engagement with native AI applications to embedded functionalities in traditional software, and fully automated workflows that operate with minimal human intervention. **Direct interaction with native AI applications:** This modality represents direct engagement with purpose-built AI tools designed for analytical, writing, or research tasks. Tools such as ChatGPT, Claude, Perplexity, or DeepSeek function as conversational interfaces that respond to queries and prompts. Beyond general-purpose chatbots, specialized research tools such as Elicit facilitate literature review and evaluation synthesis, while services such as Otter.ai provide transcription and analysis capabilities for qualitative data. These 'AI-first' applications involve primary engagement with the AI interface rather than traditional software environments.
- **Applications with embedded AI:** This category encompasses conventional software platforms that have integrated AI capabilities. Microsoft's productivity suite now incorporates Copilot functionality that offers content suggestions during document creation. Similarly, qualitative analysis software used in evaluation, such as MAXQDA or Atlas.ti, now includes algorithmic features for automated coding and thematic analysis. The distinguishing characteristic is the enhancement of established software environments with AI capabilities, which may be prominently featured as assistive tools or subtly integrated into existing functionality.
- **Automated AI tasks and workflows:** This represents the most autonomous implementation of AI in evaluation processes. Rather than continuous engagement with AI tools, evaluators establish systems that operate independently to execute routine tasks. Examples include automated systems for structured data collection (e.g., key informant interviews), stakeholder feedback mechanisms, or standardized report generation. Such implementations may involve purpose-built chatbots for information gathering or algorithmic workflows for data processing and analysis. The defining attribute is operational independence with minimal intervention, but unlike traditional automation, these AI systems can learn from data, adapt to varied inputs, and handle ambiguities beyond what rule-based procedures could manage. They continuously improve their performance through experience and can recognize patterns in complex, unstructured evaluation data without explicit programming for each scenario, allowing evaluators to allocate attention to more complex analytical aspects of evaluation.

Table 1. GenAI applications

Common AI category	Example	Brief Description
Interactive chatbots	ChatGPT, ClaudeAI, Perplexity, DeepSeek	Serve as conversational interfaces for general purpose or domain-specific inquiries. Provide instant, 24/7 access to information and assistance. Uses include aiding with tool development, translation and basic thematic analysis.

Common AI category	Example	Brief Description
AI assistants	Apple Siri, Google Assistant	Specialize in helping users with specific tasks or topics (targeted support and expertise within a particular domain) and usually operational or superficial in nature of enquiry.
Copilots	GitHub Copilot	Integrate into existing tools to provide AI-powered assistance. Enhance productivity and streamline workflows within the tool's environment.
Embedded AI tools	AI-assisted qualitative coding in Atlas.ti or MAXQDA	Incorporate AI capabilities directly into familiar platforms. Expand the functionality of existing software without requiring users to learn new tools.
AI workflows	Zapier, Salesforce Einstein GPT, Otter.ai, Reading.Ai, Zoom AI	Automate repetitive tasks through standardized sequences of AI-powered steps. Promote consistency and efficiency in routine processes (e.g., generating document summaries).
Autonomous agents	HubSpot, Zendesk	Designed to operate independently in specific contexts, such as customer service chatbots on websites. Provide intelligent, self-directed assistance to users with minimal human intervention.

The tools represent a spectrum of AI implementation, with LLMs, increasingly serving as the foundational technology enabling their core language intelligence and generative tasks.

3.2. Responsible and Ethical AI Governance

The rapid advancement of AI and algorithmic decision-making sparked a global movement towards their responsible and ethical governance. Numerous countries and organizations are now proactively addressing AI's potential impacts, often framing their approaches within an ethical context.

This has led to the development of non-binding guidelines and initiatives for responsible AI governance, typically grounded in democratic principles such as representativeness, privacy, accountability, transparency, and fairness. International bodies such as UNESCO, OECD, and the World Economic Forum issued recommendations emphasizing the importance of ethical, trustworthy, and human-centric AI development and use.[5]

To bridge the gap between principles and practical implementation, technical standards are being developed by organizations including ISO/IEC, IEEE and CEN-CENELEC. These standards aim to provide concrete definitions and frameworks for concepts such as explainability and trustworthiness, making them more accessible to AI users. On the regulatory front, the EU's AI Act (2021) represents a significant step towards binding legislation, introducing a risk-based classification system for AI applications. Yet, regulation alone is insufficient—effective AI governance must be grounded in ethical principles, ensuring AI is leveraged responsibly without reinforcing systemic inequalities. **While the prominence of ethical considerations in AI governance is encouraging, it has also raised concerns about potential 'ethics-**

[5] For a comprehensive and full discussions of the regulations as it stands in summer 2024, please refer to De Pagter et al., (2024).

washing',[6] where organizations adopt ethical language without substantive action**. This underscores the importance of a proactive stance on these issues, both for regulators seeking to prevent problematic impacts and for organizations preparing for compliance.** Many existing AI policies focus on restricting AI use without considering its transformative potential. This creates a problematic imbalance where organizations over-regulate AI applications while failing to address deeper concerns such as bias in AI-generated outputs.

Box 1. EU regulatory developments

Regulatory Developments
The EU's AI Act (2021) is a significant step towards hard regulation. It aims to address the risks of specific uses of AI, categorizing them into four different levels:

- Unacceptable risk (AI system is banned).
- High risk (AI system is subject to strict requirements).
- Limited risk (AI system subject to specific obligations, usually transparency related).
- Minimal or no risk (AI is freely allowed).

In doing so, EU's AI regulation aims to ensure that Europeans can trust the AI they are using. The regulation is also key to building an ecosystem of excellence in AI and strengthening the EU's ability to compete globally. However, regulation must be accompanied by a broader, cultural awareness of AI's implications. Focusing solely on compliance may lead to overly restrictive AI policies that stifle innovation while ignoring systemic risks such as embedded biases.

International organizations typically adopt policies that align with the applicable regulations in their member states to ensure uniform standards across operations. While formal adherence to the EU AI Act may vary, its principles often influence internal policies. For CGIAR, this represents an opportunity to take an adaptive, learning-driven approach—where AI use is continuously assessed, negotiated, and refined rather than locked into rigid guidelines.

Even for evaluators based outside the EU, understanding and considering these regulations is important because:

- **Many AI tools maintain separate EU-compliant versions**, which can affect global collaboration, tool availability and privacy/data sharing agreements.
- **Evaluations involving EU-based organizations or EU citizen data must comply with these regulations.**
- The Act **influences emerging international standards** and professional practices in evaluation.

It is imperative that CGIAR evaluations reckon with intended use of GenAI in its work-whether for automation, data analysis, or other functions—before integrating into practice.

Many of these discussions are in an ongoing process of abstract deliberation and revision, reflecting the breakneck speed and general volatility with which these developments are happening, and reflecting on

[6] Ethics washing refers to the practice where organizations feign ethical consideration or make misleading claims about their ethical practices to improve their public image, without genuinely implementing responsible actions.

the fact that in the last few years, GenAI went from being virtually unheard of to being nearly everywhere. AI is not just a technical challenge but also a governance challenge—one that requires ongoing inquiry, adaptive policy-making, and institutional learning. In this context, this Note focuses on the role of both GenAI and other forms of AI, as these technologies are most applied throughout the evaluation process and cycle.

3.3. Key Ethical Considerations for AI in Evaluation

Figure 2. Is ChatGPT Racist?

Source: Chris Lysy, Freshspectrum comic #304

As evaluators increasingly leverage GenAI's capabilities, it is crucial to **deeply understand and proactively address technology's limitations and risks through an ethical lens**. These considerations can be divided into methodological challenges directly impacting evaluation practice and broader structural ethical concerns. Two key **methodological challenges** for evaluation include:

- **Bias and lack of diversity in AI outputs due to skewed training data** which predominantly reflects historically dominant Western, white, male perspectives due to systemic factors in knowledge creation and curation as discussed more in depth in the previous section. Such bias is amplified by the global digital divide which excludes poorer populations with limited internet access. This can result in offensive stereotyping and subtle perpetuation of prejudice.[7]
- **Privacy and data rights concerns regarding personal information being ingested into AI training sets.** Personal information is often swept up from public and semi-public sources without specific consent, creating situations where AI systems may later regurgitate private details or make unauthorized inferences. Evaluators' prompts may contain sensitive test cases, proprietary information, or personally identifiable information (PII), which could be repurposed for AI training without appropriate safeguards. This creates significant ethical risks regarding confidentiality commitments to evaluation participants and raises questions about informed consent and data sovereignty in evaluation contexts.

[7] The training data used for GenAI often overrepresents Western, educated, industrialized, rich, and democratic (WEIRD) perspectives, leading to biased outputs that can perpetuate stereotypes and marginalize under-represented groups. This has also been found with testing done at CGIAR centers as shown in this blog and discussed further in Box 2.

Thoughtful experimentation with GenAI in evaluation should be grounded in core ethical principles around bias, inclusion, equity, labor impacts, creative rights, environmental costs, and resisting harmful extractive dynamics. While GenAI has the potential to enhance inclusion by enabling more people to access funding, gather survey data, automate analysis, and spur creativity, it is essential to **establish acceptable, accuracy thresholds for each use case and maintain human oversight**, especially for high-stakes decisions. The challenges of AI bias are not just theoretical concerns but are documented in CGIAR's own research. Colleagues across CGIAR centers have actively tested AI systems to understand how bias manifests in agricultural contexts, as highlighted below in Box 2.

Box 2. CGIAR research on gender bias in agricultural AI systems

Research by CGIAR colleagues (Koo et al., 2025) tested how well various LLMs respond to questions from women farmers in India. Their study revealed that while AI systems generally promote gender equality, they often:

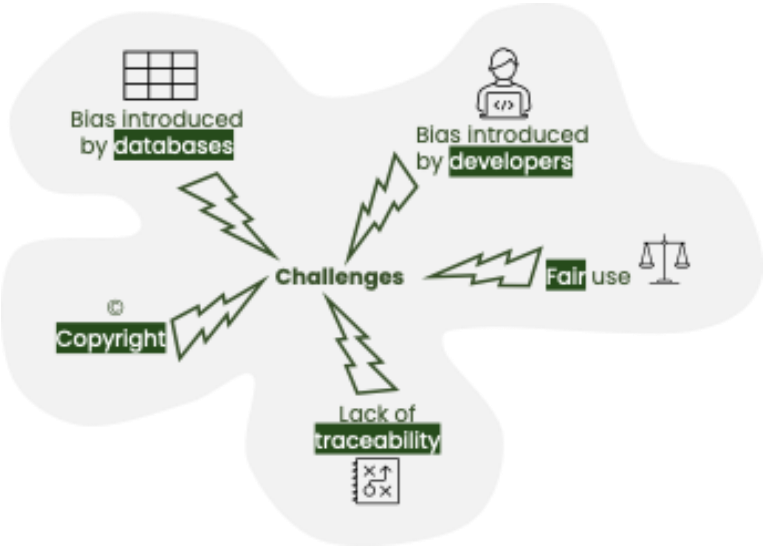
- Reinforce gender stereotypes about farming roles.
- Overlook structural barriers women face in accessing inputs and land.
- Fail to account for evolving gender roles in agriculture.
- Provide optimistic (but unrealistic) guidance that does not address real-world constraints.

The study underscores that AI systems, even those designed with good intentions, can perpetuate biases that affect agricultural evaluation and advisory services. Monitoring, Evaluation, Research and Learning (MERL) practitioners should be mindful that AI tools may require significant refinement to eliminate gender biases, address systemic inequalities, and respond effectively to diverse agricultural contexts.[8]

Beyond identifying bias within evaluation methods, evaluators should actively highlight AI bias as a substantive finding when discovered. As programs increasingly integrate AI into their operations, these biases can affect program delivery and outcomes. By documenting and reporting on AI biases, evaluations serve a dual purpose: improving evaluation methods and providing accountability for how AI is deployed in development contexts. This positions evaluation as an essential check on AI systems, ensuring they serve their intended beneficiaries equitably.

[8] Language barriers present additional challenges in agricultural contexts. Organizations like Digital Green are addressing this by developing Small Language Models (SLMs) for low-resourced languages such as Kikuyu, Bhojpuri, and Nigerian Pidgin, ensuring AI systems can accurately capture local agricultural terminology used by farmers

Figure 3. Challenges to responsible AI integration



While addressing methodological challenges in AI-enabled evaluation is critical, **the broader ethical and structural concerns that transcend methodology** must be considered. Even with perfect methodological implementation, these fundamental issues would persist:

- **Representativeness of AI:** Beyond simple bias, current AI systems reflect dominant cultural narratives, reinforcing epistemic exclusion. Addressing this is not just about reducing bias—it requires rethinking AI development itself, including challenging cultural hegemony and broadening representation.
- **Privacy and data rights:** AI training on vast amounts of data without consent raises fundamental concerns about ownership and control over personal and creative works. This goes beyond artists' copyrights to encompass broader implications for privacy and data rights in an AI-driven world.
- **Significant environmental costs:** These costs are particularly high in terms of the energy and water footprint required to train and run AI models.
- **Potential job displacement and negative labor impacts**: Especially for knowledge workers, these issues include both threats to evaluators and the exploitation of workforces helping to train AI systems, often in low-income countries.
- **Infringement on creators' copyrights:** This is particularly complicated when AI is trained on creative works without consent or compensation. [9]
- **Amplify exploitative practices:** AI can disproportionately concentrate profits and power in the hands of large tech companies.

3.4. How to Choose the Right AI Tools

The AI landscape evolves so rapidly that the question "which AI should I use?" has no definitive or static answer. Rather than spending excessive time seeking the perfect tool, **evaluators will benefit most from informed experimentation.** Most AI tools offer free tiers or trial periods specifically for exploration, making the cost of choosing a less-than-ideal option remarkably low.[10] For evaluators, it is important to understand that there are no standardized benchmarks that fully capture how these models will perform in real-world evaluation contexts, which is why hands-on testing with specific types of data and questions

[9] GenAI models trained on copyrighted material, such as artwork or text, raise questions about fair use, attribution, and compensation for content creators.

[10] When using free or personal tiers of AI tools, evaluators should be mindful of the associated usage agreements, particularly regarding data privacy and whether service providers may repurpose submitted data for their training pools. These considerations are discussed in more detail in Section 2.1.3.

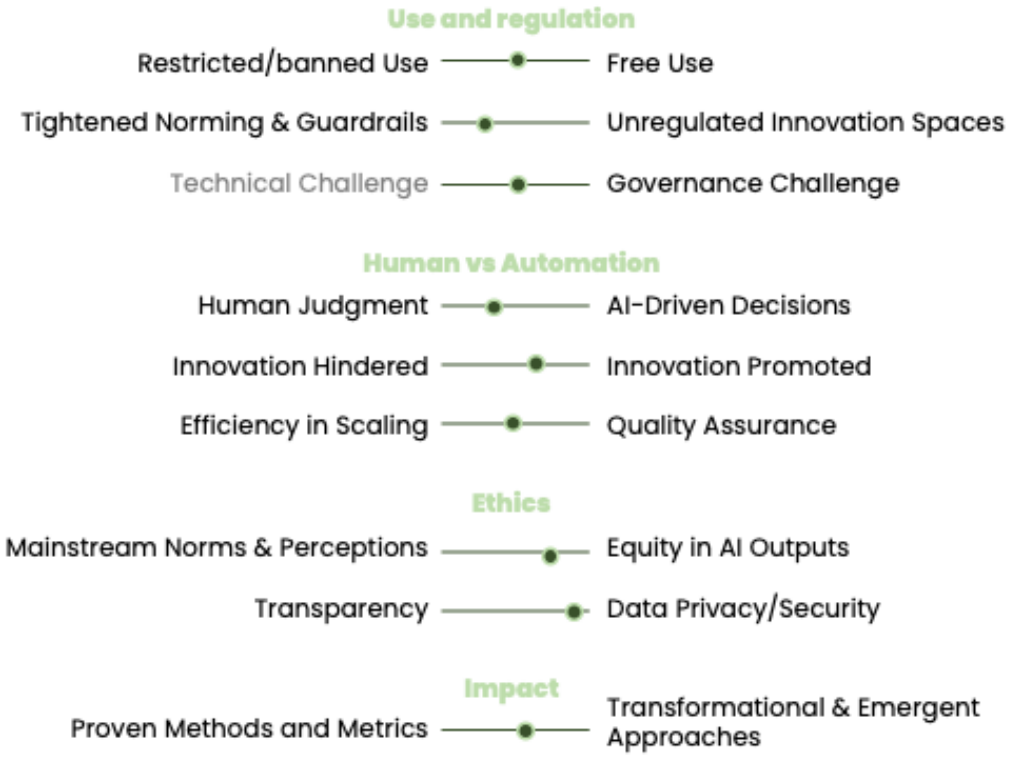
is invaluable. These models are increasingly versatile, meaning that even if the optimal model for an evaluation task is not selected, useful results will likely still be achieved. Understanding the AI ecosystem begins with recognizing two fundamental components:

Box 3. Fundamental AI components

1. **Foundation models**: Powerful AI systems that form the computational core of today's AI capabilities. They are LLMs such as ChatGPT-4, Claude, Gemini, LLaMA, Mistral, Qwen, Deepseek, and others that have been trained on massive datasets of text and code. They serve as the 'brains' behind most modern AI applications. These models constantly evolve, often with little public notice about their changes and improvements.
2. **Applications built on foundation models:** Provide accessible interfaces to powerful AI engines. These applications come in two main varieties:
   - *Enhanced familiar tools*: Software already being used that now incorporate AI capabilities while maintaining familiar interfaces (e.g., Microsoft Office with Copilot or Google Workspace with Gemini).
   - *Purpose-built AI applications*: New specialized tools designed specifically for tasks such as data analysis, transcription, or summarization.

When selecting an application, it is essential to understand that the foundation model that powers it is also implicitly being selected. This creates a multi-layered relationship where evaluation data may flow between the direct application provider and the underlying foundation model provider. This has important implications for data privacy and security that should be considered, especially when working with sensitive evaluation data.

Figure 4. Tensions to balance



The most successful evaluators approach AI adoption as an ongoing and mindful learning process rather than a one-time decision. This flexible mindset allows the toolkit to evolve alongside the rapidly advancing capabilities of AI itself. The following criteria, illustrated in Figure 3, can support informed decision-making.

3.4.1 Key Factors to Consider When Choosing an AI tool

Key factors to consider when choosing an AI are presented in Table 2.

Table 2. Key factors for choosing AI

1.AI Processing capabilities and limitations Limited number of documents? Word limits? Consider size and input/output limits.	**7.Local (offline) AI vs. cloud-based AI** Consider whether AI tools that can function offline or in low-bandwidth environments if needed.
2.Multi-functionality and supported media An all-purpose solution vs assembling a complementary set of tools. Potential uses: text processing, analysis of fieldwork inputs (images, videos, audio recordings), generating data-driven visualizations, converting reports into infographics and audio summaries.	**8. Model ownership and open-source considerations** Are they commercial models with regular updates and support? Or are they transparent open-source options)that offer greater control but may require more technical management)?
3. Cost and budget considerations Consider both immediate costs and how expenses may scale as your usage increases	**9. Explainability and AI reasoning** Is the model reasoning process transparent or does it provide only final outputs without explanation? Does it have built-in citation capabilities ?
4. Adaptability and customization Does it allow customizing its behavior for specific evaluation needs or focusing on particular types of analysis? What is the learning curve required for customization vs the potential time savings?	**10. Bias, fairness, and ethics** Has it been evaluated for biases relevant to evaluation contexts? Particularly when working with marginalized communities or sensitive social issues. Consider testing to verify that AI outputs align with ethical evaluation standards of equity and fairness.
5. Data security and privacy challenges Does the AI provider retain the content? For how long? Do they use it to train future versions of their models?	**11. Environmental and efficiency considerations** AI models demands significant energy, contributing to environmental concerns. Evaluators should weigh energy efficiency alongside performance when selecting an AI model.
6. Language and accessibility Is it fluent in the specific languages and dialects relevant to evaluation context? Does it understand cultural nuances and regional expressions that may affect evaluation findings?	

Several organizations and independent evaluation offices have been testing the use of GenAI in the MERL cycle.[11] There is growing interest in developing applications and tools specifically designed for evaluation purposes, for the broader field and/or tailored to individual organizations. For example, some institutions are exploring customized LLM implementations and analysis systems for exclusive use within their organization although this remains relatively rare due to the significant resources and expertise required for such development.

4. Building AI Integration Competencies for Evaluators

The evaluation community currently finds itself at a critical juncture. Some respond to AI's emergence by focusing primarily on restrictions, specifying what cannot be shared with AI systems or how use should be constrained. While data protection and ethical considerations are essential, this position can overlook that systems and organizations already operate within digital ecosystems where similar risks exist. Instead of merely limiting AI use, this chapter advocates developing the competencies to use it responsibly while maintaining the professional judgment that defines quality evaluation work.

This is an initial Technical Note presented based on existing work for using GenAI in evaluation. Beyond what is presented below, it is recommended that some understanding of how AI systems are developed and trained be fostered at an institutional level[12]. **By approaching AI integration as a set of competencies to be continuously developed,** rather than simply regulations and tools to be deployed, **evaluators can enhance their work in creative ways, while maintaining the human expertise, ethical considerations, and contextual understanding that remain irreplaceable in quality evaluation.**

[11] Including several UN organizations, the World Bank, ICF, the MERLTECH initiative, private consultants, and evaluators working with software engineers.

[12] At the time of finalizing this Note, CGIAR's Artificial Intelligence (AI) Strategic Roadmap and AI Strategy are in the process of being finalized and approved as integral components of the Digital Transformation Accelerator, part of the 2025–2030 CGIAR Research Portfolio.

Figure 5. ChatGPT and plagiarism

Beyond this Note, trainings on how to use AI responsibly, ethically and productively exist both inside and outside the field of evaluation including by the MERL Tech Initiative and the EES, among others.[13] Ensuring functional literacy with the various AIs at department level is encouraged.

Source: Chris Lysy, Freshspectrum comic #302

4.1 Building the 'AI Muscles'

Figure 6. Foundational AI capacities: cheat sheet for evaluators

Developing what can be called an 'AI muscle' requires consistent practice and experimentation. Rather than simply adopting tools, **evaluators need to engage in ongoing discovery of what AI can do and apply these capabilities to their specific contexts**. This means regular experimentation with new approaches within their work and accepting that trial and error will be part of the process. What makes this

[13] While these organizations offer more of an *ad hoc* event program, it is worth keeping them in mind. For example, in January 2025, EES offered a workshop on 'AI as a working companion' facilitated by Silva Ferretti, while the MERL Tech initiative hosted events such as 'Strengthening Outcome Harvesting Analysis with AI Assisted Causal Mapping' presented by Steve Powell, Heather Britt, and Gabriele Caldas Cabral in April 2025.

competency particularly valuable is that AI capabilities are still evolving. By continuously exploring possibilities, evaluators can discover novel applications unique to evaluation challenges.

The most effective approach is progressive implementation. Beginning with simple, well-defined tasks that can be easily validated creates a foundation of confidence. For example, start with using AI to generate meeting notes from transcripts, then progress to more complex tasks such as thematic analysis of interview data. It has been reported by many evaluators that after an initial intensive learning period (described by one practitioner as "three sleepless nights of experimentation"[14]), they develop a clear intuition of how AI can best support their specific evaluation context and needs. This foundational competency creates the judgment necessary for effective integration.

4.2 Remain Aware of Risks and Sensitivities

Evaluation often involves confidential information and engagement with vulnerable populations, making awareness of data sensitivities particularly crucial when integrating AI. The challenge extends beyond traditional data protection concerns to understanding how AI systems process, potentially retain, or sometimes learn from information provided to them.

Most GenAI and technical AI tools on the market do not submit to confidentiality and non-disclosure requirements, including transcription tools. Before processing any confidential data (e.g., interview data, internal reports and coded data), check the terms and conditions, location of servers, whether there is a privacy mode available, and whether there is a better alternative. For some tools (e.g., WhisperAI[15] or Audiopen[16] for transcriptions), running local instances of the software is an option that safeguards the privacy of research participants.

Particularly important is recognizing different sensitivity levels across data types: personal identifiers, contextual information that could indirectly identify participants, and cultural knowledge that requires specific handling protocols all demand different approaches. **Never share data related to vulnerable persons or sensitive personal data (as defined by GDPR) with third-party AI services**. In this case, vulnerable people include anyone who may suffer adverse consequences if their personal data became publicly available (such as refugee asylum seekers and seasonal workers and others). Sensitive personal data can be understood as personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs as well as any data that can be processed to identify a single human being.

Consider **drafting and sharing a process to be transparent with evaluators and participants about AI use in data processing.** This includes drafting an opt out process. Ensure data fed into AI tools is protected and used only for its intended purpose. Understand data handling practices before using any AI tool. **Do not use any AI tools before understanding what happens to the data shared.**

[14] See Ethan Mollick (2024): Co-Intelligence: Living and Working with AI.

[15] WhisperAI is an open-source tool which can be run locally devices, meaning sensitive audio data never leaves the device and therefore makes it ideal for confidential research interviews.

[16] Audiopen offers powerful multilingual transcription capabilities and advanced summarization features. Its ability to handle multiple languages simultaneously makes it particularly valuable for international organizations working with several languages.

When managing evaluation data for AI processing, strive to follow FAIR[17] data principles, making data Findable, Accessible, Interoperable, and Reusable. Originally developed for broader scientific data management, these principles are increasingly relevant in AI contexts for several reasons:

- **Findable and Accessible:** Ensures that both humans and AI systems can locate and access data in a structured, documented way, which supports transparency and replicability.
- **Interoperable:** Facilitates the integration of datasets across different systems and tools, a critical need when using AI, which often relies on combining diverse data sources.
- **Reusable**: Encourages proper documentation, licensing, and metadata practices that clarify how data can be used (or not used), which supports ethical reuse in AI-driven analysis.

By aligning with FAIR principles, researchers can enhance the utility and integrity of their data while upholding security, privacy, and consent. This approach helps ensure that AI tools are used in a way that maximizes value and impact without compromising trust.

The goal is not to avoid AI use but to develop nuanced judgment about appropriate boundaries and protections for different types of evaluation data. This nuanced judgment must account for power dynamics and differentials in understanding. Participants may not fully comprehend how their data could be used or stored by AI systems, creating additional ethical responsibilities for evaluators.

4.3 Building the Toolbox and Practices

When integrating AI into evaluation, thinking should shift beyond standalone software applications to a landscape of evolving practices and possibilities.

While some AI functions come as clearly defined tools, most valuable applications emerge through experimentation and workflow refinement. For example, using AI to organize workshop sticky notes or develop coding frameworks represents a practice refined through trial and error rather than simply deploying a dedicated tool.

This effort requires cultivating both practical knowledge of established workflows and openness to continuous discovery. Evaluators at the forefront of AI integration are constantly testing new combinations: perhaps using conversational AI to refine evaluation questions, then visualization tools to present findings, all connected through customized prompting techniques developed through practice.

Building a toolkit involves not just collecting technologies, but developing a repertoire of tried-and-tested practices while remaining alert to emerging possibilities. The next chapter will detail specific evaluation activities enhanced by AI, emphasizing practical workflows that connect tools into coherent, effective approaches.

4.4 Designing AI into the Whole Evaluation Process

AI can be used at any step of the evaluation process; determining at what point AI tools should operate and where human expertise should lead, represents a critical competency. This is not simply about

[17] This approach is aligned to CGIAR's AI strategy currently under development and ensures that data can be effectively utilized across different contexts and systems, maximizing its value and impact while maintaining appropriate security and access controls.

efficiency; it is also **about creating intentional workflow s that leverage the strengths of both AI and human evaluators**.

Consider these key decision points across different *evaluation phases*: In the *design phase*, should interviews remain fully human-led to preserve the relational depth of the interaction, or could there be limited, structured scenarios where AI conducts standardized interviews with pre-coded logic paths (acknowledging that this approach may raise concerns about authenticity and rapport)? While technology exists, the heart of interviews lies in people, building trust and rapport with the interviewee, participation, and meaningful human interaction. Interviews are core to the evaluations, and for the time being, this human connection is irreplaceable. However, there is a growing presence of AI-driven tools, and it is worth critically reflecting on where, if at all, they may support, not replace, this work. During the *data analysis* phase, should data be immersed in first and AI be used for pattern validation, or should AI be let to generate an initial framework for human refinement? At the *reporting stage*, should AI assist by drafting sections for review, or should the narrative be crafted by humans with AI enhancing clarity and readability?

These decisions shape not just efficiency but the entire evaluation experience. **Decisions should consider factors such as context sensitivity, required expertise, available resources, stakeholder comfort, and the importance of relationship-building.** By thoughtfully determining where AI fits, human judgment is maintained in areas requiring nuance, while leveraging AI for consistency and efficiency.

4.5 Leverage AI for Knowledge Exchange

To ensure evaluation findings are used, they must be effectively communicated. AI offers new possibilities to enhance knowledge exchange, making insights more accessible and actionable for diverse stakeholders. In a multidisciplinary environment such as CGIAR, where researchers, evaluators, and decision-makers collaborate across different specializations and languages, AI can help bridge gaps in understanding throughout the entire evaluation process.

This competency involves using AI to translate technical findings into formats that different audiences can engage with. AI can also support real-time knowledge exchange by structuring stakeholder discussions, analyzing feedback, and facilitating multilingual collaboration. Beyond communication, AI offers opportunities to reconsider how evaluative reasoning itself is conveyed—moving beyond traditional inception and final reports to more dynamic, iterative ways of sharing insights throughout the process.

Mastering this competency means recognizing both AI's potential and its limits. While it can improve accessibility, AI cannot replace human judgment in ensuring that findings are accurately framed, contextually relevant, and free from misinterpretation. Evaluators who skillfully integrate AI into their communication strategies can expand the reach of their work, foster more inclusive dialogue, and ultimately enhance the impact of evaluation findings.

4.6 Negotiate and Agree Use

The use of AI needs to be negotiated and consensual across all parties involved in the evaluation process. This means engaging in detailed conversations amongst commissioners, managers, and evaluators about specific applications-not just general principles. As the evaluation community is learning how to use AI, **it is important to collectively establish dos and don'ts based on the specific context, content, and setup of each evaluation.**

This negotiation serves as an opportunity to create a more collective understanding about what AI can do and to sensitize stakeholders about the challenges of increasingly pervasive use of AI in lives and organizations. It can also be an opportunity to demonstrate positive possibilities not already considered. By

making these negotiations explicit rather than implicit, evaluators not only establish boundaries but also **contribute to broader AI literacy among the stakeholders they serve**.

As the AI field develops, these negotiations will likely benefit from frameworks and examples; however, the core competency remains: the ability to engage in open, specific conversations about AI integration that respect all perspectives while advancing the understanding of both AI's potential and its limitations.

This agreement needs to extend beyond the evaluation team to include stakeholders involved in the evaluation. Their understanding and comfort with the use of AI affects both the process and the **perceived legitimacy of findings**: this inherently requires methodological flexibility, as stakeholder engagement may necessitate adjustments to planned AI applications even after initial agreements are established.[18]

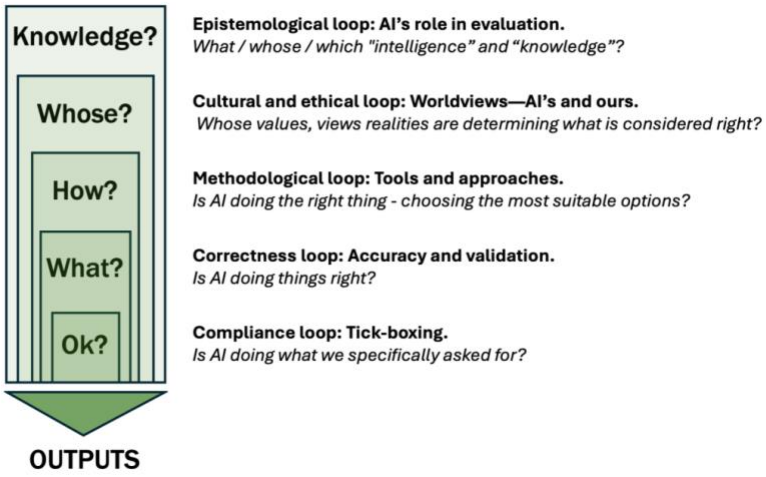
Box 4. Sample of AI use disclosure agreement for consultants to evaluation function in CGIAR

"The consultant agrees to responsibly and ethically use AI technologies in the course of his/her work under this contract, ensuring transparency, fairness, and accountability. The consultant must take all necessary measures to protect data privacy, confidentiality, and security, in compliance with applicable laws and CGIAR's policies. Any use of AI that involves personal or sensitive data must be disclosed and approved in advance. The consultant further agrees to promptly inform IAES of any significant developments, findings, or concerns related to AI use that may impact the project, data security, or ethical considerations."

4.7 Supervision at Different Levels

With AI integration requires enhanced supervision skills at every stage of the evaluation process. Accountability for quality remains with the evaluator and the evaluation manager, both requiring the ability to guide AI through effective prompting, monitor how it processes information, and critically assess its outputs. As AI becomes more integrated into evaluation practice, these supervision skills may become the most valuable competency for maintaining the quality and rigor of the evaluation.

Figure 7. Framework to highlight the different levels of supervision



Source: Silva Ferretti, 2024, AI Supervision Loops

[18] This approach suggests the need for methodological flexibility in evaluation design. Even when AI usage is agreed upon during inception phases, evaluators should be prepared to adjust approaches based on stakeholder feedback during implementation, potentially revisiting and refining methodological statements as the evaluation progresses.

Effective AI supervision operates across multiple interconnected loops:

- **Compliance Loop**: *Is AI doing what we specifically asked for?* This basic loop checks if outputs align with pre-established requirements. Despite appearing simple, AI's probabilistic nature means outputs are not always reproducible, making consistency verification important.
- **Correctness Loop**: *Is AI doing it right?* Evaluators are responsible for the accuracy and intentionality of all content, including AI-assisted work, considering that AI-generated content can be inaccurate, misleading, or lack nuance (including AI hallucinations). This loop focuses on validating outputs and catching errors, requiring strategies such as sampling and anomaly detection while avoiding automation bias.
- **Methodological Loop**: *Is AI doing the right thing?* Employ AI as a supportive tool in the evaluation process, not as a replacement for thorough research or expert judgment. This loop ensures AI-suggested methodologies and approaches are contextually relevant and appropriate. Preferably only use AI to assist the work when having enough specialist insights to properly evaluate its contributions.
- **Cultural and Ethical Loop**: *Whose values and views are determining what is right?* Recognize and actively counteract inevitable social biases in AI systems, particularly those that may affect evaluations in diverse cultural and geographical contexts. This loop assesses whether AI's decisions and outputs are culturally appropriate and whose perspective they reflect. This bias pertains both to your interaction with the AI system (e.g., prompt input) as well as in the use of AI outputs in the work.
- **Epistemological Loop**: *What/whose/which 'intelligence' and 'knowledge'?* This highest-level loop questions how AI influences the nature of evaluation itself. Is AI enhancing or undermining our understanding of change? Is it narrowing or expanding our perspectives? This loop interrogates AI's role in knowledge production and its impact on evaluative thinking.

Within each loop, consider which tasks can remain AI-driven and where human intervention is essential, the varying complexity of supervisory tasks, and the range of approaches from automated monitoring to guiding AI's decision-making frameworks.

4.8 Transparently Documenting AI Use

Systematic documentation of how AI is used throughout the evaluation process serves multiple essential functions. It creates transparency about where and how AI contributed, supports learning about effective integration, demonstrates how data sensitivities were addressed and generates more trust and accountability about the overall process.

This step unfortunately tends to be neglected. For example, there is a significant difference between using AI for editing versus using it for producing substantial write-ups. For transparency, it is recommended to acknowledge when AI helps to write content (properly supervised).

During this experimental phase of AI adoption in evaluation, documentation is particularly valuable for building collective knowledge. By tracking which tasks were AI-assisted, how outputs were verified, and where human judgment modified AI contributions, the field can develop better practices and protocols over time. Without this documentation, valuable learning is lost and opportunities for improvement go unrecognized.

The most effective documentation approaches recognize that different AI applications require different documentation standards. This presents a significant challenge. AI can be employed across a spectrum from light editing assistance to entire processes of data analysis and interpretation. Using AI for basic editing has different implications than employing it for substantive content generation or complex

analytical work. Despite this complexity, clarity about how AI was used at each stage builds essential context for interpretation of findings and maintains the integrity of the evaluation process.

It is advisable to generate a transparency policy regarding how AI has been used at each stage of the evaluation as good practice. Funders and donors may start requesting AI disclosure and other related information in the coming future. **While it may be tempting to overlook AI's minor contributions, establishing comprehensive documentation practices now will serve the field as AI integration deepens and becomes more complex.**

It is advisable to encourage an organization-wide strategy for the responsible use of AI tools at the research centers- beyond the scope of this document.[19]

5. When to use AI in Evaluation Phases

The types of AI that may be employed in evaluations are unpacked are organized by **evaluation stages** rather than by AI type, as AI capabilities often span multiple stages of the **evaluation workflow**. The table below outlines critical considerations for integrating AI into evaluation workflows, emphasizing **adaptability, scale, and human-AI collaboration**.

With promising capabilities to support evaluation work, it is important to understand that AI tools are evolving, and their efficacy varies considerably depending on specific contexts, data types, and user expertise.[20] The challenge for evaluators is not just mastering these new capabilities but developing the judgment to deploy them appropriately knowing when a particular analytical approach adds genuine insight rather than unnecessary complexity (see annexed Glossary).

Table 3. Integrating AI into evaluation workflows

Consideration	Key Points
Tools	Focus on AI functions (e.g., causal mapping, qualitative coding) rather than specific tools. The landscape evolves quickly—use general AI assistants or online searches to find current tools.
Processes and workflows	Map the evaluation needs to AI functions first, then design efficient workflows. Note that some platforms (e.g., Otter, Qualia) integrate multiple functions. Tasks may require single prompts or multi-step processes.
Scale considerations	AI works at different scales—methods that work for small datasets may not scale well. A process that works perfectly for analyzing ten interviews may become unwieldy with hundreds of documents. Technical limitations also apply—a Google Notebook may be able to handle a 'mini-RAG' for 50 papers, but larger literature reviews would require different technical setups. At the

[19] At the time of writing, an organization-wide AI strategy is being developed and distributed across the CGIAR system. Given the rapidly evolving nature of AI technologies and their regulatory landscapes, such strategies typically require regular revisions and updates. The current Note aims to provide interim guidance specifically for evaluation work while acknowledging that it will need to align with broader organizational frameworks once formalized.

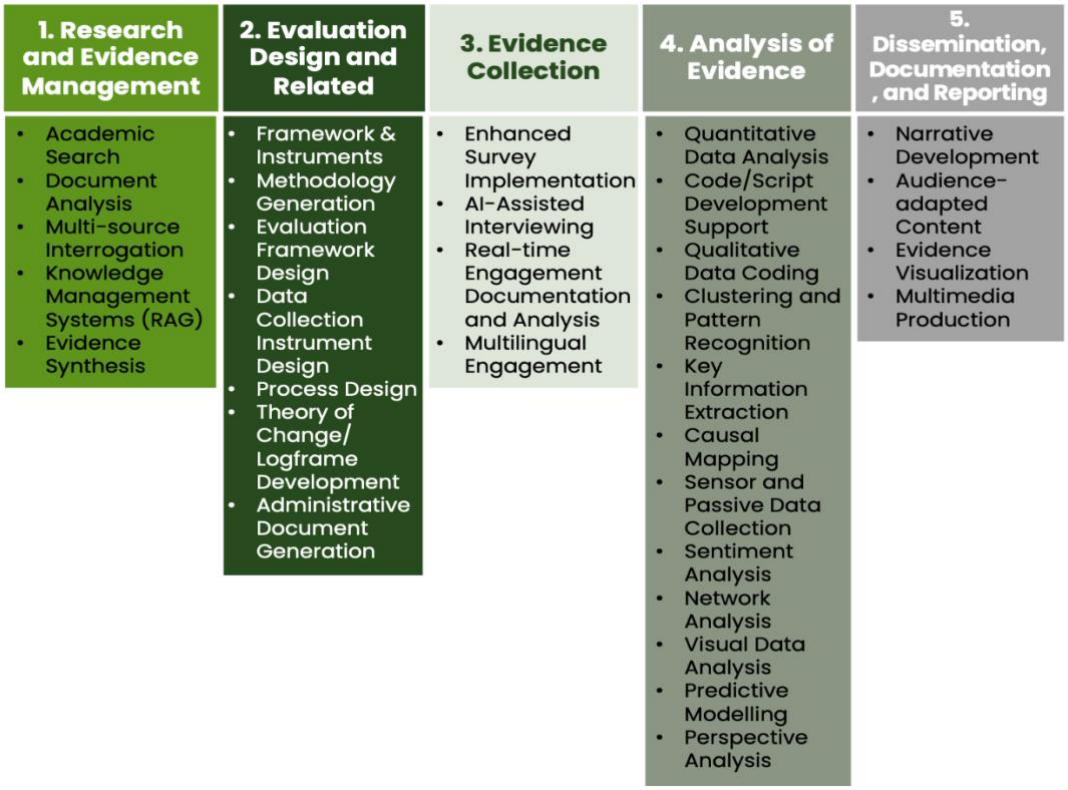
[20] There is now a fundamental expansion of the evaluator's analytical toolkit. AI now allows evaluators to rapidly iterate between analytical approaches, combine multiple methodological frameworks, and explore data through various theoretical lenses simultaneously, enabling deeper engagement with methodological pluralism. Evaluators can now test multiple analytical frames against the same data to see how different assumptions shape interpretations, design custom analytical processes that blend methodological traditions, or apply specialized analytical techniques without extensive technical training. This expansion democratizes advanced analytical approaches, making sophisticated methods accessible to evaluators working in resource-constrained contexts and potentially reducing methodological divides.

Consideration	Key Points
	highest scales, organizations may need to invest in custom or even proprietary models and substantial computing infrastructure—an approach some larger institutions are already pursuing for their work. Additionally, as processing volume increases, supervision becomes more challenging, requiring more structured quality control processes.
AI/Human collaboration	Different functions require varying degrees of human-AI interaction. Some tasks involve continuous iteration (qualitative data coding typically requires multiple refinement cycles), while others can run more independently once properly configured (transcription or standardized interviews). Understanding this interaction pattern helps set appropriate expectations for time investment and workflow design.
Reliability	AI's performance varies significantly across functions. Some applications consistently prove more reliable (text editing, summarization, basic quantitative analysis), while others remain challenging (evidence synthesis or visual design requiring conceptual integration). This variability helps you anticipate where minimal supervision might suffice versus where extensive validation will be necessary.

5.1 AI Uses per Evaluation Phase

Notably, in the tables that follow, capabilities are described in terms of what AI tools 'can' do under ideal circumstances and standard evaluation flows, rather than what they will consistently achieve. The success of these applications depends heavily on the evaluator's skill in selecting appropriate tools at the right phase, crafting effective prompts, and critically assessing outputs.

Figure 8. AI uses by evaluation phases with entry points



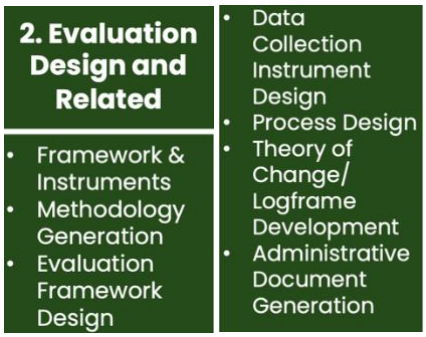
5.1.1. Research and Evidence Management

This phase focuses on using AI to enhance how evaluators identify, organize, and synthesize relevant information. The tools listed below support literature search, document analysis, and evidence integration across diverse sources, enabling faster and more comprehensive knowledge management.

Table 4. Research and evidence management

	Description
Academic search	AI locates relevant literature across multiple databases and increasingly compiles findings into structured literature reviews. Systems like DeepSeek and Gemini can connect search functions directly to analysis capabilities, helping evaluators move from finding sources to synthesizing content more efficiently.
Document analysis	AI processes documents in various formats to extract structured information.
Multi-source interrogation	Tools such as Google NotebookLM or Claude Projects allow evaluators to query across document collections, answering specific questions using multiple texts simultaneously. This helps locate evidence on topics and cross-check information between different sources.
Retrieval-Augmented Generation (RAG)	AI-powered RAG systems create organization-specific knowledge bases that maintain connections to source documents. These functions like enhanced enterprise search systems that provide direct answers while citing relevant sources. RAG combines information retrieval with text generation. Unlike standard large language models that rely solely on their training data, RAG systems supplement AI responses by first retrieving relevant documents from a knowledge base, then using that information to generate more accurate, contextual, and verifiable responses. This approach significantly reduces hallucinations and enables more transparent sourcing of information.
Evidence synthesis	An emerging capability where AI attempts to integrate findings across multiple sources, highlighting areas of consensus and disagreement. While still requiring significant human oversight, these systems help evaluators organize and compare evidence from diverse sources.

5.1.2. Evaluation Design, Research and Evidence Management



AI can assist in designing robust and context-appropriate evaluations by generating methodological frameworks, instruments, and process designs. The tools in this table help streamline planning efforts and enhance solid methodologies and conceptual clarity during the design phase.

Table 5. Evaluation design and related framework and instruments

	Description
Methodology generation	AI recommends appropriate evaluation designs and mixed-method approaches based on evaluation questions and context. It can suggest data collection techniques aligned with specific methodological needs and help evaluators consider approaches they might not have initially considered.
Evaluation framework design	AI creates structured evaluation matrices that map questions to data sources, develop indicators for key constructs, outline data collection plans, and draft analytical frameworks.
Data collection instrument design	AI develops tailored data collection instruments including surveys, interview guides, focus group protocols, and observation frameworks. Quality improves when AI is provided with clear parameters about the target population and specific information needs.
Process design	AI suggests approaches for stakeholder engagement, designs workshop agendas, and outlines facilitation techniques. This extends beyond individual instruments to comprehensive engagement strategies that encourage diverse participation.
Theory of change / logframe development	AI assists in capturing program logic. It generates theories of change and logic models based on evaluation data, allowing evaluators to compare designed versus emergent program mechanisms. This helps surface implicit assumptions, validate causal pathways, and identify disconnects between program design and implementation reality.
Administrative document generation	AI drafts project management documents including Terms of Reference, workplans, timelines, budget templates, and meeting agendas. These can significantly reduce administrative workload.

5.1.3. Evidence Collection

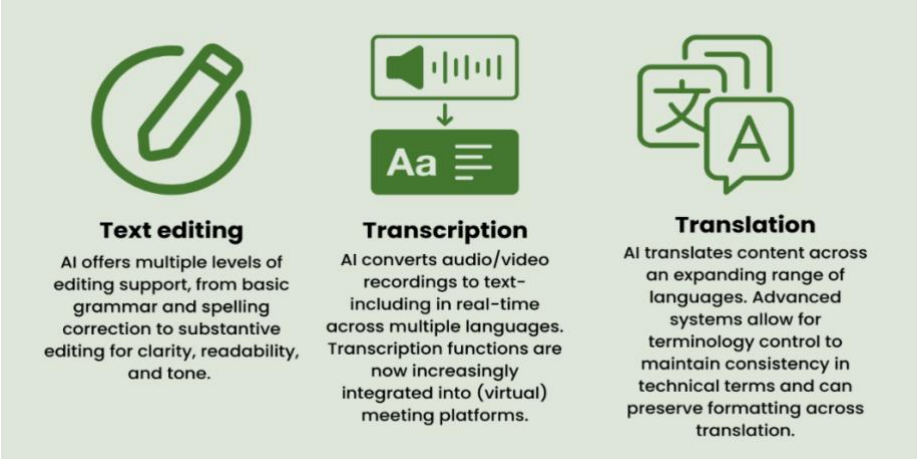
During evidence collection, AI enables real-time documentation, multilingual engagement, and intelligent data capture. The tools highlighted below can augment traditional methods, particularly in large-scale or multilingual evaluations.

Table 6. Evidence collection

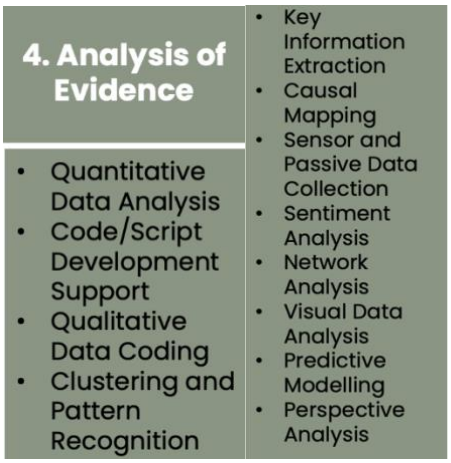
	Description
Enhanced survey implementation	AI improves survey processes through adaptive questioning, real-time validation of responses, multilingual support, and immediate preliminary analysis.
AI-assisted interviewing	AI applications now conduct interviews directly, adapting questions based on participant responses–with different interfaces (e.g., text based, avatars and video) and different input formats (e.g., text, audio).
Real-time engagement documentation and analysis	Tools such as Otter.ai or similar platforms monitor and analyze (virtual) engagements in real-time, providing transcripts, creating meeting summaries, and extracting key findings as the conversation unfolds.

	Description
Multilingual engagement	All the above can be enhanced by multilingual data collection and stakeholder engagement through real-time interpretation, while maintaining conceptual consistency.

Figure 9. Foundational AI use-text processing



5.1.4. Data Analysis



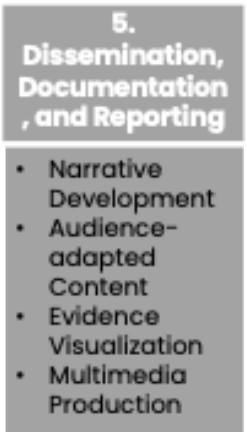
AI significantly accelerates both quantitative and qualitative analysis by supporting coding, clustering, extraction, and modeling tasks. The tools presented here help evaluators explore patterns, test assumptions, and generate insights with greater efficiency.

Table 7. Analysis of evidence

	Description
Quantitative data analysis	AI supports the entire quantitative analysis process from data cleaning and preparation to identifying trends, testing hypotheses, and generating statistical outputs.
Code/script development support	AI helps evaluators write and optimize code for specialized analysis, whether in statistical packages, data transformation tools, or visualization platforms. This is distinct from qualitative coding and focuses on technical implementation of analytical approaches.
Qualitative data coding	AI analyzes and organizes qualitative data by applying coding frameworks consistently across large text volumes. It can identify patterns in the data while allowing evaluators to refine coding approaches through iterative supervision and feedback.

	Description
Clustering and pattern recognition	AI identifies natural groupings in data without predefined codes, revealing unexpected connections across multiple data types. Effective implementation requires clear criteria for meaningful clusters and iteration to refine the analysis.
Key information extraction	AI identifies specific pieces of information across extensive datasets, extracting key facts, metrics, and statements related to evaluation questions. This supports systematic evidence mapping and focused analysis on predefined topics.
Causal mapping	AI helps visualize and analyze relationships between variables, supporting theory of change validation and contribution analysis. This provides a structured approach to understanding how program elements connect to outcomes and impacts.
Sensor and passive data collection	AI increasingly enables the analysis of data from sensors, mobile devices, and digital platforms, providing additional data streams that complement traditional collection methods. This includes processing of geospatial data, movement patterns, and usage metrics.
Sentiment analysis	AI assesses emotional tone and opinions in qualitative data, helping evaluators understand how stakeholders feel about programs or interventions. This works across text from interviews, surveys, social media, and other sources to identify affective patterns.
Network analysis	AI maps relationships between actors or concepts, calculating network metrics, generating visualizations, and interpreting connection patterns. This helps evaluators understand stakeholder ecosystems and how information or influence flows within them.
Visual data analysis	AI extracts information from images, photos, charts, and other visual materials, supporting analysis of visual documentation collected during evaluation. This includes recognizing text in images and identifying themes in photographs.
Predictive modeling	AI forecasts potential program outcomes based on current data, helping identify early warning indicators and supporting scenario planning. This assists with adaptive management and understanding potential future trajectories.
Perspective analysis	AI can analyze data through different interpretive frameworks (e.g., feminist, equity-focused, or culturally specific lenses), helping evaluators consider alternative interpretations and expand their analytical perspective beyond initial assumptions.

5.1.5. Dissemination, Documentation, and Reporting



AI promises to fundamentally transform evaluation practice in the **dissemination, documentation, and reporting** steps. By enabling rapid adaptation of content across formats, languages, and complexity levels, AI opens possibilities for broader, more inclusive communication of findings. Evaluators can now share insights through diverse channels simultaneously, reaching audiences previously excluded from traditional evaluation communications. AI facilitates more dynamic and participatory engagement with findings, allowing for the sharing of preliminary results and the gathering of stakeholder feedback throughout the evaluation process rather than only at its conclusion. This creates opportunities for iterative sense-making and collective interpretation that can deepen understanding and ownership of insights. However, the evaluator's irreplaceable skill remains understanding the specific information needs of diverse stakeholders and the contextual factors that influence how knowledge leads to action. This deeply relational and contextual judgment, discerning which findings matter to whom, how they should be framed, and when they should be shared to maximize utilization, cannot be automated. As dissemination possibilities multiply, the evaluator's role as translator and bridge-builder between evidence and action becomes even more vital.

Table 8. AI in dissemination, documentation and reporting

Narrative development AI transforms analytical outputs into coherent narratives, connecting findings to create logical storylines and developing cohesive reports from draft notes and fragmented observations. This helps evaluators move from analysis to compelling communication.	**Audience-adapted content AI adapts evaluation content for different stakeholders, converting technical reports into briefing materials, presentations, or social media content while preserving key messages. This includes adjusting language complexity and emphasis based on audience needs.**
Evidence visualization AI generates appropriate visual representations of findings including charts, graphs, diagrams, and interactive displays. It can suggest visualization approaches based on data characteristics and audience needs, making complex information more accessible.	**Multimedia production** AI supports the creation of diverse dissemination formats including podcast scripts, video storyboards, infographics, and interactive presentations. This expands the reach of evaluation findings beyond traditional written reports.

5.2. A Practical Guide to GenAI Conversations: Beyond Single Prompts to Meaningful Dialogue

Amongst all AI applications, conversational AI is a transformative force that fundamentally changed how AI is understood. These systems, from Claude to ChatGPT and beyond, shifted AI from specialized tools to dynamic partners capable of nuanced understanding and responsive engagement.

Understanding how to effectively communicate with these systems is crucial not just for getting better results, but for developing a deeper appreciation of how these models operate and respond. While there is significant emphasis on 'prompt engineering', successful interaction may not be about crafting the 'perfect prompt' (especially since, due to inherent variability, even well-designed prompts may yield different results). Instead, 'prompting' should be viewed as an ongoing conversational process with an entity that functions most effectively when approached as a capable assistant.

Box 5. The journey toward effective AI interaction

Think of it as an assistant, not a command line: Approach AI models as a knowledgeable colleague or evaluation assistant, one with extensive capabilities but who benefits from clear context and iterative guidance. These systems are not merely executing commands but interpreting requests through their understanding of language and context. It is important to remember these are computational tools built on statistical pattern recognition, not conscious beings. This mindset helps maintain appropriate expectations while still leveraging their capabilities for information processing, analysis, and content generation.

Have a conversation: The most successful AI interactions unfold as dialogues rather than one-off requests. Initial outputs serve as starting points for refinement through follow-up questions, clarifications, and redirection. This back-and-forth process allows for exploration and adaptation that a single prompt, no matter how carefully crafted, cannot achieve. The dialogue approach to AI interaction is not about equality between human and machine, but about strategically applying uniquely human cognitive abilities to guide AI. By engaging in a conversational manner, distinctive human capabilities can be exercised for:

- Critical evaluation of generated content
- Contextualizing information.
- Identifying inconsistencies that the AI can miss.
- Drawing connections between disparate domains of knowledge.
- Applying ethical and cultural nuances to frame questions appropriately.
- Making intuitive judgments about relevance and applicability.

This dialogue-based approach, combined with the techniques above, creates the foundation for integrating AI effectively into stages of the evaluation process, as explored in the following sections.

Table 9. Conversations with AI

Ingredient	What it is About	Conversation Prompts
Clear asks	Specific requests that clearly indicate what AI needs to do. Focus on being precise and actionable, using directive verbs, and limiting each prompt to a single task. Clarity reduces misinterpretation and helps direct the AI's capabilities toward specific needs.	"Identify the top three barriers to adoption based on the transcript data." "Summarize this article." "Extract the key methodological approaches from this evaluation paper." "Compare these two case studies using the criteria provided."
Contextual information	Provide relevant background, constraints, and purpose to frame AI's understanding. Consider whether enough context about why this information is needed is given, and how it will be used. Context helps the AI understand the bigger picture and tailor responses appropriately to actual needs.	"This data comes from a three-year study across five clinics-a contextual analysis for reference is attached." "These are the ToRs of the evaluation being working on...based on these, please..." "I am preparing this analysis for program managers who have limited technical background but need to make funding decisions."
Clarifications	Actively seek to resolve ambiguities and ensure mutual understanding. Regularly assess whether what the AI is saying is understood and whether the AI is interpreting requests as intended. This two-way clarification process prevents misalignment and wasted effort.	• "Are instructions clear or is more information or detail needed?" • "Please rephrase the request, so I can check your understanding". • "I notice you focused on X, but I was actually more interested in Y. Can we redirect?" • "What additional information would help you to provide a more targeted response?"
Validation and verification	Be attentive to the accuracy and soundness of the information provided. Determine how to know if information is accurate, what evidence supports the conclusions, and whether the reasoning process is sound. This critical assessment helps identify potential errors, biases, or logical flaws.	• "Please cite the specific parts of the document that support this." • "The answer to this question seems a bit off. Can you tell me what your assumptions are?" • "Make reasoning explicit at each step." • "How confident are you in this conclusion, and what factors might limit its reliability?"
Challenge and openness	Question assumptions—in the AI's responses and in own thinking—while remaining open to revising the stance. Consider what assumptions may be made that could be questioned. This mutual critical examination improves the quality of analysis and prevents confirmation bias.	• "These are my findings based on this dataset." Play the 'devil's advocate'. • "Please analyze your own answer, what are the weaknesses and strengths?" • "What alternative interpretations may be equally valid given this evidence?"

Ingredient	What it is About	Conversation Prompts
		• "I have been assuming X is true—how would this analysis change if that assumption is incorrect?"
Zooming (in/out) and scenarios	Deliberately adjusting the scope and level of detail as needed throughout the conversation, or to fit to specific contexts and scenarios. This flexible perspective-taking helps ensure you are examining the issue at the most appropriate level.	• "Can we focus more on aspect X?" • "We need to narrow our focus to just the technical implementation." • "How would this apply in context X instead?" • "What would change in this analysis if we consider a ten-year timeframe instead of two years?"
Examples (are worth 1000 words!)	Consider whether showing an example would clarify expectations more effectively. Conversely, ask AI for illustrations of the output produced. Providing examples that demonstrate what is being looking for rather than just describing it. Examples create concrete shared understanding that abstract descriptions often cannot achieve.	• "Here is an example of the format I'm looking for: [example]." • "This is a sample of a previous data analysis. I want you to replicate the output." • "Can you show me a sample of what this would look like in practice? Here is how I approached similar analyses in the past—please use this as a reference."
Clarity on preferred format and style	A conversation benefits by clarity on the desired structure, length, style, and complexity of the responses (and the same applies for the output required). Note, that AI can be used in different languages, but this can affect outputs and quality of conversation. Formatting guidance helps ensure the response is immediately useful for specific needs.	• "I prefer short answers, in plain English." • "Present the answer as a bulleted executive summary." • "Format this as a table comparing the key differences." • "This is an academic discussion, adjust the style accordingly." • "I need this analysis in both English and French, with technical terms preserved in both languages."
Stated roles and standpoints	Ensure that the conversation is informed by a clear rationale, purpose, and stance of the conversant (this may not be so important for small tasks, but can be fundamental in deeper qualitative analysis, where the stance of the analysts had an impact on the results). This requires being clear about the stance and/or defining the perspective or expertise lens the AI should adopt when approaching the conversation. Role clarity creates appropriate framing and ensures analyses reflect relevant perspectives and priorities.	• "Be aware that my priority concern is to ensure sustainability." • "I am approaching this from a position of skepticism about the underlying theory!" • "My goal is to find practical solutions, not theoretical perfection." • "Act as an XXX concerned about YYY." • "Analyze this from the perspective of a program manager." • "I need you to consider this policy through both an equity lens and an implementation feasibility lens."

Table 10. Evaluator's conversational skills and attitudes for better engagement with AI

Skill	What it is About	Inner Thinking
Human metacognition	Using uniquely human judgment, intuition, and lived experience to guide the conversation and evaluate responses critically.	• "Does this align with my experience?" • "What contextual knowledge am I bringing that's missing here?" • "Something feels off about this conclusion-what could be wrong?" • "What background insights from my may enrich this analysis?"
Ethical awareness	Bring conscious attention to ethical dimensions that AI may miss or handle poorly, including considerations of bias, impact, and representation.	• "Are there ethical implications being overlooked here?" • "Whose perspectives may be marginalized?" • "Is the AI unintentionally steering toward a particular worldview? How might this analysis impact vulnerable groups?"
Politeness	Maintain a constructive conversational approach that research shows yields better results, without unnecessary deference or formality.	• "How would I phrase this to a knowledgeable colleague?" • "Am I expressing myself clearly without being unnecessarily demanding? " • "Is my frustration affecting the quality of our exchange?"
Patience and persistence	Balance continued effort with strategic pivots when needed, recognizing when to push forward and when to change course.	• "Is this approach getting us closer to what I need, or should we try something different?" • "What small adjustments might improve our direction?" • "When should I step back and reconsider our approach entirely?"
Balance	Find the right level of detail, context, and direction for productive exchange, including when to break complex problems into manageable parts.	• "Have I provided enough context without overwhelming?" • "Should I break this down into smaller questions or maintain the broader view?" • "Am I getting lost in details when a simpler approach might work better?"
Stance awareness	Be aware of the position, intentionality, needs, objectives, and the paradigm from which the conversation is being approached.	• "Have I clarified what I'm really trying to accomplish?" • "Does this direction serve my actual purpose?" • "Am I true to my intent and approach?"

Skill	What it is About	Inner Thinking
Thoughtfulness	Take time to consider what is really wanting to be known or accomplished, recognizing that careful consideration of one's own input dramatically impacts output quality.	• "What am I truly trying to understand?" • "Is this the right question to get me closer to my goal?" • "Have I taken enough time to frame this request effectively?" • "How could a more considered approach yield better insights?"
Adaptability	Adjust the approach based on what is working and what is not, learning from the conversation patterns that emerge.	• "What is most effective in our exchange so far?" • "Which approaches are yielding the best insights?" • "How can I modify my approach based on what I am learning about this conversation's dynamics?"

Within the evolving space of AI use in the evaluation industry, a few key spaces that are recommended to monitor for continuous learning and reflection. Selected ones are included in Box 6.

Box 6. Key spaces to follow for the latest use of AI in MERL and at CGIAR

- CGIAR-wide internal and external AI guidance and policy.
- The MERLTECH Initiative (Linda Raftree, and Zach Tilton)
- Stanford University HAI–AI Index Report.
- Silva Ferretti, independent consultant
- Fteval AI working group
- VOPES (EES and AEA in particular through ongoing publication and workshops).
- Journal Space (e.g. New Directions for Evaluation Special Issue: Evaluation and Artificial Intelligence).
- ICT4D.
- UN Independent Evaluation Offices, including UNFPA, UNDP, and others.
- World Bank Independent Evaluation Group.[21]

[21] See the WB IEG 2023 Blog Series Experimenting with GPT and Generative AI for Evaluation for more information. Of particular relevance is the AI decision tree created in 2023 and illustrating the recommended uses for GPT integration into certain steps of the evaluation.

Bibliography and Further Reading

CEN CENELEC. (2020). *Artificial Intelligence*. CEN-CENELEC. https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/ Accessed: December 2024

De Pagter, S., Schuerz, S., & Lampert, D. (2024). Considerations for the Use of AI Tools at the Centre for Social Innovation. fteval. https://doi.org/10.22163/fteval.2024.648

European Commission. (2021, April 21). *Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future*. European Commission. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence Accessed: December 2024

European Commission. (2019, April 8). *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

FAO. 2023. Using artificial intelligence to assess FAO's knowledge base on the technology accelerator. Rome. https://doi.org/10.4060/cc6724en

Feretti, Silvia. 2024. LinkedIn. AI Supervision Loops. https://www.linkedin.com/pulse/ai-supervision-loops-silva-ferretti-njxof/?trackingId=Bi%2BUk4fxT4KrBnn3KC6cUQ%3D%3D

Harvard University. (n.d.). *Generative Artificial Intelligence (AI) Guidelines*. Huit.harvard.edu. Retrieved December 7, 2024, from https://huit.harvard.edu/ai/guidelinespriority_highDate publishedclose Accessed: December 2024

IEEE SA. (n.d.). *Autonomous and Intelligent Systems (AIS)*. IEEE Standards Association. Retrieved December 7, 2024, from https://standards.ieee.org/initiatives/autonomous-intelligence-systems/ priority_highDate publishedclose Accessed: December 2024

ISO. (2017). *ISO/IEC JTC 1/SC 42 - Artificial intelligence*. ISO. https://www.iso.org/committee/6794475.html Accessed: December 2024

Koo, J., Magalhaes, M. C., & Singaraju, N. (2025, February 4). Assessment of how well Large Language Models (LLMs) answer questions related to gender equality and women's empowerment. CGIAR. https://huggingface.co/blog/CGIAR/llm-gender-equality-womens-empowerment

KU Leuven. (2024). *Responsible use of Generative Artificial Intelligence*. KU Leuven Onderwijs. https://www.kuleuven.be/english/education/student/educational-tools/generative-artificial-intelligence Accessed: December 2024

Mollick, E. (2024). Co-Intelligence: Living and working with AI. Portfolio (Penguin Random House).

OECD. (2024). *Artificial intelligence*. OECD. https://www.oecd.org/en/topics/policy-issues/artificial-intelligence.html Accessed: December 2024

Raimondo, E., Anuj, H., & Ziulu, V. (2023, August 30). Unfulfilled promises: Using GPT for synthetic tasks. Independent Evaluation Group (IEG). https://ieg.worldbankgroup.org/blog/unfulfilled-promises-using-gpt-synthetic-tasks

Science Business. (2023). *Commission gears up to confront the risks generative artificial intelligence poses to science*. Science|Business. https://sciencebusiness.net/news/ai/commission-gears-confront-risks-generative-artificial-intelligence-poses-science Accessed: Dec. 2024

The Influence of LLM Inputs on Outputs (n.d.) https://shelf.io/blog/understanding-the-influence-of-llm-inputs-on-outputs/ Accessed 19 December 2024

UNESCO. (2024). *Ethics of artificial intelligence*. Www.unesco.org; UNESCO. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics Accessed: December 2024

World Economic Forum. (n.d.). *Design of transparent and inclusive AI systems - AI Governance Alliance*. Initiatives.weforum.org. Retrieved December 7, 2024, from https://initiatives.weforum.org/ai-governance-alliance/home priority_highDatepublished close Accessed: December 2024

Annex 1. Glossary of Terms

Algorithm A formula or set of rules, procedure, processes, or instructions for solving a problem or for performing a task. In Artificial Intelligence, the algorithm tells the machine how to find answers to a question or solutions to a problem. In machine learning, systems use many different types of algorithms. Common examples include decision trees, clustering algorithms, classification algorithms, or regression algorithms. (Source: AI: A Glossary of Terms)

Artificial General Intelligence (AGI): AGI refers to an AI system that could reason, learn, and adapt across multiple domains, solving intellectual problems much like a human. While AGI does not yet exist, the rapid progress in ML and GenAI has fueled speculation and debate about whether time is moving closer to it.

Artificial Intelligence (AI) or machine intelligence: The broad concept of machines performing tasks that typically require human intelligence, such as reasoning, learning, and decision-making. Systems that display intelligent behavior by analyzing their environment and taking actions—with some degree of autonomy—to achieve specific goals. AI-based systems can be purely software-based or can be embedded in hardware devices. It uses machine and human-based data and inputs to:

- Perceive real and virtual environments
- Abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and
- Use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy. (Source: AI: A Glossary of Terms & OECD AI Principles).

AI Auditing: An emerging practice of assessing AI systems against criteria such as ethical principles, standards, or laws to evaluate risks (bias, fairness, security) and ensure compliance with regulations.

AI (or algorithmic) bias: Systematic and repeatable errors in AI systems that create unfair outcomes, such as placing privileged groups at systematic advantage and under-privileged groups at systematic disadvantage. AI bias can emerge from three primary sources:

1 **Data bias**: Occurs when training datasets are unrepresentative, incomplete, or reflect existing societal inequalities. This is often considered the most challenging source of bias to address, as AI systems learn directly from the patterns in their training data. When certain groups are underrepresented or misrepresented in these datasets, the resulting models perpetuate and sometimes amplify these inequalities. For example, facial recognition systems trained primarily on light-skinned faces may perform poorly for people with darker skin tones.

2 **Algorithmic bias**: Stems from the design choices, assumptions, and technical limitations built into AI algorithms themselves. Even with balanced data, mathematical formulations and optimization criteria can inadvertently prioritize certain outcomes over others, leading to discriminatory results. This includes how algorithms weigh different features or how they measure success.

3 **User bias**: Introduced when humans interact with AI systems, influencing how technology is deployed and interpreted. This includes how users frame questions to AI systems, interpret results, or make decisions based on AI outputs. User bias can reinforce existing prejudices even when using nominally "neutral" technology.

Source: EU-U.S. Terminology and Taxonomy for AI & Ferrara 2023)

Context window: The amount of text or data an AI model can process and remember in a single interaction, typically measured in tokens (roughly corresponding to words or parts of words).

Ethics washing: The practice where organizations feign ethical consideration or make misleading claims about their ethical practices to improve their public image, without implementing responsible actions.

Generative AI (GenAI): Neural networks that can generate high-quality text, images, and other forms of content based on the data they were trained on. Differently from traditional AI, GenAI models can process inputs to produce new content by predicting the likelihood of data typically appearing together. (Source: European Commission). GenAI refers to a subset of machine learning (ML) models that generate new content—such as text, images, code, and music—based on patterns learned from large datasets. Unlike traditional AI models that primarily classify or predict based on existing data, GenAI produces original and contextually relevant outputs by predicting the likelihood of elements occurring together. This capacity stems from architectures like Generative Adversarial Networks (GANs) and transformers. Although foundational technologies were developed over the past decade, GenAI gained mainstream attention with the launch of OpenAI's ChatGPT in late 2022. Since then, numerous tools—such as ClaudeAI, Perplexity, and DeepSeek—have expanded the accessibility and application of GenAI across diverse industries through intuitive, conversational interfaces. (Source: Adapted from the European Commission and other sources)

Hallucination: AI hallucination is a phenomenon wherein a large language model perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate (Source: IBM). A phenomenon where AI systems generate content that appears plausible but is factually incorrect or entirely fabricated, often presented with high confidence.

Iterative refinement: Viewing interaction as an ongoing process of improvement rather than a one-time exchange; identifying valuable aspects of responses, addressing gaps, and building on previous responses to progressively improve outcomes through dialogue.

Large Language Models (LLMs): A class of language models that use deep-learning algorithms and are trained on extremely large textual datasets. There are two types of LLMs:

- Generative LLMs: models that output text, such as the answer to a question or even writing an essay on a specific topic (typically unsupervised or semi-supervised, predict what the response is for a given task).
- Discriminatory LLMs: supervised learning models that usually focus on classifying text, such as determining whether a text was made by a human or AI. (EU-U.S. Terminology and Taxonomy for AI)

Machine learning (ML): Branch of AI and computer science which focuses on development of systems that can learn and adapt without following explicit instructions imitating the way that humans learn, gradually improving its accuracy, by using algorithms and statistical models to analyze and draw inferences from patterns in data. (Source: EU-U.S. Terminology and Taxonomy for AI). ML is a subset of AI where machines learn patterns from data without being explicitly programmed for every decision. ML has already been used in evaluation, particularly for data analysis and modeling. However, setting up such models traditionally required significant effort—both in terms of computational resources and expertise—often making them impractical or yielding limited returns.

Natural language processing: The ability of a machine to process, analyze, and mimic human language, either spoken or written. (Source: EU-U.S. Terminology and Taxonomy for AI)

Retrieval-Augmented Generation (RAG): A technique that connects Large Language Models to external, often real-time data sources such as search engines or institutional/internal documents,

retrieving relevant information to inform text generation. This allows for more trustworthy and factually grounded responses than relying solely on the LLM's internal training data.

Small Language Models (SLMs): Compact AI language models specifically designed for limited computational resources or specialized domains, often trained on targeted datasets for specific languages or use cases.

Trustworthy AI: AI system which comprises the trustworthiness of all processes and actors that are part of the AI system's life cycle. Characteristics of Trustworthy AI systems are: valid, reliable, safe, secure, resilient, accountable, transparent, explainable, interpretable, privacy-enhanced, and fair with managed harmful bias. They have three components:

- Lawful, ensuring compliance with all applicable laws and regulations.
- Ethical, demonstrating respect for, and ensure adherence to, ethical principles and values.
- Robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. (Source: EU-U.S. Terminology and Taxonomy for AI)

Annex 2. Non-Exhaustive List of AIs for Evaluation[22]

AI	MERL Stage	Description	Advantage	Disadvantage	Cost Effective	GDPR/ Privacy
ChatGPT	Various	Large language model for text generation and analysis	Versatile, can assist in brainstorming, drafting, proofreading, summarizing. Can also generate images including graphs based on data.	May produce biased or inaccurate content	Moderate / High (free tier available). Moderate individual cost.	Depends on usage and data handling. Is licensed to operate in the EU.
ClaudeAI	Various	Large language model for text generation and analysis. Particularly strong on text.	Versatile, can assist in brainstorming, drafting, proofreading, summarising. Very good at text and language-based tasks.	May produce biased or inaccurate content. Is not well connected to the internet so may be unable to generate sources for its claims.	Moderate individual cost. Very limited free tier	Yes, with proper data handling but until recently not licensed in the EU
Perplexity	Various	Large language model for text generation and analysis. Similar to ChatGPT. Good for initial searching of information as well connected to the internet.	Versatile, can assist in brainstorming, drafting, proofreading, summarizing. Integrates software's from several models into one place with the option to choose which one to use.	May produce biased or inaccurate content. Desktop app is not great. Limited to publicly available information, requires internet connection	Moderate individual cost	Yes
OtterAI	Research/data collection	AI-powered transcription service	Real-time transcription, speaker identification	Accuracy can vary with accents. Limited hours of transcript based on license held.	Moderate / High	Yes, with proper settings

[22] Please note that some of this software is redundant in a way given CGIAR IAES Microsoft subscriptions or existing subscriptions (e.g., SurveyMonkey has some AI integrated capabilities). Realistically, a number of GenAIs operate largely in the same way with varying amount of success, and for transcribing field interviews Otter or Sonix may be good options for transcription given Microsoft Teams reliance on the internet to work. Other services will be added, as this is intended as a live document.

AI	MERL Stage	Description	Advantage	Disadvantage	Cost Effective	GDPR/ Privacy
SonixAI	Research/data collection	Automated transcription and translation	Supports multiple languages (exceptional number), fast turnaround	May require manual corrections	High	Yes
Survey CTO	Data collection/ surveys	Mobile data collection platform with built-in quality checks and automated validation	Robust data encryption Automated quality checks Integration with analysis tools. Mobile-first design	Steep learning curve Requires technical setup Limited AI features compared to newer tools	High	Yes, with set up
Caribou Digital	Data analysis	AI-powered data analysis platform	Specialized for development sector data	Limited to specific use cases		
AILYZE		AI-powered evaluation and analysis platform	Specialized for evaluation tasks	New platform, may have limited features compared to established tools	High	Yes
Microsoft Teams Transcription	Various		Integrated in CGIAR Microsoft transcription	Might not perform well in some languages		
AudioPen	Key Informant Interview (KII)/data collection	Transcription and good summary capabilities.	It will transcribe several major world languages and has capacity to give the summary in English.	Best features are on the pro version		Yes
Gemini	Various	Google's multimodal AI system supporting text, images, audio, and video	Strong multimodal capabilities for analyzing diverse data types. Integrated with Google ecosystem.	May struggle with highly technical content in specialized domains	Free tier available	Follows Google's data privacy policies
DeepSeek	Research/data analysis	Search and analysis tool with AI capabilities	Connects search directly to analytical capabilities, supporting efficient information synthesis	Newer platform with potentially less robust performance than established alternatives	xxx	Data policies may vary
Elicit	Research/data analysis	Research assistant specialized in literature review and academic search	Optimized for academic research, provides structured literature analysis	Limited to research domains, not versatile for other evaluation tasks		
MAXQDA	Data analysis	Qualitative data analysis software with AI-assisted coding	Established in qualitative research with integrated AI features that enhance	Steep learning curve, significant cost for full version	High	Yes, with proper Data implementation

AI	MERL Stage	Description	Advantage	Disadvantage	Cost Effective	GDPR/ Privacy
			rather than replace researcher judgment			
Atlas.ti	Data analysis	Qualitative data analysis software with AI-assisted coding	Rich visualization features and growing AI capabilities	Complex interface requiring training	High	Yes, with proper data implementation
LLaMA 2	Various (local implementation)	Meta's open-source language model	Can be self-hosted for greater privacy control and customization	Requires significant technical expertise and computing resources to implement effectively	Low to moderate depending on implementation	Excellent when self-hosted properly
Mistral	Various (local implementation)	Open-source language model	Efficient performance with lower computational requirements than some alternatives	Requires technical setup and management	Low to moderate depending on implementation	Excellent when self-hosted properly
Ollama	Various (local implementation)	Tool for running AI models locally	Enables local operation of various open-source AI models without data leaving your system	Requires technical knowledge to set up and operate	Free	Excellent (data remains local)
Hugging Face	Various (local implementation)	Platform hosting thousands of AI models with easy deployment options	Access to vast library of specialized AI models for different tasks	Requires development expertise to use effectively	Low to moderate (many free models available)	Depends on implementation
WhisperAI	Transcription	Open-source speech recognition system	Can be run locally for complete privacy, supports multiple languages	Requires technical setup	Free	Excellent when self-hosted

Annex 3. CGIAR IAES GenAI Prompt Record by MERL Task

Quick Guide **to Prompt Engineering**

1 **Be specific:** Instead of asking something broad e.g., "Tell me about AI," say, "Explain how AI can be used to improve customer service in retail." The more specific question gathers a more tailored response.

2 **Use direct requests**: If a list is needed, say so. For example, "List three key challenges of data integration" is more effective than "What are the challenges of data integration?" This makes it easier for the model to provide exactly what is needed.

3 **Add instructions**: Guide the model by adding instructions directly in the prompt. For instance, "Explain machine learning in simple terms, as if you're explaining to someone with no technical background." This helps adjust the complexity and tone of the response to suit the audience.

4 **Sometimes, showing is better than telling**. If the model should follow a specific format or provide certain types of answers, include examples in the prompt.

5 **Provide a template**: If a response is needed in a specific format, provide a template. For example: "Give me a summary of this article in bullet points, like this: 1) Main Idea, 2) Key Detail, 3) Conclusion." This helps the model understand exactly how the output is wanted.

6 **Illustrate with scenarios**: When a particular tone or level of detail is needed, add an example. For instance, "Explain this concept like you would if you were talking to a high school student. For example, imagine you're explaining it in a classroom setting."

Table 11. Example GenAI prompts by task

Evaluation stage/ Task	AI type / Software Sugges-tion	Example Prompt	Notes/ Comment
Meeting and Key Informant Interview (KII) note generation (detailed)	GenAI (any) ChatGPT ClaudeAI Perplexity	Please go through the provided transcript/document and split it into six equal sections. Then provide a detailed and comprehensive question and answer summary for each section, following the same format used previously. Begin each section breakdown with 'Section X/6' where X is the section number. Within each section, clearly separate questions from answers using 'Q:' for questions and 'A:' for answers. For the answers, provide thorough context, nuanced details, specific examples, and in-depth analysis of the key points and back-and-forth exchanges. Aim to capture 15% or 20% more detail and elaboration than what would be considered a standard level of detail. Do not merely summarize, but delve deeper into the underlying meanings, implications, rationales and thought processes expressed. Elucidate the interconnections between different points made. Explain acronyms, jargon or concepts that may need clarification for a general audience. Provide relevant background information to set the context where needed. Use excerpts from the original transcript judiciously to substantiate the details provided. Overall, strive to give an exhaustive and insightful account that leaves little room for further questioning	The prompt was generated using ClaudeAI and works consistently on this software. It works with relative consistency on ChatGPT as well, but when the conversation becomes too long, it reportedly does not function well. Details can be revised/changed based on what is needed.

Evaluation stage/ Task	AI type / Software Sugges-tion	Example Prompt	Notes/ Comment
		on the topics covered in each section. Only include what is in the provided document and do not guess or hallucinate.	
Desk Review – generating summaries of provided written evidence	GenAI (any) ChatGPT ClaudeAI Perplexity	Go through the provided document and split it into ten equal sections. Then provide a comprehensive summary for each section, following the same format across each section to compile a [desired number] word executive summary excluding background chapters and recommendations. Begin each section breakdown with 'Section X/[total number]' where X is the section number. Elucidate the interconnections between different points made. Explain acronyms, jargon or concepts that may need clarification for a general audience. Provide relevant background information to set the context where needed. Overall, strive to give an exhaustive and insightful account that leaves little room for further questioning on the topics covered in each section. Only include what is in the provided document and do not guess or hallucinate.	Similar prompts to this can be generated also to help speed up references to a particular theme, topic, issue of interest. Optional: ask it to reference page numbers for each piece of evidence provided in the summary.
Triangulating evidence across multiple documents	GenAI (any)	Review these [X] documents and identify supporting evidence for the following themes: [list themes]. For each piece of evidence, please indicate the source document in brackets.	If the documents are large, things can become lost. Going through each file individually with a similar prompt asking for evidence by theme can generate a better list. This list can then be run a second time.
Textual revision	GenAI (any) Calude Opus is particula rly good at this.	Enhance this paragraph with evidence from [specify documents]. Make all new additions in bold text and keep the original text intact. Please cite the source document(s) in brackets for each addition.	
Textual revision	GenAI (any) Claude Opus is particula rly good at this.	Combine these separate pieces of information into one cohesive paragraph, maintaining all key evidence while eliminating redundancy. Keep source citations and highlight additions in bold. Only include what is in the provided document/attachment/pasted material and do not guess or hallucinate.	This may need finetuning in the moment and you may want to specify a base paragraph for the additions to be carried out. Variations of this prompt work well. Always ask for it to cite.

Evaluation stage/ Task	AI type / Software Sugges-tion	Example Prompt	Notes/ Comment
Systematic validation of specific claims	GenAI (any)	Pull out the specific evidence from [document name/attached document] that supports this statement: [insert statement]? Quote directly from the source material.	
Comparing findings across documents	GenAI (any)	Analyze these documents and create a summary of key findings organized by [themes/criteria], noting where findings converge or diverge across documents. Please cite sources for each finding. Only include what is in the provided document/attachment/pasted material and do not guess or hallucinate.	This prompt may need to be repeated for each document and then be integrated at the end.

Independent Advisory and Evaluation Service
Alliance of Bioversity International and CIAT
Via di San Domenico, 1 00153 Rome, Italy
IAES@cgiar.org
https://iaes.cgiar.org/