IFRC

AI-Enabled Evaluation Workflow: Methods, Insights, and Institutional Learning from a multi-county Final Evaluation

ENYINNAYA SAMUEL

INTERNAL KNOWLEDGE NOTE

1. Introduction

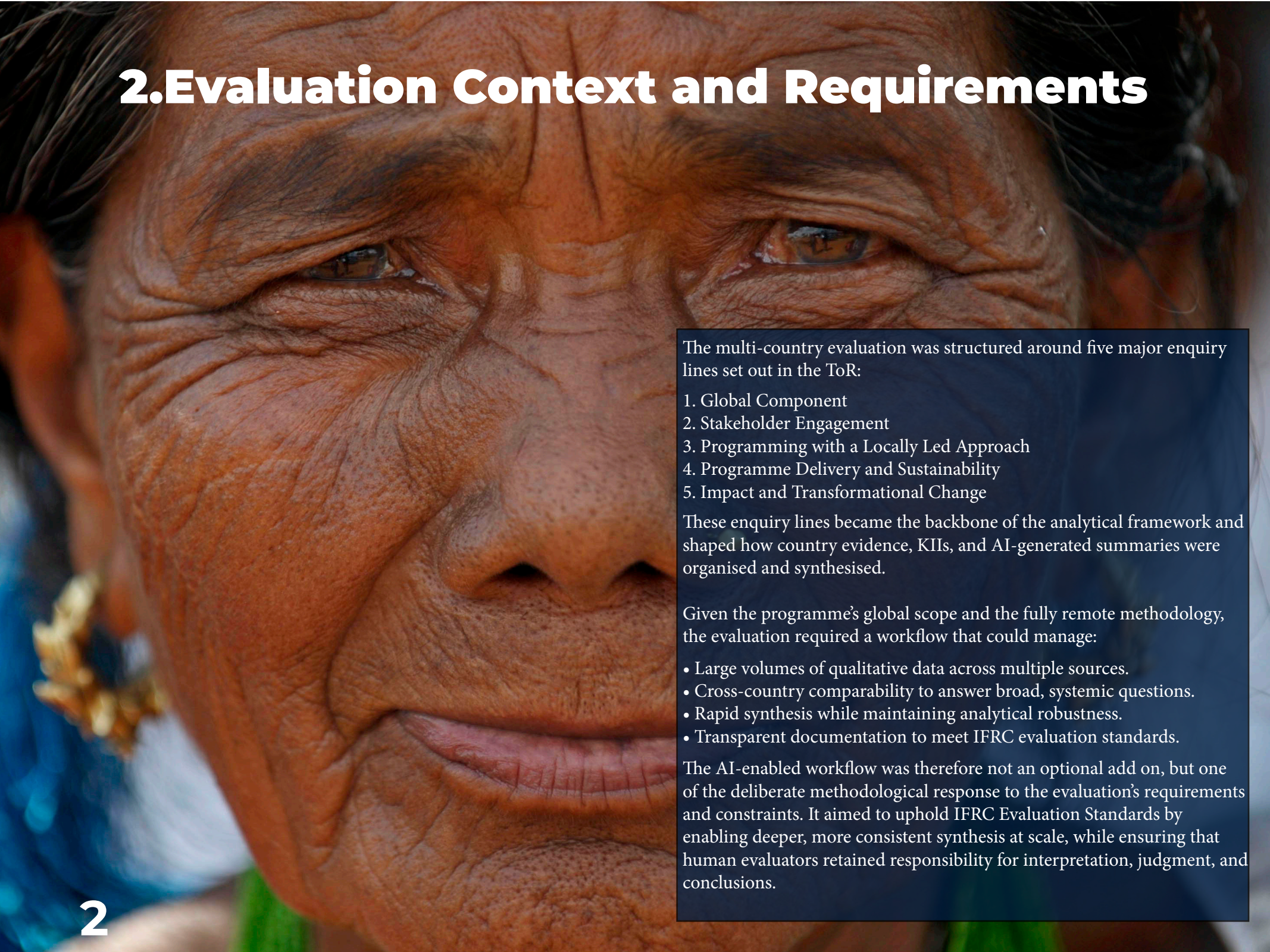
This Knowledge Note documents the AI-enabled qualitative analysis workflow conceptualised, designed, and implemented by the evaluation's Data & AI Lead (Evaluations intern) during a large multi-country evaluation. This multi-country initiative spanning 24 countries, aimed at strengthening National Society (NS) capacities, advancing locally led humanitarian action, and promoting system-wide transformation across five programmatic pillars. The evaluation assessed performance, impact, sustainability, and partnership modalities, guided by the evaluation questions outlined in the Terms of Reference (ToR).

The evidence landscape for this evaluation was both extensive and highly dispersed. With 36 interim country reports, 75 Key Informant Interviews (KIIs), and a diverse body of regional and global documentation, all totalling approximately 170 documents, the evaluation team faced the dual challenge of synthesizing large volumes of qualitative information while maintaining analytical coherence across multiple contexts. Compounding this were operational constraints: the evaluation was conducted entirely remotely, with no field visits due to resource limitations.

To address these challenges, the evaluation team developed an AI & Knowledge management -enabled mixed-tool workflow, conceptualised and executed by the team's Data & AI Lead. The workflow integrated retrieval-augmented AI systems, structured knowledge repositories, power automation, and multiple Large Language Models (LLMs) to enhance analytical depth, improve cross-country comparability, and maximize efficiency under strict time constraints.

This Knowledge Note provides a transparent account of the methodology, including the AI and data management tools used, the prompt engineering process, the power automation, evidence triangulation techniques, risks and safeguards employed, and key lessons learned. It is intended to support future evaluations within IFRC and contribute to organizational learning on responsible, ethical, and effective use of AI for complex qualitative analysis.

2.Evaluation Context and Requirements



The multi-country evaluation was structured around five major enquiry lines set out in the ToR:

1. Global Component
2. Stakeholder Engagement
3. Programming with a Locally Led Approach
4. Programme Delivery and Sustainability
5. Impact and Transformational Change

These enquiry lines became the backbone of the analytical framework and shaped how country evidence, KIIs, and AI-generated summaries were organised and synthesised.

Given the programme's global scope and the fully remote methodology, the evaluation required a workflow that could manage:

- Large volumes of qualitative data across multiple sources.
- Cross-country comparability to answer broad, systemic questions.
- Rapid synthesis while maintaining analytical robustness.
- Transparent documentation to meet IFRC evaluation standards.

The AI-enabled workflow was therefore not an optional add on, but one of the deliberate methodological response to the evaluation's requirements and constraints. It aimed to uphold IFRC Evaluation Standards by enabling deeper, more consistent synthesis at scale, while ensuring that human evaluators retained responsibility for interpretation, judgment, and conclusions.

3. AI & Knowledge Management Architecture / Workflow

The workflow consisted of five integrated stages, each contributing to a coherent evidence pipeline. Retrieval-augmented tools, structured databases, and multiple LLMs were combined to produce a transparent and auditable analytical process.

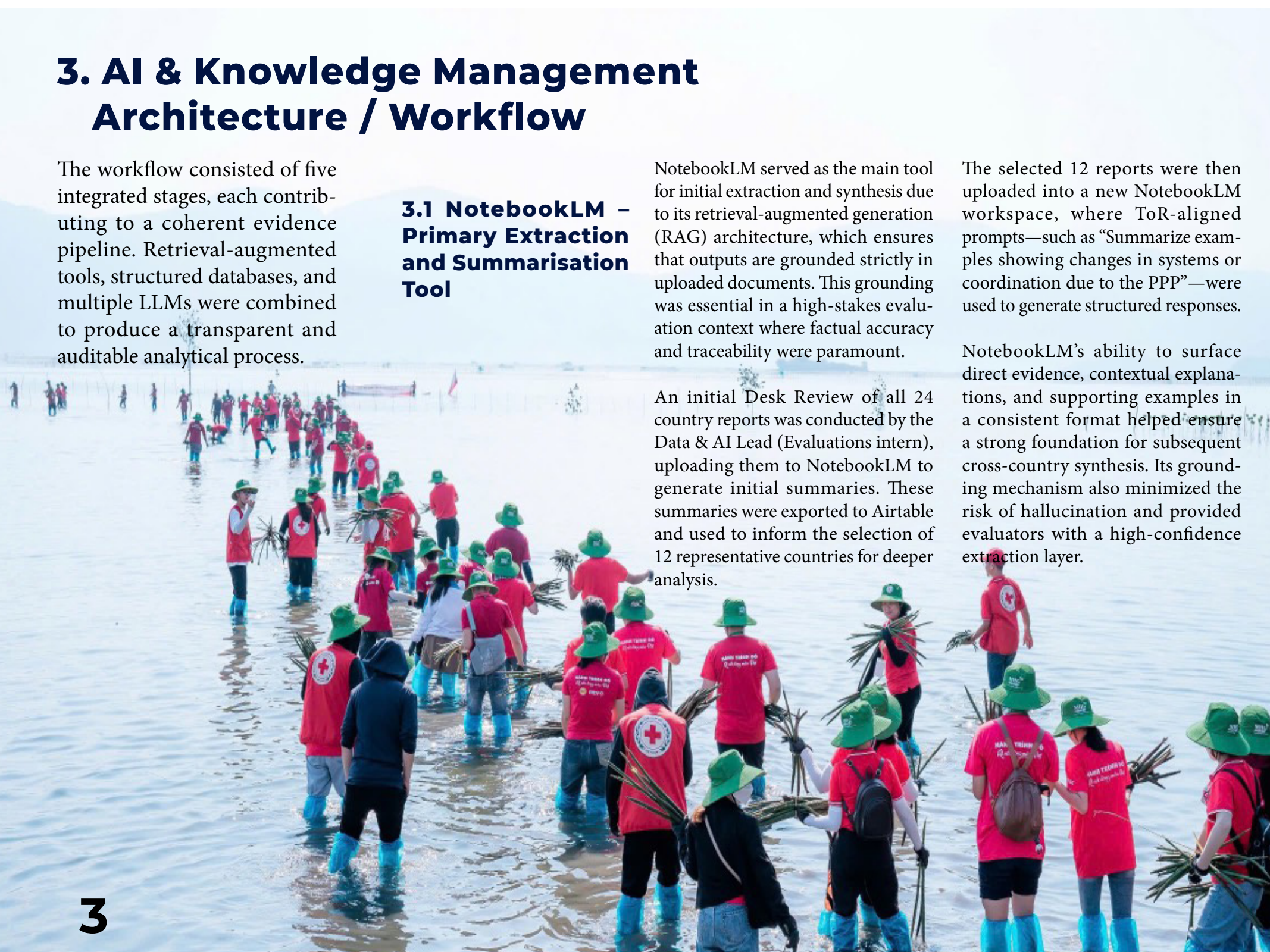
3.1 NotebookLM – Primary Extraction and Summarisation Tool

NotebookLM served as the main tool for initial extraction and synthesis due to its retrieval-augmented generation (RAG) architecture, which ensures that outputs are grounded strictly in uploaded documents. This grounding was essential in a high-stakes evaluation context where factual accuracy and traceability were paramount.

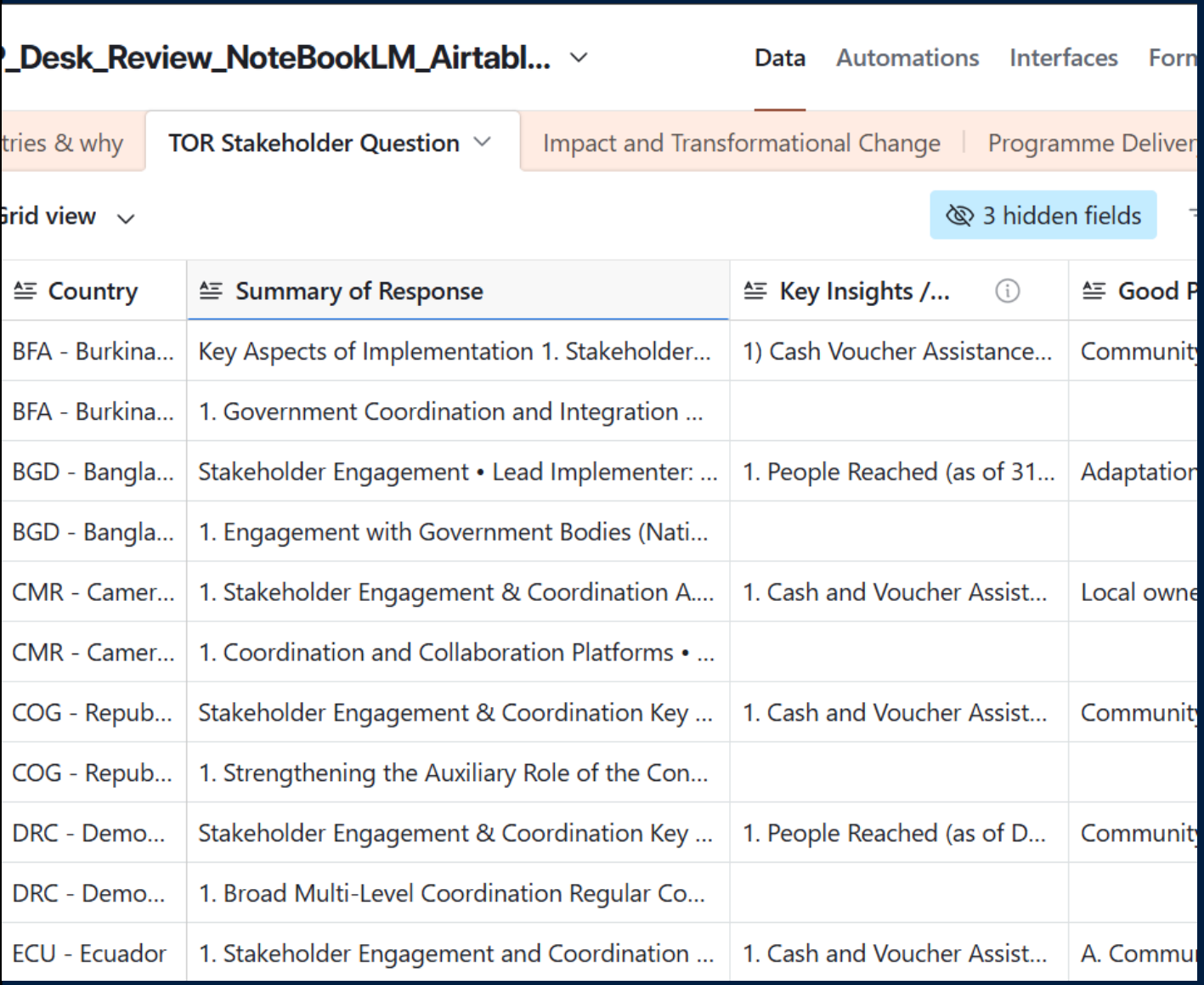
An initial Desk Review of all 24 country reports was conducted by the Data & AI Lead (Evaluations intern), uploading them to NotebookLM to generate initial summaries. These summaries were exported to Airtable and used to inform the selection of 12 representative countries for deeper analysis.

The selected 12 reports were then uploaded into a new NotebookLM workspace, where ToR-aligned prompts—such as "Summarize examples showing changes in systems or coordination due to the PPP"—were used to generate structured responses.

NotebookLM's ability to surface direct evidence, contextual explanations, and supporting examples in a consistent format helped ensure a strong foundation for subsequent cross-country synthesis. Its grounding mechanism also minimized the risk of hallucination and provided evaluators with a high-confidence extraction layer.



3.2 Airtable – Central Repository and Analytical Backbone



Country	Summary of Response	Key Insights /...	Good P
BFA - Burkina...	Key Aspects of Implementation 1. Stakeholder...	1) Cash Voucher Assistance...	Community
BFA - Burkina...	1. Government Coordination and Integration ...		
BGD - Bangla...	Stakeholder Engagement • Lead Implementer: ...	1. People Reached (as of 31...	Adaptation
BGD - Bangla...	1. Engagement with Government Bodies (Nati...		
CMR - Camer...	1. Stakeholder Engagement & Coordination A....	1. Cash and Voucher Assist...	Local owne
CMR - Camer...	1. Coordination and Collaboration Platforms • ...		
COG - Repub...	Stakeholder Engagement & Coordination Key ...	1. Cash and Voucher Assist...	Community
COG - Repub...	1. Strengthening the Auxiliary Role of the Con...		
DRC - Demo...	Stakeholder Engagement & Coordination Key ...	1. People Reached (as of D...	Community
DRC - Demo...	1. Broad Multi-Level Coordination Regular Co...		
ECU - Ecuador	1. Stakeholder Engagement and Coordination ...	1. Cash and Voucher Assist...	A. Commu

Airtable served as the central evaluation database. It stored:

- Country-level evidence extracted from NotebookLM
- KII metadata
- Thematic tags and country attributes
- Intermediate analytical outputs
- AI-generated case study building blocks

By converting qualitative evidence into a structured relational database, Airtable allowed the Data & AI Lead (intern) to:

- Query evidence across countries and themes
- Rapidly retrieve ToR-aligned findings
- Cross-reference interview data with document findings
- Maintain a transparent and auditable chain of evidence

The addition of Airtable Omni AI enabled natural language querying across the full repository, significantly accelerating the identification of cross-country patterns and reducing manual search time. This meta-layer of AI capability transformed Airtable from a storage system into a dynamic analytical asset.

3.3 Comparative Testing and Selection of Large Language Models

Recognizing that not all LLMs perform equally in qualitative evaluation contexts, the Data & AI Lead (intern) conducted a structured comparison of ChatGPT, Microsoft Copilot, Gemini, and Perplexity. Each model was tested using standardized prompts applied to the same country-level document.

The models were assessed for:
- Accuracy and grounding in source material
- Ability to process long, nuanced qualitative inputs
- Hallucination rate
- Consistency
- Structural coherence
- Analytical depth

Across repeated trials, Gemini demonstrated the strongest performance, scoring 8/10 in structured evaluation tests. It consistently produced grounded, well-organized, and balanced outputs, making it the preferred tool for generating country case studies and coding KII transcripts.

This testing phase ensured that the evaluation relied on the most reliable model for its needs—an essential step in safeguarding methodological integrity.

3.4 Case Study Production Through Structured Prompts

To ensure analytical comparability across the 12 country case studies, we developed a progressively refined prompt system.
- **Prompt 1** generated initial insights but lacked structure and consistency.
- **Prompt 2** improved structure but created sequencing issues (it required human input between chapters).
- **Prompt 3** delivered consistently high-quality outputs with explicit chapter instructions, word-count guidance, balanced analysis, and strict reliance on the uploaded file.

Prompt 3 became the standard, generating case studies with the following structure:
- A 150–200 word contextual overview
- Four analytical chapters (~1,000 words each), aligned to the ToR enquiry lines
- A 300–350 word conclusions and lessons learned section

Gemini-generated drafts were then reviewed and validated manually to ensure contextual accuracy, and eliminate inconsistencies. These case study outputs were also vetted on their fairness and accuracy by country focal points. This hybrid model retained the efficiencies of AI-assisted drafting while ensuring human-led interpretation and rigor.

3.5 AI-Assisted Coding of Key Informant Interviews

Due to limited resources, all KIIs were conducted remotely via Microsoft Teams and in some cases, Google Meet. The resulting 75 transcripts required a consistent, efficient coding process.
Using a structured coding prompt, transcripts were uploaded into Gemini (and occasionally ChatGPT) to:

- Categorize responses under the enquiry lines
- Identify and preserve illustrative quotes
- Ensure neutrality in tone while eliminating subjective interpretation
- Cross-reference responses applicable to multiple themes

This approach significantly reduced manual workload and improved the consistency of coding across interviews. Compared to traditional qualitative coding methods, this reduced manual coding time by an estimate of 70%, increasing consistency across interviews. The AI-assisted coding allowed evaluators to focus their efforts on interpretation, triangulation, and synthesis, rather than labor-intensive categorization.

3.6 Final Synthesis via NotebookLM (94 Sources)

The final synthesis stage consolidated 94 sources, including country reports, case studies, coded interviews, and other contextual documents. NotebookLM's RAG architecture ensured that all synthesis outputs were fully grounded in the compiled evidence.

Using ToR-aligned prompts, NotebookLM generated cross-country analyses, thematic insights, and stakeholder-level syntheses. These outputs were reviewed, validated, and integrated into the draft Final Evaluation Report. NotebookLM's citation and source-linking features strengthened transparency and auditability, enabling evaluators to verify claims and trace them back to the originating documents with ease.

3.7 Automating Document Notifications via Power Automate and SharePoint

To ensure the evaluation team stayed up to date with newly uploaded or revised resources, a simple automation was developed using Microsoft Power Automate connected to the shared SharePoint repository used by the Evaluation Team. The objective was to avoid manual folder checking and ensure that all three evaluators were consistently working with the latest templates, tools, and evidence.

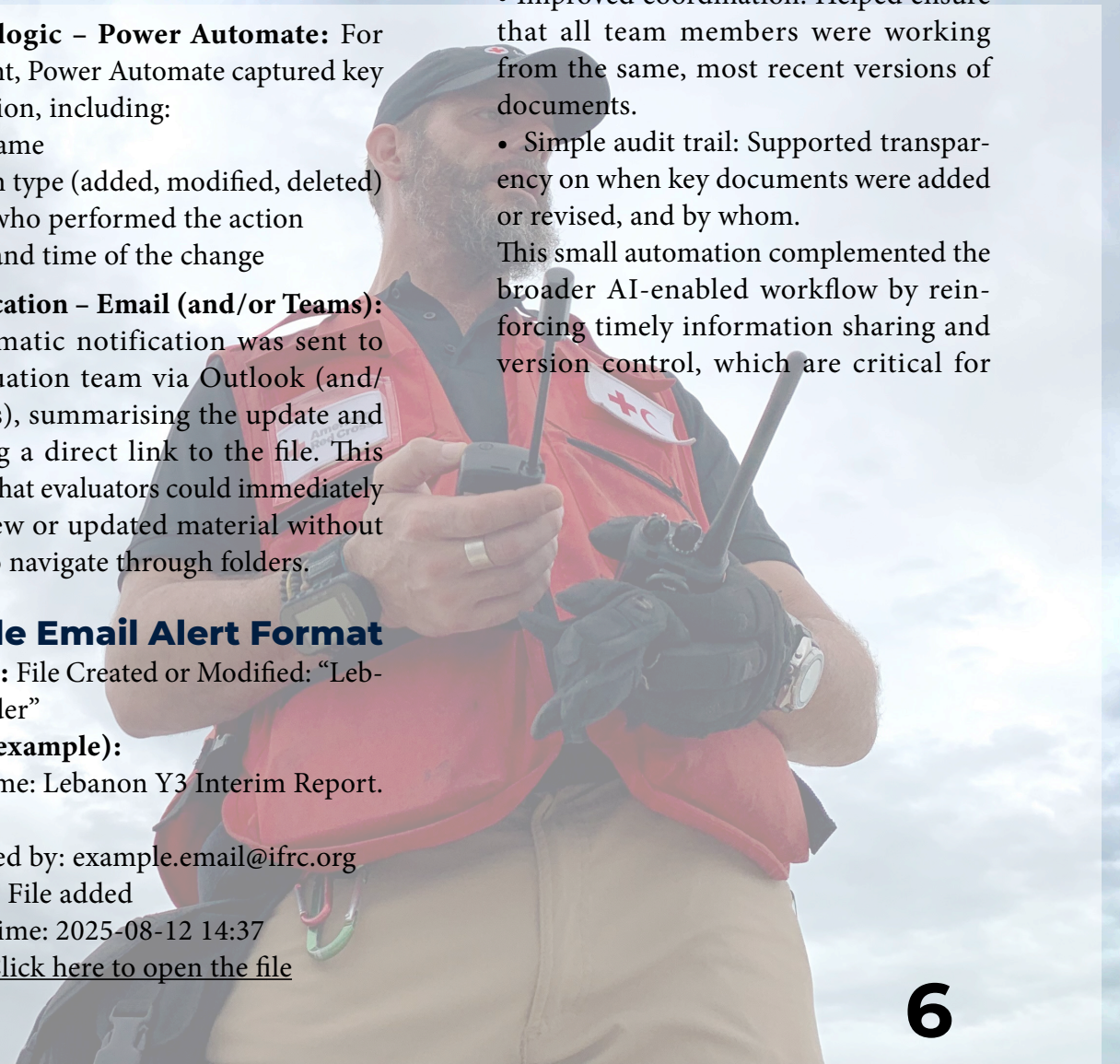
Automation workflow

1. **Trigger – SharePoint:** A flow was configured to trigger automatically whenever a file was added, edited, or deleted in the designated PPP evaluation SharePoint folder.
2. **Flow logic – Power Automate:** For each event, Power Automate captured key information, including:
   - o File name
   - o Action type (added, modified, deleted)
   - o User who performed the action
   - o Date and time of the change
3. **Notification – Email (and/or Teams):** An automatic notification was sent to the evaluation team via Outlook (and/or Teams), summarising the update and including a direct link to the file. This ensured that evaluators could immediately access new or updated material without having to navigate through folders.

Sample Email Alert Format

- **Subject:** File Created or Modified: "Lebanon folder"
- **Body (example):**

o File name: Lebanon Y3 Interim Report.docx
o Modified by: example.email@ifrc.org
o Action: File added
o Date/Time: 2025-08-12 14:37
o Link: Click here to open the file



Added value for the evaluation

- Real-time awareness: The team knew immediately when new evidence, tools or templates were uploaded.
- Reduced manual checking: Eliminated the need to repeatedly monitor SharePoint folders.
- Improved coordination: Helped ensure that all team members were working from the same, most recent versions of documents.
- Simple audit trail: Supported transparency on when key documents were added or revised, and by whom.

This small automation complemented the broader AI-enabled workflow by reinforcing timely information sharing and version control, which are critical for

5. Benefits of the AI-Enabled Workflow

5.1 Efficiency Gains

The integration of AI significantly improved the speed and consistency of qualitative analysis. First-pass coding of interview transcripts, which traditionally requires extensive manual review, was accelerated by an estimated 70. AI-assisted retrieval of cross-country evidence reduced the time required for comparative analysis, enabling faster identification of shared patterns and divergences. NotebookLM's ability to extract structured summaries from lengthy documents further reduced the manual burden associated with thematic review. Automated notifications from the Evaluation SharePoint repository (via Power Automate) reduced time spent manually monitoring document uploads and helped ensure the team consistently

5.2 Analytical Depth and Consistency

The structured prompt system ensured that all 12 country case studies followed a consistent analytical format aligned with the ToR. This allowed evaluators to conduct more rigorous cross-country comparisons and ensured that no enquiry line received uneven attention. AI-supported triangulation also facilitated deeper analysis of relationships between stakeholder engagement, localisation, sustainability, and systemic change.

5.3 Transparency and Traceability

Airtable functioned as the central audit trail, storing AI-generated outputs alongside source documents, tags, and metadata. All prompts used across AI tools were preserved, enabling replication of the workflow. NotebookLM's citation features ensured that synthesized outputs were transparently linked to original evidence, strengthening the evaluation's credibility and adherence to IFRC Evaluation Standards. Document change notifications—and optional logging of file events—contributed to a simple audit trail of when key resources were added, updated or removed from the shared workspace.

5.4 Scalability

The architecture developed for this evaluation is readily applicable to other multi-country or large-scale evaluations. Its modular nature—combining RAG tools, structured databases, and LLMs—allows adaptation to thematic reviews, rapid assessments, and other evidence-intensive exercises.

6. Risks, Limitations, and Mitigation

The team approached AI use with caution, recognizing inherent risks such as hallucination, bias, and data confidentiality. These risks were mitigated through:

- The use of RAG-enabled systems (NotebookLM, Gemini) that reduce hallucination risk.
- Human validation of all AI outputs, ensuring interpretive integrity.
- Careful management of transcript content to avoid exposing sensitive information.
- Iterative prompt testing to minimize ambiguous instructions.
- Cross-model comparisons to identify tool limitations and verify accuracy.

The workflow's transparency and documentation also strengthened ethical compliance and methodological rigour.

7. Lessons Learned

Three overarching lessons emerged from this work. First, prompt engineering is a critical analytical skill, requiring iterative refinement and testing to achieve high-quality outputs. Second, structured repositories like Airtable dramatically improve the coherence and efficiency of AI-assisted workflows, enabling rapid triangulation and traceable evidence synthesis. Third, AI should complement—not replace—human judgment. While AI excels in synthesis and pattern detection, evaluative interpretation must remain in human hands to ensure contextual nuance, ethical sensitivity, and credibility.

8. Recommendations for Institutional Adoption

To strengthen IFRC's use of AI in future evaluations, the following steps are recommended:

- Develop an AI-for-Evaluation Standard Operating Procedure (SOP) covering workflow design, ethical safeguards, and verification processes.
- Invest in database systems (e.g., Airtable) to structure and manage qualitative evidence at scale.
- Build staff capacity in AI literacy, prompt engineering, and quality assurance.
- Adopt hybrid workflows that integrate AI, structured data systems, and human oversight.
- Explore secure, IFRC-governed retrieval-augmented environments for sensitive evaluations.


G. Granada