AI in Evaluation: From interviews to insight

04 June 2026

www.glocalevalweek.org

unicef
for every child

AI assisted Survey

By Ali Safarnejad & Daniel Alonso Valckx

Problem statement

The Methodological Dilemma: Complex development settings demand mixed methods to avoid the quantitative "Cartesian trap" (missing rich context), yet traditional qualitative interviews remain vulnerable to subjective interpretation and bias.

The Integration Burden: Combining these methods often multiplies their respective representation problems, shifting the burden to the analytical stage to reconcile imperfect data.

The Operational Reality: Capturing authentic "lived experiences" across geographically dispersed, fragile/emergency settings is logistically demanding and costly to coordination.

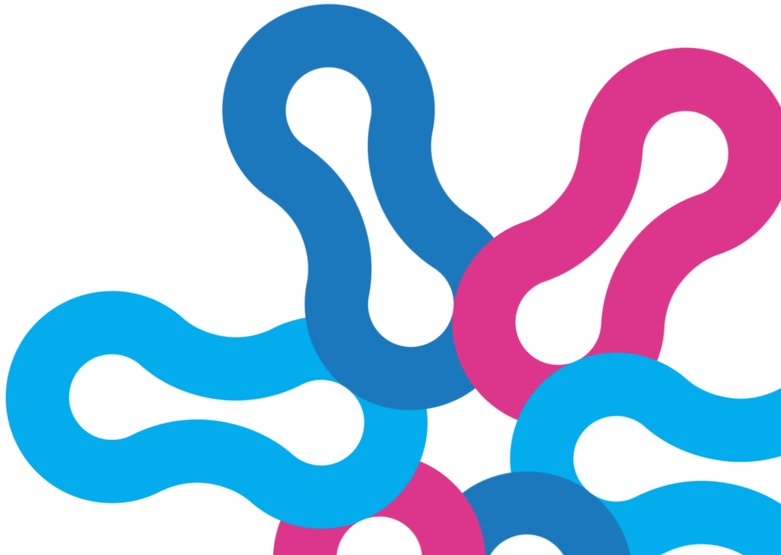
The Gap: A data collection methodology that can deliver the contextual depth of a skilled human interview, but with the procedural consistency, scale, and speed of an automated survey.

Overview of the AI solution

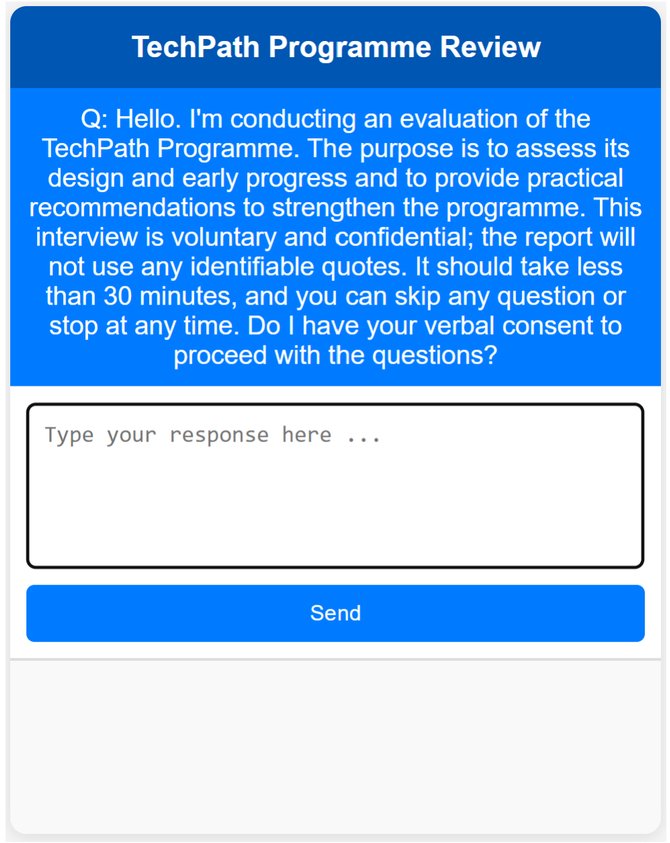
The AI Solution: An AI-powered conversational agent that conducts semi-structured interviews through natural language conversations. Successfully deployed across multiple countries and development contexts with UNICEF staff, implementing partners, youth, and community members.

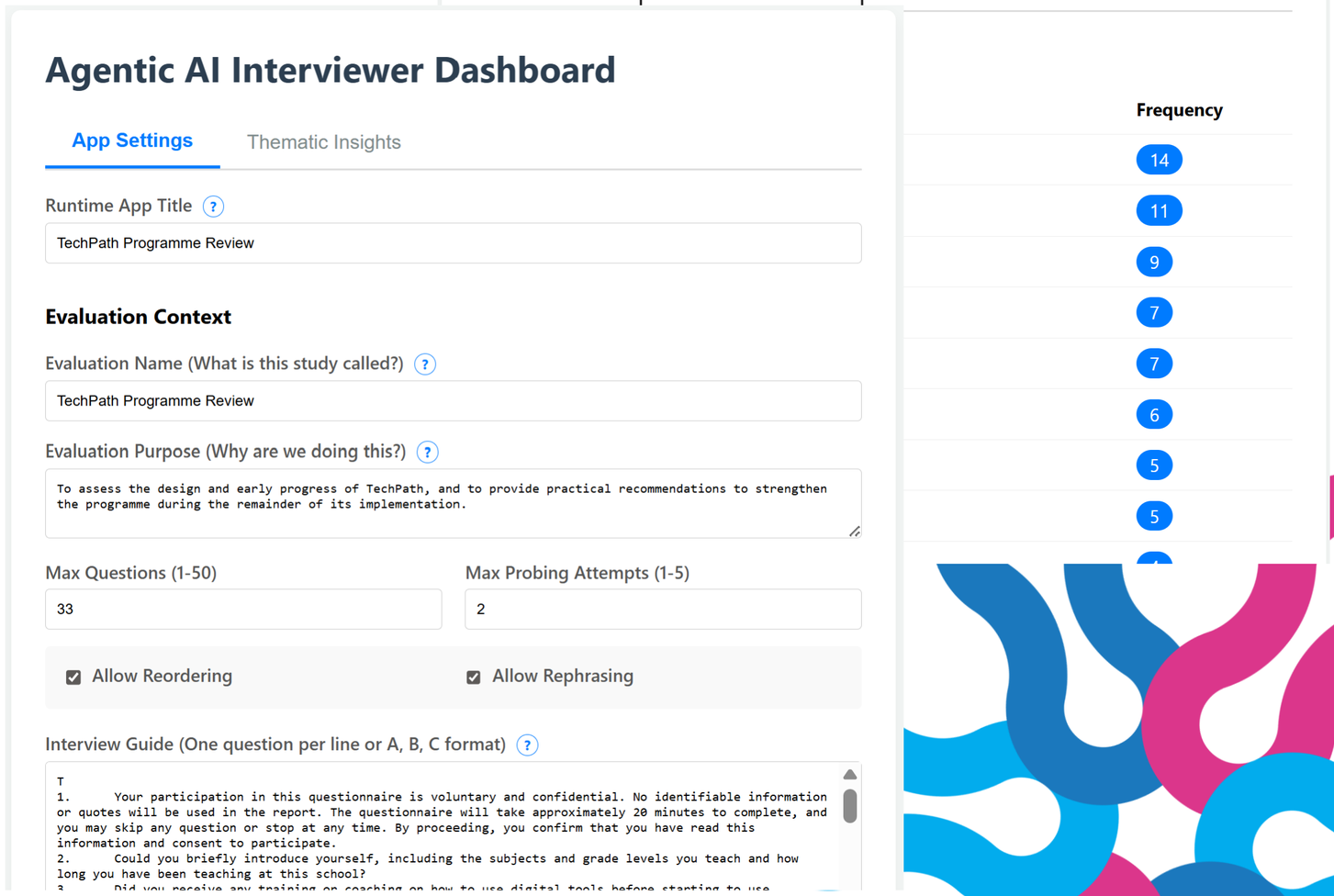
The Differentiator: Combines the depth and adaptability of qualitative interviews with the consistency, scale, and speed of an automated survey.

The Contribution: Enables rapid, procedurally consistent collection of community feedback, while providing real-time thematic analysis and operational insights to support adaptive evaluation management.



How it works in practice





TechPath: Field experiment with AI-assisted survey

What is the intervention about?

- TechPath is a joint initiative between the Ministry of Education, Google and UNICEF (2024–2026) in the Maldives.
- Implemented in 15 public schools across multiple atolls, targeting 1,500 students and supporting teachers, school leaders, and education authorities.
- Aims to strengthen foundational literacy and numeracy, digital learning, and 21st-century skills for children and adolescents through digitally enabled learning ecosystems and teacher capacity development.

How was the evaluation conducted?

- 📚 **Desk review** of implementation reports, programme documents, monitoring data, etc.
- 🏫 **Schools**: 11 of the 15 participating schools visited
- 👀 **Classroom observations**: 16 observations conducted
- 🎤 **Key informant interviews**: 28 interviews (e.g. MoE, NIE, UNICEF, etc.)
- 👥 **Focus Group Discussions:**
  - 👩‍🏫 **Teachers**: 11 FGDs involving 50+ teachers
  - 🎒 **Students**: FGDs with Grades 3 & 8 (80+ students) using game-based and drawing techniques.
  - 👨‍👩‍👧 **Parents**: FGDs with 45+ parents
- 📱 **Social listening** using of AI-enabled social and community listening (I-Hear-U)

How was the AI-assisted survey pilot implemented?

- Piloted in 4 remaining schools as an innovative complementary data collection method, implemented after fieldwork to further triangulate emerging findings
- The pilot focused on learning from the tool's use, including participant feedback on the user experience, rather than focusing solely on the data generated.
- Targeted adult participants only (principals, teachers, and parents) through AI-enabled conversational surveys with adaptive probing questions
- Dedicated helpdesk mechanism established to assist participants and troubleshoot technical issues during implementation.

Reflections & lessons learned from the experimentation

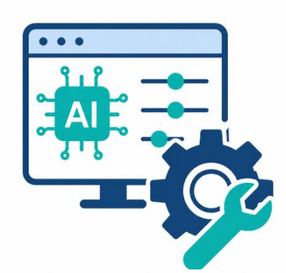
What worked well?

- Extremely **easy to deploy** and configure
- Very **low cost** per respondent with minimal logistical burden
- Broader **geographic and respondent coverage** than KIIs
- Enables **simultaneous large-scale data collection**
- Provides **flexibility for respondents** to answer at their own pace and preferred timing
- More **conversational and engaging** than standard surveys through adaptive and tailored interactions
- Improves question flow through **adaptive sequencing** and dynamic question formulation
- **Strong completion probing** to capture detailed lived experiences and concrete examples
- Generates more **nuanced qualitative data** than standard surveys
- Improves **consistency in probing** and reduces interviewer variability and bias
- **Automated end-of-conversation summaries** support validation of key discussion points
- Encourages **focused and concise responses** while limiting respondent digression

What were the key limitations?

- **Digital and inclusivity barriers**
- **Lower participation** and retention rates than KIIs
- **Frequent yes/no responses** required additional probing and follow-up compared with natural KII interactions
- **No prioritization of critical questions** when exceeding the question threshold
- Rare **limitations in conversational memory** led to repeated questions or loss of contextual continuity
- Configuration refinements needed to **improve probing logic and consent handling**
- Limited ability to identify and **explore emerging themes outside the predefined interview logic**
- Occasional **inconsistencies in conversational logic**, including references to unavailable information or inaccurate interview sequencing
- **Limited capacity to challenge** inconsistent respondent statements
- **Absence of non-verbal cues** limits interpretation compared with KII
- **Analytical outputs remain limited**, despite strong potential for automated qualitative analysis

Next steps & future considerations



System refinement

- Continued **testing & document** across diverse evaluation contexts
- **Refining system** to improve the Agent's contextual awareness
- Integrate **consistency and anomaly detection** functions

Inclusion and user experience

- Explore the **use of digital avatars** (audio/video interaction), including **integrated recording and automated transcription features**
- Develop **child-friendly AI interfaces**
- **Segment the surveys** and ask different segments from different interviewees

Smarter analytics

- Develop **smarter filtering and segmentation features**
- Improve AI-supported **auto-coding and thematic categorization features**

Governance

- Strengthen **AI governance, ethics** and safeguarding frameworks

AI for Social Listening

By Juan Pablo Arguello Yepez

Why traditional SBC messaging struggles to keep up with changing realities

SBC messaging is often too static

- Developed once, deployed broadly, adjusted late
- Limited ability to adapt while programmes are running

Limited access to real-time community insight

- Surveys and research are periodic and structured
- Miss unprompted concerns, rumours, and emerging barriers

Scale and complexity challenges

- Large volumes of feedback (digital platforms) are hard to analyse manually
- Multilingual, diverse inputs are under-utilized

Integrating Listening, Engagement, and AI for Adaptive SBC

SCL → capturing unprompted community signals

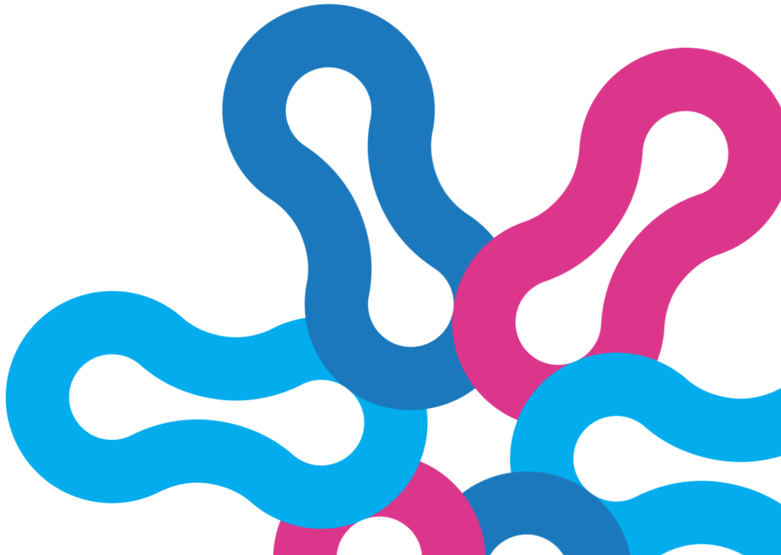
DCE → engaging communities in dialogue and co-design

AI → enabling analysis, synthesis, and pattern detection at scale + automatic insights and message recommendations

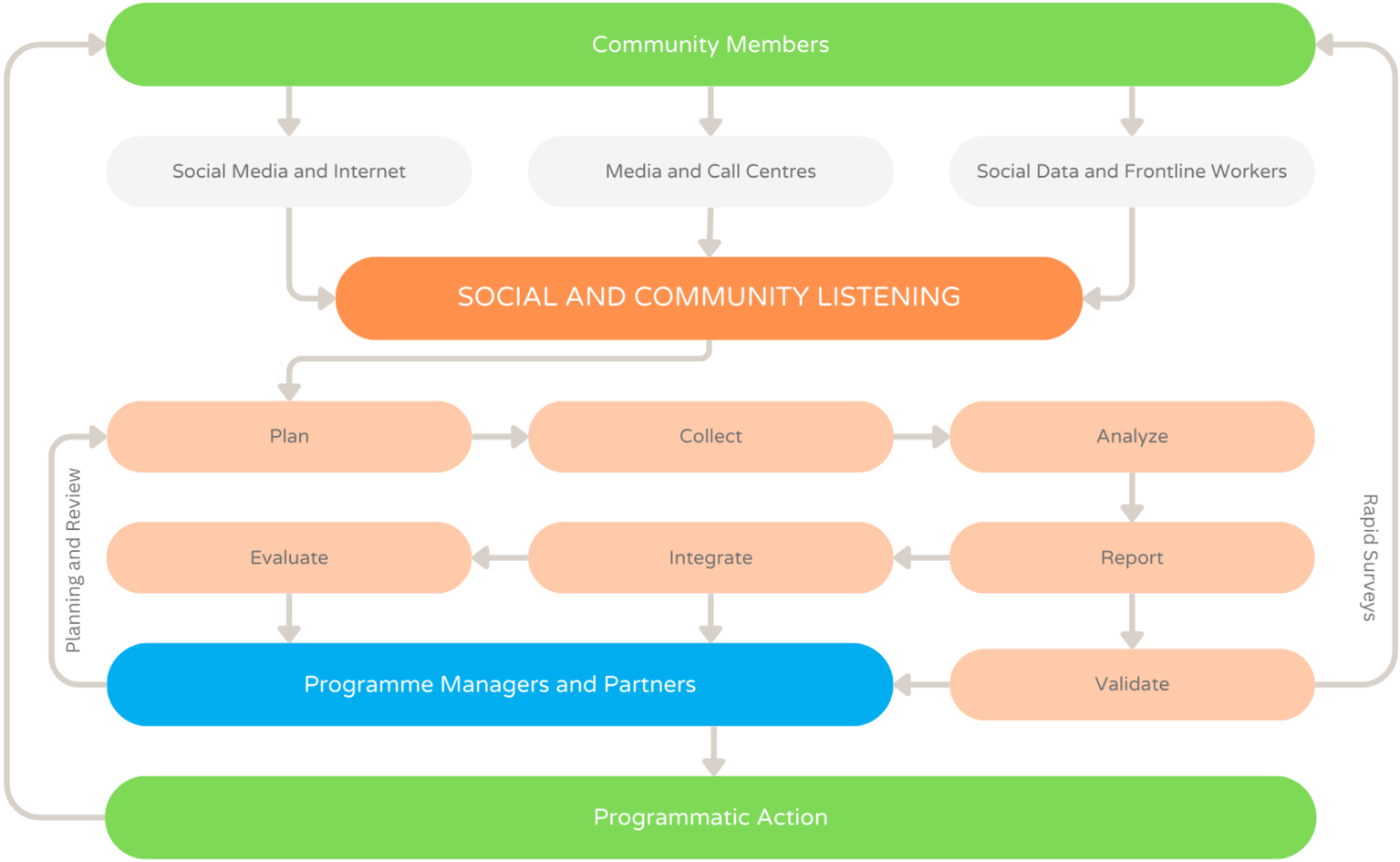
What makes it different from traditional approaches

- One-way → two-way engagement
- Periodic → continuous insight generation
- Static messaging → adaptive messaging systems

Program design informed in real-time	Integrated Data collection	Rapid synthesis of large, multilingual datasets	Surfaces diverse voices and minority perspectives	Enables course correction while interventions are ongoing



How It Works in Practice



How It Works in Practice

1

Collection

AI pours over millions of results to identify keywords people use to discuss the issues we need to report on

2

Processing

AI models organize, filter, and translate data to reduce noise and increase data quality and relevance

3

Analysis

By clustering, segmenting, visualizing, and coding data AI helps to uncover patterns that humans may miss

4

Validation

AI enables cross theme, timeline, and geographic comparisons to validate if a signal is real insight or a one-off event

5

Action

By referencing hundreds of academic papers and case studies, AI suggests proven actions based on SCL insights

Faster Insights

AI enables up to 10X faster data analysis. Reducing time to report from weeks to days or even hours

Deeper Analysis

AI enables synthesis of millions of results that humans are simply not able to process manually and fast enough

Ethics and Limitations

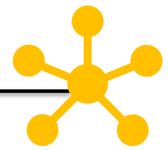
AI use does not fix SCL limitations such as overreliance on online voices, ethical and privacy concerns, or field inaction.

Human in the Loop: *AI output without human interpretation is data, not intelligence.*

Co-Design Pipeline (Mpox U-Report Example)

Communities invited to share their own mpox prevention messages and explain why they think they work

Thousands of community-generated messages consolidated into a single dataset



AI clustering surfacing themes, language patterns and emotional tones people use.

Testing patterns:
Generate Hypotheses for tone, framing, messenger, or language.

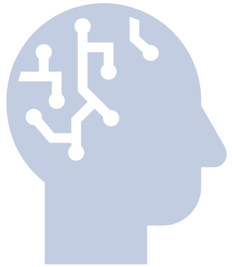
Insights informed refined messages (crowd-sourced message + AI generated messages) and recommendations.

Message variations were shared back with U-Reporters, who rated the options they found most compelling, clear, and relevant.

Disseminate Finalised Messages and Feed Learning Forward

Ongoing engagement including follow-up surveys

What worked, what didn’t, and what requires care



Added value

- Faster insight generation at scale
- Stronger grounding in community realities
- Improved relevance and adaptability of messaging
- Greater potential for accountability and responsiveness

Challenges and limitations

- Representation bias: Not all populations equally represented digitally
- Data interpretation risks: Patterns do not equal meaning without context
- Operational demands: Requires existing access to populations, time, coordination, and skills.
- Feedback–response gap: Risk

Key Insight

AI supports insights—but does not replace SBC expertise or community engagement.

From experimentation to integration and scale

The value of AI is not in replacing SBC processes—but in making listening, engagement, and adaptation possible at the speed and scale that today's programmes require.

What is being refined

- Integration of SCL and DCE into routine decision-making
- Improving multilingual and context-specific analysis
- Strengthening feedback loops and visibility to communities

Potential for scaling

- Embedding in programme cycles and SBC systems
- Standardizing workflows across country contexts
- Strengthening collaboration between digital, SBC, and sector teams

Future considerations

- Strengthening equity in participation (integrated diversification)
- Ensuring data protection and ethical use of AI
- Aligning insight generation with programmatic response capacity
- Building skills for interpreting and acting on AI-supported insights

AI-Enabled Equity Mainstreaming

By Atishay Mathur

How can evaluations serve the most marginalized

UNICEF has an imperative to respond to most disadvantaged and in-need children.

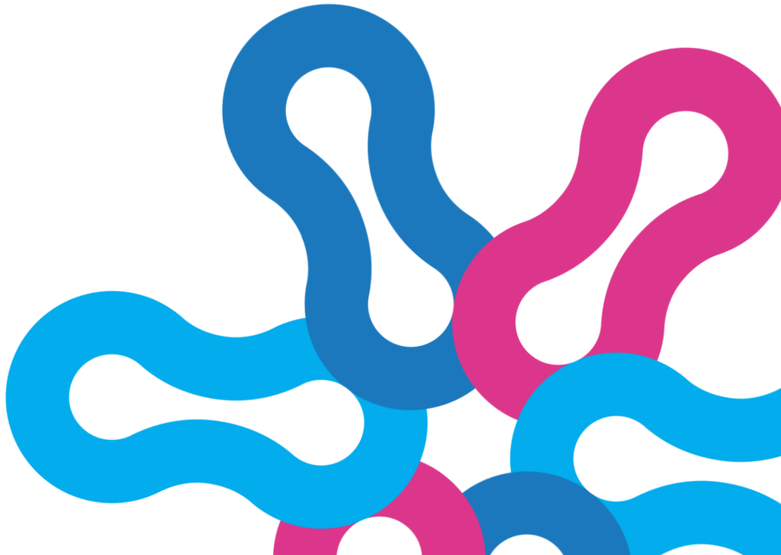
- It is essential for progress and to attain development targets and the SDGs.
- Not talking inequity means perpetuating a vicious cycle where goals remain elusive.
- Evidence work that focuses on the majority population often masks who is being left behind.

→ While UNICEF evaluations do focus on gender and equity, more can be done.

The Invisible Children

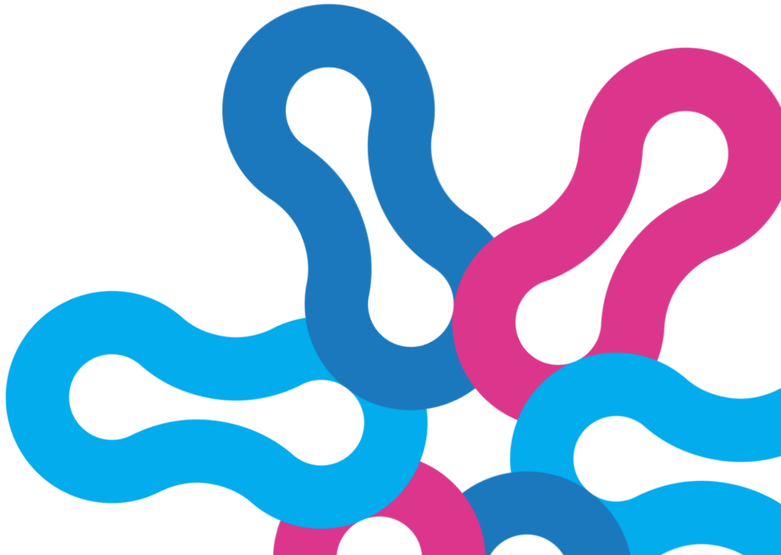
- Children of migrants
- Children engaging in labour/work; Adolescents engaging in paid work
- Children of informal workers in urban settings
- Children of sex workers
- LGBTQ+ children
- Children with learning disabilities
- Communities most at risk of (chronic) shocks and disasters
- Trafficked children
- Communities facing civil strife
- Children with parents in prison
- Children in institutions (including juvenile prisons)
- Children who have lost both parents/caregivers
- Families and children living on the streets

→ Difficulty surfacing hidden inequities, and intersectionality
→ Interpretation of inequities is subjective and often biased
→ While staff have access to norms and standards, challenges persist in their application



Can an AI Solution help Mainstream Equity in Evaluations?

Core Insight: Enabling researchers and evaluators to meaningfully incorporate equity into their work by overlooking their own **biases**, and ensuring they adhere to **best-in-class norms and standards**.



Can an AI Solution help Mainstream Equity in Evaluations?

Core Insight: Enabling researchers and evaluators to meaningfully incorporate equity into their work by overlooking their own **biases**, and ensuring they adhere to **best-in-class norms and standards**.

Brokering an AI Solution:

- "faster to re-purpose and re-orient than to re-build"
- Partnered with BelonggAI* to tailor an "advisory" tool to help UNICEF meaningfully integrate gender and equity into its evidence activities through an inclusion-focused analysis of reports.
- **Strongest asset was the curated repository of norms, standards, and benchmarks**

**BelonggAI was founded by Nirat Bhatnagar. It won the Gates Foundation AI For Equity Grand Challenge in 2023.*

Can an AI Solution help Mainstream Equity in Evaluations?

Core Insight: Enabling researchers and evaluators to meaningfully incorporate equity into their work by overlooking their own **biases**, and ensuring they adhere to **best-in-class norms and standards**.

Brokering an AI Solution:

- "faster to re-purpose and re-orient than to re-build"
- Partnered with BelonggAI* to tailor an "advisory" tool to help UNICEF meaningfully integrate gender and equity into its evidence activities through an inclusion-focused analysis of reports.
- **Strongest asset was the curated repository of norms, standards, and benchmarks**

Tailoring what the tool does:

- **Focuses on identities around – gender (LGBTQ+), disability, CSEA, caste, tribes, and religion.**
- Enables rapid QA of uploaded research, study, and evaluation reports, with a particular emphasis on research ethics, especially in contexts involving marginalized and vulnerable communities.
- Assesses documents term of references (ToR), inception reports, and final reports

**BelonggAI was founded by Nirat Bhatnagar. It won the Gates Foundation AI For Equity Grand Challenge in 2023.*

How the AI Solution Works (Intuitive and Uncomplicated)

Upload The Document To Be Analysed

Drag (or) Drop files here
or
Upload file
Size to be less than 50Mb

Which organization do you want to use?

UNICEF India

Which organizational guidelines do you want to use?

UNICEF & Belongg Combined

- Caste
- Disability
- Gender
- LGBTQIA+
- Overall
- Religious minority
- Tribal and indigeneity

26 Belo

20 min.

PART A1. Deep assessment of each identity dimension

What the CLAC already surfaces and where an evaluation ToR could go further

- Gender (beyond binary; SRHR, harmful masculinities, trans and gender-diverse persons)
    - What is covered: Chapter 4.4 Table 4.2 maps gendered measures across policies (e.g., SBM separate toilets and MHM, Mission Vatsalya's homes with "gender/transgender" options, JJM benefits for women and children, PM Ujjwala's focus on women). Section 2.6 documents girls' education disruption (care burdens, MHM constraints, early marriage risk), with Dhemaji evidence on MHM-linked absenteeism during floods (Case Study 08). Sections 2.8.3–2.8.4 acknowledge post-disaster GBV risk and child marriage surges.

CRITICAL ASSESSMENT: Priority improvements to the ToR (evaluation-focused)

Suggested insertions are framed for typical ToR sections; citations indicate why they matter, using CLAC evidence.

Methods: Intersectional Vulnerability Index and interaction analysis
Draft language: "The analysis will construct an Intersectional Vulnerability Index that overlays hazard exposure (Chapter 1.4; Figure 2.5) with identity markers and service access. Models will include interaction terms to test additive vs multiplicative disadvantages across hazards (heat, flood, cyclone, drought, air pollution)."
Rationale: Sections 1.4; 1.2.3; 2.1–2.6.

How the AI Solution Works (Intuitive and Uncomplicated)

Upload The Document To Be Analysed

Drag (or) Drop files here
or
Upload file
Size to be less than 50Mb

Which organization do you want to use?

UNICEF India

Which organizational guidelines do you want to use?

UNICEF & Belongg Combined

- Caste
- Disability
- Gender
- LGBTQIA+
- Overall
- Religious minority
- Tribal and indigeneity

26 Belo

20 min.

PART A1. Deep assessment of each identity dimension

What the CLAC already surfaces and where an evaluation ToR could go further

- Gender (beyond binary; SRHR, harmful masculinities, trans and gender-diverse persons)
  - What is covered: Chapter 4.4 Table 4.2 maps gendered measures across policies (e.g., der"
    ction

(Case
arriage

CRIT
(eva

Sugge ng
CLAC

Meth
Draft language: "The analysis will construct an Intersectional Vulnerability Index that overlays hazard exposure (Chapter 1.4; Figure 2.5) with identity markers and service access. Models will include interaction terms to test additive vs multiplicative disadvantages across hazards (heat, flood, cyclone, drought, air pollution)."
Rationale: Sections 1.4; 1.2.3; 2.1–2.6.

Nature of feedback:

- Background-Context (data sources)
- Objectives to Questions (framing)
- Methodological (aptness)

- Analysis (intersectionality)
- Crafting recommendations

→ While AI tools are capable of re-writing the report, we wanted to have the human in the loop do so (capacity building)

Reflections and Lessons Learned

Risks	Safeguards
Data Protection	Data safeguarding measures in-line with UNICEF data protection policies and laws in India (encryption, data storage etc.). It has been cleared by UNICEF HQ.
Bias Amplification	Controlled and curated repository that the AI-tool is trained on. Additionally, a layer of human review (since the tool does not make direct changes).
Context Blindness	
False Confidence	The repository of material that the AI tool is trained on is not only closed and curated, but also contextually rooted and updated regularly. Thereby introducing some operating costs.
Human-in-the-loop	We stopped short of operational offloading, i.e., getting the AI tool to re-write reports because true and sustained change can only come if the 'human' in-the-loop (our staff) builds capacity and sees value.
Lessons Learned	
AI Output	Both the AI output as well as the general user experience is now benchmarked against the more popular LLM tools such as ChatGPT, Claude, CoPilot. Given staff members use these tools even in their personal lives, the expectation of AI tools is to mimic their user interface, experience, and processing time. This has been a useful insight.
User Experience	
Testing Core Assumptions	Most AI tools lead to efficiency gains, this is one tool that aims to bring effectiveness gains by improving the quality (not just the speed) of equity-integration in evidence work.

Next Steps and Future Considerations

1. The tool is in a pilot-at-scale stage at UNICEF India with the aim to adopt it for its evidence work

Dimension		Assessment Criteria	Indicators / Questions
Efficiency	Cost	Reduction in resource requirements for quality assurance	Has the AI tool reduced the need for external reviewers or expert consultants?
			What is the cost per review pre/post AI use?
	Time	Faster review and feedback cycles	Average time taken to conduct quality checks manually vs. by AI
			Has turnaround time for reviews decreased?
	Personnel	Lower human resource burden or redeployment to higher-value tasks	Does staff capacity increase with the use of this tool?
Effectiveness	Relevance	AI accurately identifies missing or weak equity dimensions	Does the tool flag products that lack attention to key equity areas (e.g., caste, gender)?
			Are the AI suggestions contextually appropriate?
			Does the AI suggestions flagged for bias or lack of representation?
	Depth	AI goes beyond keyword spotting to assess meaningful integration	Does the AI evaluate the depth of equity analysis (e.g., disaggregation, intersectionality)?
	Consistency	Reduces subjectivity and variation across reviewers	Are AI assessments consistent across similar products?
			Do different users get similar results?
			Does the tool produce consistent result at different time interval?
	Uptake	Findings from the AI tool are used to improve drafts and final versions	Are flagged issues being addressed in revisions?
	Accuracy	Correctness of AI outputs compared to expert reviews	How often does the AI correctly identify or miss equity gaps (false positives/negatives)?

2. Explore potentially how this tool can inform programme planning not just evaluations

3. Collaborate with other agencies and institutions interested to use this tool

4. Continue finding new ways to strengthen UNICEF's work on equity – always open to ideas

global evaluation initiative

unicef for every child

Thank you!

EVIDENCE FOR CHILDREN