CENTER FOR GLOBAL DEVELOPMENT

The Agency Fund

IDinsight

GENERATIVE AI EVALUATION PLAYBOOK:

POLICY BRIEF

APRIL 2026

INTRODUCTION

From math tutors to farmer advisory tools, generative AI (GenAI) is rapidly expanding in low- and middle-income countries. While some evidence shows development gains, other findings point to harm. Evaluations can assess and help address these risks, but there is little agreement on what they should include. Tech teams prioritize product performance, often overlooking impact, while impact evaluators focus on outcomes but may neglect the underlying technology.

The Center for Global Development convened 30 experts across computer science, economics, gender studies, and development to bridge this gap. Experts aligned on a standard set of evaluation practices captured in the Generative AI Evaluation Playbook, released in April 2026. The practices are organized into four levels that each address a key question:

- **Level 1** – Model evaluation: Does the AI system perform as intended?
- **Level 2** – Product evaluation: Does the overall product engage and retain users?
- **Level 3** – User evaluation: Does the product impact users' thoughts, feelings, and behavior towards the development outcome?
- **Level 4** – Impact evaluation: Does the product improve development outcomes?

This document summarizes the playbook's core concepts—what to evaluate at each level, why it matters, who should do it, and how. It also outlines Minimum Viable Evaluations (MVEs) which are the most basic set of practices organizations should pursue at each level.

HOW TO USE THE PLAYBOOK

- **Builders of GenAI tools:** The primary audience includes development practitioners, engineers, product managers, data scientists, behavioral researchers, and impact evaluators. They can use the playbook to plan evaluations and revisit it to check progress.
- **Funders and policy makers:** While they may not use it daily, they can reference the playbook to define high-quality evaluations and encourage grantee adherence.

The Playbook is a living document, updated as new practices emerge. Builders of GenAI tools are encouraged to contribute amendments and case studies. Beyond detailed implementation guidance for each level, the Playbook discusses cross-cutting themes such as risk mitigation.

Scan here for the Playbook

L1 MODEL (SYSTEM) EVALUATION

Level 1 – Does the AI system perform as intended?

What is being evaluated?

The AI system which encompasses the full collection of data, models, and software, to process inputs (like text or images) and produces outputs (like predictions or advice) to support decisions or actions.

Why Evaluate?

GenAI tools can sound fluent yet be inaccurate or harmful. In sectors like education and health, unverified outputs risk real-world harm. Level 1 ensures that the full AI system performs reliably for its intended use and prevents costly misalignment downstream.

How to Evaluate?

The playbook recommends 6 steps to evaluate Level 1:

1. **Decide on an evaluation rubric:** Define the characteristics the AI system must exhibit—for example, a mental health bot may prioritize empathy and medical accuracy.
2. **Decide on metrics:** Define metrics they use to track performance along dimensions such as “accuracy”, or “empathy” that the rubric has identified.
3. **Develop a golden dataset:** Track improvement against the rubric and create a Golden Dataset that consists of the ideal user interactions with the system.
4. **Execute scoring and error analysis:** Once a golden dataset is developed, compare the system’s output against it.
5. **Automate evaluations:** Compare the system’s performance against the golden dataset, which should be automated to save time at scale.
6. **Conduct red-teaming:** Beyond the steps above, engage in red teaming, which is the practice of systematically testing a system by simulating adversarial or edge-case inputs.

Who Evaluates?

AI/ML engineers typically lead this level. Domain experts, product owners, and user researchers support them in rubric design, metric validation, golden dataset creation, and safety evaluation. Note that these roles may be filled by consultants, and a single person may hold multiple roles within an organization. The listed roles reflect the required proficiencies rather than headcount.

What is the Minimum Viable Evaluation?

1. Establish 2-3 rubrics for model success with at least one robust safety/guardrail metric computed on your Golden Dataset.
2. In consultation with product and business owners, set a success criteria or threshold for each rubric/metric that needs to be passed before it is ready for deployment.
3. Develop a Golden Dataset with at least 30-50 items representing key, diverse user interactions.
4. Establish a process for expert review of AI system responses for inputs in the Golden Dataset, as you iterate on your system configuration.

L2 PRODUCT EVALUATION

Level 2 – Does the overall product engage and retain users?

What is being evaluated?

Uptake of the AI product by users, e.g. a patient's level and regularity of interactions with a health chatbot.

Why Evaluate?

An AI product that produces perfect responses is worthless if it fails to engage users. Builders must track critical user signals, like activation, engagement and retention.

How to Evaluate?

The playbook recommends 5 steps to evaluate Level 2:

1. **Define the user funnel and metrics:** Define the stages of the user journey (e.g., acquired, activated, engaged, retained) and select metrics for each, such as frequency or duration of usage.
2. **Instrument the product:** Identify and track key user actions that signal progress through the funnel, ensuring each event is tied to a unique user ID.
3. **Automate metrics and analyze trends:** Use event data to calculate metrics, visualize them in dashboards, and monitor trends and drop-offs over time.
4. **Identify frictions and design improvements:** Analyze where users drop off or struggle, and use qualitative insights to diagnose issues and propose product improvements.
5. **Test product upgrades:** Use A/B testing or experimentation to evaluate whether product changes improve key metrics and user progression through the funnel.

Who Evaluates?

Product managers lead this level. Data scientists support them in metric construction, dashboard design, and experiment analysis.

What is the Minimum Viable Evaluation?

1. Instrument the product to capture events automatically.
2. Use the events data to produce two metrics: activation (used once), and retention (used repeatedly).
3. Look for patterns in the data, and talk to users to identify opportunities for improvement.
4. Test these ideas for improvement against these metrics with an A/B test.

L3 USER EVALUATION

Level 3 – Does the product impact users' thoughts, feelings, and behavior towards the development outcome?

What is being evaluated?

Users' knowledge, attitudes, behaviors, and decision-making e.g., a patient's subjective assessment of their symptoms and behaviors relevant to managing their condition.

Why Evaluate?

Engagement at Level 2 may leave behaviour unchanged. Students can use tutoring apps heavily without true learning, or an unhealthy eater may engage a chatbot daily, yet not change diet. Level 3 identifies and measures specific changes in thoughts, feelings, and behavior that the theory of change expects to lead to improved development outcomes. Intermediate indicators enable fast product iteration and signal if an GenAI system is on track before committing an impact evaluation.

How to Evaluate?

The playbook recommends 5 steps to evaluate Level 3:

1. **Generate hypotheses based on TOC:** Based on the theory of change, define and validate intermediate cognitive, affective, or behavioral outcomes that link to targeted impact.
2. **Identify outcome metrics:** To understand meaningful user interaction, combine quantitative log and conversation data with surveys or interviews to capture self-reported experiences and behavioral change. Validate them with on- and off-platform research.
3. **Define guardrail metrics and measure potential harm:** Define safety guardrails to avoid unintended consequences. Privacy safeguards are crucial throughout this level given the sensitivity of this personal information.
4. **Consider constructing proxies for long-term development outcomes:** No single lower-level metric can reliably predict long-term development outcomes, but you can build a "Surrogate Index" (Athey at al, 2025) that combines Level 2 and 3 metrics as a proxy for Level 4 outcomes.
5. **Consider conducting experiments to improve the selected key metrics and running process evaluations:** Experiment with L3 measures via product changes—such as A/B tests, MABs, or holdouts—to learn why and when those metrics fail to improve as planned.

Who Evaluates?

User researchers and behavioral scientists lead this level. Data scientists and AI engineers support them in survey deployment, NLP analysis, and experiment infrastructure.

What is the Minimum Viable Evaluation?

1. Use the theory of change's most decision-relevant cognitive or behavioral outcome and include at least one early-warning indicator of harm.
2. Combine one behavioral or trace metric with a brief self-reported measure to capture change.
3. Include a minimal external check to ensure link of on-platform measures with real outcomes.
4. Consider testing product changes on selected outcomes using simple experimental methods.

L4 IMPACT EVALUATION

Level 4 – Does the product improve development outcomes?

What is being evaluated?

The impact of the AI product on real-world outcomes such as verified learning, morbidity, income or productivity.

Why Evaluate?

For funders and policymakers, credible evidence that a product improves development outcomes—beyond engagement or self-reports—is essential for informed decisions about scaling. Level 4 evaluations isolate causal effects by comparing the outcomes of a group that has used the AI product to a counterfactual that has not.

How to Evaluate?

The playbook recommends 5 steps to evaluate Level 4:

1. **Choose your Methodology:** There are multiple impact evaluation methods including Randomized Controlled Trials, Propensity Score Matching, and Regression Discontinuity.
2. **Select the right counterfactual:** You must define what "the world without the AI product" looks like. In AI evaluations, the comparison isn't always "nothing"—it might be a static chatbot, a human teacher, or a traditional paper-based process.
3. **Account for product dynamism:** AI products evolve continuously—so maintain rigor by tagging versions, preserving a hold-out group, and aligning with engineering to avoid breaking the evaluation design.
4. **Measure true outcomes, not proxies:** Measure real welfare or capability gains, avoid tests that can be gamed and use validated tools like standard assessments or administrative data.
5. **Manage spillovers and attrition:** AI tools can spread to control groups, risking contamination—mitigate with cluster randomization (e.g., by school or village) and monitor drop-outs using Level 2 and 3 data, as attrition can undermine statistical power.

Who Evaluates?

Policy researchers, economists, and social scientists lead this level, ideally working with independent, external evaluators. AI engineers ensure stable product behavior and version control throughout the evaluation.

What is the Minimum Viable Evaluation?

1. Implement an impact evaluation with counterfactual and enough sample size to measure the key outcomes of interest, including for sub-populations of interest (e.g. gender, geography).
2. Implement strong version control, testing either with a single, frozen version or a limited number of product versions.
3. Cost data collection.

LINKAGES ACROSS LEVELS

Beyond providing recommended practices across the four levels, the playbook also describes additional assessments that can be deployed repeatedly at each level. These include methods like:

1. **Process evaluations:** A process evaluation examines whether the right actors are doing the right things at the right time in response to the AI tool. It doesn't estimate impact but checks if the intervention is implemented as intended and identifies areas for improvement.
2. **User research:** User research is the systematic study of users' needs, behaviors, and experiences to inform better design and decision-making. This can include interviews to design golden datasets (Level 1), workflow observation to develop hypotheses (Level 2), and cognitive interviewing to inform survey design (Level 3–Level 4).
3. **Risk assessment and mitigation:** Risks range from user safety and mental health to hallucination and dependency. Builders can use a host of mitigation strategies across all four levels to address them.

CONTRIBUTORS

Working Group & Core Contributors

Han Sheng Chia (Co-Chair)
Director, Artificial Intelligence Initiative and Policy Fellow, Center for Global Development

Markus Goldstein (Co-Chair)
Vice President & Senior Fellow, Center for Global Development

Temina Madon (Co-Chair)
Co-Founder & Chief Executive Officer, The Agency Fund

Becky Faith
Research Fellow and Leader of the Digital and Technology cluster, Institute of Development Studies

Bilal Mateen
Chief AI Officer, PATH

Bilal Zia
Head of Data Science & Analytics, Duolingo

Brian DeRenzi
Head of Research and AI, Dimagi

Chenai Chair
Director, Masakhane African Languages Hub

Crystal H. Huang
Senior Economist & Director, IDinsight

Dean Karlan
Professor of Economics and Finance, Kellogg School of Management, Northwestern University

Denys Sementsov
Chief Technology Officer, Dalberg Data Insights

Donald Lobo
Founder, Project Tech4Dev

Edmund Korley
Member, The Agency Fund

Emmanuel Olang'
Senior Machine Learning Engineer, Jacaranda Health

Engy Saleh
Director of Behavioral Research and Academic Engagements, Busara

Gabriel Demombynes
Manager of the Human Capital Project, The World Bank

James Walsh
Member, The Agency Fund

Katrin Tomanek
Lead AI and Voice Technology, Centre for Digital Language Inclusion

Kelly Zhang
Member, The Agency Fund

Leah Rosenzweig
Senior Fellow, Center for Global Development & Director, University of Chicago's Market Shaping Accelerator

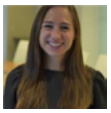
Niall Keleher
Senior Director of Global Research and Data Science, Innovations for Poverty Action

Paul Atherton
Founder, AI-for-Education.org

Robert Korom
Chief Medical Officer, Penda Health

Robert On
Member, The Agency Fund

Ryan L. Boyd
Assistant Professor of Psychology, School of Behavioral and Brain Sciences, University of Texas at Dallas



Sam Carter
Senior Policy Manager (AI), J-PAL Global

Shi Feng
Assistant Professor of Computer Science, George Washington University

Sid Ravinutala
Chief Data Scientist, IDinsight

Stanslaus Mwongela
Senior Machine Learning Engineer, Audere

Tarunima Prabhakar
Co-Founder, Tattle Civic Tech

Tim Ohlenburg
Research Fellow, Center for Global Development

Umar Saif
Founder and Chief Executive Officer, aiSight.ai

Vineet Singh
Chief Technology Officer, Digital Green

Zezhen (Michael) Wu
Member, The Agency Fund

Asim Fayaz
Member, The Agency Fund

Implementers Consulted

GENERATIVE AI EVALUATION IN THE DEVELOPMENT SECTOR: A LIVING PLAYBOOK

APRIL 2026

CGD CENTER FOR GLOBAL DEVELOPMENT
The Agency Fund
IDinsight