Pre-Workshop Form

The Agency Fund

https://tinyurl.com/aes-2026-pre-workshop-form

Evaluating AI for Social Impact

A 4 Level Framework

Africa Evidence Summit, July 2026

Kelly Zhang & Edmund Korley, Core Team at The Agency Fund

The Agency Fund

The Agency Fund

Overview

- Background on the 4-level evaluation framework
- Level 1 Demo: Model Evaluation on Calibrate
- Level 2/3 Demo: A/B Testing on Evidential
- Feedback Form and Q&A

The Agency Fund

Researchers are often focused on impact evaluation

*But there are **multiple levels** to consider before evaluating the impact of an AI product*

The 4-Level Evaluation Framework

The Agency Fund

L1

Model Evaluation

Does the AI system perform as intended?

AI Engineers
Domain Experts

L2

Product Evaluation

Does the overall product engage and retain users?

Product Managers
Data Scientists

L3

User Evaluation

Does product usage change how users are thinking, feeling, and behaving?

Behavioural Scientists
UX Researchers

L4

Impact Evaluation

Does product usage improve development outcomes?

Economists
Impact Researchers

Today's workshop focus

The Agency Fund

LEVEL 01

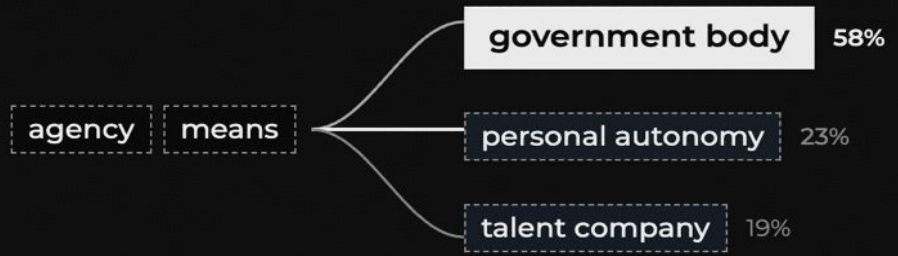
Model Evaluation

Does the AI system perform as intended?

How do LLMs really work?

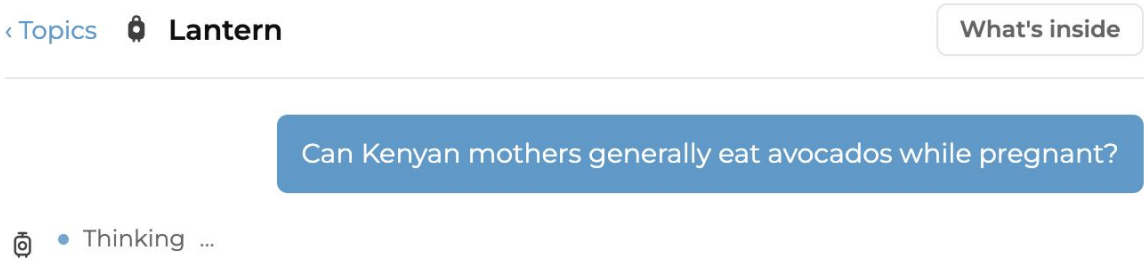
The Agency Fund

From these patterns the model learns to predict the most plausible next word. It doesn't retrieve facts. It generates whatever continuation fits the pattern best.



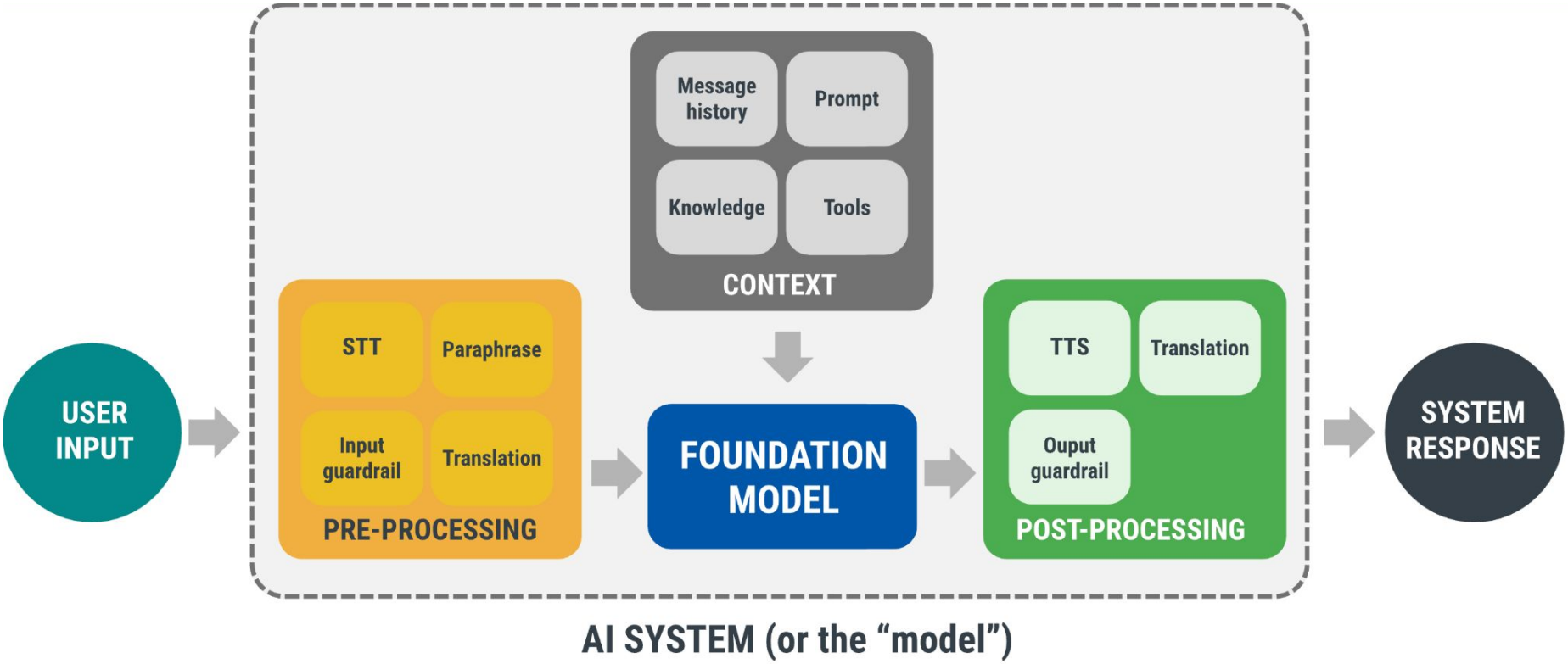
How do LLMs really work?

The Agency Fund



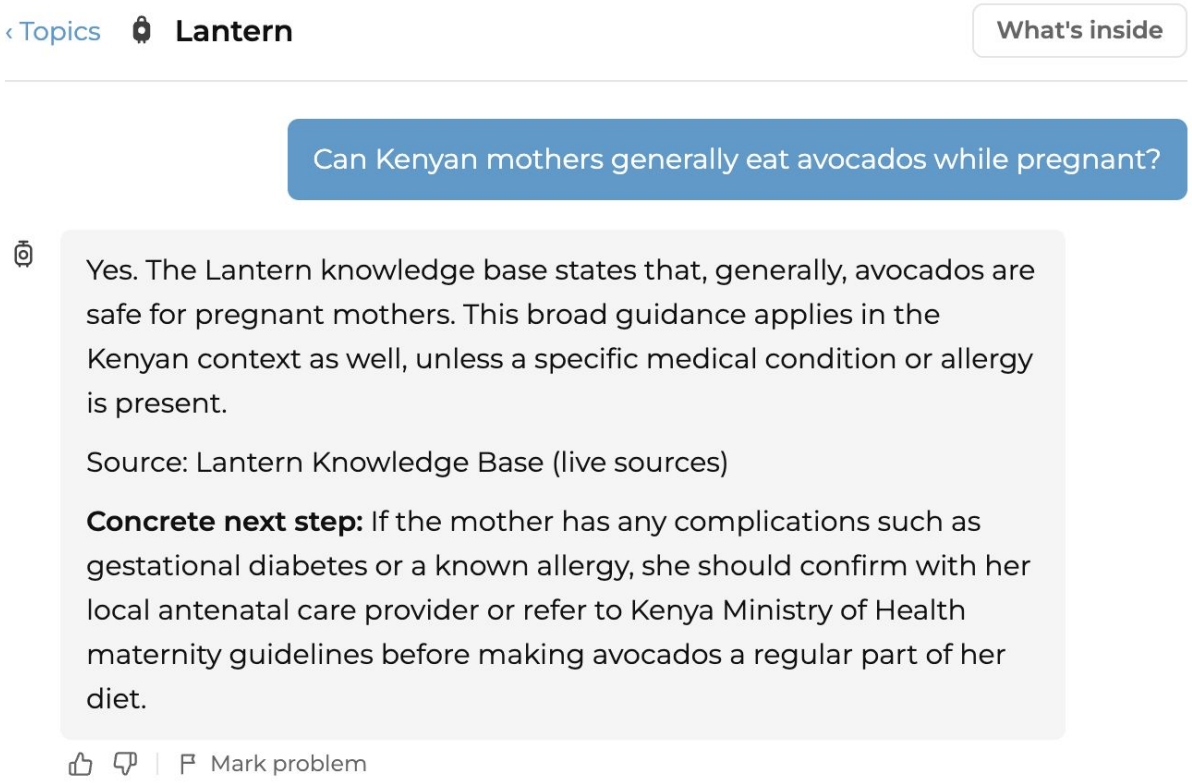
What is model in model evaluation?

The Agency Fund



How do LLMs really work?

The Agency Fund



The Agency Fund

Where have you seen AI fail?

The Agency Fund

Why AI can fail your users

- **Biased** training data, **misinformation**, and **unverified claims** baked into the model
- Poor performance in **low-resource languages** — your users may be underrepresented in training data
- **Knowledge cut-off** — the model doesn't know what happened after its training ended
- No access to **local context** — norms, policies, real-time conditions on the ground
- **Inconsistent instruction-following** — complex or ambiguous prompts often produce the wrong output
- **Wrong tool** for the task — AI may confidently give wrong answers for a complex multiplication that any calculator can easily do

Evaluation creates reliability & safety

- Are the AI responses grounded in **reliable and credible sources**?
- Are failures rooted in the model itself, or how we are prompting and integrating it?
- Are users in LMICs getting **accurate, safe, and culturally-attuned responses**?
- Why does AI produce **different outputs for the same inputs**?
- How to identify when to discard the AI response and **fall back to humans**?

Without structured evaluation, you cannot know. In social sector, this gap is not acceptable.

The Agency Fund

How to evaluate

UNTIL NOW

How we judged human performance

Throughput

How many people were served?

End-user surveys

Did people find it helpful?

Outcome tracking

Did the programme work?

None of these measure the interaction itself

THE SHIFT

The Agency Fund

What AI demands instead

You cannot deploy first

Similar to how we train frontline workers before they interact with end users, we need to **interview** AI to ensure it is ready

Define good upfront

Clearly define **how to judge** whether a **single response** given by a human is **good.**

We have relied on **proxies** to measure human performance but now we need to define what makes a single exchange good. Not the programme. Not the outcome... The interaction.

EXAMPLE: NOORA HEALTH COPILOT

The Agency Fund

Examples of model evaluation criteria

01 **Factual accuracy**

Is every claim medically correct?

Do they avoid overstating confidence where the evidence is still uncertain?

02 **Comprehension**

Are they easily understandable?

Do they avoid using clinical jargon? Do they use the same cultural references and vocabulary as the end users?

03 **Grounded in protocol**

Does it follow our internal guidelines?

Is their advice grounded in your specific protocols? Are they adding information from their own knowledge outside the protocol?

04 **Empathetic tone**

Do they respond with empathy?

Or does it feel clinical and transactional when the situation calls for warmth?

Minimum viable evaluation (MVE)

The Agency Fund

The perfectionism trap

Waiting for the perfect golden dataset before moving deploying with real users

→ **Delays** real-world learning indefinitely

→ Perfect datasets are **never truly finished**

→ Evaluation without deployment **teaches you little**

Iterative deployment

→ Level 1 needs to be **good enough**, not perfect

→ The **minimum** degree of evaluation that gives you confidence to safely deploy to real users

→ Evaluation is not a one-time task. It is a **continuous** process that continues post deployment

MVE checklist

The Agency Fund

01

Golden Dataset

30–50 items

Represents key, diverse user interactions

The core test bed for evaluation

02

Expert review

Domain experts create ideal response or rubrics

Experts assess AI responses

Not automated

03

Safety metric

At least one robust safety/guardrail metric

Ensure AI accuracy improvement does not compromise safety

Offline vs online evaluations

The Agency Fund

Offline evaluations

Testing your AI system in a **controlled setting** on your golden dataset

Augment golden dataset with edge cases from real user data from deployment

Analyse errors in your AI system and identify **root causes**

Iterate on your AI system and **validate** if errors are fixed

Online evaluations

Evaluating your AI system on **real user queries** after launch

Use LLM judges and human review to **monitor** response quality, latency and cost

Identify samples where AI responses are **incorrect or need improvement**

Deploy new changes **incrementally**, verify if they work and then launch for all

The Agency Fund

LEVEL 02

Product Evaluation

Does the overall product engage and retain users?

The Agency Fund

A perfect model that no one uses has zero impact

EXAMPLE: DIGITAL GREEN FARMERCHAT

The Agency Fund

One metric per stage

Acquisition	Activation	Engagement	Retention
Recruitment into the app	First activity that shows value for user on the app	How much users are engaging with the app	How often users continue to engage with the app over a longer period of time
A farmer creates an account on the app	**The farmer logs in and asks their first crop question on the app**	**The farmer logs in daily to ask questions**	**The farmer continues to ask questions over 3, 6, 12 month period**

From gut feeling to data-driven

The Agency Fund

Without user metrics

"I think users dropped off after onboarding."

"Maybe the interface is confusing?"

With user metrics

- Only 10% of users are retained on the app after Day 3
- Activation rate drops from 71% to 34% on low-connectivity devices

The Agency Fund

How to experiment?

Some places to experiment

The Agency Fund

Product Features

Test new features (e.g. model updates or changes) before full release to validate they deliver value

Changes to User Interface

App interface or workflow improvements that affect how users interact with your product

Program Activity

Vary the programming on the platform (e.g. weekly vs. biweekly reminders, behavioral nudges)

Bug Fixes

Even small fixes can have unexpected effects in improving engagement and retention

A/B Testing: Youth Impact

The Agency Fund

Table 1: Description of A/B tests and program modifications

Innovation Type	Version A Status Quo Model	Version B Modification
	Panel A: Cost-reducing tests	
Dosage Distribution	Weekly calls + SMS Weekly calls + SMS	Bi-weekly + SMS Bi-weekly + SMS
Implementer Type	Bi-weekly + SMS (same tutor) Bi-weekly + SMS (same tutor)	Bi-weekly + SMS (different tutor) Bi-weekly + SMS (different tutor)
Scheduling Assignment Mechanism	Bi-weekly + SMS (facilitator-led decentralized scheduling)	Bi-weekly + SMS (call center centralized scheduling)
	Panel B: Effectiveness-enhancing tests	
Tech Package: Add-ons	Bi-weekly calls Bi-weekly + SMS Bi-weekly + SMS	Inclusion of SMS Complementary WhatsApp videos Complementary homework SMS
Motivational Nudges	Bi-weekly + SMS Bi-weekly + SMS	Caregiver involvement nudge Alumni caregiver involvement testimonials
Caregiver Engagement	Bi-weekly + SMS Bi-weekly + SMS	Caregiver co-lead tutorial via encouragement Caregiver co-lead tutorial via encouragement

Iterative A/B testing of a phone-based tutoring service aimed at improving learning outcomes

Source: Angrist, Cullen, and Magat (2025) Cheaper (and more effective) by the dozen: Evidence from 12 randomised A/B tests optimising tutoring for scale.

The Agency Fund

LEVEL 03

User Evaluation

Does the product change user thoughts, feelings, knowledge and behaviour towards the development outcome?

The Agency Fund

Intermediate Outcomes

The Agency Fund

INTERMEDIATE OUTCOMES

Is the user thinking, feeling, or acting differently in ways might lead to the development outcomes of interest?

The Agency Fund

Example dimensions of user change

Cognitive

e.g., Comprehension, knowledge acquisition and retention, belief updating, complexity of reasoning

Affective

e.g., Mood, frustration, confusion, confidence, sense of agency, sense of safety, sense of belonging, perceived empathy, trust

Behavioral

e.g., Shifts in platform behavior (i.e. tool downloads), application of information to new queries, asking for additional information or resources

The Agency Fund

Three ways to measure intermediate outcomes

A

Interaction data

Analyze log data:

- Number of follow-up questions by user
- Engagement with a specific set of features

LEADING INDICATORS

B

Survey data

Ask users directly:

- Short in-chat surveys
- Longer off-platform surveys
- Knowledge quizzes

SELF-REPORT

C

Conversation logs

Mine the text itself:

- Sentiment analysis
- Topic modeling

NLP & TEXT ANALYSIS

The Agency Fund

Measure not just if it works, but also does it harm

User agency is a critical guardrail in the social sector, is the AI product **building capability** or **creating dependency**?

SUBJECTIVE AGENCY · INTERNAL-FACING

Do users believe they can solve the problem on their own now?

Users' beliefs about their own capabilities captured through surveys, interviews, and reflection prompts

OBJECTIVE AGENCY · EXTERNAL

Can they actually plan and execute the goal without the help of AI?

Do users have the skills to act? Are they learning the underlying logic, or just copy-pasting answers?

Sources: Albert Bandura's Social Cognitive Theory, Amartya Sen's Capability Approach

The Agency Fund

LEVEL 04

Impact Evaluation

Do users with access to the product have improved development outcomes?

The Agency Fund

Traditional Randomized Controlled Trial (RCT): Experiment measuring the development outcomes

Level 4 depends on Levels 1/2/3

The Agency Fund

Level 1: Model evaluation

Is the AI model is providing accurate information?

Level 2: User engagement

People are using the product through engagement and retention?

Level 3: User research

Are users behaving in ways that will lead to the development outcome of interest?

The Agency Fund

DEVELOPMENT OUTCOMES

If the product is inaccurate (L1) or if no one is using the product as intended (L2/L3) , then it is unlikely that there will any impact (L4)

The Agency Fund

Technical Demos

Demos

The Agency Fund

1. Model Evaluation with Calibrate
    - Goldens
    - Knowledge Base
2. A/B Testing with Evidential
    - Feature Testing

Demo: Lantern Chatbot

The Agency Fund

https://tinyurl.com/lantern-chatbot

Example prompts:

The Agency Fund

https://tinyurl.com/eval-workshop-test-prompts

Session Feedback + Q&A

The Agency Fund

https://tinyurl.com/aes-2026-workshop-feedback

The Agency Fund

Appendix

Evidential

The Agency Fund

https://tinyurl.com/evidential-announcement