# A playbook on AI evaluation for the social sector

eval.playbook.org.ai

The Agency Fund

The Agency Fund

*Before we begin...*

# How do you know if your AI system working?

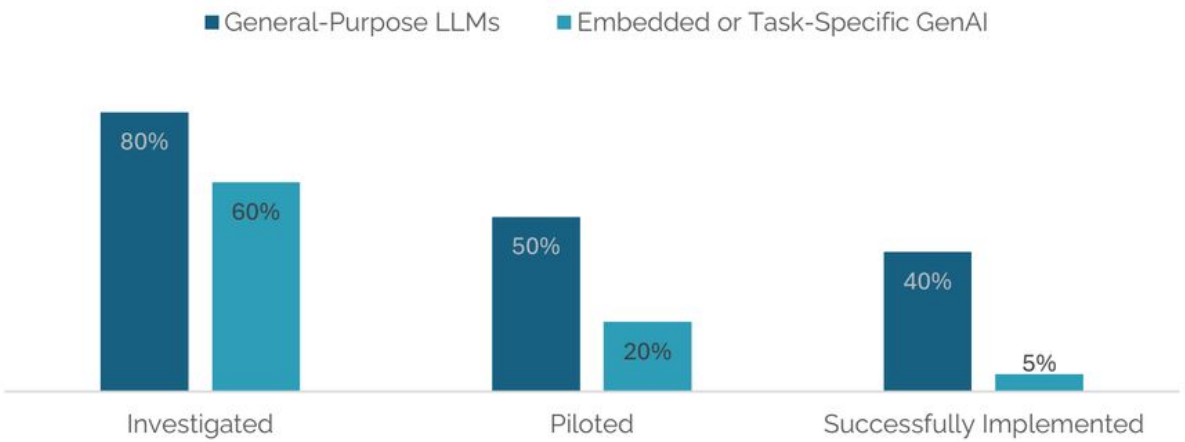
Not just technically — but for the people whose lives it is meant to improve

The Agency Fund

# AI Capability Is Accelerating

- Models are improving rapidly
- Costs keep falling
- Capabilities are expanding into new tasks
- **Risk is increasing alongside capability**

The Agency Fund




# 95% of AI pilots in enterprise fail to deliver measurable impact

— MIT, July 2025 (150 interviews, 350 surveys, 300 deployments)

> The problem isn't the models. It's integration, workflows, and alignment.

The Agency Fund

# Common AI Challenges in Development

- Lack of benchmarks that reflect the context of LMICs
- Higher risk of unsafe and inaccurate AI responses
- Limited bandwidth from domain experts
- Premature AI optimisation before validating user adoption
- Decisions taken based on intuition instead of data
- Measuring real world impact beyond engagement
- Staff capacity and change readiness

# Two worlds. One product.

The Agency Fund

### What tech teams measure

→ Hallucination rate

→ Response latency

→ Accuracy

→ Daily active users

→ User retention

→ Time spent on the app

≠

*The Gap*

### What governments & funders need

→ Behaviour change evidence

→ Education/economic/health outcomes

→ Cost per beneficiary

→ Causal impact on outcomes

→ Continued efficacy during scale-up

→ Equity across subgroups

**Neither side is wrong, they are measuring different things.**
**The 4-Level Framework bridges both.**

The Agency Fund

THE 4-LEVEL AI EVALUATION FRAMEWORK

# A shared language from model to mission

# 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 the product move the needle on user thoughts, feelings, and behavior?

Behavioural Scientists
UX Researchers

L4

## Impact Evaluation

Do users with access to the product improve development outcomes?

Economists
Impact Researchers

CASE STUDY: FARMER.CHAT

The Agency Fund

# What This Looks Like in Practice

LEVEL 1

## Model

AI gives correct, local farming advice

**Accuracy**

LEVEL 2

## Product

Farmers come back to ask more questions

**Retention**

LEVEL 3

## User

Farmers act on the recommendations

**Behavior change**

LEVEL 4

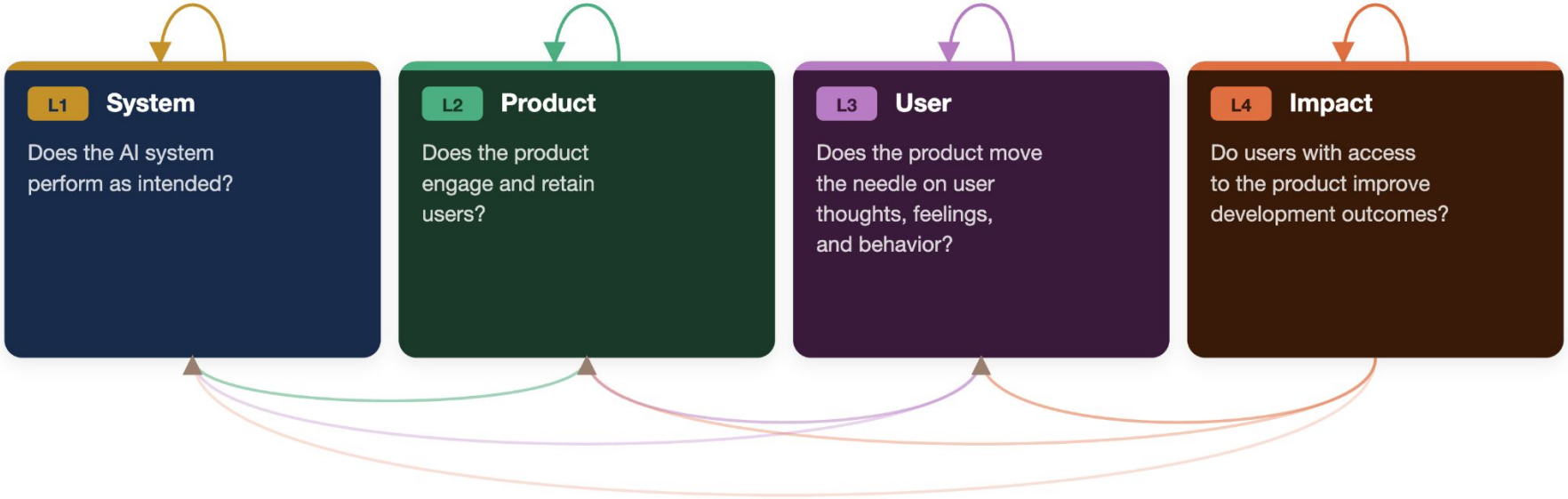
## Impact

Crop yields and farmer incomes improve

**Outcome**

# Learnings feed back to lower levels

The Agency Fund




Greater marginal **Cost**

Slower **Speed** of iteration

Increased **Risk** to user or org

The Agency Fund

LEVEL 01

# Model Evaluation

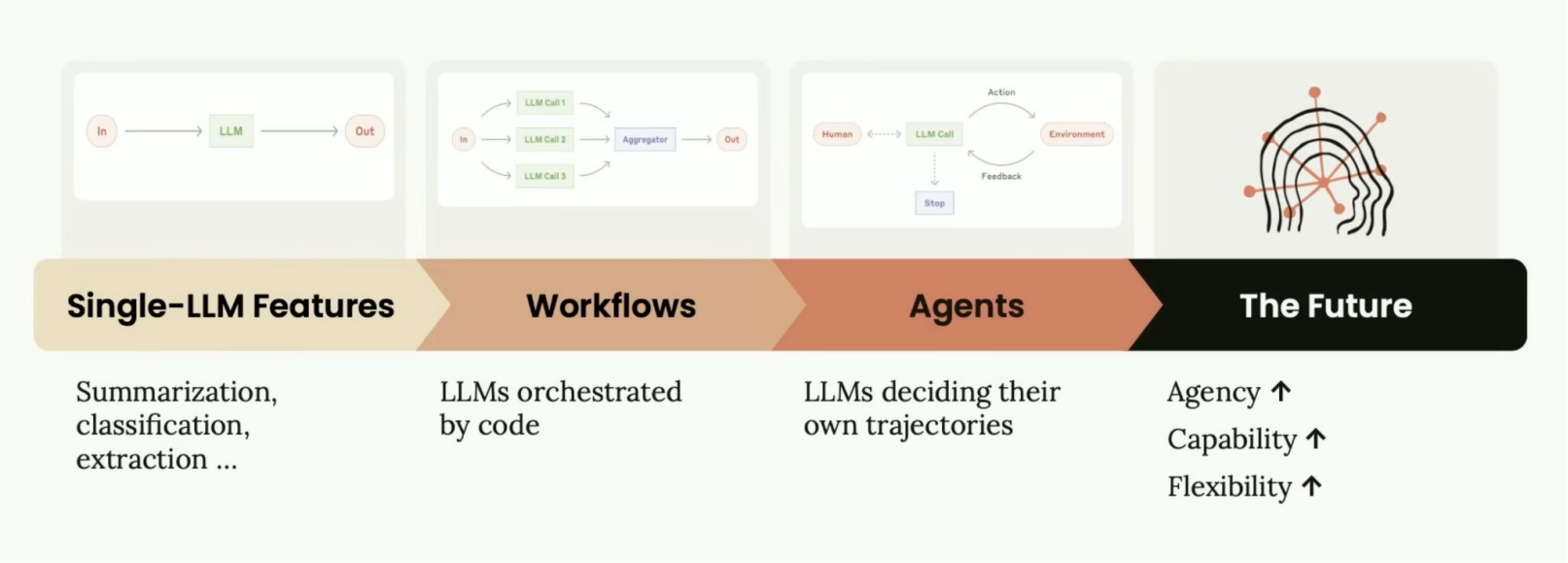
Does the AI system perform as intended?

The Agency Fund

# What is the model in model evaluation?

# What is the model in model evaluation?

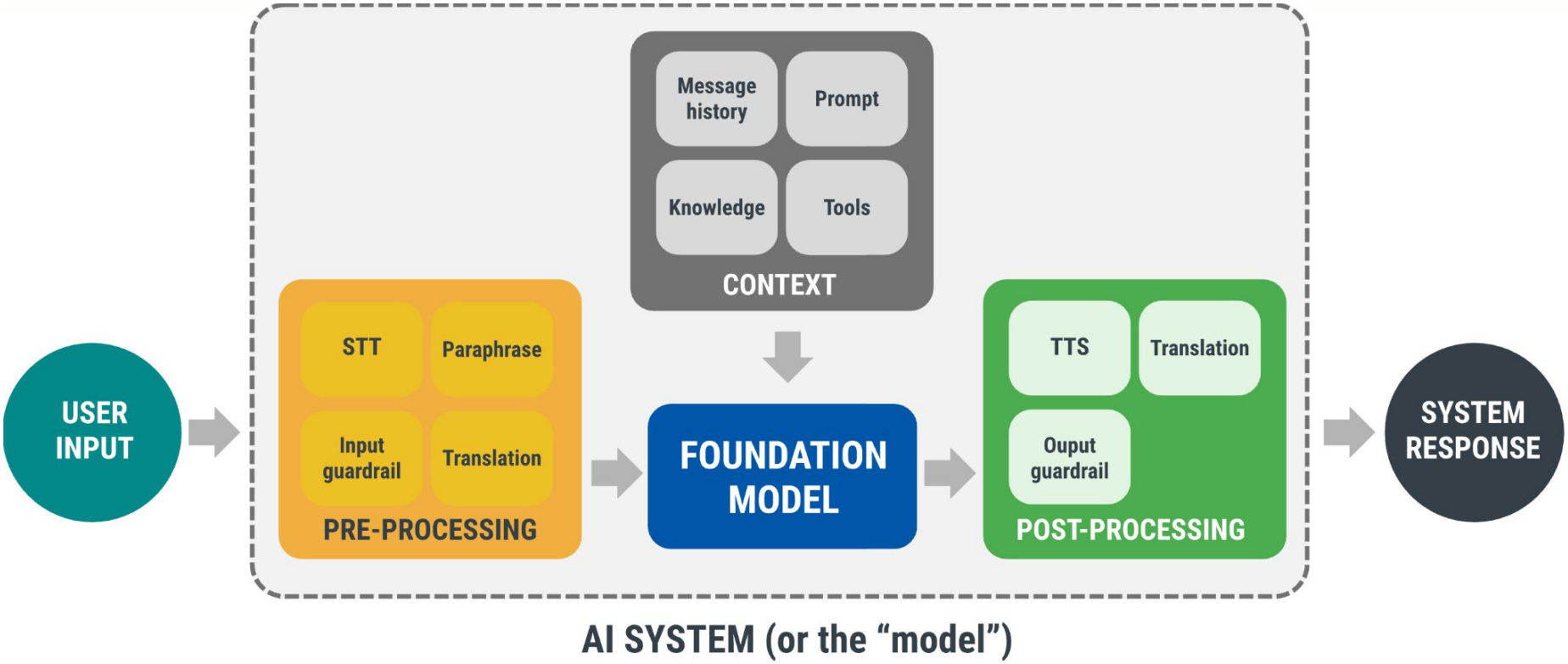
The Agency Fund




Source: Claude Code SDK, Anthropic

# What is model in model evaluation?

The Agency Fund



# It means the system, not just the LLM

The Agency Fund

## Pre-processing

**Transforms user input before it reaches the model**

→ Speech-to-text
→ Translation
→ Guardrails

**EXAMPLE**

**Detect malicious or off-topic content**

**Translate query from Pulaar to English**

→

## LLM Context Preparation

**Assembles everything the model needs to respond**

→ Conversation history
→ Knowledge base
→ External tool access

**EXAMPLE**

**Retrieve relevant agricultural content from knowledge base**

**Gather context about the farmer from the previous chats**

→

## Post-processing

**Final checks before the user receives the output.**

→ Hallucination detection
→ Text-to-speech
→ Safety checks

**EXAMPLE**

**Verify the answer is grounded in the knowledge base**

**Translate the response back to Pulaa**

The Agency Fund

# Why is model evaluation needed?

The Agency Fund

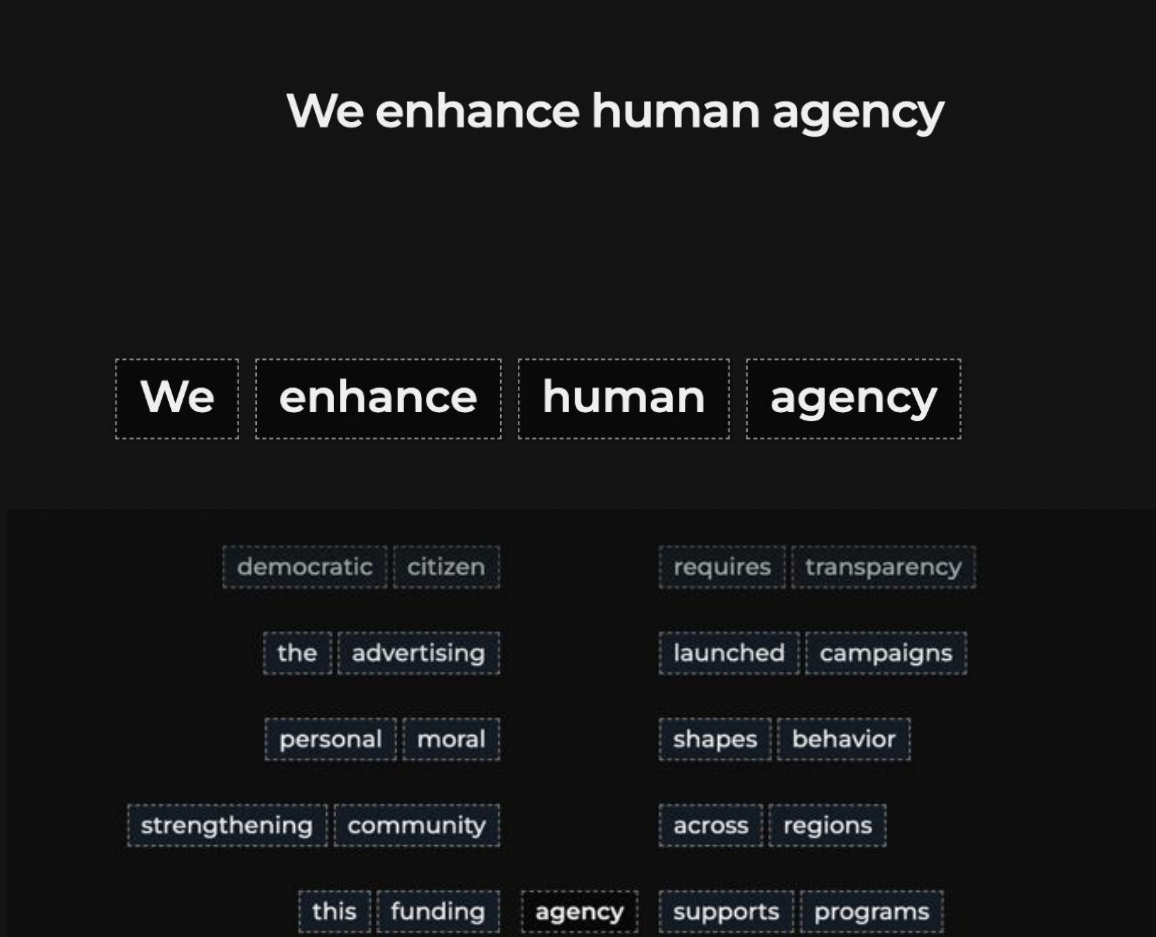
# Let us understand how these models work

https://agenticworkflowsclub.org/ai-101/

To generate text, LLMs must first translate words into a language they understand.

First a block of words is broken into tokens, the basic units that can be encoded.

The model learns meaning by observing which words appear near each other in training data.

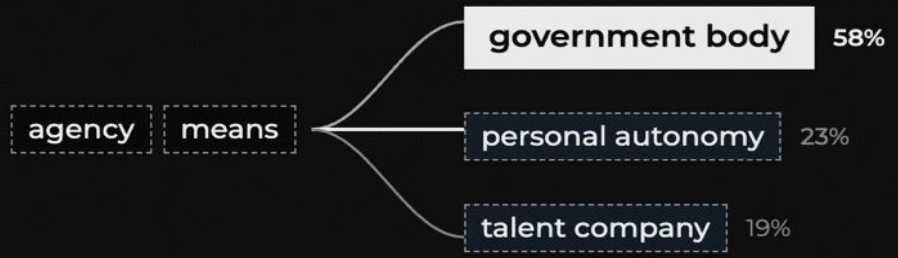



The Agency Fund

# How LLMs really work?

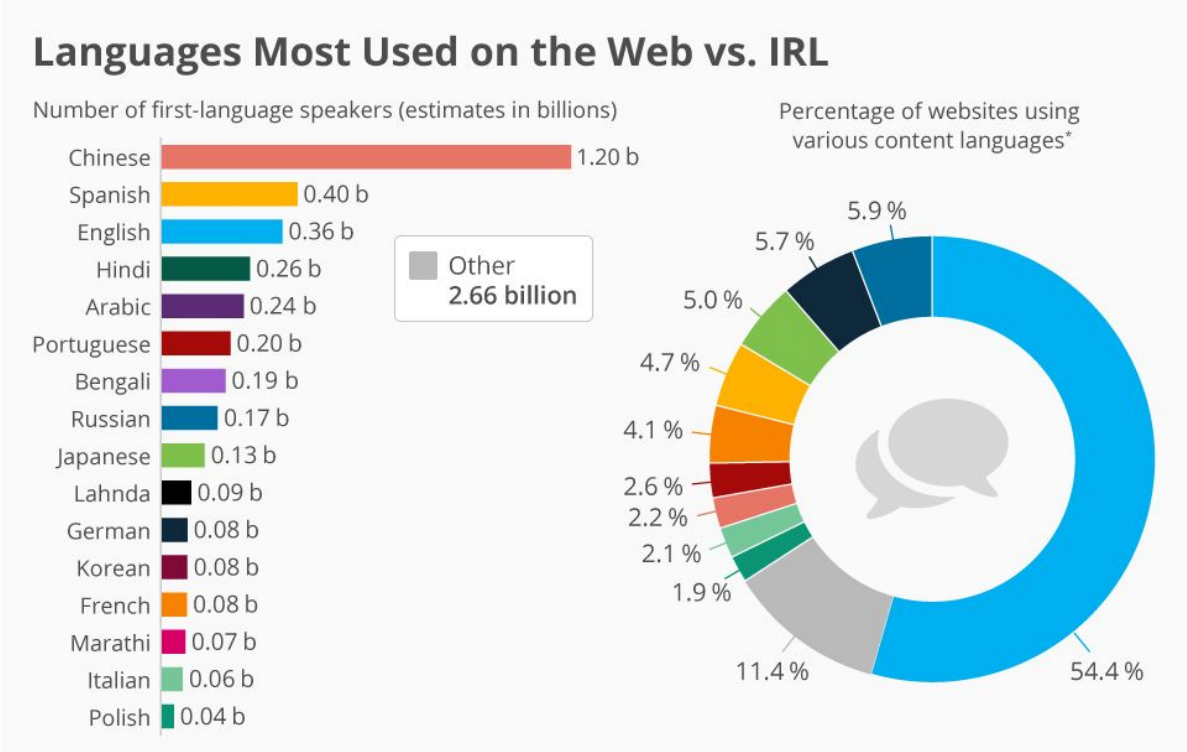
The Agency Fund

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.




https://agenticworkflowsclub.org/ai-101/

# Languages of LLMs vs humans




Source

Disparity between the distribution of languages spoken by humans compared to the content available on the internet

Hindi is the 4th most spoken language but not even on the chart

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

# Unlocking capability that exists

The Agency Fund

WHEN YOU DEMAND AN IMMEDIATE ANSWER

**Q: What is 23 − 20 + 6?**

Like asking someone to multiply 847 × 63 instantly in their head — even capable people stumble when rushed

**A: 27 ✗**

**Rushed. No reasoning. Wrong.**

GIVE IT TIME TO REASON

**Q: What is 23 − 20 + 6?**

"23 minus 20 is 3. Then 3 plus 6 is 9." — The same model, working through it step by step

**A: The answer is 9 ✓**

**Same model. Same question. Better answer.**

# Other failure modes for AI

- The **instructions** given to AI might be **vague, confusing, conflict or incorrect**
- The full **capability** of the AI model might **not be utilised**
- The best-performing model might be **too slow or too costly**

# 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

# Why has model evaluation been hard for you?

RESPONSES FROM REAL LEADERS

The Agency Fund

# Voices from the field

> Evaluation ends up on the CTO's plate and nothing gets done. There's no clear point when it is finished or even where to start.

OWNERSHIP

> If you don't define what 'good' looks like upfront, you can't evaluate. But defining what 'good' means for GenAI outputs has been a challenge in itself

STANDARDS

> The same prompt gives different answers every time. If results aren't consistent, what are we actually measuring?

RELIABILITY

> With what confidence do I deploy a health chatbot? How do I convince a client or user it is trustworthy?

DEPLOYMENT CONFIDENCE

> How do I evaluate AI responses at scale without relying only on humans? If I use an LLM to judge outputs, how do I know it isn't hallucinating too, especially with long, complex prompts for subjective responses?

LLM JUDGES

The Agency Fund

# How to evaluate

The Agency Fund

How would you judge a **human** doing the same task?

# What people said in the room

> We built a checklist and SOP upfront — defining what queries should be handled how. Evaluation happened through **learner surveys** over time, not individual interactions.

ASK THE END USER

> Check whether they followed the right process and logical framework — not just the final answer. If the process is right, the answer is likely in the right zone.

PROCESS OVER OUTPUT

> I'd interview the farmer about what they understood — not the trainer. The user experience is what matters, not the trainer's self-assessment.

ASK THE END USER

> We used historical throughput data as our benchmark — beneficiaries served, escalation rates. Detailed quality and satisfaction data simply wasn't available.

PROXY METRICS ONLY

> What if users weren't happy with the human agent to begin with? AI can sometimes exceed what any single human provides on precision, understandability, or feasibility.

QUESTIONING THE HUMAN BENCHMARK

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: HEALTH COPILOT

The Agency Fund

# What makes a human excellent at this job?

### 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 advise 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?

The Agency Fund

If a human can do it well,
**why** do you need AI?

The Agency Fund

# Name the real reason AI matters

Not **funding pressure**. Not **hype**. What is the **genuine core value** that only AI can deliver for your end users?

**EXAMPLES**

### Reach at scale

**Is the goal to serve 10× more people than any human team could?**

Will AI assisting the human help each human do more with their time and reach more people?

### Cost effectiveness

**Is AI the only way to make this viable at the unit cost required?**

Doing everything manually can be costly. Does using AI bring the overall unit cost down?

### Reduce admin load

**Does it take admin work off teachers?**

AI can handle the routine tasks so that teachers can focus on providing support and attention to students

### Always available

**Can a student ask a question at midnight, the only time they get access to a device?**

Teachers are available sparingly for 1:1 support and often not at the time of day when a child actually needs

**That answer is your north star. Keep that at the centre while making decisions and tradeoffs.**

# Define a rubric

The Agency Fund

What are the **characteristics** that your AI solution must exhibit

EXAMPLE: MATERNAL HEALTH CHATBOT

## Accuracy

The response must address the user's specific medical query instead of giving a generic answer and it must be medically accurate

## Safety

The AI system must never recommend medications or diagnose

3 broad **buckets**

## Quality

Does it work?

## Cost

How much does it cost?

## Time

How long does it take?

# A comprehensive rubric

The Agency Fund

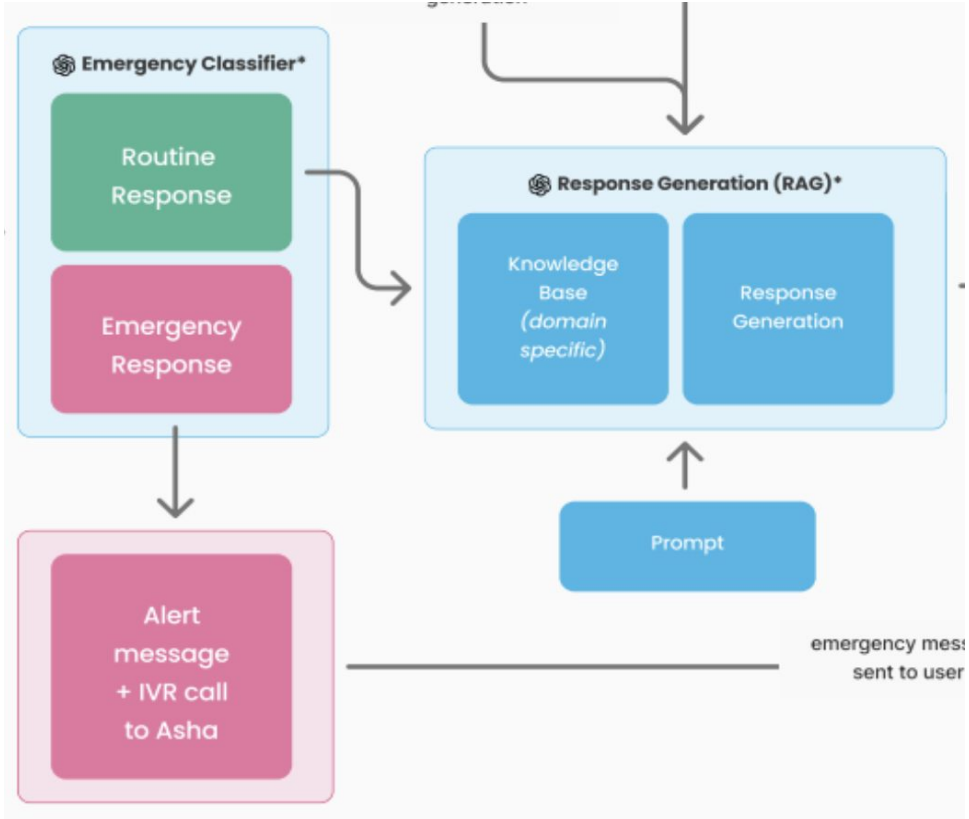
| | | |
|---|---|---|
| Accuracy/ Usefulness | The quality of the AI's response and whether it sufficiently addresses the task at hand | "The response must address the user's specific question instead of giving a generic answer and it must be medically accurate." |
| Qualitative / Branding | The "personality" and tone of the AI. | "The response must be professional and never use jargon." |
| Safety & Sensitivity | Identifying sensitive issues specific to your use case and specify any unacceptable behaviours. | "The AI system must never provide legal advice or comment on [Sensitive Topic X]." |
| Robustness & Stability | The system's ability to remain consistent when the same question is asked in different ways. | "The core answer should not change if the user uses different phrasing or synonyms." |
| Linguistic Consistency | For multi-language apps, ensuring performance doesn't drop across languages. | "The Swahili and Sheng question must receive the same level of detail as the English version." |
| Service-Level Performance | The "cost of doing business." | "The end-to-end response time must be less than 2 seconds at a cost of <$0.01 per query." |

The Agency Fund

CASE STUDY

# Noora Health

# The problem




## For every query

→ Identify if it is an emergency

→ If yes, raise an alert

→ Otherwise, generate a response using the knowledge base

# The tradeoffs

The Agency Fund

Detect emergency using strong model. Generate response if not emergency!

cost ✓ quality ✓ time ✗

Detect emergency using weak model. Generate response if not emergency!

cost ✓ time ✓ quality ✗

Generate response and detect emergency together. Discard response if emergency!

time ✓ quality ✓ cost ✗

**Neither approach were perfect** but decisions were taken based on the **context** and the **stage** of development

# Define metrics

The Agency Fund

## Statistical measures

Benchmark AI predictions against gold-standard answers for narrow, specific tasks

Precision/Recall/F1: For measuring classification accuracy

Word Error Rate (WER): For accuracy of transcription in speech recognition

✓ Fast, cheap

✗ Not very accurate

## Human-as-a-judge

Ad-hoc vibe checks

Rate AI responses as pass/fail

Provide qualitative feedback through comments

Create rubrics to evaluate AI responses

✓ Accurate, fast at low scale

✗ Costly, slow at scale, biased

## LLM-as-a-judge

Use a strong model to evaluate a weaker model's responses

Strong models often not feasible for production but great for evaluation

Needs alignment with human labels to ensure reliability as LLM judges can make mistakes

✓ Accurate, fast

✗ Not cheap

Source: HuggingFace Evaluation Guidebook

THE ROLE OF LEADERSHIP

# Velocity > speed

Help with clarity on the balance across the rubrics

## Ask the right questions

The Agency Fund

**01**
Is this even the right thing to build? Is AI even needed here?

**02**
Is 100% AI accuracy the core challenge? What if we achieved it?

**03**
Have we tested simpler approaches exhaustively?

**04**
Is our dataset representative of our users?

**05**
5% accuracy improvement or 10X cost reduction — what is better?

**06**
Should we buy an existing tool or build one?

**07**
Do we wait for a new model update that might fix the error or solve it ourselves knowing that our solution might be outdated in 3 months?

# 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

Human 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

# The golden dataset

The Agency Fund

## What it entails

→ Expected **user personas** and **interaction modes** with **ideal responses**

→ **Out of Scope** requests that should be ignored

→ **Malicious or unsafe** requests to test safety and moderation

## Why is it needed?

→ Compute metrics to **compare AI variants**

→ **Track performance** over time

→ **Prevent regressions** as you make changes to your AI system

# Creating the golden dataset

The Agency Fund

## What it entails

→ **Past interactions** between user and humans

→ **Generated** by humans or LLMs by imagining usage patterns

→ Sourced from **public benchmarks**

## Mistakes to avoid

→ **Avoid using AI** to prepare the ideal responses

→ Ask experts to **generate responses/rubrics from scratch** instead of reviewing AI responses

→ AI responses can **bias** the expert's own thinking

# 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

Human experts create ideal response or rubrics

Experts assess AI responses

Not automated

## 03

### Safety metric

At least one robust safety/guardrail metric

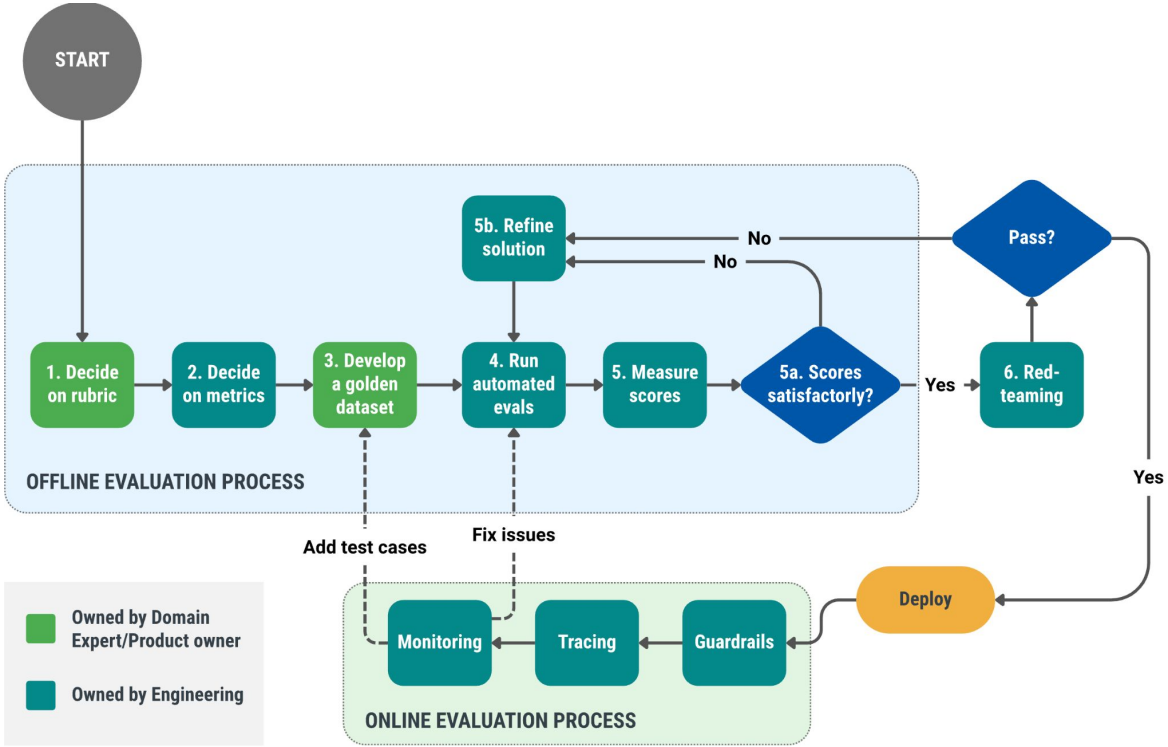
Ensure AI accuracy improvement does not compromise safety

The Agency Fund

# Test, analyse and iterate

# The AI evaluation flywheel

The Agency Fund



# 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

# Adversarial testing for high-stakes

The Agency Fund

## Structural risks

**Access to PII** — adversarial prompts can manipulate the system into leaking restricted information

**Fine-tuned models** — custom training can silently weaken built-in safety guardrails and introduce new biases

**Agentic systems** — the more autonomy a system has, the more pathways exist for failure

## Deployment risks

**Long conversations** — a small misunderstanding in turn 1 can amplify and drift into unsafe territory by turn 10

**High-risk domains** — in maternal health or financial planning, failure causes severe harm

**Population-scale** — at scale, even improbable edge cases will occur with real frequency and you do not know how users will interact with the system

# Red teaming

The Agency Fund

## Plan

Define scope & team

→ **Set redlines** by defining what must the system never do?

→ **Assemble a team** of diverse mix of engineers and domain experts

→ **Choose the method**, set clear guidelines and processes

## Probe

Adversarial mindset

→ **Attack** by trying to distract, exploit, and stress-test the system

→ **Hunt failures** like long conversations and sensitive domains first

→ **Log everything** including the inputs, outputs, and the contexts for every failure found

## Prioritize

Not all failures are equal

→ **Rank risks** by severity (impact) and likelihood (frequency)

→ **Assign owners** for vulnerabilities with the highest priority

→ **Re-test** to confirm the fix worked and nothing else broke

# Resources

The Agency Fund

LLM Evals: Everything You Need to Know – Hamel’s Blog

Your AI Product Needs Evals – Hamel's Blog

What We’ve Learned From A Year of Building with LLMs

Using LLM-as-a-Judge For Evaluation: A Complete Guide – Hamel's Blog

A Field Guide to Rapidly Improving AI Products – Hamel’s Blog

An Open Course on LLMs, Led by Practitioners

Galileo blog

Jason liu blog

AI engineering book

Shreya Shankar blog

Eugene Yan blog

AI.engineer (youtube channel)

Interconnects (open AI research)

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

# Key drivers

The Agency Fund

## Engagement and retention

Are users using your product? Are they coming back?

## How are decisions made?

Are decisions made based on intuition or are they grounded in data?

## What is the impact of any change?

How will users react? How can you confidently experiment with ideas while protecting your downside?

The Agency Fund

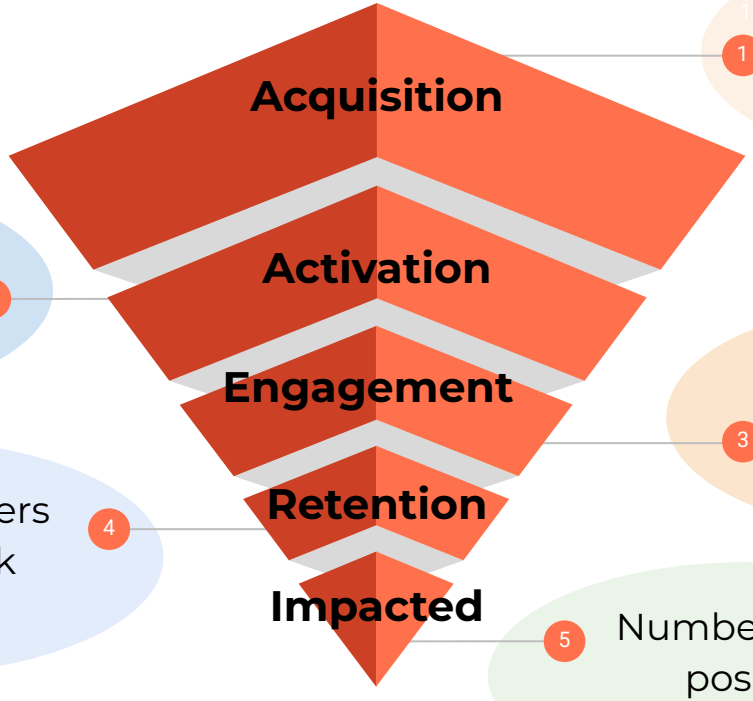
# Where would your users first get stuck?

The Agency Fund

# The User Funnel

# Find exactly where your users leak

The Agency Fund




1 Number of users brought in

2 Number of users who tried

3 Number of users finding some value

4 Number of users coming back regularly

5 Number of users positively impacted (L3/L4)

# One metric per stage

The Agency Fund

| Acquisition | Activation | Engagement | Retention |
| --- | --- | --- | --- |
| Not just app installs, active recruitment | Those who got some value vs those who only installed it | Frequency (how often) and depth (how meaningful) of usage | One-time use vs habit, whether you solved something real or just satisfied curiosity once |
| **A farmers creates an account on an agri-support bot** | **The farmer asks their first crop question** | **If the farmer logs in and leaves in 10 seconds, they are not engaged** | **The farmer keeps coming back to ask more queries regularly** |
| **What does it cost to bring one real user in?** | **% of new users who hit their first value moment within Day 1–3** | **What share of monthly users are actually active each day** | **The ratio of users active last week who returned this week** |

The Agency Fund

# What action by a user signals they actually received value?

# Define what active means

The Agency Fund

✗ User opens the app

✓ Specific action that signals real value was received

**EXAMPLES**

→ Pregnant mother asks a medical question

→ Farmer asks a specific crop query, not just opens the chatbot

→ Student completes a practice problem, not just watches a video

→ Worker submits a job application, not just browses listings

# From gut feeling to data-driven

The Agency Fund

## Without a user funnel

"I think users dropped off after onboarding."

"Maybe the interface is confusing?"

"Maybe they are not returning because the content isn't useful."

## With a user funnel

- 68% of users never return after Day 3
- Activation rate drops from 71% to 34% on low-connectivity devices
- Users who complete onboarding step 3 retain at 3× the rate of those who skip it

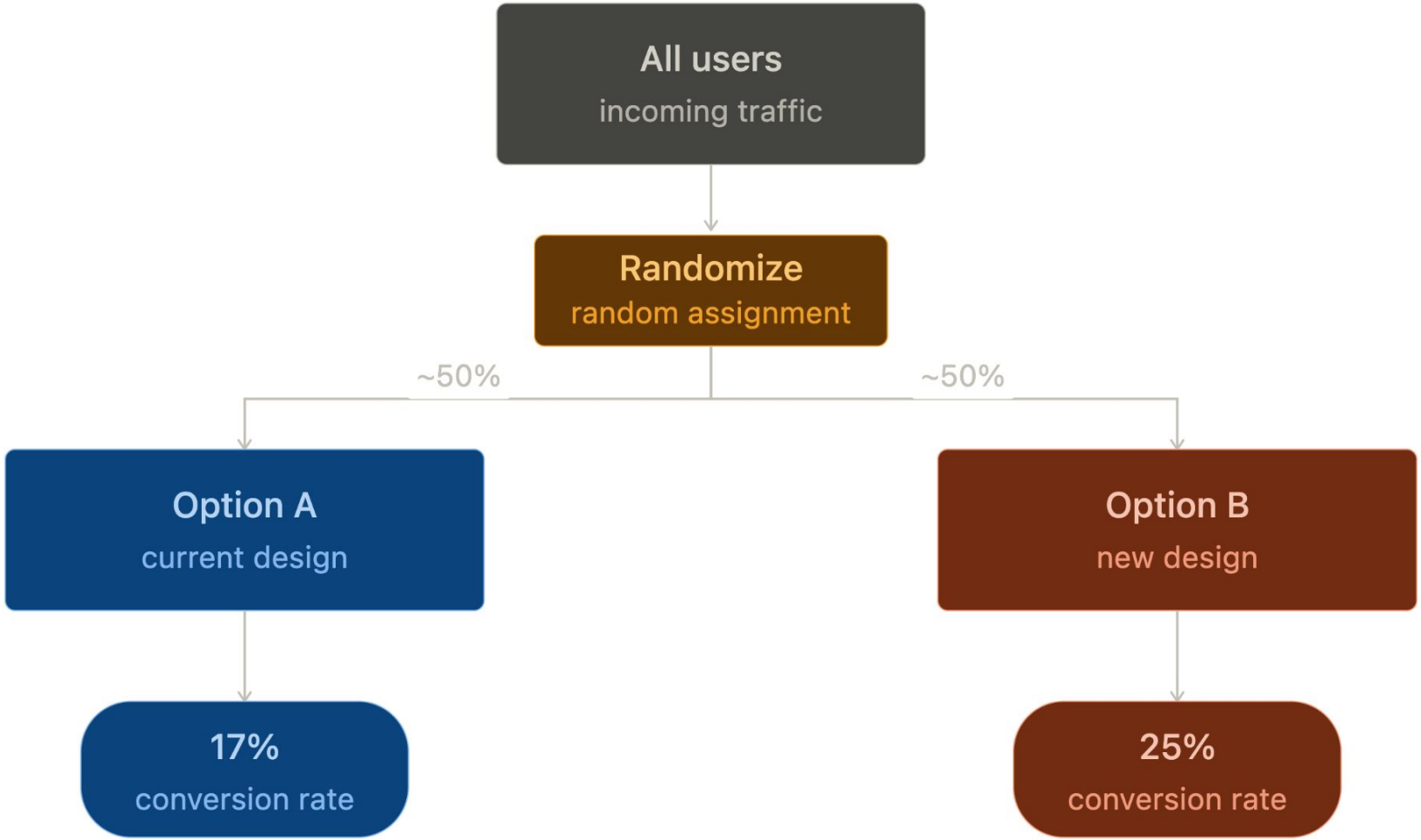
The Agency Fund

# How do you currently test ideas?

The Agency Fund

# What is an experiment?

The Agency Fund



# Example

The Agency Fund

**Prompt A — baseline**

Simple, generic instruction

SYSTEM PROMPT

```
You are an agricultural assistant. Answer the farmer's question.
```

MODEL OUTPUT

Yellow leaves on maize can be caused by nitrogen deficiency, overwatering, or disease. Consider applying fertilizer or adjusting irrigation. Consult a local agronomist if the problem persists.

LEVEL 1

| | |
|---|---|
| Actionability | 34% |
| Local relevance | 21% |
| Hallucination-free | 61% |

LEVEL 2

**3 / 10** 7-day return rate
farmers who asked a follow-up within a week

Each dot = 1 farmer out of 10

**Prompt B — enhanced** winner

Context + grounding added

SYSTEM PROMPT

```
You are an agricultural assistant supporting smallholder farmers in East Africa. Answer the farmer's question, grounding your response in the provided knowledge base. Give one specific, actionable first step.
```

MODEL OUTPUT

**First step:** Check the lower leaves for a pale yellow color starting at the tip — this pattern typically signals nitrogen deficiency. Apply **CAN fertilizer at 50kg/acre** if you haven't top-dressed in the last 3 weeks. Yellowing that starts at the edges instead may be a potassium issue — bring a leaf sample to your nearest agro-dealer.

LEVEL 1

| | |
|---|---|
| Actionability | 81% |
| Local relevance | 74% |
| Hallucination-free | 89% |

LEVEL 2

**7 / 10** 7-day return rate
farmers who asked a follow-up within a week

Each dot = 1 farmer out of 10

USER QUERY

"My maize leaves are turning yellow from the bottom up. What should I do?"

For the same user query, how do two different system prompts compare with each other

The Agency Fund

# Why do we experiment?

# Motivation

Improving **cost-impact** at **scale**

01 CAUSALITY

## Confidence that a change is positive

Be more than just data-driven

02 DERISK

## Derisk changes before you scale them

Avoid harm to users and wasting resources

03 CULTURE

## Democratization of ideas

Sets a culture where ideas are tested on merit over authority

04 COMPOUNDING

## Continuous innovation

“Minor” enhancements compound over time

EXAMPLES

The Agency Fund

# Promise and Risks

ITERATION

## Rocket Learning

*Early childhood education, India*

A WhatsApp-based ECE program successfully engaged parents yet showed no learning gains after 9 months. The team incorporated iterative learnings to deploy an intensified variation.

### Result

**Significant positive learning outcomes** within 6 more months.

Early evaluation + willingness to **iterate** can lift up a null program

PROMISE

## Living Goods

*Community health workers, Uganda*

A program to reduce child mortality shared performance data with two implementing partners mid-study, allowing them to adapt their approach.

### Result

Partner who adapted: **28% reduction**
Partner who didn't: **insignificant**

Willingness to **act** on evidence made the difference

CAUTION

## Reach Digital Health

*Maternal health SMS service*

Used AI to optimize onboarding for an SMS service supporting new mothers. Institutional approval delays left no time for proper model or product testing (Levels 1 and 2).

### Result

Survey completion **decreased**, despite users sending more messages.

Skipping early evaluation stages has real **risks**

The Agency Fund

# When should we experiment?

# What big tech suggests

Industry leaders treat experimentation as a core operating principle

## Duolingo

*"Test everything"* is one of our operating principles — we're constantly iterating on different features to see how they impact our most important metrics."

*Behind the Metric, 2024*

## Microsoft

"I'm very clear that I'm a big fan of test everything, which is any code change that you make, any feature that you introduce has to be in some experiment. Because again, I've observed this sort of surprising result that even small bug fixes, even small changes can sometimes have surprising, unexpected impact."

*Ronny Kohavi, 2023*

## Google

"We evaluate *almost every change* that potentially affects what our users experience."

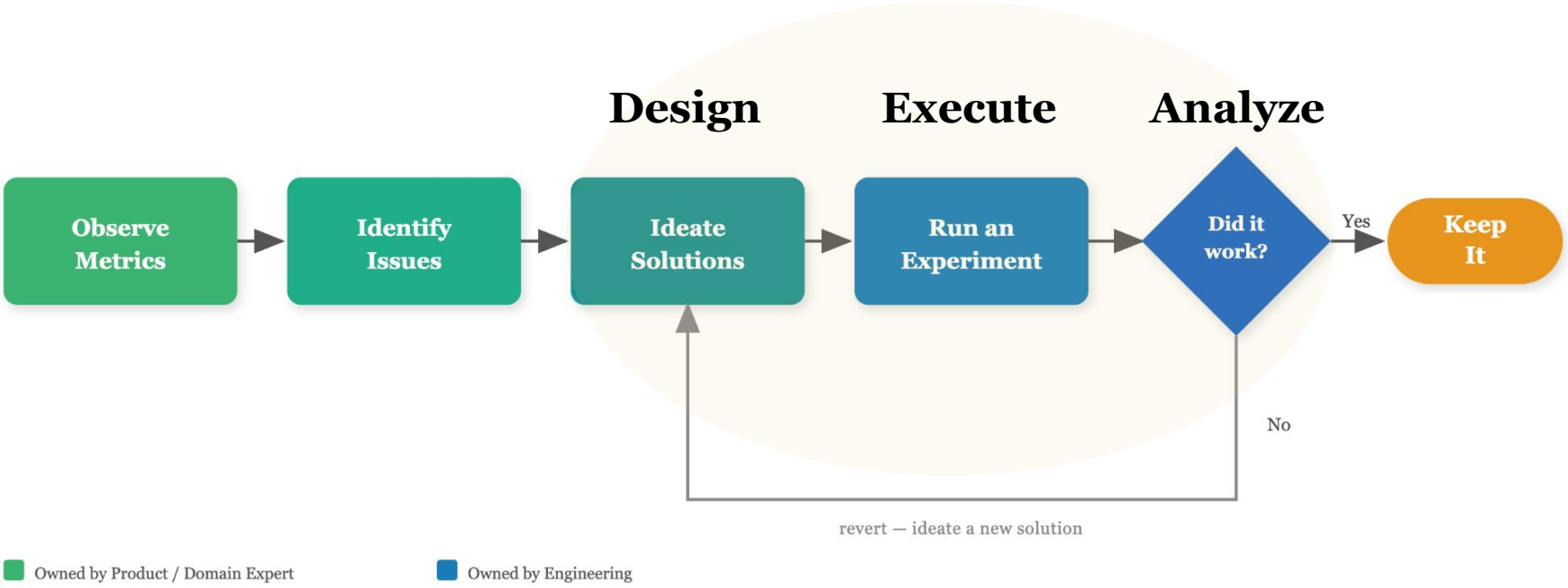
*Overlapping Experiment Infrastructure, 2010*

*Modern experiment platforms can bring these practices to any organization.*

The Agency Fund

# How to experiment?

# The L2 evaluation flywheel



# Some places to experiment

The Agency Fund

## New Features

Test new capabilities before full release to validate they deliver value

## UX Changes

Interface or workflow improvements that affect how users interact with your product

## Incremental Rollouts

Slowly ramp up new features to catch issues early before full deployment

## Bug Fixes

Even small fixes can have unexpected effects

# But first: Qualitative Research

The Agency Fund

## Why qualitative first?

**Hypothesis generation** — Understand why something should work before building it

**Diagnosis** — When an experiment shows an effect, why did it work (or not)?

**Save resources** — Avoid building the wrong thing. These techniques are designed to work with small samples.

## Key methods

**User Interviews** — Talk to users about their experience and needs

**Journey Mapping** — Map the full user experience to find friction points (reference)

**Persona Development** — Build profiles of key user types to guide design (reference)

**Usability Testing** — Watch 5 users interact, catches up to 85% of issues

**Surveys** — Open-ended questions to surface unexpected insights

IDEO Design Kit for Human-Centered Design

# Problem & Hypothesis

The Agency Fund

## 01 Define the Problem

What are we trying to solve or learn? Why does it matter? What data supports our understanding?

## 02 Form Your Hypothesis

What specific change do we expect? What assumptions are we making? What variants will we test?

## 03 Design the Experiment

Which users will we target? What metrics will we track? What does the experience look like?

## 04 Plan for Action

Who are the stakeholders? What are the launch and rollback criteria? What are the next steps?

See also the Experiment Design Template as used by The Agency Fund

The Agency Fund

**“When a measure becomes a target, it ceases to be a good measure”**

- *Goodhart’s Law*

# Guardrail metrics

What **quality** or **counter-metrics** exist to ensure you are **benefiting users**, not just **hitting your target** number?

## Health SMS Service

**Target**

Improve onboarding completion

**Guardrail**

*Rate of medically relevant questions doesn't drop*

## Digital Advertising

**Target**

Revenue growth

**Guardrails**

*Click-through rate, conversion rate, time on page*

## Any Product Team

**Target**

Team-specific metric

**Guardrail**

*"Each team had target metrics, and each was a guardrail for the other teams" — Jorge Mazal, former CPO of Duolingo,* User Funnels for the Social Sector

The Agency Fund

# Examples of experiment methods

PRE / POST

## Before and After

Deploy your change for everyone

Compare metrics before your change with the same metric after

**High-impact, “no-regrets” changes; unavoidable spillovers; treatment can’t be parallelized**

RECOMMENDED

## A/B Testing

Randomly split users into two groups

One group sees the original product while the other sees your change

Compare metrics across both groups

**Rapidly test multiple ideas reliably with control**

The Agency Fund

# Things to remember

## 01 Randomisation

Any observed difference can be **attributed to your change**, not to pre-existing differences between users.

A **biased comparison** produces a misleading result

## 02 Sample Size

It should be large enough to **detect real effects**

A small sample **amplifies noise**, making it impossible to separate a genuine improvement from random chance

**Real improvements** in small groups will seem insignificant and be **missed**

# Questions before experimenting

The Agency Fund

## Technical

How do we build this?

Can we filter/identify the target users?

Are metrics available in the data warehouse and computed over the right time frame?

Does the user need to update their app to see the change?

## Operational

Which stakeholders need to coordinate?

Do we need to recruit or train?

Are the materials and content ready?

# Use Feature Flags

The Agency Fund

## Safer Deployment

Roll out changes gradually. If something breaks, turn it off instantly without a new release.

## Personalization

Show **different experiences** to different user groups based on their needs or context.

## Experimentation

Route users to different variants (test vs. control) to measure what works

More: LaunchDarkly Guide | Open Source Tools Compared | OpenFeature Standard

# Do all users update their apps?

The Agency Fund

## 01
### The Challenge

- Not everyone **updates** immediately
- Running the experiment only on early updaters introduces **bias** as they tend to be power users.

## 02
### Pre-Check

- Do early updaters' **metrics differ** from late updaters?
- Check your **time-to-update curve** to decide how long to run the experiment

## 03
### Recommendation

- Analyze all users according to their **original group assignment**, regardless of whether they updated
- This "intent-to-treat" approach **minimizes bias**.

**Trade-offs:** **decision speed** vs. **population representation** vs. **analysis complexity**

# What next

The Agency Fund

01

Launch, scale, or rollback

02

What did we learn?

03

New hypotheses?

04

Next experiment?

The Agency Fund

# Wrap up

# Building Capacity: Experiments Council

The Agency Fund

**Idea**: A small team reviews and approves all experiments through a lightweight **checklist** to ensure **quality** while building **organizational capacity**

## Teach Newcomers

First-time experimenters **learn** proper **experiment design**, sizing, and implementation details

## Interpret Results

The council can help teams make sense of their data after experiments have run, reducing **misinterpretation**

## Experiment Registry

Checklists also serve as an **archive** building **institutional memory** of what's been tested and learned

## Spread Best Practices

Checklists are a practical way to **disseminate updated experimentation standards** across the organization

# Building Capacity: Other Suggestions

The Agency Fund

## 01 Build a Data Culture

Set key organizational **metrics** and **dashboard** them
Build your user **funnel**
**Warehouse** your data as a single source of truth for metrics

## 02 Adopt Accessible Tooling

Use tools and processes that **lower the barrier** and enable non-experts to run experiments

## 03 Create an Experiment Registry

Track what's been tested and what was learned. Prevents **duplication** and builds **institutional knowledge**.

## 04 Leverage Your M&E Experts

**Share knowledge** across teams, learn from each other's practices, and parallelize insights + innovations.

TAF Blog: Bringing a Culture of Experimentation to Any Organization | Youth Impact's A/B Testing Toolkit

# Minimum Viable Evaluation (MVE)

The Agency Fund

## Safer Rollouts

- Test on a **small subset of users** before releasing to everyone
- **Monitor key metrics** on the subset and flag anomalies
- Have your MEL team act as an **experiment council** to others
- A bad rollout you **catch early** costs a fraction of one you catch late

## MVE checklist

- Build your **user funnel**
- **Track two metrics**
  e.g. activation rate + 7-day retention
- **Talk to 5 users** who dropped off
- Ideate, run an **experiment**, repeat

**Remember**: **No data collection required**. All data is already being collected. You can start right away!

# Iterative Learning: Youth Impact

The Agency Fund

| Innovation | | Efficiency gain* |
|---|---|---|
| **Cost- reducing test** | Weekly vs. biweekly dosage | 0.11 |
| | Same vs. different tutor | 0.5 |
| | Appointments vs. call centre scheduling | 0.10 |
| **Effectiveness-enhancing test** | Call with/without SMS | 0.32 |
| | WhatsApp Videos | 0.00 |
| | Homework assigned | 0.29 |
| | Caregiver nudge | 0.00 |
| | Testimonials | 0.00 |
| | Caregiver Co-tutors | 0.22 - 0.30 |

Improvements in cost or efficacy for a phone-based tutoring service

*Defined by authors as $e = \Delta/\alpha$ where $\Delta$ refers to a change either in impact or cost, with $\alpha$ the baseline.

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

# Parting questions

The Agency Fund

**01** What's your organization's journey with experimentation so far?

**02** What experiment are you working on or aiming for right now?

**03** What practical concerns crossed your mind about designing, running, and analyzing experiments?

**04** What do you envision for the future of experimentation within your organization? What will it take to get there?

The Agency Fund

LEVEL 03

# User Evaluation

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

The Agency Fund

# Why is this level important?

The Agency Fund

# Engagement does not mean life-changing

People may use a product a lot but it may not change their life meaningfully

EXAMPLES

The Agency Fund

# Liking is not the same as life-changing

## AI tutoring app

- ✓ Student uses the app to help them with homework
- ✗ Does not actually master the curriculum (e.g., when they use AI as a crutch)

## AI health chatbot

- ✓ Patient consults the chatbot a lot
- ✗ Does not act upon recommendations given by the chatbot

# Liking is not the same as life-changing

The Agency Fund

## What L1 & L2 tell you

- The model responds accurately
- Users open the product
- Sessions are increasing
- Users are coming back
- People say they find it helpful

≠

*The Gap*

## What L3 asks instead

- Are users learning, gaining new knowledge or updating beliefs?
- Do users report feeling supported, motivated, and capable after interactions?
- Does the product energize or deplete a users' drive to pursue goals independently?
- Are there indications of confusion, anger, or emotional distress?
- Does using the product provide broader effects on users' overall quality of life beyond momentary feelings?

The Agency Fund

# Intermediate Outcomes

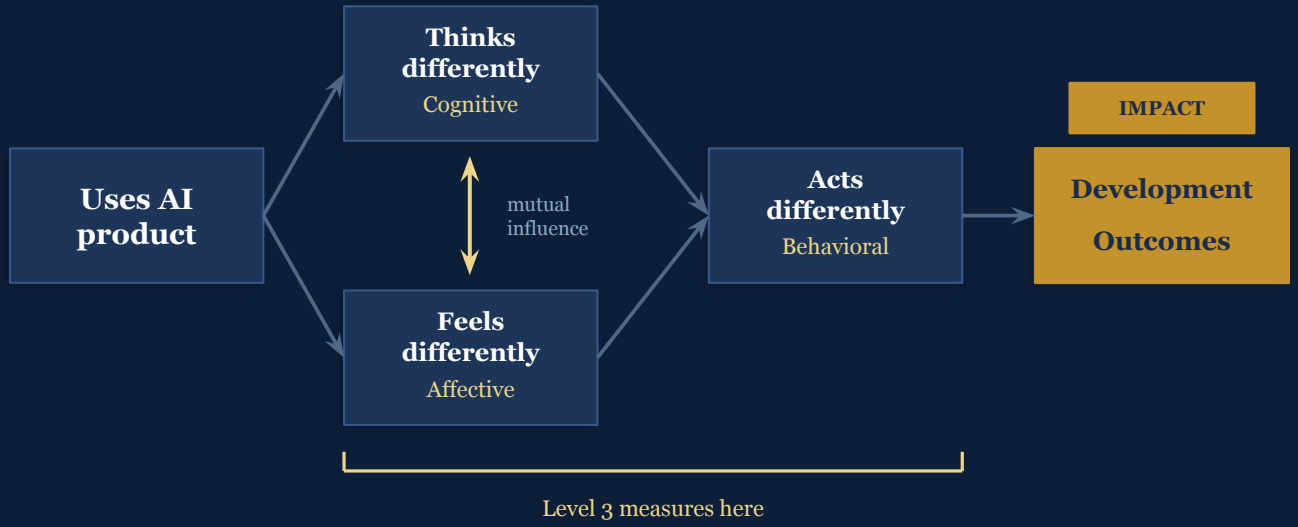
The Agency Fund

INTERMEDIATE OUTCOMES

# Is the engaged user thinking, feeling, or acting differently as a result of the product in ways that predict improved life outcomes?

# Measure the stepping stones to impact instead

Cognitive and affective pathways interact — either or both can activate behavior change.




**WITHOUT LEVEL 3**

**Years**

to see if impact outcomes actually improve

**WITH LEVEL 3**

**Weeks**

to see if users are thinking, feeling, or acting differently

# Key issues to address to enable this

The Agency Fund

## MEASURES

**Which specific user-level changes matter to our Theory of Change?**

Can we measure these short-term changes relatively cheaply and frequently?

## ATTRIBUTION

**Can we claim these changes are caused by our AI product?**

Establish the causal chain before investing in a Level 4 RCT.

## TRAJECTORY

**Are metrics heading in the right direction?**

Do sub-groups behave differently? Track direction, not just magnitude.

## MALLEABILITY

**Can we shift metrics by altering the product experience?**

Do users show increased drive to act when we intervene with product improvements?

## PERCEPTION

**Do users feel more empowered to act?**

Do they have a clearer understanding of their next steps, even if they have not yet taken action? User perceptions can predict outcomes.

The Agency Fund

# Six dimensions of user change

## Cognitive

Are users learning? Are they gaining new knowledge or updating beliefs? Do they demonstrate improved skills or decision-making ability as a result of engaging with the product?

## Affective

How does the product make users feel? Do users report feeling supported, motivated, and capable after interactions, or are there indications of confusion, anger, or emotional distress?

## Behavioral

Is the user doing something different? Are users taking small but meaningful actions that predict longer-term development?

## Motivational

Does the product energize or deplete a users' drive to pursue goals, learn, or act independently?

## Social / Relational

Does use of the product affect the user's human-to-human relationships and broader social functioning?

## Well-being

Does using the product provide broader and more distal effects on users' overall quality of life and psychological health beyond momentary feelings/mood?

EXAMPLES

# Example measures for each dimension

## Cognitive

e.g., Comprehension, knowledge acquisition and retention, belief updating, critical evaluation of information, metacognitive awareness (e.g., accurate calibration of what one does and does not know), perceived clarify, complexity of reasoning during/following interaction.

## Affective

e.g., Mood, emotional valence and arousal, frustration, confusion, emotional granularity, felt support, sense of safety, sense of belonging, perceived empathy, trust, or comfort interacting with AI.

## Behavioral

e.g., Application of new information, intent to try recommended behaviors, observable shifts in interaction patterns(e.g., asking more complex questions, prompt sophistication) that proxy for longer-term development outcomes, help-seeking behavior.

## Motivational

e.g., Intrinsic motivation, curiosity, self-efficacy, perceived autonomy, goal commitment, persistence (not the same as perseveration), dependency (i.e., reduced willingness to attempt tasks without AI assistance).

## Social / Relational

e.g., Social displacement (substitution of AI for human interaction), loneliness, perceived social support, quality of interpersonal communication, willingness to engage with others, trust in human versus AI sources of social support and information.

## Well-being

e.g., Life satisfaction, meaning and purpose in life, flourishing, burnout (i.e., in professional contexts), perceived agency/control over one's environment.

The Agency Fund

# Evaluating the right users

**Who is being evaluated?**

Level 3 evaluates users who have received **adequate product dosage**, not first-time visitors or one-off interactions

CASE STUDY: FARMER.CHAT

Farmers who have **asked 3 or more agronomic queries** and received recommendations.

The Agency Fund

# How to evaluate at this level?

# The L3 evaluation workflow

The Agency Fund

## Generate hypotheses based on a theory of change

Define intermediate cognitive, affective, or behavioral outcomes that could be linked to your targeted social impact and validate with quantitative or qualitative methods

## Identify outcome metrics

Identify outcome metrics that reflect psychologically and behaviorally meaningful user interaction through surveys, interviews, conversation logs and other interaction data

## Define guardrail metrics and measure potential harm

Measure potential harm and track unintended consequences

## Construct proxies for long-term development outcomes

Construct proxies by collecting indicators of links between the intervention, its adoption, underlying mechanisms, and ultimate development outcomes and validate them through RCTs

## Experiment to improve key metrics and run process evaluations

Assess how model or product changes influence Level 3 metrics and run process evaluations to understand why and when they don't change

STEP 01 · HYPOTHESIZE

The Agency Fund

# Generate hypotheses based on a theory of change

- **Start from impact:** Identify why you are measuring level 3 outcomes first. How are they linked to your targeted social impact? Measure what comes up instead of whatever is easiest to measure.
- **Link to product choices:** If your product is improving certain cognitive, affective, or behavioral outcomes, how does that inform your product innovation or revamp
- **Validate early:** Test the hypothesis through user interviews and a light literature review before committing engineering effort to instrumentation.
- **Ground it in theory:** Use psychological and behavioral literature (cognitive, affective, behavioral theories) to justify why a shift in that outcome should matter.

The Agency Fund

# Three ways to identify outcome metrics

Triangulation using more than one metric makes **intermediate outcomes trustworthy**

## A

### Interaction data

Analyze logs and Level 2 metrics for psychologically meaningful patterns like frequency of queries, follow-up rates, feature uptake, query complexity.

PROXY SIGNALS

## B

### Primary data

Ask users directly. Short in-chat surveys for subjective experience and longer off-platform surveys, quizzes, and interviews for long-drawn behaviour changes.

SELF-REPORT

## C

### Conversation logs

Mine the text itself. Sentiment analysis, topic modeling, LLM-based text analysis can score user utterances for cognitive and emotional change at scale

NLP & TEXT ANALYSIS

The Agency Fund

# What constitutes a psychologically and behaviorally meaningful user interaction for you?

The Agency Fund

# Look for on-platform behaviors that proxy for change

| | |
|---|---|
| **Frequency of queries** | How **often** do users **query** your AI service? |
| **Depth of queries** | How **deeply** do users **query** your AI service? |
| **Changes in language** | How **complex** are user queries? With expertise, their vocabulary, syntax, and linguistic sophistication may advance. |
| **Follow-up question rate** | Are users **persisting** in a line of questioning? When a user asks a follow-up question, they are likely engaging in deeper thinking. |
| **Feature utilization** | Do users **follow recommendations** or **use** suggested tools? Can signal levels of trust and motivation. |

EXAMPLES

The Agency Fund

# Look for on-platform behaviors that proxy for change

| | |
|---|---|
| **Frequency of queries** | Increased frequency of interaction with an AI tutor can signal **confidence, curiosity and learning gains** |
| **Depth of queries** | Students progress from asking **basic factual questions** to more **advanced specific inquiries**, it indicates **knowledge growth** |
| **Changes in language** | Students tend to write **lengthier, more complex sentences** when engaging in **in-depth learning tasks** |
| **Follow-up question rate** | Longer conversational exchanges often indicate **intellectual curiosity, active learning**, and **higher-order thinking** |
| **Feature utilization** | When learners **consistently** accept AI suggestions and **complete optional exercises**, this can signal trust in the AI and a high **motivation** to learn |

The Agency Fund

# Just ask users

## IN-CHAT SURVEYS

### Short, validated, naturally placed

- Adapt or use validated, well-tested psychological scales, don't write from scratch.
- **≤ 3 items**. Avoid fatigue or annoying users. Mix multiple choice, rating scales and open-ended questions.

  *e.g. "How did you feel during the conversation?" (use emojis)*
- Ensure feedback is tied to a concrete task. Avoid pop-ups or making it feel an **instrusive add-on**

  *e.g. ask for feedback naturally within the chat flow*
- Capture constructs beyond habituation to distinguish user adaptation to the tool from meaningful gains

  *e.g. "I understand why the recommended solution by AI works" can assess self-efficacy and cognition*

## OFF-PLATFORM SURVEYS

### Decoupled from the product to avoid halo effects

- Longer surveys, interviews and focus groups to delve into how users' attitudes or habits have changed over time

  *e.g. a student might say, "I never liked math before, but now I find myself challenging myself with problems for fun"*
- Observer reports from teachers, supervisors, family can validate self-reported survey items

  *e.g. "I noticed my child now approaches homework more confidently"*
- Link AI usage to objective outcomes external to the tool

  *e.g. Level 2 metrics correlation with exam scores, task completion rates, and even health indicators*

  *e.g. writing assessment before and after using AI writing assistant reviewed by blind graders*

The Agency Fund

# Mine conversations for signals of change

## Sentiment analysis

**Score sentiments of user messages over time**

Trends from negative to positive tone can indicate growing comfort

Negative sentiment spikes flag frustration

## Topic modeling

**Track emergent themes over sessions**

Moving from fundamentals to advanced concepts indicates cognitive growth

Keyword analysis to surface unexpected themes (“exam anxiety”)

## Linguistic Inquiry & Word Count (LIWC)

**Dictionary-based text analysis tool** maps **words** to **psychological categories** (like emotion, social words, cognitive processes), correlated with psychological states

*e.g. an increase in first-person plural pronouns (“we, us”) might indicate users feel more socially connected*

## LLM-based analysis

**Prompt an LLM to score text to detect psychological constructs**

e.g. *track self-confidence, loneliness from conversations over time*

*Be **careful** when using an LLM to evaluate LLM-user interactions*

The Agency Fund

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

AI agents can empower or **disempower** users. In the social sector, user agency is a critical guardrail. Is it **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

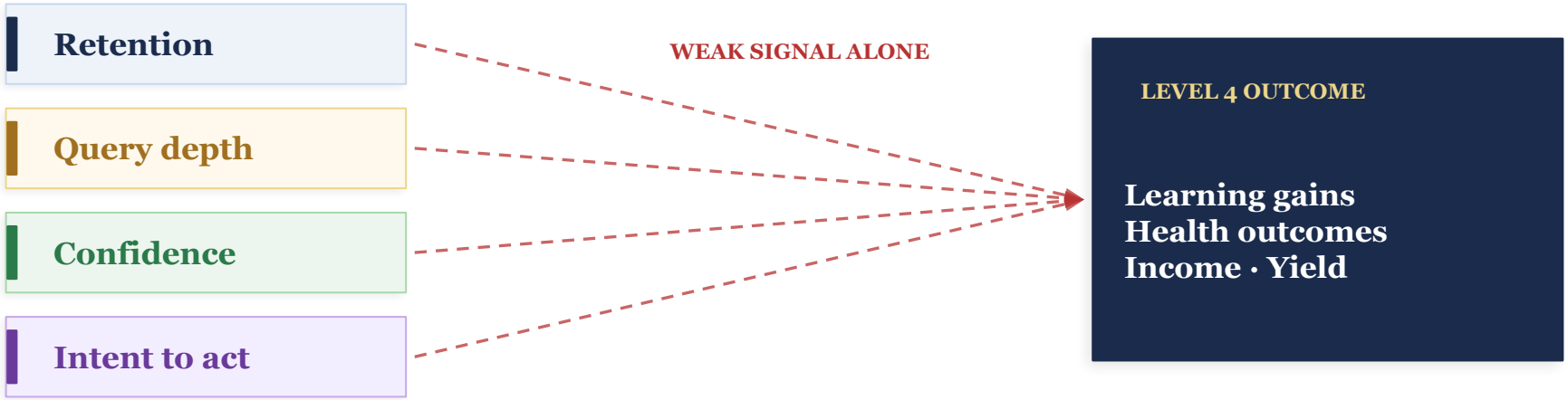
### Can they actually plan and execute the goal without the help of AI?

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

# No single metric predicts Level 4




**Each metric captures a sliver**

**None captures the whole**

STEP 04 · PROXY

The Agency Fund

# Combine them into a surrogate index


**Caveat:** assumes unconfoundedness, surrogacy, and comparability. Validate via L4 RCT.

Source

STEP 05 · EXPERIMENT

The Agency Fund

# Experiments tell you what moved. Process evaluations tell you why

## EXPERIMENTAL METHODS

### A/B testing

Feature A vs. B on L3 outcomes

### Multi-armed bandits

Adaptive allocation based on performance

### Holdout testing

AI vs. non-AI to isolate product effect

## PROCESS EVALUATION · WHEN METRICS WON'T MOVE

### Interpretation

Are users receiving and interpreting AI outputs as intended? Does this differ by sub-group?

### Opportunity to act

What social norms, competing demands, or structural constraints are blocking users from acting on AI recommendations?

### Externalities

Has the tool shifted the roles or behaviors of others in the program ecosystem, intentionally or not?

# MVE Checklist

01 OUTCOME METRICS

## Are your outcomes tied to the theory of change?

Define 1-2 decision-relevant cognitive or behavioral outcomes. Include at least one early-warning indicator of harm — e.g., over-reliance or disengagement.

02 BEHAVIORAL + SELF-REPORT

## Does your measurement combine trace data and survey?

Pair at least one behavioral or trace metric with a brief self-report measure (3 items or fewer) to capture meaningful user change.

03 EXTERNAL VALIDATION

## Do on-platform measures reflect real-world outcomes?

Include a minimal external check: a focused group discussion, offline data, or stakeholder validation to anchor platform metrics to reality.

04 EXPERIMENTAL TESTING

## Are you testing what actually moves the needle?

Consider simple experimental methods — such as A/B tests — to assess the effect of product changes on your selected outcomes.

The Agency Fund

# Examples

# **Example application:** GPT-based tutors

The Agency Fund

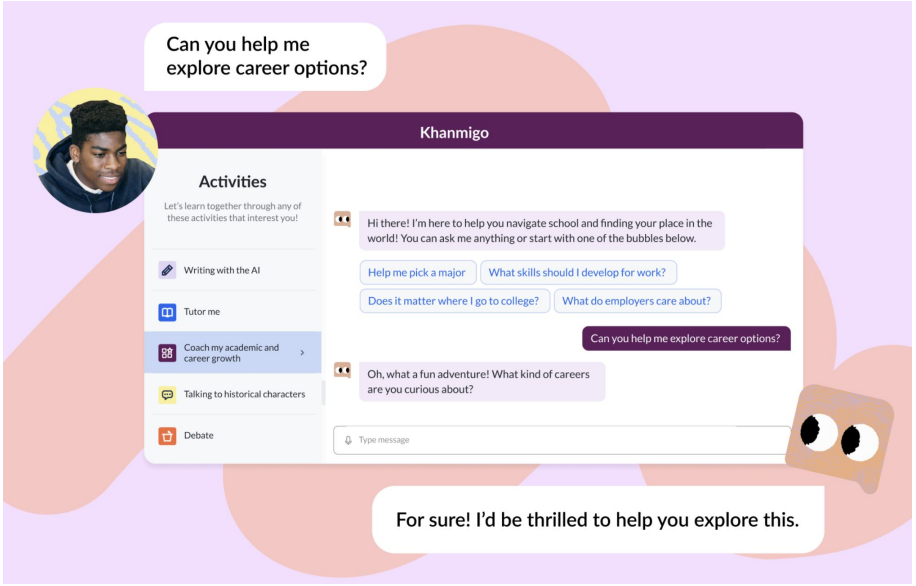
## Generative AI Can Harm Learning

*The Wharton School Research Paper*

59 Pages • Posted: 18 Jul 2024

Hamsa Bastani
University of Pennsylvania - The Wharton School

**"Our results suggest that students use generative AI as a "crutch" and simply copy answers instead of engaging substantively - a concern anecdotally raised with Khanmigo's generative AI tutoring bots."**

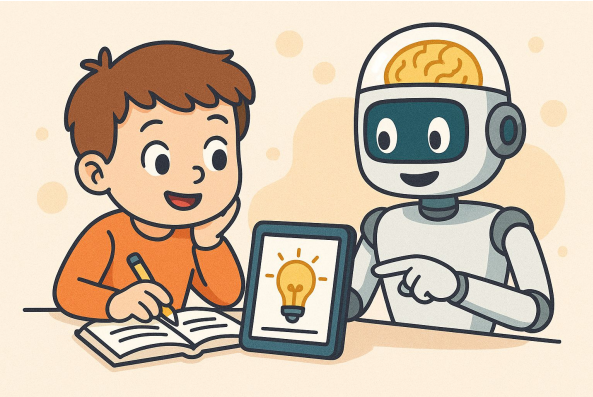



Source: Generative AI can Harm Learning

The Agency Fund

# Measuring the cognitive construct of agentic learning in GPT-based tutors

***using the self-regulated learning theory***



### Task Definition

e.g., Does the student understands the task (e.g., seeing how multiplication is a relevant capability)?

### Goal Setting & Planning

e.g., Does the student set clear goals (e.g., solving 10 problems in one session)?

### Enacting Tactics

e.g., Does the student use effective strategies and tracks progress (e.g., using notes, visual aids or chatbot quizzes)

### Adaptations to Metacognition

e.g., Does the student reflect on their performance to improve (e.g., recognizing challenges and seeking targeted practice)

Source: Supporting students' self-regulated learning in online learning using artificial intelligence applications

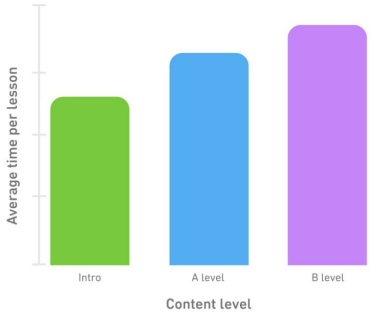
# Duolingo: proximal metric for Learning

The Agency Fund

*"Our Efficacy Lab conducts rigorous research to measure the efficacy…but these studies aren't helpful in conducting A/B tests on a daily basis.* ***It's extremely important that we have a proxy metric for learning****"*

## Total Sessions?

NO; biased towards easy lessons




## Time Spent Learning?

But skewed to studious, competitive learners

*"a) we were motivating a small group of learners who b) were already doing more than enough!"*

⇒ new goal: optimize for a higher percentage of learners spending at least 15 minutes/day

## Time Spent *Learning Well*

Because not all learning time is equal

$$\text{TSLW} = \text{Minutes Learning}_{Path} + \frac{\text{Minutes Learning}_{Other}}{2}$$

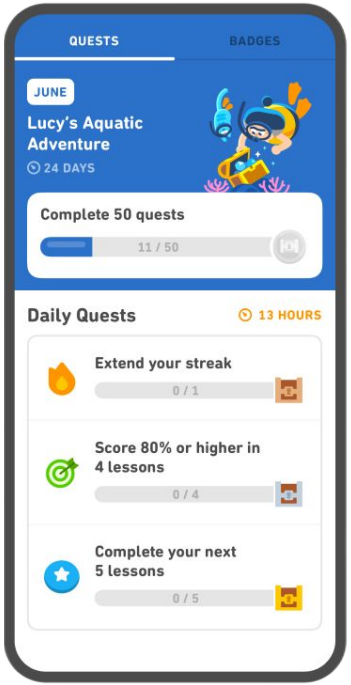
Source: Behind the metric: how we developed "Time Spent Learning Well"; see also User Funnels for the Social Sector: Lessons from Duolingo

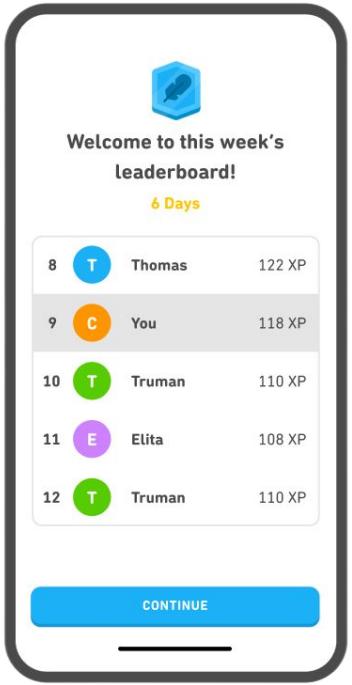
# Evaluating features that affect learning engagement on Duolingo

The Agency Fund

**Through A/B tests, Duolingo found that the following features that positively impact TSLW:**

- **Daily Quests:** Designed to build momentum – starting easy and getting harder – these quests encourage meaningful progress like completing a unit or reading the next story.
- **Monthly Challenges:** Now quest-based instead of XP-based to promote consistent learning, not last-minute XP grinding.
- **Leaderboards:** Still popular, but now better aligned with real learning. XP on the learning path has been increased to reward effort and progress more fairly.



The Agency Fund

# Appendix

# Rocket Learning: Adapting AIM-ECD

Anchor Items for the Measurement of Early Childhood Development framework (World Bank)

⇒ **phone-based** caregiver survey with **minimal bias**

⇒ **cheaper**

⇒ more **scalable** (from hundreds quarterly, to many thousands on short demand)

**Can it be adapted to WhatsApp? 🤔**

Work by: RL's Bilal Afroz, TAF's James Walsh and researchers Aarti Malik and Ronak Jain
Read more: https://theagencyfund.substack.com/p/ab-testing-for-the-social-sector

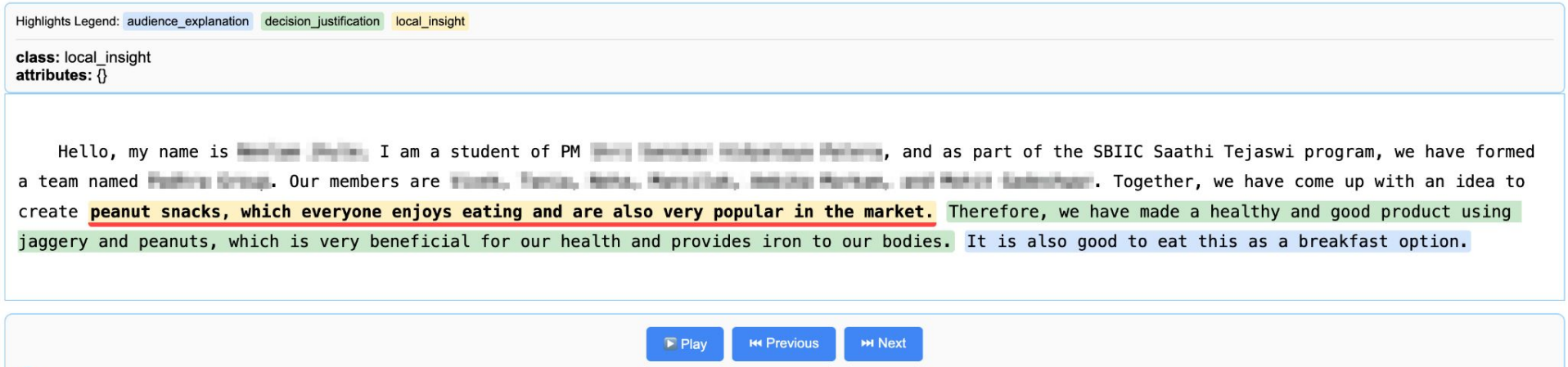
Udhyam
Learning Foundation

# Udhyam: Measuring agency

Goal: **auditable, scalable way** to measure *whether* and *how* students demonstrate agency-related skills and capabilities

Input: students' pitch transcripts, Udhyam's agency rubrics

Tools: `langextract` with `gpt-4o-mini` with examples in prompt
extract observable traces ⇒ higher-order constructs ⇒ aggregate into signals



# Example measures

The Agency Fund

## On-platform measures (for in-app studies):

- **AI-session-specific trace data, e.g.,**
  - Frequency and depth of queries (do users ask deeper, more advanced questions over time?)
  - Changes in vocabulary and complexity of responses
  - Follow-up questions (indicating curiosity and engagement)
  - Session duration (longer sessions may indicate sustained engagement)
  - Repeated interactions (do users return to learn more?)
  - Chatbot suggestions followed (does the user report acting on AI recommendations?)
  - Confidence improvement (fewer basic or repetitive questions over time)
  - Accuracy in user queries (does the user refine their questions over time?)
  - Reduced reliance on fallback responses (fewer "I don't understand" cases)
  - User sentiment and engagement trends (positive reinforcement = successful learning)
- **Short self-report survey measures**
- **Metrics from text analysis using NLP methods (e.g., sentiment analysis, topic modeling, BERT, etc.)**

## Off-platform measures (for field studies):

- **Short and long self-report, other-report, or observer-report measures of cognitive, emotional, and behavioral constructs**
- **Self-report, other-report, or observer-report measures of behaviors**
- **Qualitative synthesis**
- **…**


Source: Towards investigating the validity of measurement of self-regulated learning based on trace data