

AI Evaluation: From Model to Mission

A clear framework to turn AI pilots into
real development outcomes.

CONTEXT

AI Capability Is **Accelerating**

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

**WHY IT MATTERS**

Act now with structured evaluation.

REALITY CHECK - MIT, JULY 2025

Most AI Projects **Fail** to Deliver Impact

The issue isn't model quality.

It's integration and alignment with work.

Many pilots stall after the demo.

- Real users and messy data break assumptions.



WHY IT MATTERS

Build evaluation into delivery.

PAIN POINTS

Common AI **Challenges** in Development

- Measuring real impact beyond demos.
- Integrating into existing workflows.
- Staff capacity and change readiness.
- Data quality and governance.
- Choosing benchmarks that reflect context.

**WHY IT MATTERS**

Success requires solving these, not just better models.

THE FRAMEWORK

AI Evaluation: From Model to Mission

A structured path to move beyond technical metrics and measure what actually matters.



WHY IT MATTERS

Links tech capability to mission results.

01

MODEL

Are outputs accurate and safe?



02

PRODUCT

Is it usable and adopted?



03

USER

Does it change behavior?



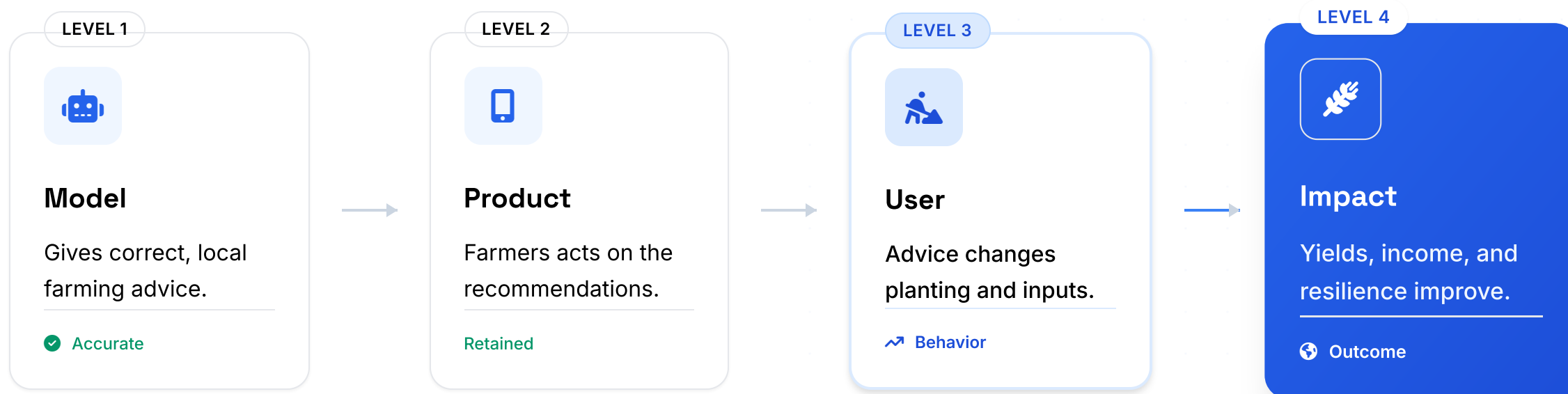
04

IMPACT

Do outcomes improve?



What This Looks Like in Practice



WHY IT MATTERS

Moves from technical demo to real development gains.”

STRATEGY

Why Evaluation Must Be **Phased**

- Perfecting the model first slows learning.
- Optimizing engagement alone misses impact.
- Jumping to RCTs too early wastes resources.
- **Evaluation should evolve across all four levels.**

**WHY IT MATTERS**

Faster progress, fewer dead ends.

PRACTICAL APPLICATION

Minimum Viable **Evaluation**

- Start small at each of the 4 levels.
- Log key interactions and decisions.
- Track basic usage, drop-off, and returns.
- Review real conversations with users.
- Define one clear user outcome (behavior change).



PRACTICAL STEP

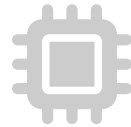
Don't overbuild. Start measuring today.

Different Levels, Different Expertise

LEVEL 1: MODEL →

AI Engineers

Focus: Accuracy, Latency, Safety



LEVEL 2: PRODUCT →

Product Managers

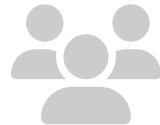
Focus: Adoption, Retention, UX



LEVEL 3: USER →

Behavioral Researchers

Focus: Decision Making, Trust, Psychology



LEVEL 4: IMPACT →

Economists & Policy Experts

Focus: ROI, Livelihoods, Scale



WHY IT MATTERS

Align the right people early to avoid bottlenecks.

LESSONS

What We Learned from the Accelerator

- Skipping early evaluation creates costly rework.
- Over-investing in models delays field learning.
- Structured, phased evaluation speeds progress and trust.

**WHY IT MATTERS**

Do less, learn faster.

NEXT STEPS

From Model to Mission:

Turn the evaluation framework into immediate action.

01

Pick one high-value use case

Don't boil the ocean. Choose one pilot to test.

02

Set Minimum Viable Evaluation

Define one metric for each of the 4 levels.

03

Convene a cross-functional team

Engineers, product, and impact experts together.

04

Share learnings

Transparency accelerates sector-wide progress.

