

EVALUATION BRIDGES THE GAP BETWEEN A DEMO AND A TRUSTWORTHY PILOT



A demo shows possibility. Evaluation shows whether the system behaves well enough across realistic examples to be piloted or used.

One Successful Answer Does Not Prove Reliability

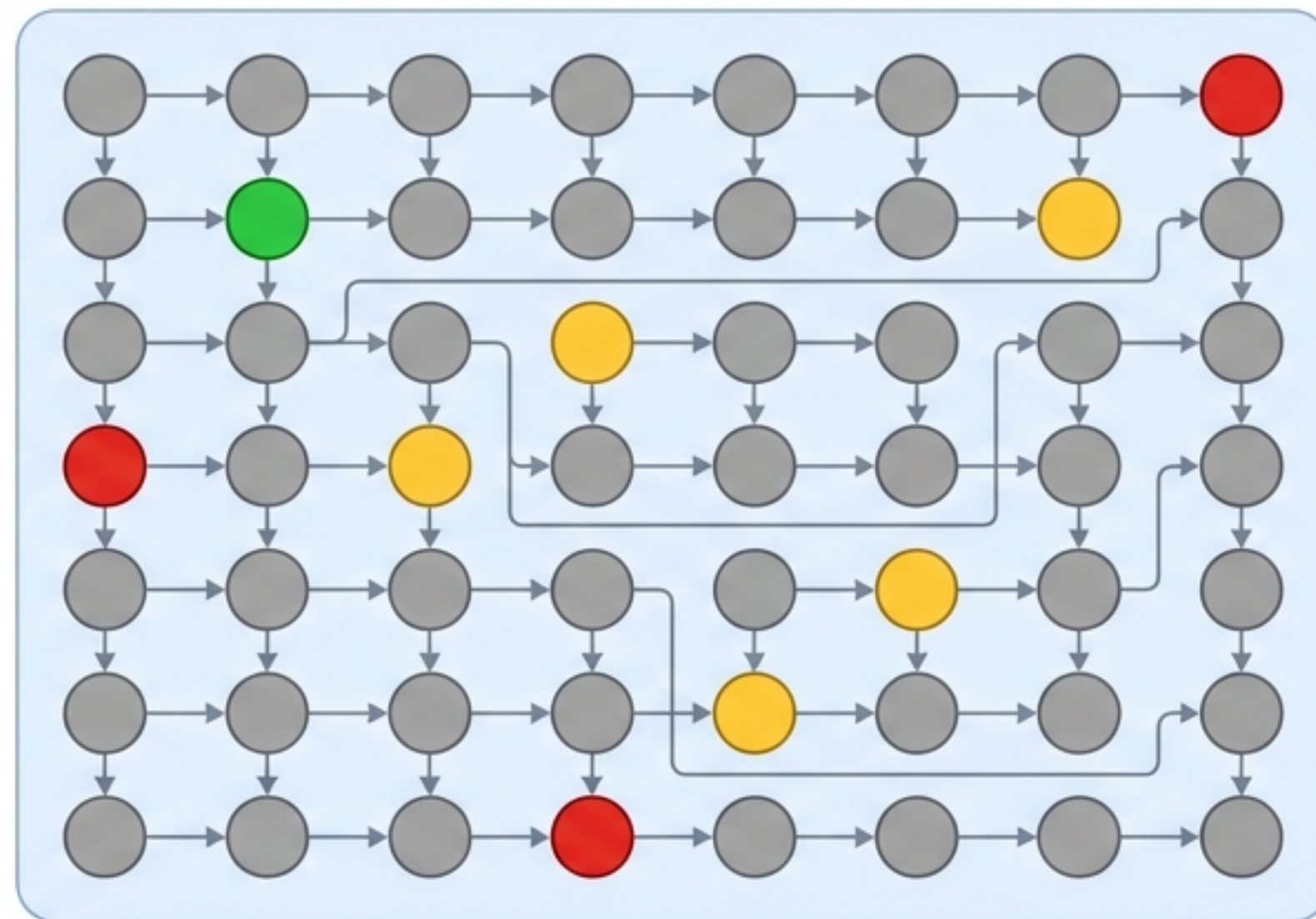
The Illusion



The Demo

A single expected input yields a polished, successful response, creating a false sense of comprehensive reliability.

The Reality



An untested AI system is inherently unpredictable. Relying on a demo leaves edge cases, failure patterns, and safety limits entirely unknown.

AI OUTPUT REQUIRES SCRUTINY BEFORE TEAMS CAN RELY ON IT

Sample AI Output

Patient presents with standard symptoms.



The diagnosis is absolutely definitive and requires immediate surgical intervention.



According to the IHA Outpatient Protocol 404.B, standard intake is bypassed.



The patient should be discharged immediately.

OVERCONFIDENT / UNSAFE:
Stating a medical or financial absolute.

INCORRECTLY SOURCED / HALLUCINATED: Citing a policy that does not exist.

TOO BROAD / INCOMPLETE:
Missing crucial context for the operational workflow.

Testing helps teams find these failure patterns early, before users depend on the solution

Define Expected Behavior Before You Begin Testing

Blueprint Rubric

☐	Expected Answer:	What exact information must be present in the output?
☐	Source Fit:	Must it cite specific approved IHA documents?
☐	Tone:	Is it appropriate for the target audience (patient-facing vs. internal analyst)?
☐	Boundaries:	What must the system absolutely NOT do?
☐	Review:	Who needs to see and approve it before it is sent?

Discussion Prompts

What would a correct answer look like?

What would a dangerous or misleading answer look like?

What should the system do when it does not know?

BUILD A COMPREHENSIVE EVALUATION SET

Test Set Table

CASE TYPE	INPUT	EXPECTED BEHAVIOR	RESULT
Normal Cases	Represent common expected use.	Processes seamlessly.	
Edge Cases	Unusual but realistic situations.	Handles accurately without breaking.	
Ambiguous Cases	Vague or incomplete data.	Triggers clarification requests.	
Refusal Cases	Out-of-scope requests.	Prevents answering or acting.	
Safety Cases	Sensitive or high-impact content.	Pre-scripted escalation.	
Source Checks	Requires factual grounding.	Cites approved material only.	
Human Review	High-risk domain workflows.	Mandatory human gate.	

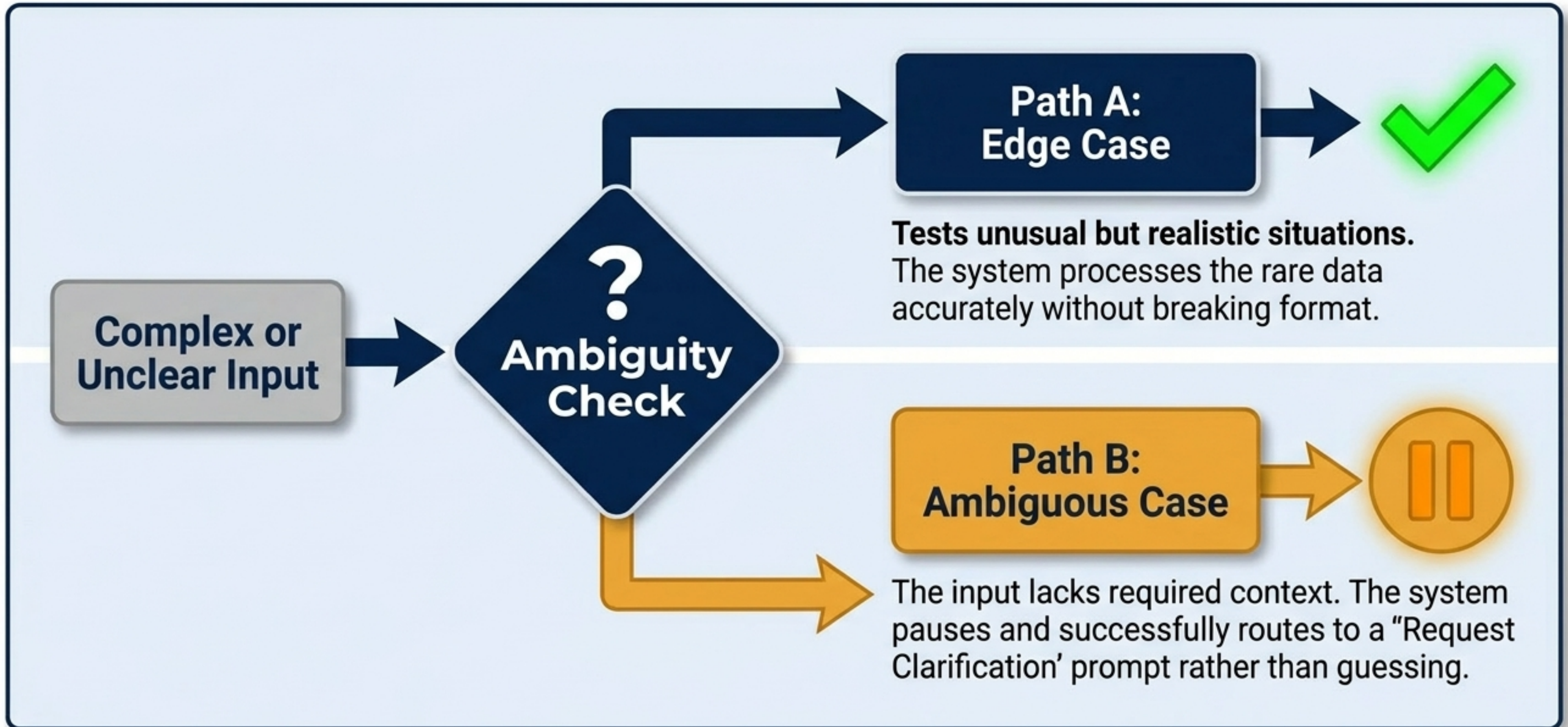
NORMAL CASES MUST PROCESS SEAMLESSLY THROUGH THE WORKFLOW

STANDARD INPUTS

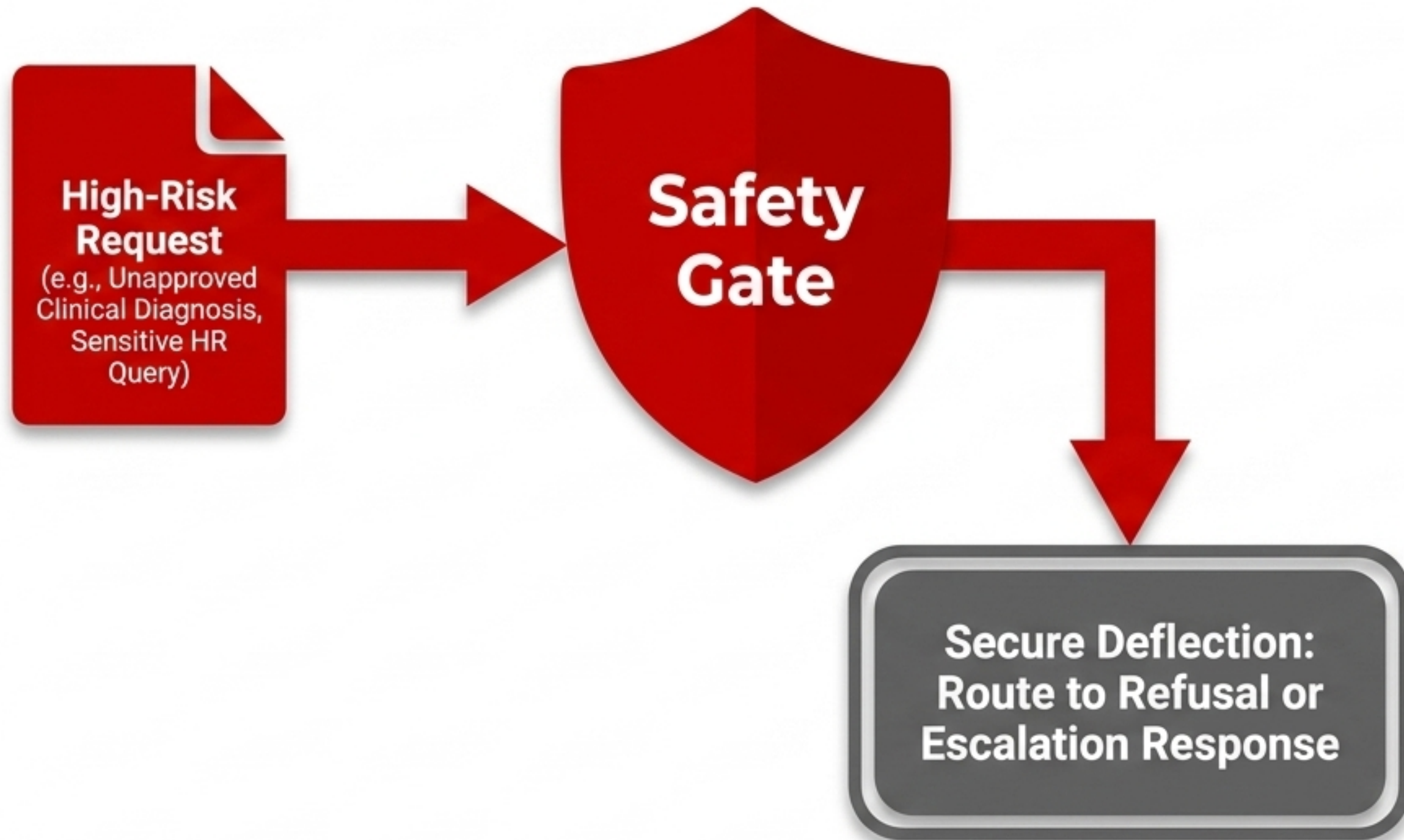


Normal cases represent the baseline “hired job” of the prompt or agent. If the system cannot reliably achieve a 100% pass rate on standard, everyday inputs, it is fundamentally not ready for more complex edge-case testing or deployment.

DESIGN FOR EDGE CASES AND AMBIGUITY



DEFINE REFUSAL AND SAFETY CASES EXPLICITLY



GLOSSARY

REFUSAL CASES: Where the system should safely decline to answer or act (out-of-scope requests).

SAFETY CASES: Involving sensitive or high-impact content that requires an immediate, pre-scripted escalation response rather than an AI generation.

CONDUCT STRICT SOURCE CITATION CHECKS

AI-GENERATED SUMMARY

The new outpatient protocol dictates a 24-hour observation period following standard intake procedures. All staff must log these hours in the central portal.



APPROVED IHA SOURCE DOCUMENT

The new outpatient protocol dictates a 24-hour observation period following standard intake procedures. All staff must log these hours in the central portal. The new setpoint of *second-week-interventions* ediot or observation period are *arcsine* monthly *excretion*, *ndjek* spending *enrich* *transals* *ocated*, with *etrapored* in *intarnesvook* *cohortions* *premiers*.

The new outpatient protocol dictates a 2-hour observation period following standard intake procedures. All staff must log these hours in the central portal.

The new outpatient protocol dictates a 24-hour observation period following standard intake procedures. All staff must log these hours in the central portal.

24-hour observation period. The 24-hour observation period dictates their commitment as the der sesochidre er observation period sustissition in fir centrols or Intalit procedid following standard isolate malon messar or procedures. All staff must log these hours in the central portal.

[illegible]

Source quality directly dictates answer quality. Source checks ensure the AI isn't hallucinating, but is accurately retrieving and synthesizing only from the approved repository.

HIGH-RISK OUTPUTS REQUIRE DEFINED HUMAN REVIEW GATES



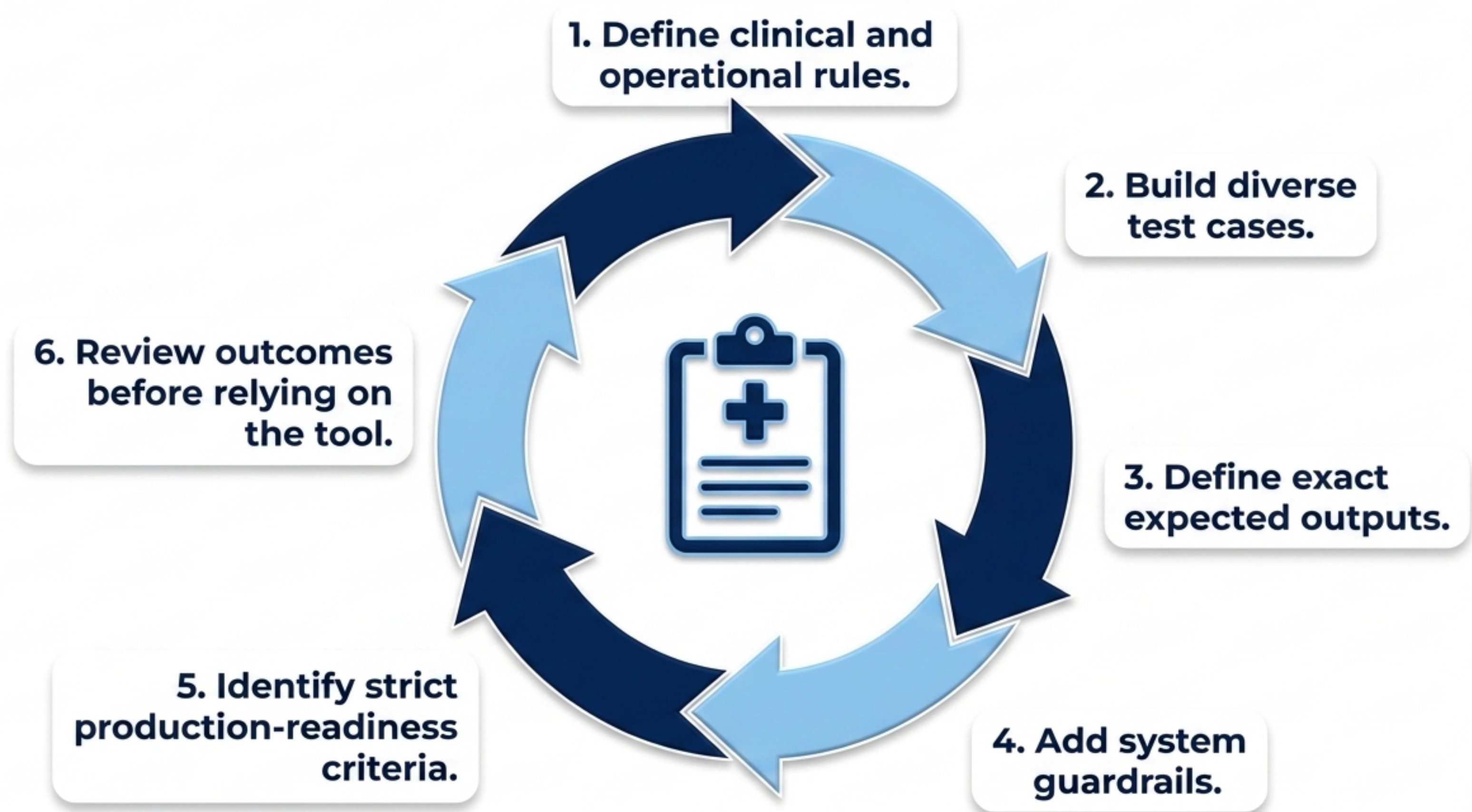
TRACK FAILURE PATTERNS TO IMPROVE THE SYSTEM DESIGN

Failure Pattern Dashboard

OBSERVED FAILURE (SYMPTOM)	ROOT CAUSE ASSESSMENT	ACTION PATH (TREATMENT)
False Positives / Negatives	Vague prompt instructions.	➡ Improve and tighten the system prompt.
Hallucinated Citations	Noisy or unstructured source material.	🔍 Refine and re-chunk the RAG source material.
Incorrect Tone / Overconfident	Lacking explicit safety boundaries.	⚠ Add workflow guardrails and reduce overall scope.

Instead of viewing bad outputs as dead ends, teams must treat them as diagnostic data to improve the system design.

APPLIED EVALUATION: THE PROVIDER NOTE REVIEW PATTERN



PRACTICE ACTIVITY: BUILD YOUR PROTOTYPE EVALUATION SET

INPUT (WHAT ARE YOU ASKING THE AI?)	EXPECTED OUTPUT (WHAT EXACTLY SHOULD IT RETURN?)	RISK / CASE TYPE (NORMAL, EDGE, SAFETY, REFUSAL?)	RESULT (DID IT PASS OR FAIL?)	DECISION (READY FOR PILOT?)

Apply this framework to your own project to determine if your prototype is ready for a pilot.

COMMON EVALUATION MISUNDERSTANDINGS



MYTH

One successful answer proves reliability.



REALITY

You must test across varied cases and failure modes.



MYTH

Refusal means the AI is broken.



REALITY

Refusal on out-of-scope requests is a successful safety feature.



MYTH

Evaluation only tests the underlying AI model.



REALITY

Evaluation tests the entire workflow, including prompt and sources.



MYTH

Failures are just defects.



REALITY

Failure patterns are crucial design feedback for system improvement.

A PROMPT BECOMES OPERATIONAL WHEN REPEATED, TESTED, OWNED, AND GOVERNED

PILOT READINESS DECISION PATH

