The AI Observability Loop

Deployment is the starting line, not the finish line. Observability is what keeps running after launch.

WHAT TO MONITOR IN PRODUCTION

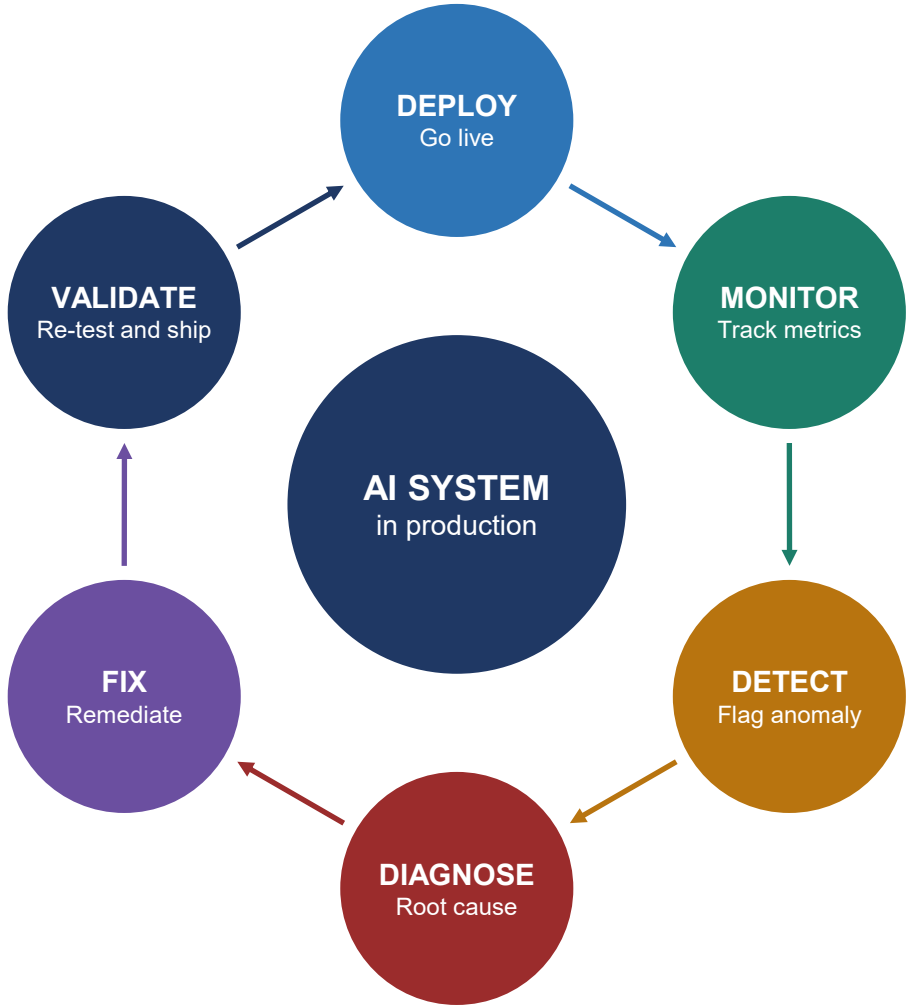
Input drift
Are user queries shifting away from what was tested?

Response latency
Degradation under load signals system stress

Refusal rate
Sudden increases signal a prompt or knowledge base failure

Escalation rate
Users giving up on the AI: a real-time quality signal

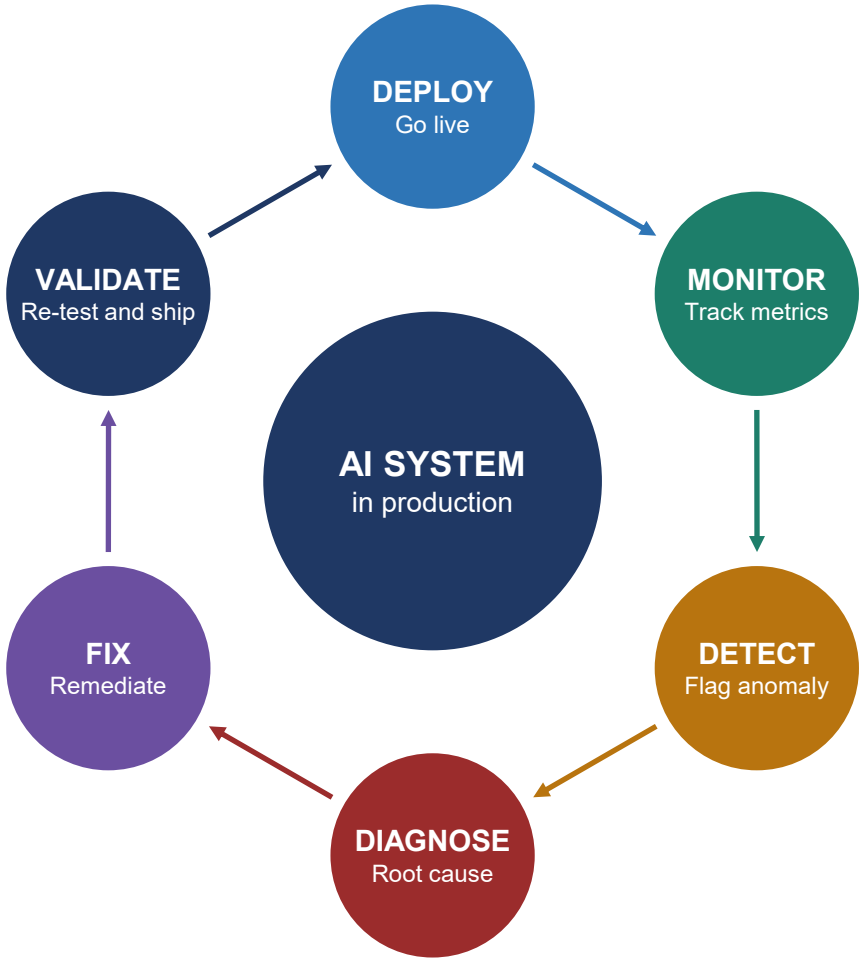
Cost per interaction
Token spikes point to prompt failure or retrieval waste



Source: [AI in Production: Knowing When Your AI Stops Working]

Set up your observability loop

For each stage, decide what triggers it, who owns it, and which tool or process you use.



Stage	What triggers it	Owner	Tool or process
Deploy			
Monitor			
Detect			
Diagnose			
Fix			
Validate			

Source: [AI in Production: Knowing When Your AI Stops Working]

About this diagram

The AI Observability Loop

When this fits

- Any team with an AI system live in production, or about to launch one.
- Setting up ownership after launch, so that detecting a problem leads to someone fixing it.
- Explaining to leadership why go-live is the start of a measurement commitment.

When it doesn't fit

- Systems still in prototype. Start with a golden dataset first (see the Eval Maturity Ladder).
- It is not a monitoring tool. It shows the process; you still need logging and dashboards to run it.

How to use this file

- Every element is a native PowerPoint shape: edit the text, recolour it, or delete what you don't need.
- Use the blank template slide (the one headed Map your or Plan your) to fill in your own version.
- Copy the diagram straight into your own deck. It is sized for a standard widescreen slide.

Source: AI in Production: Knowing When Your AI Stops Working