CHAPTER 2: MAPPING THE DEPTHS | TOOLKIT

Canary Test and Kill Switch Builder

Know when your AI starts drifting, and know exactly what to do about it.

Tools don't fail during business hours. They fail at 4am on a holiday weekend, mid-crisis. Any AI you bring into a workflow needs two guardrails: a canary test that tells you when it starts to drift, and a kill switch that says exactly what happens when it does.

AI tool

Task it supports

PART 1 Build your canary set

Pick a small batch of samples you already know the answers to. Aim for 10 to 20, with a mix of clean examples and confirmed hits. Run it every time the vendor pushes an update.

#	Sample ID / description	Known answer	Clean or hit?
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			

PART 2 Run log

Date	Version / update	Correct	Missed hits	False alarms	Accuracy %	Pass?

PART 3 Your kill switch policy

Trigger: the result that flips the switch (example: accuracy drops below 90%, or any confirmed hit is missed)

What happens next: the fallback process the workflow reverts to

Who gets notified, and who escalates to the vendor

What happens to outputs already produced since the last passing run

Conditions to turn it back on

Policy owner

Approved by

Date