

Shadow Run

FIELD SHEET

Run the candidate beside the live workflow, using only what was known at request time. Compare outputs against evidence. The human answer is a reference.

No customer messages or account changes. Use only permitted request-time context.

Workflow: _____ Owner: _____ Case / run: _____

Version + sample window: _____ Prewritten acceptance bar / link: _____

1 Score each case

		PASS	FAIL	N/A
Answer quality	Correct, complete and usable for this case?	☐	☐	☐
Source grounding	Material claims supported by current, relevant evidence?	☐	☐	☐
Escalation judgment	Stops or escalates when this case requires it?	☐	☐	☐
Tone / format	Clear, respectful and code-safe?	☐	☐	☐
Permissions	Proposed actions and data access stay within permissions?	☐	☐	☐
Handoff quality	Enough facts, risks and next steps for a human?	☐	☐	☐

Evidence / reason for failure or N/A: _____

2 Five stop conditions

- 1 Invents material policy, pricing, security or product claims.
- 2 Repeats failed advice or misses a required escalation.
- 3 Attempts actions or data access outside its permissions.
- 4 Cannot support a material claim with valid evidence, even if the claim is correct.
- 5 Cannot be audited, stopped or corrected by an owner.

Test-set coverage: easy, hard, missing docs, frustrated customers, action / permission boundaries and security-sensitive cases. Categories can overlap.

Critical failure: hold the affected scope until fixed and retested. A shared defect may block the whole system.

4 Start on Monday

- 1 Pick one workflow and name its owner.
- 2 Write limits, thresholds and stop conditions.
- 3 Mix routine and risky cases; keep fresh cases aside.
- 4 Block customer messages and account changes.
- 5 Review pairs; score and tag failures by case type.
- 6 Fix and retest; test routing, shutoff and fallback.
- 7 Launch only the scope that passes. Keep reviewing.

3 Give the failure an owner

Knowledge / docs	Docs or knowledge owner
Product / promise	Product; GTM for promise gaps
Routing	Support operations / workflows
Tone / format	Guidance or content owner
Policy	Support + Legal / Security
Tool / action	Support engineering

Show the team what failed and what should have happened. Agree who will fix it, then rerun the test.

5 Report counts, not just percentages

Acceptable drafts: sendable drafts / reviewed drafts.

Missed escalations: missed required escalations / cases that required escalation.

Unsupported claims: drafts with an unsupported material claim / reviewed drafts.

Docs defects: cases with a confirmed source defect / reviewed cases.

Report x/n by case type. Retain N/A counts. A clean sample is not proof of zero risk or customer resolution.

Decision: ☐ Ready ☐ Hold ☐ Constrained scope

Scope / owner / date: _____