CST · PAM · FOR ROSS'S REVIEW BEFORE ANY BUILD STARTS

Pam → 95%: The One Plan

One merged plan from Anthony's Fable review and Ross's "Fix the Instrument First" research. Every load-bearing claim was independently re-verified before it went in.

Date: Thu 2026-07-02 **Supersedes:** pam-path-to-95-plan-2026-07-02.md + the 7/2 session review

Build window: Fri 7/3 – Tue 7/7 (Fable's last day) **Detailed MD:** attached on the Basecamp card

TL;DR

- The dashboard says 84%; most of that gap is the **measuring stick, not Pam**. The grader never sees the email subject line, some brands are mislabeled, and Pam's backtest lookups have been near zero since 6/14 with nothing alarming on it.
- Ross's research was right on the big call and it survived a hard challenge. One correction: **there is no broken lookup harness to restore**. Every live probe passes; the fix is to make lookups deterministic, not to repair infra.
- Real fabrication by Pam exists (one proven case), so a **live-safety tripwire ships this week**, not in a later phase.
- Nothing gets approved from the fix queue until a lift gate exists: **5 of 5 previously approved fixes made Pam worse**.
- Three decisions (D1–D3) are needed from Anthony + Ross by Friday; then Phase 0 builds over the weekend while Fable is still available.

Decisions needed first (by EOD Fri 7/3, on the Basecamp card)

#	DECISION	RECOMMENDATION
D1	What "95%" means.	Per-topic, real-miss-free pass rate: (match + minor) / testable tickets, at least 30 graded per topic. Not a blended headline;

		blends hide weak topics. Fix the trend chart's "90" label to match.
D2	Approve Phase 0 scope for the Fable window (items 0.1–0.6).	Yes. All six are build items; each is accepted by an on-disk artifact, not a promise.
D3	Freeze the 👎 approval queue until the lift gate exists.	Yes. Five of five previously approved fixes had negative lift (receipt R4).

Alternatives considered, and the cost of delay: (a) keep approving rules to chase the nightly score: proven net-negative; (b) treat it as an infra outage and "restore the harness": every live probe passes, so the window gets spent debugging instead of measuring; (c) this plan. Deciding nothing keeps the gate untrustworthy, blocks the go-live ramp, and forfeits the Fable window that expires Tue 7/7.

Where the numbers stand (dashboard, 7/2 4:20 AM)

84%
Last night (37/44) and 7-day (n=477)

75%
30-day (n=1,255), contaminated in both directions

95%
Goal (per D1: per topic, not blended)

0
Live lookups on all graded tickets since 6/29

TOPIC (7-DAY PASS)	WISMO	RETURN ISSUE	ORDER ISSUE	Q&A	PAYMENT
Pass rate	89%	87%	79%	82%	100%

The blended 84% appears only to size the gap. All gating happens per topic per D1.

What we verified (challenged, not accepted)

CLAIM	VERDICT	WHAT THE EVIDENCE SHOWED

The grader never sees the email subject; Pam does.	CONFIRMED	4 of last night's 7 "fabricated order number" fails had the number sitting in the Gorgias subject line Pam legitimately read. The grader is handed the body only. (R1)
Brand mislabeling fails correct answers.	CONFIRMED	ITAM tickets stamped "ihr" get graded against iHeartRaves policy; the 7/2 exchange "fail" was a correct ITAM answer. (R2)
"Lookup harness dead since 6/24, restore it."	PARTLY WRONG	Lookups collapsed from **6/14** (not 6/24) and are zero since 6/29. But the connectors, the backtest's exact invocation, and the model all pass live probes today. Nothing identified to restore; cause unknown; the fix is deterministic pre-fetch plus an alarm. (R3)
All five approved fixes had negative lift.	CONFIRMED	Lift log: −15.0pp, −0.6pp, −17.5pp, −1.7pp, −7pp. The approval gate counts support, not lift. (R4)
The grader flips verdicts on identical input.	DIRECTION ONLY	Two independent probes agree it flips on borderline tickets; the κ=0.52 magnitude is from an older grader model and gets re-measured in 1.1. (R5)
Pam's prompt copies have drifted (new finding).	CONFIRMED	The live prompt is missing the NEVER-FABRICATE and tools-down rules; the backtest prompt is missing the constitution with the "use your live tools" mandate. (R6)
"0 of last night's 7 fails are real."	MOSTLY	Most were instrument artifacts, but real fabrication exists: one proven invented-store-credit case, and one team note asserting tracking movement the customer never pasted. Fabrication stays guarded. (R7)

Bottom line: the instrument owes Pam most of the visible gap. Her true level is plausibly near the bar but unproven in both directions: only fails were audited so far, and every grade is an LLM judging an LLM. Fix the instrument, protect live customers now, then prove the level with checks the instrument cannot game.

The plan

PHASE 0 · BUILD FRI 7/3 – SAT 7/4, VERIFY THROUGH TUE 7/7

Fix the instrument + protect live customers

Every item is accepted by a **persisted artifact on disk**. A step without its artifact is not done, no matter what the chat says.

0.1 Deterministic order-fact pre-fetch. Wire the already-built `fetch_order_facts()` into the backtest replay: Shopify → ShipStation → Loop facts injected on every ticket with an identifier. Removes "did the model feel like calling a tool" from the measurement entirely.

First mover: Fable · Artifact: nightly `prefetch-coverage.json` · Accept: ≥95% coverage on a full night + 20-ticket held-out spot-check

0.2 Grader sees what Pam saw. Same context (email + subject + body) to the grader; brand stamped from Gorgias tags, mismatches flagged.

First mover: Fable · Ross signs off a 10-ticket smoke test; real acceptance is 1.4's full re-adjudication

0.3 Enforce "not testable" + flatline alarm. Tool-less tickets with an identifier leave the denominator; dashboard shows "not testable: N"; Slack alert on exclusions >30% or a zero-lookup night. An outage must read as "we measured nothing", never as a clean score. Also fixes the 90→95 label.

First mover: Fable · Artifact: per-night excluded count + one fired test alert

0.4 Pin the 4 AM cause. Instrument one nightly run (MCP startup log + per-ticket tool-error capture). 0.1 makes the measurement immune either way; this names the cause instead of guessing.

First mover: Fable · Artifact: `tool-runtime-log.jsonl`, cause named or ruled out

0.5 One canonical prompt. Single source feeding backtest and live. Backtest regains the tools mandate; live regains NEVER-FABRICATE + tools-down routing, extended to cover the Team note line.

First mover: Fable · Artifact: nightly prompt-sync diff check, alarms on drift

0.6 Live safety now. Live Pam talks to reps today without the fabrication rule, and one real fabrication is proven. Ship the prompt fix (0.5) plus a deterministic tripwire in the guardrail: any order#/tracking#/refund-$ that came from neither the customer nor a tool this turn forces a regeneration, logged.

First mover: Fable · Artifact: tripwire firing on a seeded test in `guardrail_misses.jsonl`, zero silent releases

PHASE 1 · STARTS SUN 7/5, RUNS INTO THE WEEK OF 7/8

Make the score decision-grade

1.1 Majority-vote grading. 3 grader votes per ticket + rubric disambiguation on the two known flip patterns; re-measure grader consistency before/after (target $\kappa \geq 0.8$). Cost stays an order of magnitude under the Opus-grader experiment that was reverted for cost.

First mover: Anthony (primary grader) · Ross mirrors on the audit grader + owns the κ report

1.2 Rubric re-weight toward customer harm. Fabricated specifics = hard fail even when phrased politely; honest tools-down deferrals never penalized. One rubric change at a time, after 1.1.

1.3 Lift gate on approvals; queue stays frozen (D3). No rule ships without a passing full-topic re-grade on both graders attached to the approve action. Default lever is remove/simplify. The pending email-first proposal is not approved; 0.1 makes it moot.

1.4 Re-adjudicate the contaminated window (6/14–7/2) with subject + brand fixed, including the 29 "fabrications" from 6/26. Runs on both graders, Ross hand-checks a 10-ticket sample, and old vs corrected numbers publish side by side, never silently rewritten.

PHASE 2 · WEEK OF 7/8 ONWARD, NO FABLE DEPENDENCY

Prove it, then ramp

2.1 Weekly pass-side audit (~15 random passes, adversarial re-grade, reported next to the pass rate). False passes are proven possible; this keeps the corrected score honest in both directions. *Owner: Ross.*

2.2 Human calibration set. Ross hand-grades ~50 tickets; both LLM graders are scored against them; agreement ≥90% or the grader gets fixed, not the number. Ross pulls the sample Sun 7/5; grading can follow. *Owner: Ross.*

2.3 Re-baseline, then gate per topic (D1). A topic goes live at ≥95% real-miss-free over ≥30 graded. **Standing rollback:** a live topic reverts if the weekly pass-audit finds >2% real misses in it, or it drops below the bar 2 nights running; reversals are announced where the flip was.

2.4 Live parity. The same deterministic pre-fetch goes into live Pam once 0.1 proves out. Then the manager-expansion track (cost to serve, Yuma 60%, Loop replacement) proceeds on earned trust.

Sequencing (single first mover per step)

WHEN	ITEM	FIRST MOVER
Fri 7/3	D1–D3 decided on the card	Anthony + Ross
Fri 7/3 – Sat 7/4	0.1, 0.2, 0.3 build	Fable
Sat 7/4	0.5, 0.6 build	Fable
Sat–Sun night	0.4 instrumented run	auto
Sun 7/5 – Tue 7/7	1.3 + 1.4 build and re-adjudication	Fable, Ross sign-off
Sun 7/5	Pull the 50-ticket calibration sample (2.2 prep)	Ross
Week of 7/8	1.1, 1.2	Anthony + Ross
Week of 7/8+	2.1, 2.2 grading, then the 2.3 gate	Ross

Post-7/7 fallback: every Phase 0/1 item is ordinary Python/config once written; nothing depends on Fable to run. If a build item slips past 7/7, Opus/Sonnet sessions finish it from the spec. Only the deep re-adjudication (1.4) meaningfully benefits from Fable, which is why it sits inside the window. Two build days are padded across five calendar days; if 0.4's cause turns out deep, 0.4 slips first.

What NOT to do

- Don't quote the 30-day 75% (or any nightly number since 6/14) to leadership; contaminated in both directions.
- Don't approve anything from the 👎 queue before the lift gate exists; 5 of 5 prior approvals regressed.
- Don't declare "Pam is already at 95%" from fail-side audits alone; that claim waits for 2.1 + 2.2.
- Don't spend the week debugging the 4 AM mystery instead of shipping 0.1; the pre-fetch makes the measurement immune to the cause.
- Don't rewrite historical numbers silently; side-by-side only.

Glossary (for anyone reading cold)

Backtest	Each night Pam re-answers real, already-resolved tickets; a grader compares her draft to the approved playbook.
Match / minor / material	Right and complete / right outcome with a small omission / wrong outcome (the only real fail).
Testable	Pam had working lookups, or needed none. Tickets where lookups were impossible test the harness, not Pam.
Fabrication	Stating a specific (order#, tracking, refund $) that came from neither the customer nor a lookup this turn.
Lift	The change in a topic's pass rate after a rule ships, measured by re-grading the same sample.
κ (kappa)	Grader self-consistency on identical input; 1.0 = deterministic, ~0.5 = coin-flip on borderline calls.

Receipts (condensed; full detail in the MD on the card)

- **R1:** The backtest hands Pam "email + subject + body" but hands the grader body-only. Ticket 583397363's subject is literally "Replacement Item Requested order number 50041336233", the number the grader called fabricated. Same on 583681627, 583199039, 583287037.

- **R2:** Record 583199039 stamped "ihr" while the customer message quotes INTO THE AM Support; the 7/2 exchange fail applied iHR store-credit-only policy to an ITAM order.
- **R3:** Tool use by night: 242/275 (6/12) and 242/266 (6/13), then between 5-of-27 and 0-of-44 from 6/14 on, exactly zero since 6/29. Live probes 7/2: connectors return real data; the backtest's exact invocation returns a true order total; the tool counter counts it; the production model runs an email-first lookup chain unprompted. The "not testable" detector only fires if Pam says tools failed, which her prompt forbids.
- **R4:** Lift log: −0.150, −0.006, −0.175, −0.017 (+ exchange −0.07). Approval gate = support count; 96 proposals applied, 1 pending.
- **R5:** κ=0.52 over 8 identical reruns (older grader model); ~20% flip rate independently observed. Re-measure is step 1.1.
- **R6:** "NEVER FABRICATE" appears in the constitution and backtest prompt but zero times in the live prompt; the backtest prompt lacks the constitution's tools mandate.
- **R7:** One proven invented-store-credit fabrication (ticket 582519873, graded "minor" by the nightly grader); last night's WISMO/lost team note asserted tracking movement the customer never pasted.

Level 1 — Static Doc · Generated by Claude (Fable 5) for Anthony + Ross · 2026-07-02 · Full plan MD attached on Basecamp card "Fable use cases"