PAM TRAINING · CST HUB

What Got Built Overnight

Built Jun 12-13, 2026 · For Anthony + Ross · Plain-language walkthrough of the deep-grading run and the three new tools on the CST hub

---

TL;DR

- We graded Pam on **10 of her highest-volume topics** (about 500 real tickets) in one night instead of the ~98 nights the old slow method would have taken.
- Pam is **not ready to take over any of these topics yet**: every one scores 67-78% "ok" and the readiness bar is 90%.
- The big new insight: her misses are **systematic, not random**. The same six patterns repeat on every topic, which means they are fixable.
- Those patterns are now **32 ready-to-approve fixes** you can tick and approve in one click, and we can measure whether each fix actually improves her.
- Nothing changes Pam automatically. Every fix waits for a CS lead to approve it.

The headline finding

Across 10 of Pam's biggest topics, graded against how each ticket was actually resolved (a human agent or Yuma), **every topic lands 67-78% "ok," and none is close to the 90% readiness gate.**

So the honest answer to "can Pam take over a topic?" is **not yet, on any of them.** But the misses are not scattered noise: the **same six patterns repeat across every topic**, so this is a fixable problem, not a rebuild.

Grading against **Yuma is consistently a tougher bar than grading against humans**, which lines up with the view that Yuma already outperforms the human floor.

Where each topic stands

"Ok" means Pam matched the resolution or was only slightly off. A real miss is when she resolved it worse. The gate to go live is **90% ok over at least 30 graded tickets.**

Topic	Ok rate	Graded
RETURN ISSUE / return item	**77%**	53
ORDER ISSUE / wrong item	**77%**	52
PAYMENT / discount code	**78%**	50
ORDER ISSUE / cancel	**72%**	54
ORDER ISSUE / out of stock	**72%**	36
ORDER ISSUE / missing item	**72%**	51
RETURN ISSUE / exchange item	**70%**	53
WISMO / invalid address	**70%**	50
WISMO / mdnr (where's my order)	**68%**	59
WISMO / lost package	**67%**	60

All ten sit 12 to 23 points under the 90% gate. Consistent, which is itself the signal: the same gaps drag every topic down.

The six recurring failure patterns

This is the heart of it. Every place Pam diverged from the team got clustered into named patterns. The same six show up again and again.

HIGH RISK

1. Fabricating tracking numbers or delivery status

Pam invents a tracking number or claims a package "shipped" or "delivered" with no actual order lookup to back it up. This is the most dangerous one because the customer is told something false.

MEDIUM

2. Missing the standard timeline or canned promise

She leaves out (or gets wrong) the expected timeline: refund in 3-5 vs 2-4 business days, when the store-credit email arrives, the follow-up window. The customer is left without the answer they came for.

MEDIUM

3. Re-asking a question the customer already answered

She asks the damage/tampering gate question, or for a discount code, that the customer already provided earlier in the thread. Reads as not paying attention.

MEDIUM

4. Defaulting to a refund instead of the offer ladder

The playbook says offer a reship or store credit first; Pam jumps to a refund the customer never asked for, which costs more and skips the recovery path.

HIGH RISK

5. Skipping a required step or escalation

Missing the mandatory address confirmation, the apology on an address hold, the Final Sale acknowledgment, or a manager escalation on a trigger (a 35-day-old order, a mystery-item case). She resolves things she should have routed up.

TONE

6. Brand-voice and sign-off gaps

Missing the agent name, the brand signature block, or the warm emoji-forward tone the playbook calls for. Low risk, but it is the difference between on-brand and robotic.

What got built

Three tools, all on the CST hub. Here is what each does, why, what it achieves, and what to watch for.

A. "Why Pam misses" analysis

WHAT On the Backtest tab, every place Pam diverged is clustered into named failure modes per topic, each with how many tickets it hit, a severity (red / amber / gray), and a one-line fix.

WHY The prior run distilled 52 possible rules but only 1 was strong enough to queue. We could not tell if Pam's misses were a few fixable holes or many scattered ones. This answers that.

ACHIEVES The answer is "a few fixable holes." It turns the strong, repeated patterns into **32 ready-to-approve fixes**.

WATCH FOR The ticket counts are estimated groupings of the real per-ticket reasons. The auto-naming normally runs on each deep grading run; the cloud account was throttled overnight, so the clustering was done by hand this time. The numbers and fixes are sound; treat the counts as "roughly this many."

B. Faster approval + proof it worked

WHAT On the Today tab you can now tick many proposed fixes and approve them in one click. Separately, a re-grade tool re-answers the SAME graded tickets after fixes are approved and reports the before/after "ok" lift.

WHY Approving 32 fixes one at a time is friction. And approving a rule without measuring whether it helps is flying blind.

ACHIEVES One-click bulk approval (it reuses the existing, proven approval path, so no new moving parts), plus a way to **prove a fix actually moved the needle** before trusting it.

WATCH FOR The lift number cannot appear until you approve some fixes and a re-grade runs. It is a tool for after approval, so that panel stays hidden until then. That is by design, not a bug.

C. Deep-grade the next tier of topics

WHAT	A one-command mode that auto-picks the 5 highest-value topics still under-measured and grades 60 fresh tickets in each.
WHY	The first deep run covered 5 topics. This widened it to the next 5: cancel, out of stock, exchange item, lost package, missing item.
ACHIEVES	All 10 top topics now have a solid readiness read in **one night** instead of the ~98 nights the old slow trickle would have needed.
WATCH FOR	It confirmed the same "not ready" pattern, which was expected. It runs on the cloud account, which was throttled overnight; that slowed it down but did not break it.

Why this matters

We now know exactly **where** Pam fails and **why**, the fixes are queued and bulk-approvable, and we can measure whether each fix actually improves her. That turns "is Pam ready?" from a guess into a loop you can run:

approve fixes → re-grade → watch the lift → re-deep-grade → a topic clears the 90% gate → that topic becomes Pam's first live takeover

First live takeover stays a mid-impact topic, never the riskiest, and never the lowest. We expand one topic at a time as each clears the gate.

Things to know

Nothing changes Pam automatically. All 32 fixes are proposals a CS lead approves. Until approved, Pam behaves exactly as she does today.

There is still no "first live" pick. Nothing has cleared 90% yet. The fixes are the path to getting there.

Where to look: the Backtest tab's top table shows the most recent run; the per-topic readiness lives on the **Pam Takeover** tab.

Under the hood: a safety lock was added so overnight grading runs cannot corrupt each other. None of this touched the Work Hub itself (it is mid-edit); it publishes to the same files the Work Hub's CST tab reads, so it shows up there automatically.

Level 1 · Static Doc · Generated by Claude · Jun 13, 2026