PMs set the launch bar for AI evals; product leaders shape decisions instead of owning them

PM Daily Digest

2026-11-24

PMs set the launch bar for AI evals; product leaders shape decisions instead of owning them

By PM Daily Digest • November 24, 2026

Aakash Gupta lays out how PMs should run AI evals, Emily Andrews argues against three habits of product leadership, and OpenAI explains why it bought a product harness instead of only training a better model.

A PM's working model for AI evals

Aakash Gupta's guide, built on examples from a guest named Mahesh, makes the eval suite a PM artifact. It has four steps:

- **Define "good" as yes/no checks.** Write specific tests for helpful, honest and harmless. For a contract agent, "helpful" means it struck out only what it needed to [1]. Judgment calls should be binary because "you fix it or you don't." Anything code can check should be a code-based pass/fail test [1].
- **Set the bar.** The floor is your competition and the ceiling is about 70–80%, which is where top AI benchmarks tend to top out. On the contract agent, plain Cowork scored 45% (the floor), the bar was 65%, and the best model reached 70% [1]. If you come in under the bar, guardrail the product so it only answers the kinds of questions it gets right 80% of the time [1].
- **Scale with LLM judges.** Human experts label first; the example cost was $25,000 for legal experts. Then tune an LLM judge to agree with them 70–80% of the time, because aiming for overall agreement "might actually be a trap" [1].
- **Keep humans grading samples.** For a large contract agent like Legora or Harvey, that could mean 200–1,000 outputs a week [1].

On ownership, the PM writes the yes/no questions, sets the launch bar and stage

gates, checks the judge, and decides what ships. Engineers build the judges and the pipeline, and push eval scores up and costs down [1].

A r/ProductMgmt thread raises a gap this framework doesn't measure after launch. If an agent finishes a task but the user spends 15 minutes checking or fixing the output, is that a success? The poster argues adoption and engagement metrics miss whether the job actually got easier [2].

"Decision architect," not decision owner

At a Mind the Product talk, Emily Andrews challenged three beliefs product leaders hold dearly, which she calls "sacred cows." Her opening story: her team spent hours building a RACI and "left with clearer letters, but we did not leave with a better decision." A RACI can't create shared context, judgment or trust [3].

1. **Product should own prioritization.** Applied literally, this turns ownership into territory [3]. Instead, product should bring customer evidence, make trade-offs explicit, and agree success measures with business partners. "The goal is not to win the decision"; it is to prevent an uninformed one [3].
2. **Product should be the voice of the customer.** No single function has the whole answer. Product's value is connecting customer truth, business reality (including what each leader is accountable for) and what the tech makes possible into a set of options [3].
3. **Clear decision rights solve friction.** When trust is low, organizations add process: more governance, approvals and slides [3]. Trust comes from steady competence, delivery and transparency [3].

In exchange for giving up sole ownership of the roadmap, product gets involved earlier and shares accountability. She stresses this is not a lower bar on craft [3].

More from Atlassian's Sherif Mansour

Last brief covered Atlassian's failed "everyone's a builder" experiment. Lenny's takeaways from the episode add several points:

- **Junior hiring.** Most companies have stopped hiring juniors because AI can do their work; Atlassian is doing the opposite. Juniors help a company unlearn old ways of working, and seniors "can pick AI slop a mile away," so the two complement each other [4].
- **PM archetypes.** Atlassian uses three in hiring and calibration: scientist (hypothesis-driven, fits growth teams), general manager (thinks in TAM, revenue and profit) and artist (starts from a UX vision). Mansour says artist PMs are gaining value as features converge [4].
- **Disagreeing with a founder.** First work out what led the founder there. If you still disagree, ask: "Are you the decision-maker or am I?" [4]

Hiten Shah makes a related point: as AI lets people express judgment through software at scale, the software will only be as good as the judgment behind it [5].

OpenAI buys a product harness

On a16z, OpenAI explained why it acquired Orchestra Bio. GPT Rosalind did well on benchmarks, but scientists didn't know what prompts to use or what questions to ask. That takes "an exceptionally good product or good harness around the model" [6]. Two product lessons came out of the conversation:

- **Build for whole workflows, not single tasks.** That means capturing tacit reasoning such as "why did we select that molecule?" [6]
- **Make evidence traceable.** Compile the supporting evidence for each conclusion so a human can review it quickly, rather than amplifying errors [6].

OpenAI also told founders to build plugins and MCP connectors, because work is moving into ChatGPT and Codex as the "front door" [6].

Smaller signals

- **Activation rate across channels.** Casey Winters warns that one channel can show a better activation rate simply because it filters out low-intent users before they're measured. His advice: improve each funnel on its own and rank channels by total activations, not rate [7].
- **Pulling in customer feedback.** PMs describe using MCPs where they exist and Claude browser control where they don't [8]. Another says the hard part is knowledge-graphing feedback so it can be queried cheaply; they pay for Enterpret [9]. A third says to tag 90 days of one marketplace's reviews by product and recurring complaint before wiring anything up [10].
- **A/B testing.** One PM says most of their tests didn't produce a clear winner and is surveying how teams pick what to test next [11].

Sources

1. A FAANG PM's Guide to Practical AI Evals
2. r/ProductMgmt post by u/Far_Breadfruit_1137
3. The three sacred cows of product leadership that are holding you back (Emily Andrews)
4. X post by @lennysan
5. X post by @hnshah
6. Why OpenAI Acquired a Biotech Startup
7. X post by @onecaseman
8. r/ProductManagement comment by u/5hredder
9. r/ProductManagement comment by u/mik5062

10. r/ProductManagement comment by u/lacartrad
11. r/ProductMgmt post by u/django-8