Measure What Agents Own, Not How Often They're Used

PM Daily Digest

2026-09-04

Measure What Agents Own, Not How Often They're Used

By PM Daily Digest • September 4, 2026

This brief focuses on the period's clearest PM shift: evaluating AI products by evidence, workflow ownership, and reduced operator burden rather than interaction volume alone.

Big Ideas

Make AI evals a discovery habit, not a late QA step. Teresa Torres defines evals as methods for measuring whether an AI product or workflow performs as intended; they help teams maintain quality, catch issues before users do, and create a feedback loop alongside interviews and assumption testing. For PMs, the practical shift is to specify expected behavior and test it while the product hypothesis is still changing. [1]

For agents, measure ownership rather than engagement. Hiten Shah argues that a more useful agent may generate fewer messages and sessions because the customer is spending less time supervising it. A successful first run proves capability, not ownership: run a real job for at least three cycles, log every intervention, and separate valuable human judgment from accidental handbacks such as rebuilding context, restarting work, or checking whether the task finished. The target is declining accidental operator work while consequential decisions and exceptions remain visible. [2]

Tactical Playbook

Score evidence per assumption before scaling. Strategyzer's readiness framework separates **desirability** (do customers need and want it?), **feasibility** (can it be built and delivered?), and **viability** (can it create business value

profitably). A board-approved business case is still a hypothesis; score the evidence for each aspect rather than claiming the whole idea is validated. [3]

Use the evidence ladder to choose the next test: business plans and high-level research are level 0; statements and reactions are levels 1–2; low-stakes actions such as signups or sales-call requests are level 3; pilots, letters of intent, deposits, and pre-orders are level 4; real market behavior or a Wizard-of-Oz test reaches level 5. More interviews do not strengthen evidence if prospects never commit. [3] In one example, Fireflies.ai charged $100 per month while founders manually joined meetings and sent summaries: payment and continued use tested desirability and viability, but the non-scalable delivery left feasibility unproven. [3]

Case Studies & Lessons

Murmur productized the bottleneck around coding agents. Macroscope says an internal tool orchestrated 90% of the code it shipped over two months, leading it to release Murmur as an early preview. Its diagnosis was not primarily model weakness: engineers were bottlenecked by local development environments and the post-PR lifecycle of review comments, failed CI, rebases, and merge conflicts. [4]

Murmur gives each agent a dedicated cloud VM, lets it test and verify its work, and keeps it moving through GitHub events until human approval. A local Claude Code or Codex session can act as a director for a fleet, while Slack, Linear, GitHub, REST, and MCP make existing work systems entry points for bounded agent tasks; engineers stay focused on work requiring deeper context or judgment. [4] The product lesson is to own the workflow around the model—not just expose model capability—and make access profiles, deployment choices, and audit logs part of the product for enterprise use. [4]

Career Corner

AI fluency is becoming a hiring signal for PMs. Aakash Gupta reports that 76% of 113 PM job postings he reviewed asked for AI knowledge, and lists evals, RAG, agents, MCP, observability, context engineering, and LLM-as-judge among the emerging vocabulary. [5] A community discussion raises the unresolved consequence for entry-level roles: AI can now accelerate documentation, research, analysis, SQL, prototyping, and workflow creation, potentially shifting the bar toward independently identifying problems and showing judgment. [6] Build evidence of both: a small AI-enabled product workflow with explicit evals and instrumentation, plus a clear explanation of trade-offs and failure modes.

Tools & Resources

Explore local AI-assisted product analytics. A community-built open-source MCP server lets an agent run data-science analysis on tabular event data

to examine retention, churn differences, underused features, segment behavior, drop-off points, and pre-conversion actions; computation stays local and the model receives analysis results rather than the full CSV. Treat it as a hypothesis generator, validating event definitions and conclusions against raw data and customer conversations. [7]

Sources

1. X post by @ttorres
2. X article by @hnshah
3. How close is your new business project to success
4. X article by @kayvz
5. substack
6. r/prodmgmt post by u/RecentAttitude1618
7. r/ProductManagement post by u/No_Incident_6009