

AI products hit two limits: the reliability layer and model costs

PM Daily Digest

2026-10-01

AI products hit two limits: the reliability layer and model costs

By PM Daily Digest • October 1, 2026

Several sources argue that AI product work now depends on who supplies reliability, how much the model layer costs, and whether evals measure real work. Also covered: agent loops for PMs and how mature products can compete with AI-native demos.

Who supplies the reliability?

Kendra Vant, CPO at Tapi, cites Princeton research finding that models have become more accurate but not more consistent: ask the same task five times and the answers differ [1]. When she uses Claude Cowork, she corrects it and re-prompts. In her words, the user is “providing the reliability layer” [1]. Her own customers, busy property managers, “just want my software” to work the first time [1].

So she says “can the model do it?” is the wrong question. The right one is where the reliability layer sits and whether you can afford to build it [1]. Her checklist: - What are you quietly fixing when you use similar AI yourself? - Will your users do that work for you? - Who owns the layer: engineering, product, or nobody (in which case it’s still a demo) [1]? - What does it cost in trust when the layer fails, and does it fail gracefully [1]? - What does the layer cost to run at scale? She has seen evals, model checkers, and human reviewers that worked at demo scale cost more than the product takes to run [1].

Model dependence is vendor risk

On r/ProductManagement, one post pointed to reporting on Anthropic’s IPO prospectus and asked what PMs will do if the models behind their “wrapper” features become unaffordable [2]. The top advice was to treat this as vendor

risk. Put the model behind your own interface so changing providers is a config change, and don't let one vendor's pricing become a single point of failure [3]. One B2B team uses "bring your own API key": the customer owns the provider relationship and pays the token bill, which also makes data privacy easier [4]. Another commenter warned that switching isn't free, because any change of model or configuration changes agent behavior and quality at scale [5].

An a16z episode adds data on cost. Databricks' router, which picks a model for each task, solved more problems at 35% lower cost than the strongest single model. Elise AI fine-tuned a smaller model that was 60% cheaper and had low enough latency for live audio [6]. The speakers argue the unit that matters is the cost of getting the customer's job done, not the cost of using the top model for every step [6]. Their adoption figures: 69% of S&P 500 companies have live AI deployments, 30% report quantifiable impact, and only 2% track a metric over time [6].

Evals that look like real work

Hiten Shah traces how Harvey went from wrapper to legal product. It built legal process into the product, then created BigLaw Bench, which grades real legal tasks. That let the team test model, workflow, and retrieval changes against the same tasks, and turn failures lawyers spotted into tests [7]. On Harvey's Legal Agent Benchmark of longer tasks, frontier models completed under 10% end to end under strict scoring. Shah says that kind of result shows whether the bottleneck is context, workflow, tools, or the model [7].

Aakash Gupta makes the same argument for personal workflows: "A prompt is text. A skill is behavior." He runs 25 skills against 75 tests. He added the tests after an exec-update skill filed a Stripe bug that was corrupting a metric under "wins" [8]. Separately, Shah open-sourced 16 competitive-intelligence skills, covering positioning, pricing, battlecards, win/loss, and more [9].

Agent loops for PMs

Gupta lists six recurring loops: feedback digest, deal intelligence, competitor brief, metric anomaly flag, onboarding friction, and a customer call list [10]. He adds that one agent should rank the backlog while a second checks the ranking against strategy [10]. Roadmap bets and reading stakeholders stay with the PM [10]. His rules: put every correction into the skill file, not the chat, and retire any loop whose output you've stopped editing [10].

Also worth noting

- **Competing with AI-native demos.** A mature B2B team keeps losing the first impression to demos built on one clean prompt [11]. One suggestion: run the same appealing prompt, then have the buyer change the

source data after the output was approved, so your controls show up as part of real work [12].

- **Outcome-based pricing.** HP’s Faisal Masud describes pricing that is either seat-based with outcome commitments or purely outcome-based, tied to reductions in support tickets. His reason: CIOs no longer want to pay for dashboards [13]. He also says AI compresses work from idea to PRD to demo, but getting to production is “a whole different ballgame” [13].
- **Cutting interruptions from QA.** One PM added a “what happens if” section to every spec (empty states, no connection, wrong input, permissions), and QA questions mostly stopped after two sprints [14].
- **Decision-making as the core craft.** Lenny Rachitsky highlighted Robby Stein’s line that PM craft is “above all else... decision-making” [15], and Shreyas Doshi endorsed it [16].

Sources

1. Four questions to ask before building AI into your product—Kendra Vant (Chief Product Officer, Tapi)
2. r/ProductManagement post by u/Common_North_5267
3. r/ProductManagement comment by u/trickytoddler
4. r/ProductManagement comment by u/Afton11
5. r/ProductManagement comment by u/michinya
6. AI, Infrastructure, and the Next Investment Cycle
7. X article by @hnshah
8. substack
9. X post by @hnshah
10. substack
11. r/ProductMarketing post by u/OwlZealousideal4779
12. r/ProductMarketing comment by u/Visible_Speed8843
13. HP President on Replacing Middle Management with AI Agents | Faisal Masud | E314
14. r/ProductManagement comment by u/CauliflowerAsleep700
15. X post by @lennysan
16. X post by @shreyas