AI Evaluation Loops, Outcome-Led PM, and Growth Beyond the Funnel

PM Daily Digest

2026-03-19

AI Evaluation Loops, Outcome-Led PM, and Growth Beyond the Funnel

By PM Daily Digest • March 19, 2026

This issue centers on four shifts in PM practice: evaluation-driven AI product management, outcome-led operating models, growth strategies built on access and community, and the rising premium on human judgment. It also includes concrete playbooks for metric interpretation, stakeholder influence, service triage, and career differentiation in an AI-heavy market.

Big Ideas

1) PM work in AI is shifting from specs to evaluation systems

At Shopify, the move to AI products like Sidekick changed the PM job from hardening every requirement in a spec to defining 'what good looks like' and encoding that judgment in an internal LLM-based evaluator called Ace. PMs review fake or sanitized real conversations, score them across dimensions such as accuracy, grounding, merchant sentiment, goal fulfillment, execution quality, errors, language quality, and safety, and then approve aligned examples into a ground-truth set used to create a judge. The same pattern is applied to generated workflows in Flow, not just chat experiences. [1]

Why it matters: Non-deterministic products cannot be fully managed through traditional specs alone; quality depends on whether the team can repeatedly evaluate open-ended outputs. [1]

How to apply: Treat the eval stack as part of the roadmap. Define dimensions of quality early, create annotated examples, require agreement across raters on subjective calls, and reuse the approach anywhere your product generates content, queries, or workflows. [1]

2) Outcomes are a better organizing unit than outputs

> “When we measure success by what you ship, you get a feature factory. When you measure success by the impact of what you ship, you get a product team that learns, adapts, and creates real value.” [2]

Teresa Torres draws a clean distinction: an output is the thing you build, while an outcome is the customer or business change it creates. Teams given outputs have little latitude to explore; teams given outcomes have more freedom and more responsibility to find the best path. [2]

Kate Tarling’s service-organization guidance pushes the same logic up a level: list services from the outside in, set measures for what good looks like, map current work against those services, and reject the assumption that the status quo is a neutral or risk-free baseline. [3]

Why it matters: This is not just a metrics tweak. It changes portfolio decisions, intake discipline, and how teams defend capacity. [2, 3]

How to apply: Rewrite incoming requests in problem-and-outcome language, choose metrics tied to value creation rather than shipment, and expect to iterate on the first version of the metric. If the work sits inside a broader service, map it to the service outcome before you prioritize it. [2, 3]

3) Growth is moving away from funnel tuning and toward access, trust, and community

Elena Verna argues that growth used to be an optimization problem focused on acquisition, activation, and monetization, but is now shifting toward strategic product giveaways, outcome-aligned monetization, and deliberate community building. She also argues that feature differentiation is weakening, making trust, brand, and human connection more central to adoption. [4]

Her distribution view is similar: paid marketing is harder, organic app-store style discovery is crowded, and the clearest current channel is organic, creator-led social. At Lovable, that shows up as employee-led social, building in public, and feeding social feedback directly back into the product. [4]

Why it matters: PMs can no longer assume that better funnel copy or pricing-page tweaks will do most of the work. Growth is increasingly tied to lowering barriers, delighting users, and creating visible feedback loops. [4]

How to apply: Test barrier-removal moves, instrument how users discover and talk about the product, and align monetization with customer outcomes rather than short-term extraction. Also protect room for innovation: Verna warns that private companies can damage long-term competitiveness when quarterly pressure crowds out experimentation. [4]

4) Agentic products and AI-native teams still need visible human judgment

Scott Belsky's product strategy point is simple: enterprise products that help humans get credit for work done by supervised agents may adopt better than products that simply do the work instead of humans, because credit affects ego, adoption, and accountability. [5]

That complements a warning from Shreyas Doshi: founders and executives are seeing AI-drafted proposals collapse under pushback because the person presenting them has not thought them through deeply enough. AI-prepped ideas can fail the moment an unexpected question appears. [6, 7]

Why it matters: The failure mode is not just bad output. It is weak ownership—products that obscure human credit and teams that outsource reasoning. [5, 6]

How to apply: Design human review, attribution, and decision checkpoints into agent workflows. Internally, use AI for drafts, but require the owner to defend trade-offs, assumptions, and edge cases without the model in the room. [5, 6, 7]

Tactical Playbook

1) Build an AI quality loop before you scale the feature

A practical loop from Shopify looks like this: define quality dimensions, assemble fake or sanitized real examples, have multiple PMs annotate them, approve aligned examples into a ground-truth set, and use that set to create or refine a judge. The same structure can be reused across different generative surfaces, including workflow generation in Flow. [1]

1. Define what good looks like in dimensions, not just one score. [1]
2. Review examples that are close to real usage, including sanitized live conversations where possible. [1]
3. Have at least two PMs rate the same examples to surface subjectivity early. [1]
4. Approve only the aligned examples into ground truth. [1]
5. Use the approved set to create or update the judge. [1]

2) Interpret metrics with orthogonal context, not one impressive number

Julie Zhuo's example is a useful habit check: 5M MAU and 80% DAU/MAU looks strong on its own, but a 30-second average daily session length can make the same product look shallow—until you learn it is a payments app, where a short session may be exactly right. Her recommendation is to add orthogonal context: independent facts that reduce ambiguity from different directions. [8]

Use this sequence:

1. Start with the headline metric. [8]
2. Add a behavioral counterweight, such as session length or task-completion pattern. [8]
3. Add use-case or business-model context that changes what 'healthy' should mean. [8]
4. Delay conclusions until you have multiple independent facts pointing in the same direction. [8]

3) Run a deep-dive sprint when you lack enough context to help

Vanessa Lee describes an unorthodox but practical leadership move: if something feels off, run a sprint with the team and meet daily. In her example with Shopify's inventory team, the work included examining current bugs, data structures, UX, and next-day experiments until she had a much better understanding of the domain and the team had a clearer shared aim. [1]

Why it matters: Leaders often try to guide work from too far away. A short period of intense involvement can replace vague opinions with specific judgment. [1]

How to apply: Use a time-boxed daily cadence, stay close to the real artifacts, and push one concrete next step each day. The goal is not permanent micromanagement; it is to earn enough context to give better guidance later with less involvement. [1]

4) For hard stakeholder debates, bring proof and a lightweight artifact

The API versioning story at Shopify is a strong template for hard influence. The PM had spent 12 months working with teams that were inadvertently breaking the API, entered the meeting with four concrete examples, and brought a hand-drawn proposal for quarterly versioning with one year of backward compatibility. The drawing was deliberately simple so the CEO could engage with the idea rather than react to a polished deck. [1]

Why it matters: Senior stakeholders are easier to move when you combine superior domain knowledge with a proposal that still leaves room for their input. [1]

How to apply: Do the slow homework before the meeting, arrive with evidence instead of opinions, and use artifacts that clarify the decision without signaling that every implementation detail is already fixed. [1]

5) Triage new work by service and outcome before you discuss priority

In complex organizations, Tarling recommends mapping existing work to services, comparing each item to what good looks like, and sorting work into strong fit, needs investigation, or candidate to stop. For new requests, a triage function should require the requester to describe the problem and the outcome before the work gets any further. [3]

Why it matters: This moves prioritization away from who asked loudest and toward evidence, overlap, and strategic fit. [3]

How to apply: Start in one cross-functional service area rather than changing the whole organization at once, learn where the real barriers are, and then expand the operating model. [3]

Case Studies & Lessons

1) Shopify Sidekick and Flow: product judgment became infrastructure

Shopify's response to AI product quality was not just to improve prompts. The team built an internal evaluation workflow where PMs define the dimensions, score conversations, align across raters, and create ground-truth data that powers an internal judge. The same logic is used beyond Sidekick conversations and applied to generated flows in Flow. [1]

Lesson: For AI products, your evaluation system is part of the product. If it is weak, quality work stays anecdotal. [1]

2) API versioning at Shopify: courage works best when it is evidence-backed

The starting point was operational pain: teams were breaking an unversioned API. After a year of working directly with those teams, a more junior PM entered a meeting with the CEO knowing the domain in depth, carrying multiple proof points, and proposing quarterly versions with one year of backward compatibility. The CEO started from a 'not doing that' position, but the decision changed in a 45-minute meeting. [1]

Lesson: Courage alone is not enough. The persuasive sequence was domain mastery, concrete evidence, and a proposal simple enough to discuss collaboratively. [1]

3) Lovable Free Day: removing friction can be a growth event

Lovable made its platform free on March 8 to lower barriers to experimenting with AI and tech. Verna says the result was more than 100,000 new users and 120 events across 40 countries. Her broader framing is that growth is increasingly about giving access, building community, and creating delightful experiences that later connect to monetary outcomes. [4]

The operating model around that is equally notable: build in public, encourage employees to post organically, watch social feedback closely, and turn that feedback into rapid product changes—she describes fixes going live in 20 minutes. [4]

Lesson: Sometimes the best growth move is not another funnel optimization. It is a visible barrier-removal moment backed by fast feedback loops and a clear product story. [4]

Career Corner

1) If AI tools are widely available, your edge shifts to judgment, strategy, and domain depth

Sachin Rekhi's answer to the 'how do PMs differentiate?' question is a useful checklist: stay at the frontier of AI fluency, keep high standards and taste, deepen domain expertise, invest in product strategy, and keep building design skill—especially interaction design, where he says AI still lags world-class practitioners. [9]

How to apply: Pick one domain to know unusually well, use AI aggressively but discard low-quality output, and make sure you are still practicing strategy and design decisions rather than only prompting for them. [9]

2) Do not let AI do your thinking for you

Shreyas Doshi's warning is career-relevant: competent executives are repeatedly seeing people bring AI-assisted proposals they cannot defend once the conversation gets messy. The failure shows up when a client pushes back, an investor asks an unplanned question, or an unexpected condition breaks the logic. [6, 7]

How to apply: Before circulating a doc or proposal, pressure-test it yourself. List what would change your mind, what edge case breaks the recommendation, and what question you hope nobody asks. If you cannot answer those without the model, the work is not ready. [6, 7]

3) Career signal: proof and courage compound faster than credentials

> "PMs are the, I call it courage as a service to the team." [1]

The Shopify interview frames PMs as the people who make hard calls, surface broken team dynamics, and sometimes kill months of work when it is wrong. The same conversation also notes a preference for people who have been founders for at least a year, and describes Shopify as a place where strong storytelling and decision-making can create real autonomy at scale. Paul Graham makes the adjacent point from startups: users care whether they like the product, not what is on your resume; a resume is only a predictor of performance. [1, 10, 11]

How to apply: Build a track record of visible decisions, not just shipped tickets. Keep artifacts that show how you framed a problem, changed course, or defended a tough call; those are better career signals than generic claims of ownership. [1, 11]

Tools & Resources

- Shopify VP of Product on Building a $100B+ Agent-Led Commerce — strong on AI evaluation systems, leadership sprints, and stakeholder management under disagreement. [1]
- How to fix broken systems - Kate Tarling (CEO, The Service Group) — practical for service mapping, portfolio cleanup, and problem-and-outcome intake triage. [3]
- Teresa Torres on outputs vs outcomes — short refresher on choosing outcome metrics and iterating when the first version is wrong. [2]
- Lovable Free Day, 100K New Users, Here's Why It Worked — useful for PMs rethinking growth around access, community, and building in public. [4]

Sources

1. Shopify VP of Product on Building a $100B+ Agent-Led Commerce
2. X post by @ttorres
3. How to fix broken systems - Kate Tarling (CEO, The Service Group)
4. Lovable Free Day, 100K New Users, Here's Why It Worked
5. X post by @scottbelsky
6. X post by @shreyas
7. X post by @DrDominicNg
8. X post by @joulee
9. X post by @sachinrekhi
10. X post by @paulg
11. X post by @paulg