

AI skills need A/B evals, not install counts, and win/loss needs the buyer's own words

PM Daily Digest

2026-10-07

AI skills need A/B evals, not install counts, and win/loss needs the buyer's own words

By PM Daily Digest • October 7, 2026

Hiten Shah on testing whether AI skills actually change agent output, ThoughtSpot on evals as the new PRD, Gamma's rebuild against generic AI output, practitioner advice on win/loss, and a PM job market that keeps tightening.

A skill being available doesn't mean the agent uses it

Hiten Shah's team has built 16 skills for competitive product marketing, and he asks a basic question: does a skill change the work once the model has it? He notes that Skills.sh reached one million skills and nearly 280 million installs in seven months, so availability "tells us very little about quality" [1]. He cites a Vercel Next.js eval. The baseline agent passed 53%. Making a skill available left the score at 53%, partly because the agent skipped the skill in 56% of cases. Telling the agent explicitly when to use the skill raised it to 79%, and a compressed docs index in AGENTS.md reached 100%. He says one test can't settle how every skill should be designed [1].

His test is one PMs can reuse. Give the same model the same job and the same evidence, once with the skill and once without. Define what good looks like before the run, and keep the failures. If the model catches up, shrink the skill or remove it, and rerun the test whenever the model changes [1]. He also writes each skill around the job it serves. For example, a CRM loss reason records what someone typed into the CRM, not why the buyer left. If nobody asked the buyer, the buyer's reason is unknown [1]. In a demo, Claude without the skills treated an old source as recent and stated a guess as fact. With the skills, it dated every claim and noted what it couldn't see. The skills are MIT-licensed and work across Claude Code, Codex, Cursor and others [2].

ThoughtSpot: the eval is the new PRD

Francois Lopitaux (SVP Product, ThoughtSpot) told Mind the Product that “your eval system is almost becoming your PRD.” It defines what good looks like and what to avoid, because a prompt-based product has no fixed UI to spec [3]. Analytics answers have to be the same every time. So he uses LLMs only where they’re needed and grounds them in a semantic layer that defines terms like “new customer” and “revenue”. In his words, without that layer an LLM is “a new intern” that makes poor judgments [3]. A few other points:

- Teams are organized by product “track,” not by feature, so the people closest to customers decide what to build next [3].
- Pruning matters more now that code is cheap to generate [3].
- He expects smaller teams and a lower PM-to-developer ratio, as deciding what’s worth building becomes the bottleneck [3].
- For a recent APM hire, he asked candidates for a Git repo and a video explaining the problem their code solves [3].

Product moves

- **Gamma 5.** Gamma concluded that its output looked “too similar to all the other AI tools.” It started over and rebuilt for visual variety and brand fidelity across presentations, docs, social assets, and graphics [4].
- **AI simulations.** Lenny Rachitsky is following teams that use AI simulations to test product ideas and flow tweaks. He names Simile, Primitive Labs, Synthetic Users, Tenera, and Seldon, and asks how they’ve worked for people. No results are reported [5].

Craft: win/loss is not competitor research

A r/ProductMarketing thread argued that knowing a rival’s pricing, features, and positioning doesn’t tell you why a customer chose them [6]. Advice from the thread:

- Run unbiased win/loss interviews, not run by sales, and do 10–20 before looking for patterns [7].
- Ask buyers what they expected to go wrong six months after signing. This tends to surface implementation risk, internal politics, and trust issues [8].
- Commenters disagreed on price. One said it’s almost always the reason [9]. Another said C-suite buyers often pick more expensive competitors for ROI reasons [10].

Teresa Torres made a related point about records. Her transcripts kept showing that her confident memories were wrong. Notes and transcripts each add interpretation, and a record can be used to collaborate or as a weapon [11].

Job market

A 50-year-old senior PM with 22 years in tech has applied to more than 200 roles in nearly a year and gotten few interviews [12]. Another commenter missed one of six technical requirements and was passed over for a role that had been open eight months. They described a “buyer’s market” in which managers hold out for a perfect, cheaper PM [13]. A third said the market never recovered from its 2022 peak and hiring has shifted toward senior roles [14]. These are personal accounts, not market data.

Sources

1. X article by @hnshah
2. X post by @hnshah
3. How to escape the feature factory: Francois Lopitiaux (SVP Product Management, ThoughtSpot)
4. X post by @thisisgrantlee
5. X post by @lennysan
6. r/ProductMarketing post by u/Opposite_Butterfly89
7. r/ProductMarketing comment by u/AlternativeIll2946
8. r/ProductMarketing comment by u/Significant_Elk6305
9. r/ProductMarketing comment by u/emimimama
10. r/ProductMarketing comment by u/Athenawize
11. X post by @ttorres
12. r/ProductManagement post by u/Basic_Instinct_79
13. r/ProductManagement comment by u/Up-in-the-Ayre
14. r/ProductManagement comment by u/JimOfSomeTrades