Cold-Start Evals, the AI Smile Curve, and Google's Prototyping Interview

PM Daily Digest

2026-07-29

Cold-Start Evals, the AI Smile Curve, and Google's Prototyping Interview

By PM Daily Digest • July 29, 2026

A six-step method for building AI evals with no production data, Andrew Chen's framework for why AI retention rises over time, Cutler on surviving strategic incoherence, and Google's new live AI prototyping interview round.

Big Ideas

The AI smile curve: usage rises as models improve

Andrew Chen identifies a new "smile curve" for AI products where retention and usage increase over time as foundation models improve. The pattern: a user tries an AI app, finds it lacking; a new model release improves performance; eventually the user incorporates it into their workflow, driving up both usage and spend. What seemed "meh" at first eventually becomes indispensable. [1]

Historical smile curves came from network effects (social), supply growth (on-demand), or workplace spread (SaaS). The AI version is driven by model capability improvement rather than user-side dynamics. [1]

Why it matters: products that seem underwhelming at launch may still be "must fund" bets if their value depends on model trajectories. This reframes early retention metrics—low initial usage isn't necessarily a kill signal, but the product must be positioned to capture the uplift when models cross a capability threshold.

The Border Collie's Faustian Bargain

John Cutler maps how organizations react when leaders present something "objectively not a strategy" that everyone experienced in the room knows isn't one. He sorts attendees into archetypes: Believers (can't see the gap, tolerate it),

Players (see it, tolerate it, even relish the game), Purists (can't tolerate it but miss the social function), and Coherence Checkers who see the gap and struggle. [2]

The "Border Collie" watches closely, sees the mismatch between presentation and reality, and feels driven to herd everyone toward coherence. [2] The Faustian Bargain: a Border Collie can become a Fox—gaining influence by surrendering the right to challenge the fiction—but risks gradually becoming the person who preserves the ambiguity and disciplines those who point it out. [2]

Why it matters: PMs who see strategic incoherence face a real choice—herd locally by translating ambiguity into something their team can use, backchannel, name the contradiction, or leave. Cutler makes the slow cost visible: the drift from "staying close to power to improve things" to "protecting the ambiguity as your job." [2]

Tactical Playbook

Build a cold-start eval set in one sitting

Daniel McKinnon (ex-PM at Meta and Google) outlines a six-step method for building an offline eval set for an AI feature before any production data exists. The premise: "evals are the new PRD"—the detailed if/then product statements that used to live in a PRD now live in specific evals. [3]

1. **Write the problem in one sentence**—e.g., "Extract sender's name from support e-mails." If you can't, you don't understand the feature yet. [4]
2. **Validate domain expertise.** If you've never shipped an AI feature, find someone who has to walk you through your first eval. [4]
3. **Set up the workspace.** Create a Project in Claude or ChatGPT, upload sample data, and use a structured prompt that returns the input, correct answer, and a grader line. [4]
4. **Find your floor.** Test the easiest genuine case. If the model can't do it, your floor is above the model's ceiling—make it easier. [4]
5. **Find your ceiling.** Test a case you expect to fail. If nothing fails, your eval is already saturated and will never tell you anything. [4]
6. **Fill in the middle.** Binary-search between floor and ceiling, generating roughly 100 cases that vary difficulty. Use AI to build the test for the AI—this is what makes it 90 minutes instead of two weeks. [4]

Scoring: binary pass/fail judge with three criteria—substantive correctness, format, and scope. Calibrate the judge once before trusting it. [4]

Reading the score: if the eval comes back at 50%, slice by dimension. Put guardrails on the product so it only handles cases it gets right ~80% of the time. Everything below the line goes to engineering with a specific target, not a feeling. [4]

Why it matters: McKinnon estimates fewer than 100 PMs worldwide build frontier-model evals, and the analytical parts of the PM role are being commoditized by AI. Evals are built purely on judgment—making them a place to go deep and differentiate. [4]

Cluster feedback before it reaches the roadmap

Treating every raw user comment as a direct roadmap input degrades decisions. Requests for CSV export, Sheets sync, email reports, and API access may all reflect the same underlying need: users moving data into another workflow. Raw feedback is evidence, not a product decision. [5] Public reviews become a signal when multiple users describe the same pain after a release. The rule: cluster first, decide second. [5]

Career Corner

Google adds live AI prototyping to PM interviews

Google has added a 45-minute live AI prototyping round to its PM interview loop for 2026. Candidates receive a prompt and must build a working prototype using any AI coding or prototyping tool. The format is expanding from AIPM roles to all Google PM interviews. [6]

Interviewers evaluate three things: problem framing before building (candidates who jump straight into building consistently fail), technical execution (prompting, debugging, working around limitations), and communication (narrating decisions while building). [6]

The coaching framework: clarify the prompt (users, goals, constraints), scope aggressively to prove one core interaction, build with a simple stack, test and acknowledge what's broken, and close by defining success metrics. Be completely fluent in your tool before interview day—all cognitive energy should go to the product problem, not the tool's interface. [6] Even if told the round isn't part of your specific loop, practice it anyway—AI fluency will come up somewhere in the interview. [6]

Sources

1. X post by @andrewchen
2. TBM 433: The Border Collie Faustian Bargain
3. substack
4. How to Build Frontier-Lab Quality Evals with Daniel McKinnon, ex-PM at Meta, Google
5. r/prodmgmt post by u/Careful-Stay-7112
6. The New Google Interview Round Screwing PM Candidates in 2026