CASE STUDY 06 · GENERATIVE AI · EVALUATION & MEASUREMENT

Evals Are the New PRD: Measuring a Generative Recommendation Product

A practical framework for evaluating LLM-powered experiences — relevance, coherence, brand safety, taste — written for PMs who need to define “good” before engineering can ship it.

Why Traditional Product Measurement Breaks

For a decade, PMs had a reliable playbook: define a metric, ship an A/B test, read the dashboard. Generative products break that playbook in two ways. First, the failure modes are qualitative — a hallucinated product claim, an off-brand tone, a stylistically incoherent set — and no engagement metric catches them before customers do. Second, the output space is effectively infinite: you cannot QA every response the way you QA every screen. Shipping an LLM-powered recommendation experience taught me that evaluation is no longer a QA activity that follows the spec. It *is* the spec.

The Core Reframe

> *“A PRD describes what the product should do. For generative products, an eval suite is the only enforceable definition of what the product should do.”*

If “good” isn't written down as a measurable rubric, then “good” is whatever the model happens to produce that day — and every model upgrade, prompt tweak, or retrieval change silently redefines your product. The PM's job is to make quality legible: to translate taste, trust, and brand into criteria a system (or a human reviewer) can score.

The Framework: Four Layers of Evaluation

- **Layer 1 — Grounding & truthfulness (automated, blocking).** Every factual claim must trace to source data: products exist, are in stock, and attributes match the catalog. These evals run on every change and block release. Hallucination is a launch blocker, not a tuning parameter.
- **Layer 2 — Safety & brand (automated + spot-check, blocking).** Toxicity, bias, off-brand tone, and category-sensitive content (e.g., children's products) get automated classifiers plus human spot-checks. The key PM decision is defining the severity taxonomy: which failures are ‘fix in next sprint’ vs. ‘roll back now.’
- **Layer 3 — Quality & taste (human rubric, sampled).** Is this inspiration set coherent? Would a stylist endorse it? We wrote rubrics with anchored examples — a 1, a 3, and a 5 for each dimension — so that different reviewers score consistently. Anchors matter more than scales: without examples, ‘4 out of 5’ means nothing.
- **Layer 4 — Outcomes (behavioral, in production).** Engagement, add-to-cart, return visits — the classic metrics, now interpreted carefully: a generative surface can win clicks through novelty while losing trust. Pair short-term engagement with longer-horizon holdouts.

What This Looks Like in Practice

Before any build	The eval rubric is drafted alongside the PRD — engineering sees the definition of ‘good’ before writing a line of code.
Every iteration	Prompt or model changes run the full automated suite plus a sampled human review; scores are compared against the current baseline, like a regression test for taste.
At launch	Release criteria are eval thresholds, not vibes — layers 1 and 2 at target, layer 3 at parity or better with editorial benchmarks.
Post-launch	Layer 4 metrics feed back into rubric revisions: when behavior and rubric disagree, the rubric is wrong.

Hard-Won Lessons

- **LLM-as-judge is a force multiplier, not an oracle.** Using a model to grade model outputs scales beautifully for layers 1–2, but calibrate it against human judgment regularly — judges drift, and they share blind spots with the systems they grade.
- **Evaluate the retrieval, not just the generation.** In grounded architectures most ‘generation’ failures are retrieval failures wearing a fluent disguise. Separate eval suites for candidate quality and composition quality tell you which system to fix.
- **Budget for evals like a feature.** Rubric writing, review tooling, and human grading time are real costs. Teams that skip them don't save the money — they pay it later, in production, with interest.

What I'd Tell Other PMs

> *“You can't A/B test your way out of hallucination. If you can't articulate what ‘good’ looks like in a rubric, you can't ship, iterate, or defend your product.”*

Owning the eval framework is the highest-leverage thing an AI PM can do. It's where product judgment, customer empathy, and technical fluency converge — and it's the artifact that outlives any individual model version.

Sneha Chitte · Senior Product Manager – ML & Personalization · Greater Boston, MA · Metrics are directional and anonymized to protect confidential information.