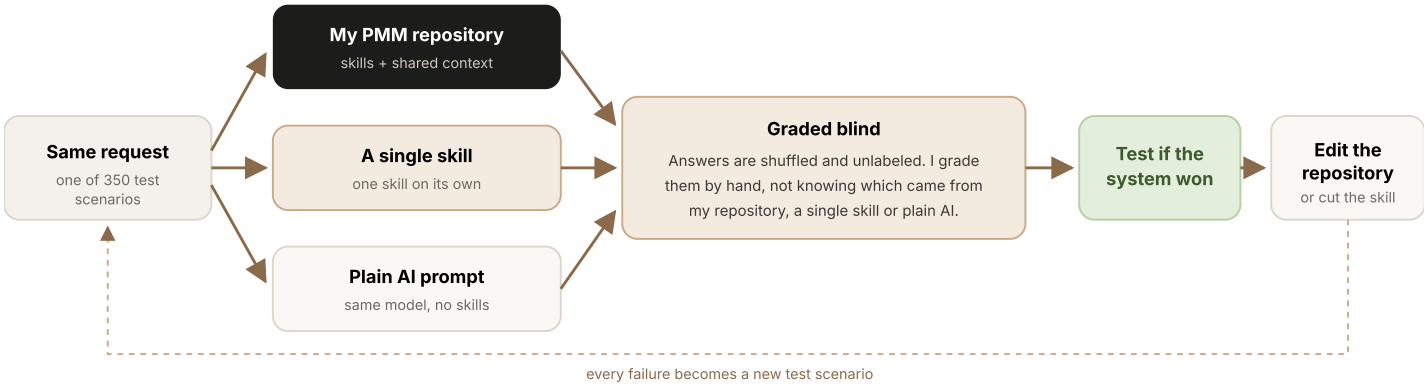
02 · EVALS

How and why I built evals for my PMM repository

It is important to test if the PMM repository I've built beats individual skills or general AI prompts. This is how I tested that and built my benchmarks.

HOW I RUN THE TEST



WHAT I GRADE EVERY ANSWER ON

Grounded
Stays true to the input and says what it assumed

Decisive
Makes a clear call on reasoning that holds up

Usable
A senior PMM could act on it with light edits

Sharp
Says something general best practice would not

RESULTS: WITH THE SYSTEM VS A PLAIN AI PROMPT



BENCHMARKS BEHIND THE NUMBERS

45	350	2,000	19	14%
Connected skills, all passing validation	Test scenarios	Quality checks	Skills with "never invent" checks	Format-only checks, down from 19%

HOW EVERY CHANGE IS TESTED

Smoke test first
Every change runs once in session and is reviewed.

Weekly benchmark
I retest my library and rescore it often.

Preserve the baseline
Edits follow the baseline I set, based on feedback.

No edits mid-run
Finish the run, then report the skill problem.

WHAT I LEARNED

Rules can fight
"Size the market" plus "never invent" let it size nothing.

Labels become theater
An unranked gap list met the rule and helped no one.

Tests miss what they skip
A rebuild passed every check but lost proof points.

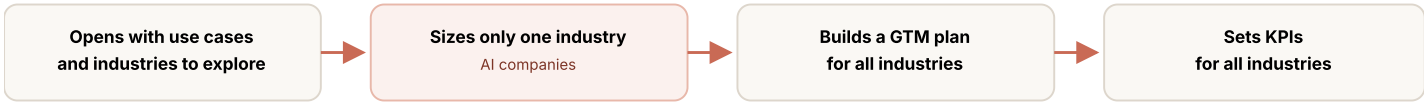
Run paid outputs
Full paid runs catch what quick tests miss.

02 · EVALS

When two rules compete, which one wins?

An example of a real failure test. On a contract project, my launch-strategy skill had two goals pulling against each other: pick the best market, and only trust positioning that already had social proof behind it.

WHAT THE PLAN DID



DIAGNOSIS

Proof was lopsided
All my stats and logos came from AI, our most dominant sector.

No proof read as not relevant
The skill dropped every other industry from the sizing.

It picked the wrong lead
E-commerce was the real winner. AI won on evidence alone.

THE CORE RULE THAT CAME OUT OF IT

I have to tell my agents how to prioritize when rules compete. Here, the rule was "no social proof, deprioritize the segment." **No social proof should mark a claim as not yet proven, never decide its priority.**

HOW I SETTLE CONFLICTS NOW

Opportunity sets priority
Pain and reach pick the lead. Proof sets confidence.

Size all markets, then prioritize
Build sizing across a range of verticals, then decide which rule holds.

One plan for the whole launch
If an industry is picked, sizing, GTM plan and KPIs all reflect it.

THE RULES I ADDED, AND WHAT EACH ONE STOPS

1. **Label every claim**
   Sourced, Assumed with a check, or Gap with how to get proof
   Prevents: unproven segments being deleted
2. **Never stall**
   When facts are missing, proceed and state the assumption
   Prevents: agents blocking on missing inputs
3. **Derived is not invented**
   A sourced total always yields a full-market range, ratios labeled
   Prevents: labeling everything a Gap to avoid sizing
4. **Figure trace**
   Every number is sourced, derived or assumed with a basis
   Prevents: manufactured facts
5. **Plausibility check**
   Derived counts show their share; above 50% needs a reason
   Prevents: 93% of an anchor "qualifying"
6. **Segment selection**
   Score pain, reach and proof separately, then pick a lead
   Prevents: proof acting as the only filter
7. **Real wedge**
   Year one is a subset, with a reason for each segment left out
   Prevents: a "wedge" that includes everything
8. **Consistency check**
   Same segments in sizing, positioning, Phase 2, KPIs and copy
   Prevents: hero copy drifting back to one segment
9. **Ranked gaps**
   Top five gaps by impact in the body; the rest in an appendix
   Prevents: gap lists as compliance theater
10. **Inputs and assumptions note**
    Open with what was used, missing and assumed
    Prevents: hidden input guesses

HOW I TESTED THE FIX

I made up a product, Tollgate, so the AI couldn't use what it already knew. Its proof came only from AI companies. To pass, the plan had to:

- [x] 2+ segments outside the proven one, each labeled
- [x] Full market sized before the lead segment is chosen
- [x] 2 to 3 sentence "why this segment first"
- [x] Same segments in sizing, Phase 2 and KPIs, both ways
- [x] Gaps ranked, top five in body, owner on every row
- [x] No invented sizes for unproven segments

WHAT HAPPENED

Previous version
Dropped every industry without proof. Failed the test.

New skill, first try
Kept the industries we were prioritizing, but sized none of them.

After one more rule
Sized the full market at $2B to $6B and picked 3 of 6 industries to start, with a reason for each one left out.