



From working to launched.

The four steps that take an AI-built prototype the last mile to real users. It's the part where almost every team stalls.



THE LAST MILE

AI got you to working. Launch is where it stalls.

A working prototype is a smaller achievement than it feels. It behaves because you're the one holding it: you know the happy path, you avoid the inputs that break it, and nothing is trying to abuse it. Real users do none of that.

The gap between "it works on my screen" and "it works for a customer" is not the model. It's the unglamorous last mile: auth, permissions, cost control, observability, rollout. None of it demos, and all of it decides whether anyone can depend on the thing.

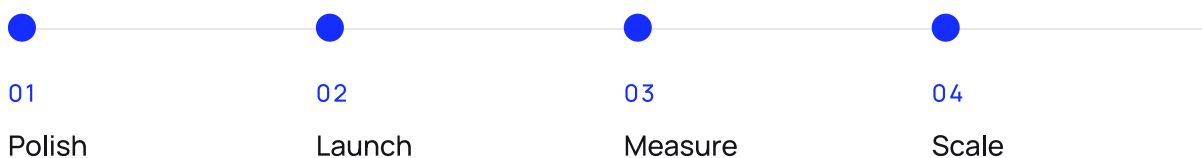
This playbook is the route we walk teams through: four steps, in order, from a prototype that works to a product that's launched and measured.

~6
wks

is the window in which most AI prototypes stall after the "it works" moment. Momentum fades before the last-mile work is scoped.

ILLUSTRATIVE

THE FOUR STEPS





01

STEP ONE · POLISH

Make it survive contact.

A demo forgives everything a production environment punishes. Polish is the work of closing the distance: turning the parts you faked or skipped into the parts a stranger can lean on. It's rarely hard. It's just invisible, so it gets deferred until it's the thing blocking launch.

WHAT'S USUALLY STILL MISSING

- ☐ **Auth & permissions** enforced, not mocked
- ☐ **Rate limits & cost caps** on every model call
- ☐ **Monitoring** that pages a human, not a log
- ☐ **Secrets** out of the codebase and rotated
- ☐ **Input validation** for the paths you didn't test
- ☐ **A graceful failure** for when the model is wrong

"If a wrong answer is expensive and nobody would notice it, you haven't finished Polish. You've just moved the risk onto your users."

1 → 5

A single vibe-coded endpoint typically hides **five** distinct production gaps once you look, most of them in this list. (Illustrative.)



02

STEP TWO · LAUNCH

Own the rollout.

Launch isn't a moment. It's a controlled sequence you can reverse. The teams that ship AI calmly are the ones who decided, before the lights came on, exactly how they'd turn them back off. Treat the first real users as a dial you can turn, not a switch you flip.

Environments

Staging that mirrors production closely enough that "it worked in staging" means something. Config lives in the environment, not the code.

Deploys

A repeatable, boring deploy anyone on the team can run. Ship behind a flag; release to a slice of users before everyone.

Rollback

One command, rehearsed, that returns you to the last good state. A rollback you've never tested is a hope, not a plan.

THE LAUNCH-DAY QUESTION

"If this goes wrong at 2pm, what's the exact sequence that gets us back, and who runs it?"

A staged rollout to **5–10%** of traffic for the first 48 hours catches most launch-day surprises before they reach your whole user base. (Illustrative.)



03

STEP THREE · MEASURE

See what it's actually worth.

An AI feature you can't measure is a feature you can't defend, not in a cost review and not in a roadmap fight. Instrument it from day one so the conversation is about evidence, not opinion. Four things are worth watching from the first real user.

WHAT TO INSTRUMENT

ILLUSTRATIVE SIGNAL

Usage

who does what, how often



Feedback

was the answer good?



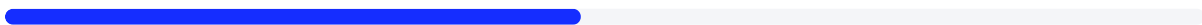
Cost per action

tokens & infra, per outcome



Errors & escapes

what went wrong, unseen



"Cost per successful outcome is the number that ends most 'is the AI worth it' debates. Track it before someone asks."



04

STEP FOUR · SCALE

Meet production with confidence.

Scaling isn't only more traffic. It's more surface area, more edge cases, and more people depending on something that used to live in one person's head. The goal of this step is a product that can grow without a rewrite, and a team that can carry it without you.

Harden

Close the edge cases the first users found. Add the tests that stop them coming back.

Extend

Grow it deliberately, internally or externally, on the evidence Measure gave you rather than on guesses.

Document

Write down how it works and how it fails, so the knowledge outlives whoever built it.

Hand over

Leave a team that can run, extend and trust it. That's the real end of the last mile.

Done well, Scale is quiet. The product keeps working as more people rely on it, and the team that owns it stops needing the people who built it.

BUILD WITH AI · LAUNCH WITH SUBWORKFLOW

The prototype was the easy part.

We bring battle-tested engineering to the last mile:
polish, launch, measure, scale. It's how the software
you built with AI actually ships to real users.

Stuck on the last mile?

Book a 30-minute discovery call. We'll tell you
honestly what's between your prototype and a
launch, and exactly where we'd start.

[Book a discovery call →](#)