The use-case hypothesis

ar003 · 1 July 2026

This document is the outline of the use case for testpath, it is to become the basis of the blog post that gets sent out in the email campaign.

What testpath is for

Regression testing for stochastic LLM agents, gated in CI. You change a prompt, a model version, or a tool on a production agent and need to know **did I break anything?** before you ship. Because the agent is non-deterministic, and the LLM grading it is too, the same eval gives different scores on reruns, so a green score on a once-run eval can't separate a real regression from the model just rolling differently. testpath measures the pass rate **with an error bar**, turns it into a 🟢/🟠/🔴 verdict, and blocks the merge on a real regression without flaking on noise.

The concrete failure it prevents: a team tweaks a prompt on its support/refund agent, the eval goes green, they ship; a week later a customer reports the refund flow has been quietly wrong since Tuesday. Bad answers, escalations, churn, a frantic rollback.

Three types of potential customer

The use case only bites for one of them. The decisive filter is **who owns the agent's quality** and **what a silent regression costs them**.

1. Small — rejected

Solo builders and very early teams whose agent is a prototype or carries little production traffic. A silent regression costs them almost nothing yet, the budget isn't there, and a CI gate is premature. Real pain hasn't arrived; there's nothing to sell against.

2. Medium — our target

A team running a **customer-facing agent in production that they built themselves**, typically on a builder platform (Botpress, Voiceflow, Rasa, …) or in-house. They **own the agent's quality**, they're technical enough to wire a verdict into CI, and a silent regression is a genuinely bad day. They can pay ≈ $500/mo without a procurement cycle. Pain × ability-to-pay × reachability all line up here.

The orthogonal qualifier: they must have *built* it. Customers of managed / turnkey services (Ada, Intercom Fin, Decagon) don't own quality (the vendor does), so they feel no pain. Builder customers are targets; managed customers are noise. The full split is in the market landscape (ar001).

3. Enterprise — rejected

Large orgs likely already run internal eval/ML teams and bespoke tooling, so the gap is smaller. Security review and procurement kill the near-zero-CAC cold-outbound motion, and the land is slow. Deprioritize until *inbound* pulls us there: not a cold-start target.

So the bet is

The medium segment (builder customers with a production agent they own) feels the regression-vs-noise pain acutely, can pay the anchor price, and is reachable by cold email. That is the audience the lead-gen plan (ar002) targets and the design-partner questions (ar004) are meant to validate.