RESEARCH

The 90-Day AI Pilot Playbook

How to test AI in a single workflow before committing — and what to measure before you scale.

By **Alan Babbitt** · Intelligent Systems Architect · May 2026 · 5-minute read

The premise

Most AI pilots fail not because the technology was wrong, but because the pilot was never **designed to succeed.**

The dominant pattern: a vendor demo lands. Leadership greenlights "a pilot." Someone picks a workflow, installs the tool, gives access to a few enthusiastic users. Three months in, the question gets asked: ***did this work?*** Nobody can answer cleanly. Usage exists but value is fuzzy. The conversation drifts. Six months in, the pilot is "ongoing." A year in, it quietly dies — or quietly scales, with nobody really sure if the scale is justified.

The pilot didn't fail to produce a tool. It failed to produce a **decision.**

A well-designed pilot has exactly one job: produce a clean, defensible decision on day 90 — ***scale, kill, or iterate.*** Anything that doesn't support that decision is overhead. Anything that prevents that decision is sabotage.

The pattern

Five failure modes show up in nearly every undisciplined pilot:

No pre-committed success metric. The pilot launches without anyone defining, in writing, what "worked" looks like. Three months later, success becomes whatever the champion needs it to be. The pilot is judged on enthusiasm, not outcome.

No kill criteria. Nobody pre-commits to ***stopping.*** So the pilot runs until political will runs out — not until evidence resolves. Tools accumulate. Budgets compound.

Scope drift. What started as "AI in the discovery conversation" becomes "AI across the sales cycle" by month two. When the pilot tested everything, it learned nothing about anything specific.

Vanity metrics. Usage gets measured, not outcomes. ***"95% of advisors are using the tool weekly"*** — true and meaningless. The question is whether the tool changed conversion, cycle time, or accuracy. Usage is necessary, not sufficient.

Champion-driven survival. The pilot's most enthusiastic user becomes its evaluator. Of course it succeeded — they need it to succeed. A real pilot has at least one skeptic empowered to call it.

The pattern across all five: pilots designed to *generate evidence* outperform pilots designed to *prove correctness* every time. The first treats the outcome as a question. The second treats it as a foregone conclusion. The first scales the right things and kills the wrong things. The second scales whatever survives the political fight.

> *Pilots designed to generate evidence outperform pilots designed to prove correctness — every time.*

The 90-day frame

Ninety days isn't arbitrary. It's the minimum window to see real signal in most workflows, and the maximum window before momentum stalls. Inside that frame, the playbook breaks cleanly into four phases.

- **Days 1–7 · Define.** Write down the single workflow, the success criteria (quantified), the kill criteria (also quantified), and the day-91 decision-makers. If you can't write these in a single page, you're not ready to start. The week of definition is the cheapest insurance you'll ever buy.
- **Days 8–30 · Baseline + build.** Measure the current-state metric on the workflow *before* the tool touches it. Most pilots skip this and lose the ability to claim any lift afterward. Install the tool, train the pilot users, start measuring.
- **Days 31–75 · Run + adjust.** The tool runs in production on the pilot workflow. Weekly review against success and kill metrics. Small adjustments allowed (configuration, training). No scope changes. No new use cases. The workflow is the laboratory.
- **Days 76–90 · Decide.** The team reviews the evidence. The decision is one of three: **scale** (commit beyond the pilot scope, with a new defined scope), **kill** (the tool stops on day 91, license is canceled), or **iterate** (one more focused 90-day cycle with a clear hypothesis about what's different).

The decision is the deliverable. Not the tool. Not the usage. **The decision.**

The 5-question pilot brief

Before kicking off any AI pilot, fill in this one-pager. If you can't answer all five, you don't have a pilot — you have a software trial.

THE 90-DAY PILOT BRIEF

1. **What single workflow are we piloting?** Not a category, not a tool. A workflow. *"Inbound lead qualification, from form-fill to first call booked."*
2. **What does success look like on day 90?** Quantified, time-bound, observable. *"30% reduction in time-to-first-call, with no drop in show rate."*
3. **What does failure look like on day 90?** The kill criteria. Quantified. *"No measurable lift over baseline, or show rate drops more than 5%."*
4. **Who runs the pilot day-to-day, and who has authority to decide on day 91?** Name names. Including the skeptic.
5. **What happens on day 91 for each outcome — scale, kill, iterate?** Pre-define all three paths. The decision should not feel novel when it arrives.

A pilot designed against these five questions resolves cleanly. A pilot without them drifts indefinitely.

Bottom line

The expensive AI mistakes in 2026 won't be the failed pilots — they'll be the pilots that *never resolved.* A clean "no, this didn't work, we're killing it" is a successful pilot. A clean "yes, this worked, here's exactly what we're scaling and why" is a successful pilot. The only failure mode is the pilot that runs for a year, produces neither answer, and quietly sets the budget for next year.

Pilot to decide. Decide on the evidence. Move.

The 90-second version of the diagnostic that tells you which workflow to pilot first is at **isaadvisory.com →**