



AI Simulation Pilot Kit

Updated July 11, 2026

This kit is the worksheet version of my guide to AI training simulations. Print it or keep it open in a second window. Fill it in as you plan the pilot, score your use cases, and grill vendors.

1. The 90-day pilot timeline

Ninety days is enough to prove or kill an AI simulation pilot. Assign one owner per row before the pilot starts.

Weeks	Milestone	Owner
1 to 2	Pick one use case with the scoring rubric below. Get the business sponsor to sign off on the before/after metric.	
2 to 3	Run the vendor evaluation checklist. Shortlist two vendors, max.	
3 to 4	Security and procurement review (SOC 2, SSO, data retention). Sign the contract or pilot agreement.	
4 to 6	Build 3 to 5 scenarios with SME input. Set pass-rate targets with the worksheet below.	
6 to 7	Dry run with 5 to 10 friendly users. Fix broken scenarios. Capture baseline business metric.	
7 to 10	Live pilot with the real cohort (30 to 100 people). Monitor completion and pass rates weekly.	
10 to 11	Pull results: pass rates, attempts to pass, learner feedback, manager feedback.	
11 to 12	Compare before/after business metric. Write the one-page readout for the sponsor.	
13	Go/no-go decision: scale, extend the pilot, or kill it.	

Failure mode: the drifting pilot

Pilots that skip the week 13 decision meeting run for six months and die of neglect. Put the go/no-go meeting on the sponsor's calendar in week 1, before anything else.

2. Use-case selection worksheet

Score each candidate use case from 1 to 5 on the criteria below. Pilot the highest total. Ties go to the use case with the strongest business metric.

Scoring rubric

Criterion	1 (weak)	5 (strong)
Conversation-shaped skill	The skill is a process or software task	The skill is a live conversation (sales calls, coaching, difficult feedback, support)
Practice gap today	People already get lots of live practice	People go live with little or no practice
Measurable business metric	No metric anyone tracks	A metric the business already reports (close rate, CSAT, ramp time, escalations)
Cost of failure in the real world	Mistakes are cheap and reversible	Mistakes lose deals, customers, or trust
Audience size and access	Under 20 people, hard to reach	50 or more people you can enroll this quarter
SME availability	No SME can give you time	An SME has committed hours in weeks 4 to 6

Your candidates

Use case	Skill	Practice gap	Metric	Cost of failure	Audience	SME	Total

Failure mode: the pet project

The most common bad pick is a use case a stakeholder loves but that scores low on the business metric row. If you cannot name the metric and the person who owns that number, do not pilot it.

3. Vendor evaluation checklist

Run this against every vendor demo. Ask them to show, not tell. The first five groups are the criteria from the guide. The last group is what security and procurement will ask you anyway, so ask it first.

Criterion 1: Scenario realism

- ☐ The AI character stays in role when I go off script or act difficult
- ☐ Responses reflect my industry and the personas I need, not generic small talk
- ☐ Voice option sounds natural enough that learners will take it seriously
- ☐ I tried to break it in the demo myself, not just watched the vendor drive

Criterion 2: Feedback and coaching quality

- ☐ Feedback references specific things the learner said, with quotes or timestamps
- ☐ Feedback maps to a rubric or competency model I can see and edit
- ☐ Feedback tells the learner what to do differently on the next attempt
- ☐ Learners can retry immediately after feedback

Criterion 3: Scoring and analytics

- ☐ Scoring criteria are visible and adjustable, not a black box
- ☐ Two similar performances get similar scores (ask for consistency evidence)
- ☐ Dashboard shows pass rates, attempts, and score trends by cohort
- ☐ I can export raw data, not just view dashboards

Criterion 4: Authoring and control

- ☐ My team can build and edit scenarios without vendor services hours
- ☐ Time to build one new scenario is hours, not weeks
- ☐ I can control what the AI will and will not say (guardrails, banned topics)
- ☐ I can test a scenario as a learner before publishing it

Criterion 5: Integration and rollout

- ☐ Works in the browser without installs or IT tickets for learners
- ☐ Completion and score data flow to my LMS or LRS (SCORM, xAPI, LTI, or API)
- ☐ Vendor supports a pilot-sized contract, not just annual enterprise deals
- ☐ Realistic go-live estimate for my cohort is inside my week 7 target

Security and procurement

- ☐ SOC 2 Type II report available under NDA (ask for the actual report, not the badge)
- ☐ SSO via SAML or OIDC included, and I know at which pricing tier
- ☐ Written data retention policy: what is stored, for how long, and how deletion works
- ☐ Learner conversation data is not used to train models outside my tenant, in writing
- ☐ Data residency options match my legal requirements
- ☐ Subprocessor list available, including which AI model providers they use
- ☐ Vendor completed my security questionnaire before contract, not after

Failure mode: the demo-day mirage

Vendors demo their best scenario with a trained driver. Insist on hands-on access with your own use case before you sign. If a vendor will not give you a sandbox, that answers your question.

4. Pass-rate target worksheet

Set targets before the pilot goes live. Targets set after you see the data are not targets, they are rationalizations.

Question	Your answer
What score counts as a pass on one attempt?	
What percent of the cohort should pass by the end of the pilot?	
How many attempts is a learner allowed before we flag them for coaching?	
What first-attempt pass rate is too high (scenario too easy)?	
What final pass rate is too low (scenario too hard or broken)?	
Who reviews flagged learners, and how often?	

Tip

Set targets that match the stakes and design of the scenario. A certification gate may be built for near-universal first-attempt passes, while a stretch scenario is built so most people need a few attempts. What matters is committing to the targets before the data arrives and reviewing flagged learners on a schedule.

5. The before/after business metric

This is the section your sponsor actually reads. Complete it in week 1 and again in week 12.

Field	Your answer
Business metric (one number the business already tracks)	
Who owns this number	
Baseline value (captured week 6, before the live pilot)	
Value at week 12	
Other changes during the pilot that could explain the movement	

Sponsor's verdict: scale, extend, or kill

Failure mode: measuring only learning data

Pass rates prove people can perform in the simulation. They do not prove the business changed. Pilots that report only completion and pass rates get polite nods and no budget. Anchor the readout to the metric your sponsor already answers for.

Have any questions? Email me at devlin@devlinpeck.com