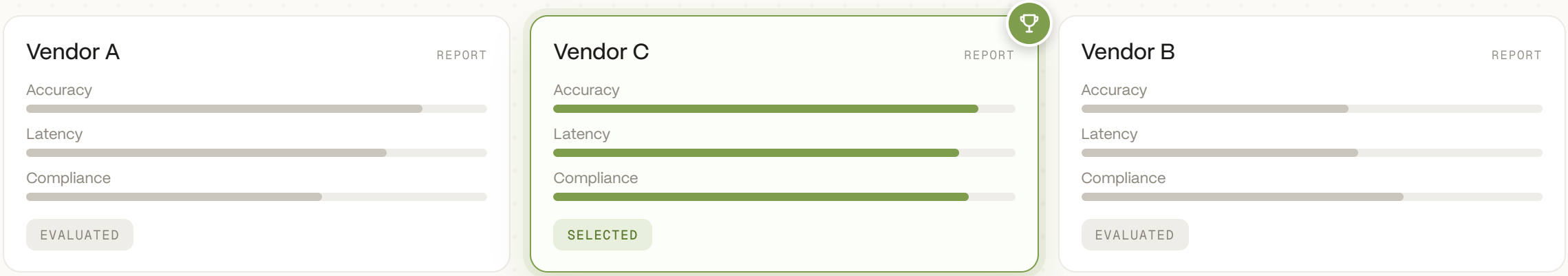
COVAL

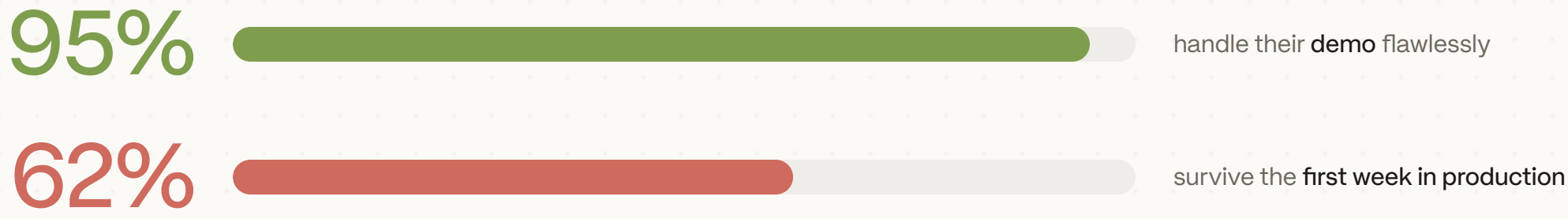
THE PLAYBOOK · VOICE AI

The Voice Agent Vendor Testing Playbook

How to pick the voice AI vendor your callers actually experience, beyond the demo.



The demo is not the deployment.



You buy the curated demo; you deploy the long tail of accents, interruptions, noise, and edge cases, expensive to unwind and slow to detect. **A bake-off moves the measurement from the demo to the deployment before you sign.**

Run every vendor through one shared test, the way Waymo runs simulated miles.

Same callers, same scenarios, same scoring, for every vendor. A demo is the ride-along; a bake-off is the simulation. Three ingredients make the shared test suite:

THE HOW

Personas

Synthetic callers: accents, interruptions, emotion, noise.

+

THE WHAT

Test sets

What each caller is trying to get done.

+

THE GRADE

Metrics

How every call is scored, the same way.

=

The Outcome

One shared test suite

Every vendor runs it. Reusable into production.

Three ways vendor testing misses the mark on production performance.

01 Demo theater

Vendors demo cherry-picked base cases on clean audio. None of it predicts the caller with a thick accent interrupting from a moving car.

02 Subjectivity wins

One senior stakeholder dislikes a voice on a single call and overrides weeks of structured scoring with a thirty-second gut reaction.

03 The invisible 10,000 conversations

A real run generates thousands of calls. Nobody reviews them by hand, so the long-tail failures stay hidden until a customer hits them live.

Every bake-off is one of four shapes. Name it, then pick the decision rule.

SHAPE 01

Formal enterprise RFP

A multi-phase funnel narrowing a wide field to a few finalists; the bake-off is the last phase.

WHEN Large, regulated, or high-spend buys that need a paper trail.

TRAP Scoring the paperwork, not the agent, a rubber stamp on a slideware decision.

SHAPE 02

Build vs. buy

An internal or incumbent build measured against an external challenger.

WHEN You already run an agent and want to know if a vendor beats it.

TRAP Grading the build on sunk cost and familiarity instead of the same suite.

SHAPE 03

Head-to-head

Two or three finalists decided on the merits, with no incumbent.

WHEN You've narrowed to two or three credible platforms.

TRAP Letting one dimension settle a decision that should be weighted.

SHAPE 04

Incumbent displacement

A challenger must clear a bar set by an entrenched, deployed vendor.

WHEN A live vendor underperforms and you're weighing the migration cost.

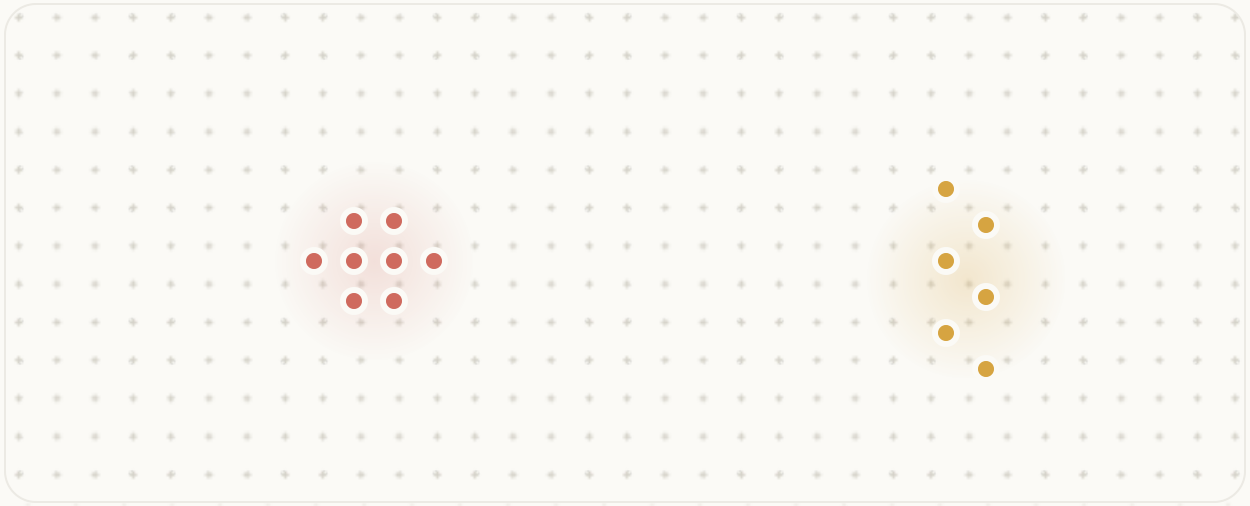
TRAP Holding the challenger higher than the incumbent ever passed.

Seven steps to a defensible bake-off. The order matters.

1. **Define "good" first**
   Write and weight your criteria before any demo.
2. **Mine production calls**
   Build the suite from what really happens on the line.
3. **Build one shared suite**
   Personas, test sets, and metrics every vendor runs.
4. **Translate to metrics**
   Turn each requirement into something you can score.
5. **Run identically**
   Same suite, same iterations, in parallel.
6. **Score honestly**
   Grade against the rubric, not the pitch.
7. **Decide on your weights**
   The scorecard plus the weights you locked first.

→ SEE FULL METHOD IN THE PLAYBOOK

Mine your production calls into the test suite.



~10,000 calls graded Dropped readback Latency spike under load

- **Scenario taxonomy:** the requests that actually come in, and where money and patience get lost.
- **Behavior taxonomy:** where agents fail, like multi-item edits and confirming the final total.
- **Persona realism:** accents, interruptions, emotion, and engine noise on every call.

That shared suite then covers the long tail that a manual sample of a few dozen calls would otherwise miss.

Cull the field in rounds, then watch how finalists respond.

Between rounds, hand finalists their failures and score how fast they fix them. Responsiveness and pace of improvement are the strongest predictors of the partnership, and they are invisible in a one-shot test.

No single vendor wins every row.

Vendors	Order Accuracy	Readback	Latency	Naturalness
Vendor A	94%	96%	2.9s	70%
Vendor B	86%	90%	3.6s	92%
Vendor C	91%	41%	1.4s	58%

- **Vendor A:** accurate, but drops the complex orders that drive margin.
- **Vendor B:** best experience, but quotes prices the chain can't honor.
- **Vendor C:** fast and reliable, but skips a required readback.

The decision follows **the weights you set first**, not whichever vendor gave the best demo.

Your selection suite becomes your monitoring suite.

SELECT

The bake-off

Build one shared suite of personas, scenarios, and metrics to choose a vendor.

→

SIGN

The chosen vendor goes live

The same suite ships into production with the agent.

→

SCALE

Production monitoring

Regression, adversarial, and human review, scored against the bar you set.

When you future-proof with a vendor- and model-agnostic evaluation platform, your vendor evals compound as you scale.

Run your next bake-off with rigor.

READ

The full playbook

Every failure mode, the four shapes, and the seven-step method in depth.

coval.ai/blog/voice-agent-vendor-testing →

RUN IT YOURSELF

Get the Claude Code skill

Load Coval's vendor-eval skill into Claude Code and scope your bake-off in minutes.

coval.ai/vendor-eval-skill →

TALK TO US

Discuss your vendor testing with a Coval solutions engineer

Book a strategy call →