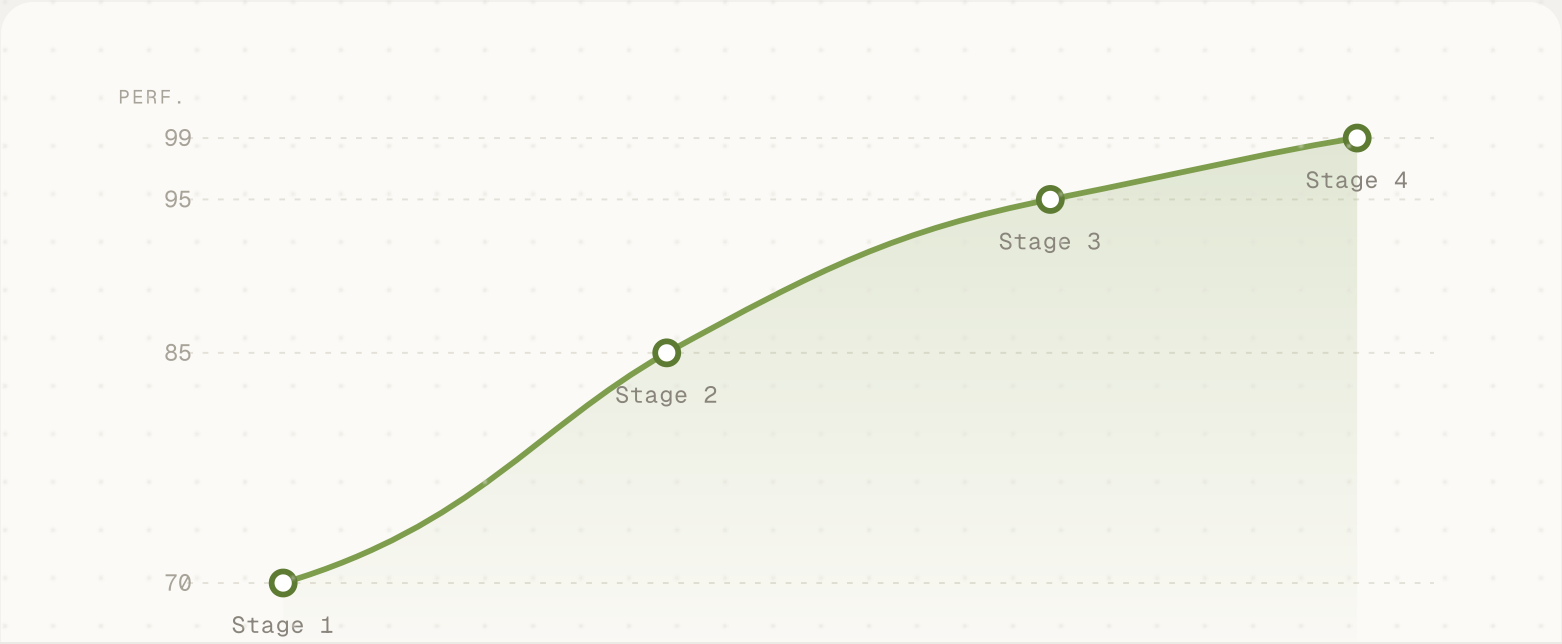
COVAL

THE PLAYBOOK · VOICE AI

The Voice AI Eval Strategy Playbook

Getting a voice agent from 70% to 99% in production, on purpose.



The climb from 70% to 99%.

Each stage demands a different approach. The first gains are easy; the last mile is where evaluation infrastructure earns its keep.

01 — 70%

Out of the box

Base cases work. Demos feel magical. Edge cases are invisible because nobody is testing for them.

HOW TEAMS GET HERE

Pick a platform, write prompts, ship.

02 — 85%

Manual effort

Prompt tuning, listening to calls, ad-hoc fixes. Quality climbs, but every fix is manual and nothing is systematic.

HOW TEAMS GET HERE

Hire QA. Listen. Tweak. Repeat.

03 — 95%

Programmatic evals

Automated test suites and regression detection. The team shifts from firefighting to quality engineering.

HOW TEAMS GET HERE

Build voice AI evaluation infrastructure.

04 — 99%

The frontier

Production feedback loops, cross-functional buy-in, cost-optimized orchestration. Every failure becomes a test.

HOW TEAMS GET HERE

A mature eval platform, calibrated and continuous.

Your first evals are two lists: ALWAYS and NEVER.

MUST ALWAYS

- Verify caller identity before accessing account information
- Offer a path to a human agent when the caller requests one
- Confirm critical actions before executing them — transfers, cancellations, bookings
- Follow the required disclosure sequence (HIPAA, TCPA, state-specific)

MUST NEVER

- Provide medical, legal, or financial advice
- Repeat back sensitive information — SSNs, card numbers, full account numbers
- Continue a conversation after the caller has asked to stop
- Skip required compliance disclosures or make promises it cannot guarantee

Pass them at **100% against easy-mode callers** before you add a single hard persona.
That is your regression floor.

Real callers don't sound like easy mode.

Easy

Clear speech, neutral tone, single intent, no background noise, standard dialect.

TESTS

Baseline — does it work at all?

Medium

Slight accent, minor background noise, two intents in one call, occasional filler words.

TESTS

Readiness for real human speech.

Hard

Strong regional accent, heavy background noise, an emotional caller, mid-conversation pivots.

TESTS

Resilience under real conditions.

Adversarial

Non-native speaker, loud environment, an angry post-IVR caller, social-engineering attempts.

TESTS

Stress limits — where it breaks.

Lock the floor first, then broaden. And remember the **LLM people-pleaser trap**: simulated callers won't misbehave unless you make them.

Calibrate the judge before you rank anything.

HUMAN REVIEW VS. THE AI JUDGE — EVERY DISAGREEMENT BECOMES GROUND TRUTH

Identity verified before account access

AI JUDGE Yes

Human Review

Yes | No | N/A

✓ you agree with the judge

No medical, legal, or financial advice given

AI JUDGE Yes

Human Review

Yes | No | N/A

● disagreement saved as ground truth

Order read back before confirming

AI JUDGE Yes

Human Review

Yes | No | N/A

✓ you agree with the judge

One dimension per metric. A bundled "quality" score tells you nothing when it fails. Aim for above 85% agreement before you act on a result.

Structure the suite by cadence.

Not every test runs every time. Keep the per-deploy gate fast; let the broad sweeps run when they can.

PER-DEPLOY	Core regression suite — must-always, must-never, easy-mode personas.	Block bad releases. This is the gate. Minutes, not hours.
NIGHTLY	Extended regression plus medium-difficulty personas.	Catch regressions that slip past the fast per-deploy checks.
WEEKLY	Full suite including hard and adversarial personas, plus edge cases.	Comprehensive coverage sweep. Not blocking a deploy.
ON-DEMAND	Targeted tests for a specific change — new prompt, tool, or model.	Validate a hypothesis before it graduates to regression.

Four pillars of voice observability.

Not whether the call completed, but how the conversation flowed, where it struggled, and why.

01

Conversation content

Full transcripts and original audio. "Yes, I'd like to cancel" reads one way and sounds very different when the caller is crying, angry, or matter-of-fact.

02

Context signals

Who called, from where, at what time, with what history. A caller's third attempt at the same issue should be handled differently from a first-time call.

03

Outcome data

Did the task complete? Resolved on first contact? Any call-back within 24 hours? Resolution rate beats containment as a north-star metric.

04

Performance metrics

Turn-by-turn latency, component breakdowns (ASR, LLM, TTS), and confidence scores. When a call goes bad, the traces tell you where in the stack.

Every production failure makes the suite stronger.

The simulation-to-monitoring loop Waymo pioneered: a failure never ships twice.

1

Capture

Monitoring flags calls that deviate — unusual failures, edge cases, anomalies.

2

Route

Flagged calls route automatically to a human review queue for annotation.

3

Analyze

The team finds patterns across flagged calls — recurring failure modes, clusters.

4

Generate

Each pattern becomes a new test case. The unknown unknown becomes a known scenario.

5

Prevent

New cases join the regression suite, so the same failure is caught before it ships again.

↩ CONTINUOUS — EVERY PRODUCTION FAILURE MAKES THE NEXT TEST SUITE STRONGER

The one-page eval strategy checklist.

- [ ] I know my stage on the maturity curve.
- [ ] The floor is locked before any hard personas were added.
- [ ] Every metric measures exactly one dimension.
- [ ] Evals run on every deploy, with threshold gates.
- [ ] Production failures feed back into regression.
- [ ] ALWAYS / NEVER lists pass at 100% in easy mode.
- [ ] Personas broaden (accents, noise) and deepen (tools, load).
- [ ] The judge is calibrated vs. human review (above 85%).
- [ ] The same suite runs on production traffic.
- [ ] One named owner maintains the suite and reports results.

Run your eval strategy with rigor.

READ

The complete guide

The maturity curve, calibrated metrics, CI/CD gates, and the monitoring loop in depth.

coval.ai/blog/voice-ai-agent-evaluation-guide →

RUN IT YOURSELF

Get the Claude Code skill

Scaffold your eval strategy from this playbook, right in your terminal.

coval.ai/eval-strategy-playbook →

TALK TO US

Master your maturity curve with a Coval solutions engineer

Book a strategy call →