LIVE AMA · JULY 2026

The future of agentic voice

with Brooke Hopkins

Founder & CEO, Coval

YOUR HOST

Brooke Hopkins

FOUNDER & CEO, COVAL · FORMERLY EVALUATION INFRASTRUCTURE AT WAYMO

At Waymo, shipping meant proving a system through rigorous simulation and evaluation, not intuition. Voice AI is now crossing that same line, from impressive demo to production system. That is the vantage point for today: **what changes when the demo is no longer the hard part?**

A FRONT-ROW SEAT

Voice AI, beyond the demo.

Coval works with teams putting voice agents in front of real customers at real volume. That gives us a front-row seat to what improves quickly, what still breaks, and which lessons travel across models, vendors, and industries.

perplexity

StubHub

zoom

Hippocratic AI

toast

Upstart

FROM THE FRONT ROW

What actually breaks in production.

- **Failures live at the edges of conversation.**
  Interruptions, silence, turn handoffs, and recovery separate the demo from the real system.
- **Many failures happen outside the model.**
  Telephony, integrations, tool calls, and permissions can break an otherwise good interaction.
- **Regressions arrive quietly.**
  A model, prompt, or workflow change shifts behavior somewhere the team was not watching.
- **Scripted demos underrepresent reality.**
  Real callers bring accents, noise, ambiguity, urgency, and emotion that happy paths do not rehearse.

Four patterns we see repeatedly, across industries, architectures, and model vendors.

THE MARKET · NEXT 12–24 MONTHS

Five forces reshaping voice AI.

1 **Native audio reaches production.**

Speech-to-speech becomes a real architectural option. Cascaded and hybrid systems do not disappear.

2 **Agents talk and act at the same time.**

Tool use moves inside the conversation, turning voice from a channel into an operating interface.

3 **Naturalness gets cheaper.**

Low latency, expressive speech, and interruption handling become expected, not differentiating.

4 **Success shifts from conversation to outcome.**

The question becomes less "did it sound good?" and more "did it finish the job?"

5 **Trust becomes a product requirement.**

Identity, consent, permissions, evidence, and human override move into the core architecture.

Signals: OpenAI Realtime API, Gemini Live API, FCC artificial voice ruling, NIST AI Risk Management Framework.

WHAT DOES NOT CHANGE

The demos get better.
The job stays the same.

Understand the customer. Complete the work. Recover when reality gets messy. Hand off well. Prove what happened.

HOT TAKE 01 · ARCHITECTURE

Hybrid wins before native audio wins everything.

CASCADED

Control

Observable stages, mature tooling, and predictable behavior.

HYBRID

Choice

Native expression where it matters, explicit control where it counts.

NATIVE AUDIO

Expression

Richer timing, emotion, and turn-taking with a new observability burden.

Current platforms expose both native audio and transcription, making architecture a tradeoff, not a one-way migration.

HOT TAKE 02 · MEASUREMENT

The transcript survives. Transcript-only evaluation does not.

Native audio systems can still produce transcripts. But text cannot tell you whether the agent interrupted at the wrong moment, sounded uncertain, missed a barge-in, or completed the promised action.

The new ground truth: what was said + how it sounded + what the system did.

HOT TAKE 03 · ORCHESTRATION

The best voice agents do work while they talk.

A useful agent does not merely conduct a fluent conversation. It verifies identity, reads the right systems, takes action, confirms the result, and knows when to hand off.

Voice is becoming the front end to **a real-time workflow engine.**

HOT TAKE 04 · DIFFERENTIATION

Naturalness gets commoditized. Reliability becomes the moat.

EXPECTED

Sounds natural

Fast, expressive, and easy to interrupt.

VALUABLE

Finishes the job

Uses tools correctly and produces the intended outcome.

DEFENSIBLE

Recovers well

Handles edge cases, explains itself, and escalates safely.

ONE COVAL BET

Voice agents will improve themselves.

The evaluation system will become the improvement loop: find the failure, propose the next test and a targeted tweak, run simulations, grade the results, and report what changed.

With Coval, teams can already close the measurement loop: rerun the same scenarios before and after a suggested change, quantify the exact delta on specific metrics, and expose any regressions.

`Every tweak gets a measurable before-and-after.`

THE NON-NEGOTIABLE

Trust moves from a principle to a product surface.

Identity. Consent. Disclosure. Permissions. Evidence. Human override. These cannot live in a policy document beside the system. They have to live inside the system.

The future is not an agent that never fails. It is an agent whose limits are **visible, testable, and recoverable.**

COVAL

The winning agent won't sound the most human. It will finish the job.

MODELS CHANGE. THE OBLIGATION TO KNOW WHAT HAPPENED DOES NOT.

NOW, YOUR QUESTIONS.

THE AMA · NINE QUESTIONS FROM YOUR SIGNUPS

Where should we start?

01 Architecture	02 Beyond transcripts	03 Voice eval fundamentals
04 Outcome vs. quality	05 Production reliability	06 Eval operations
07 Hard audio problems	08 Adoption	09 Choosing a provider

Drop a number in the chat to vote for where we start.

QUESTION 01 · ARCHITECTURE · THE MOST-ASKED TOPIC

Will speech-to-speech models ultimately replace cascaded pipelines as the default way to build voice agents, and on what timeline? Where does hybrid land?

Also asked: where do multi-loop architectures fit, with background agents working long tasks alongside the live conversation?

QUESTION 02 · EVALUATING SPEECH-TO-SPEECH

Native audio can still produce transcripts, but they no longer tell the whole story. What should teams measure beyond the transcript?

QUESTION 03 · VOICE EVAL FUNDAMENTALS

What's genuinely unique about evaluating voice versus text agents, and which parts of voice eval are still understudied or misunderstood?

QUESTION 04 · OUTCOME VS. QUALITY

As deployments mature, how does evaluation shift from "was the conversation good" to "did the agent achieve the business outcome"? What becomes table stakes?

QUESTION 05 · PRODUCTION RELIABILITY

What are the biggest reasons voice agents fail in production, especially in high-stakes domains like healthcare, and what would you prioritize to make one reliable from day one?

Also asked: where do today's testing and simulation tools themselves break down in real deployments?

QUESTION 06 · EVAL OPERATIONS

Voice tests are slow and expensive. What should run on every change, what should gate a deployment, and what can run less often?

QUESTION 07 · THE HARD AUDIO PROBLEMS

For turn-taking, interruptions, and background noise, where is off-the-shelf good enough versus building it yourself?

QUESTION 08 · ADOPTION

What rollout strategies are actually working with initially cautious CX teams?

QUESTION 09 · CHOOSING AN EVAL PROVIDER

There are now many platforms offering voice simulations and evaluation models. How should a team choose an eval provider?

KEEP THE CONVERSATION GOING

Thank you.

RECORDING AND SLIDES LAND IN YOUR INBOX

coval.ai

Brooke Hopkins on LinkedIn