LIVE AMA · AUG 13, 2026 · 3:00 PM PT

Improve agent performance with voice benchmarks

Brooke Hopkins
CEO & FOUNDER

Kobi Hudson
LEAD ENGINEER

Brooke Hartley
PARTNERSHIPS LEAD

Cale Payson
ENGINEER

Madina Seribay
ENGINEER

Cooper Rubens
ENGINEER

HOW THIS HOUR WORKS

How and when benchmarks matter.

01 **The framework**
Benchmarks as decision instruments.

02 **Live leaderboard reads**
STT, TTS, and speech-to-speech.

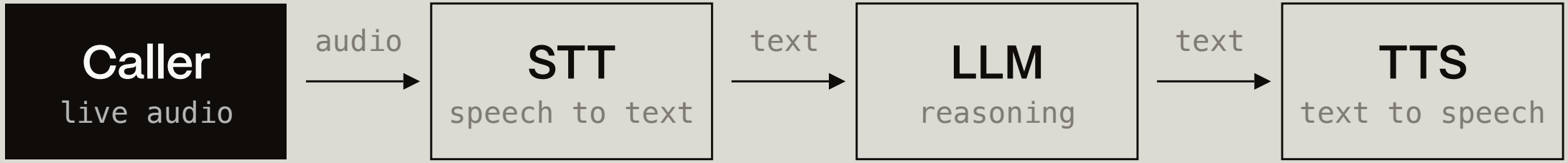
03 **First look**
Coval Voice Arena.

04 **Your questions, live**
Drop them in the chat any time.

Recording, show notes, and links shared after call.

HOW A CALL MOVES

Each layer is a different question.



Can it hear?
WER · TTFS

Can it speak?
TTFA · intelligibility

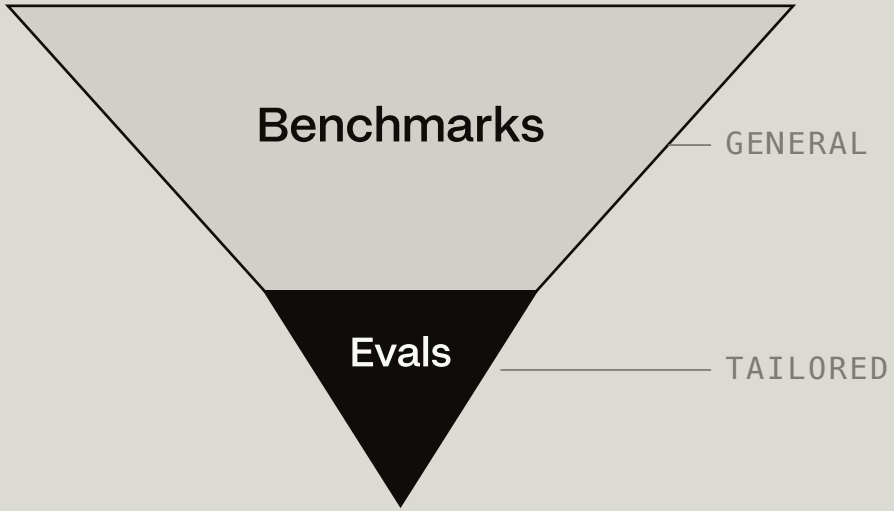
S2S **Can it converse?**
voice-to-voice latency · interruptions

Components diagnose. End-to-end benchmarks decide.

BEST MODEL FOR YOUR AGENT

Start with benchmarks, then use evals.

- Choose a provider
- Set a launch baseline
- Investigate a regression
- Validate a stack change



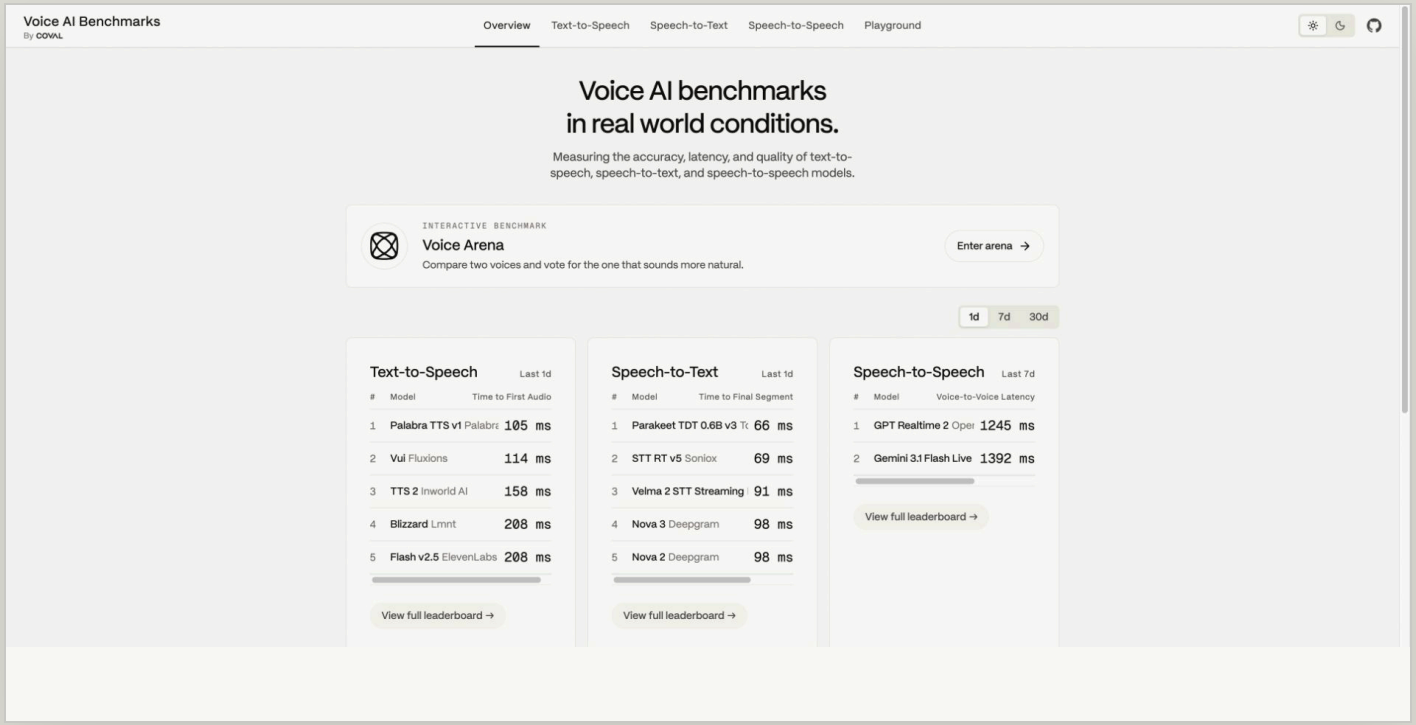
WHAT COVAL BENCHMARKS

See the voice stack in one view.

TEXT-TO-SPEECH
SPEECH-TO-TEXT
SPEECH-TO-SPEECH

Plus: Coval Voice Arena

Source: benchmarks.coval.ai/overview · captured Aug 10, 2026



WHY COVAL BENCHMARKS

Coval makes the comparison inspectable.

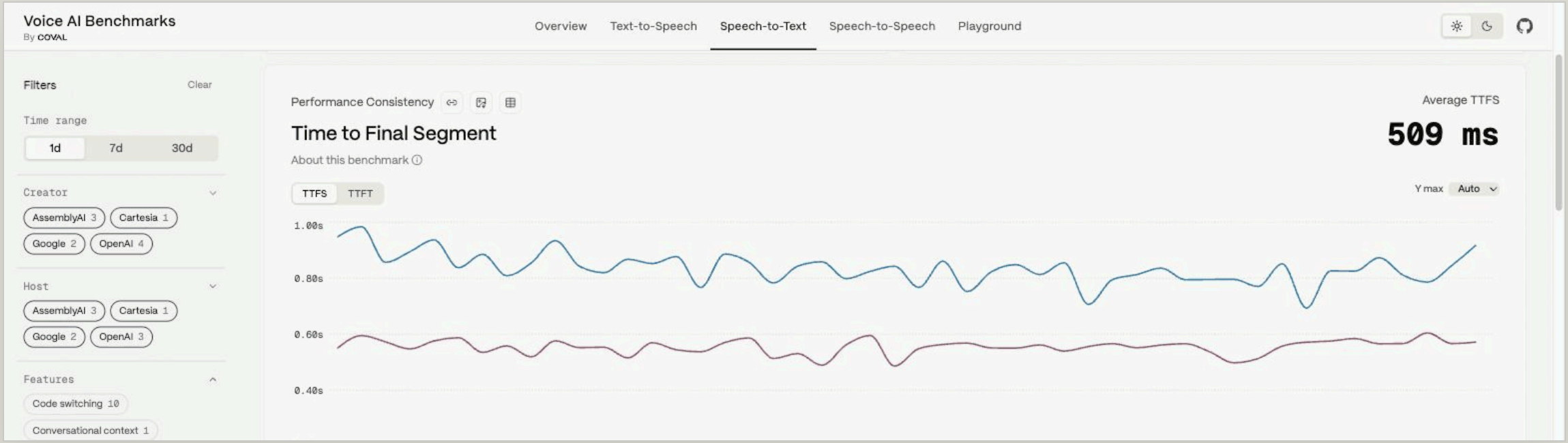
- Continuous provider drift
- Test infra as well as the models
- Open-source methodology

Research benchmark · validate on your workload

SPEECH-TO-TEXT

Use STT benchmarks to validate "hearing."

WER BY CONDITION · TTFS / TTFT · ERROR TYPE MIX



Source: benchmarks.coval.ai/stt · captured Aug 12, 2026

HOW TO READ STT

One WER score hides production failures.

Production PipeCat voice-agent clips

Clean studio-quality baseline

Accents accented speech

Clipping distorted signal

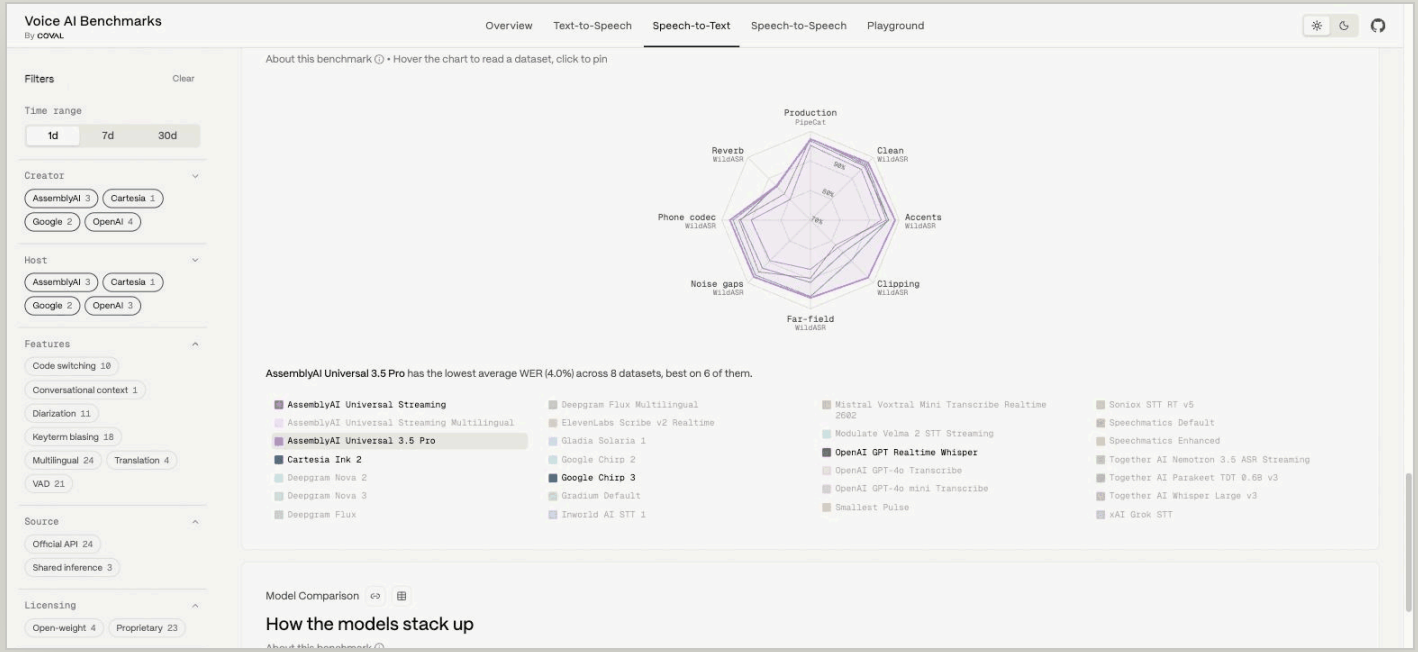
Far-field distant microphone

Noise gaps background noise dropouts

Phone codec 8 kHz telephony

Reverb room echo

Then rerun the shortlist on production audio.

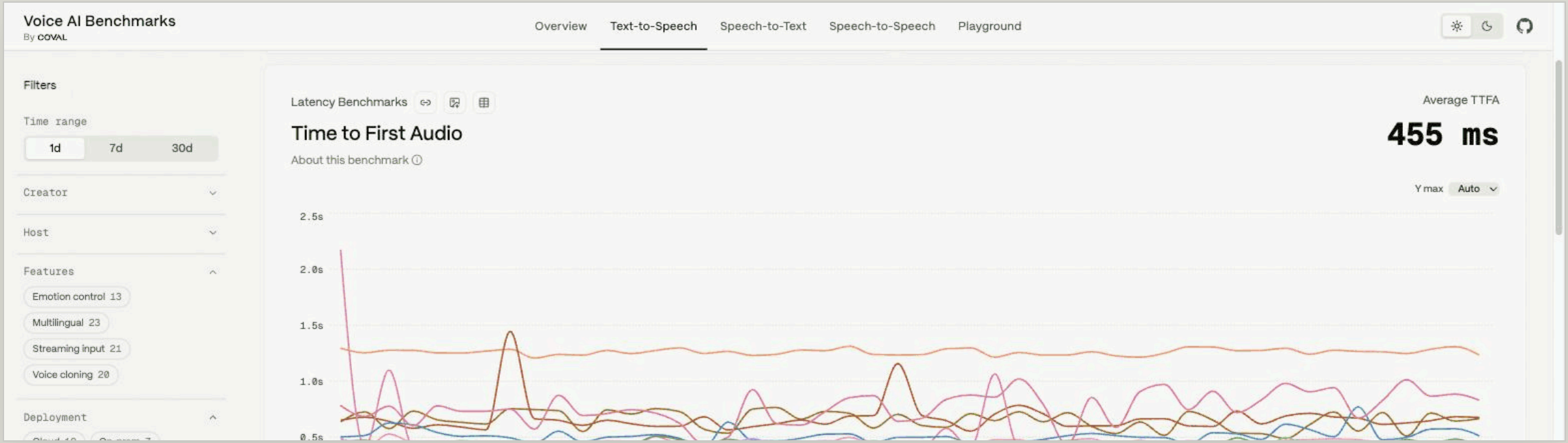


Source: benchmarks.coval.ai/stt · Accuracy Across Datasets · live interaction · Aug 12, 2026

TEXT-TO-SPEECH

Use TTS benchmarks to validate "speaking."

TTFA AND TAILS · OUTPUT INTELLIGIBILITY · DEPLOYMENT FIT



Source: benchmarks.coval.ai/tts · captured Aug 12, 2026

HOW TO READ TTS

Latency is a distribution and a pipeline.

START

Median TTFA

Typical wait before audible output.

TAIL

p95 under load

Rare stalls define the experience.

DELIVERY

Streaming smoothness

Isolate model, network, infrastructure.

Profile the whole path under realistic load.

HOW TO READ S2S

Speech-to-speech changes the unit of evaluation.

The transcript survives. Transcript-only evaluation does not.

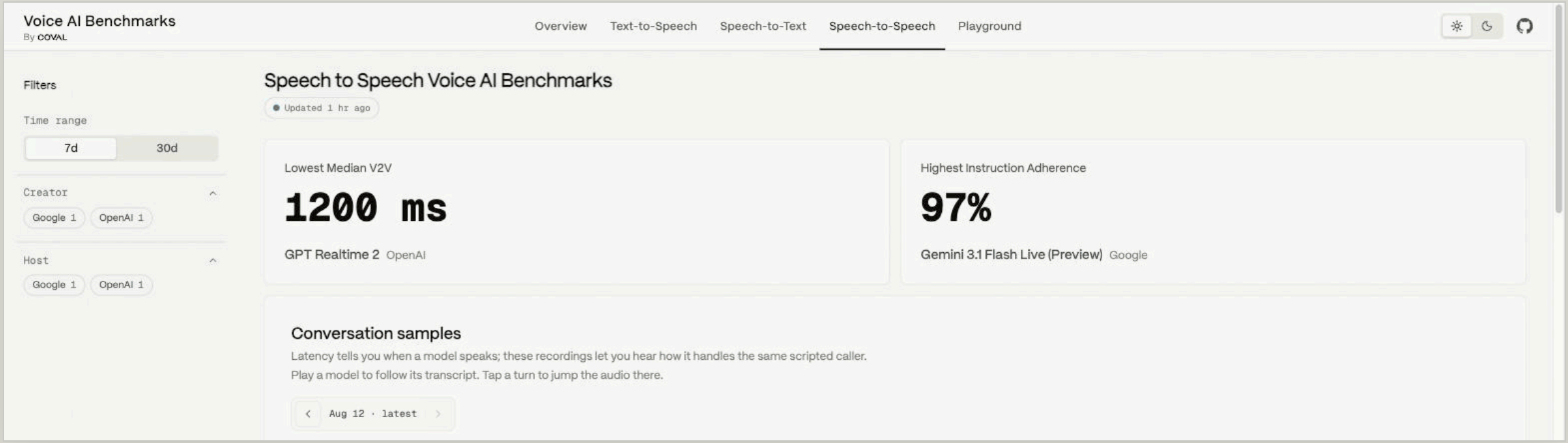
- Instruction adherence vs. drift
- Interruptions and recovery
- Caller realism (background noise, accents, etc.)

`Timing · overlap · recovery · naturalness`

SPEECH-TO-SPEECH

Use S2S benchmarks to validate "conversing."

VOICE-TO-VOICE LATENCY · INTERRUPTIONS · NATURALNESS



Source: benchmarks.coval.ai/s2s · captured Aug 12, 2026

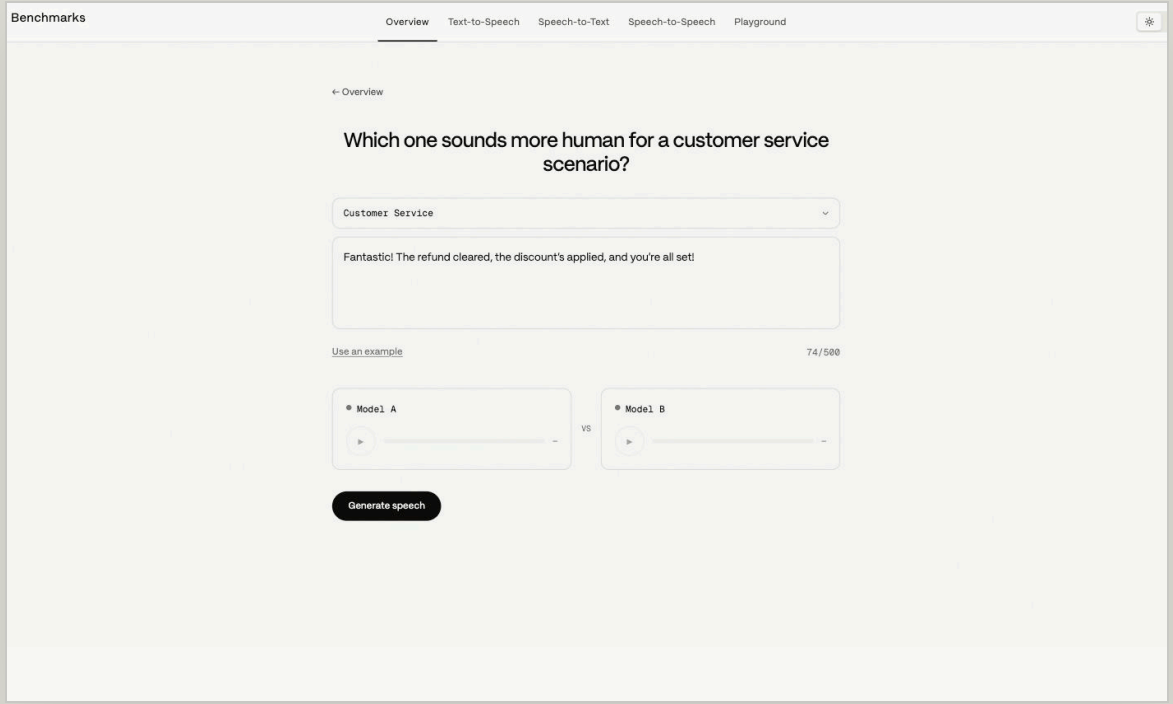
FIRST LOOK: COVAL VOICE ARENA

Naturalness requires humans. Vote in the arena.

- **Two voices, one prompt**
  Same text, side by side, Model A vs Model B.
- **Vote blind**
  Pick the one that sounds more natural. No model names shown.
- **Votes become signal**
  Preference results accumulate into a leaderboard alongside the measured benchmarks.

`Try it: benchmarks.coval.ai · Enter arena from the overview page.`

Source: benchmarks.coval.ai · Voice Arena · captured Aug 12, 2026



BENCHMARK YOUR DATA

Public evidence narrows. Your traffic decides.

- **Shortlist from the public leaderboards**
  Filter STT, TTS, and S2S rankings by the conditions that match your traffic.
- **Replay your test cases against the shortlist**
  Same prompts, personas, and audio conditions, held constant across every candidate.
- **Score with the same metric stack**
  Task outcome, latency distribution, and recovery, then rerun before every release.

Coval is an agent improvement platform that automates this workflow.

FROM YOUR REGISTRATIONS

Benchmarks are too generic. How do I benchmark on my data?

FROM YOUR REGISTRATIONS

S2S, cascade, or hybrid? And how do you evaluate without a transcript?

FROM YOUR REGISTRATIONS

How do we keep ground truth current while we iterate?

FROM YOUR REGISTRATIONS

How should a provider replicate your results?

FROM YOUR REGISTRATIONS

What should we track in production?

THE OPERATOR TAKEAWAY

A benchmark is useful when it changes a decision.

USE PUBLIC EVIDENCE TO NARROW. USE YOUR AGENT TO DECIDE.

benchmarks.coval.ai The live leaderboards

github.com/coval-ai/benchmarks Methodology · reproduce any number

Voice Arena benchmarks.coval.ai · Enter arena

benchmarks@coval.dev Add a model · dispute a result

Recording and links land in your inbox this week.


Brooke Hopkins


Kobi Hudson


Brooke Hartley


Cale Payson


Madina Seribay


Cooper Rubens

NOW, YOUR QUESTIONS.