



COMMUNITY DAY

DMV

LEVEL 200 · INTERMEDIATE

Building Production-Ready Voice AI Agents on AWS

From Prototype to Production October 2nd 2026



COMMUNITY DAY

A REAL TRANSCRIPT FROM THIS RIG

The agent rebooked a passenger,
confirmed it, and said goodbye.

There was nobody on the line.

It had been talking to **itself** for four turns — because it could hear its own voice through the speaker, and nothing in the pipeline could tell the difference.





COMMUNITY DAY

Who is telling you this



**Wanjohi
Christopher**

Data & AI Engineer
AWS Community Builder

“These lessons came from
an agent
that took real phone calls.”

I rebuilt it on AWS for this talk.





COMMUNITY DAY

What you will leave with

01

Where the latency actually is

A measured breakdown of every hop

02

What each production fix costs you

Endpointing, gating, echo suppression

03

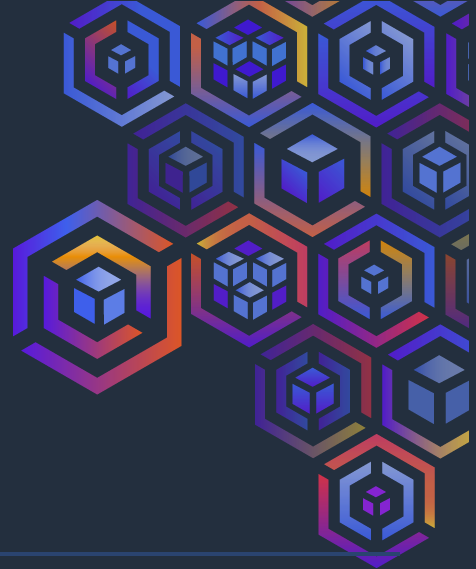
How to instrument what you cannot see

CloudWatch EMF from a Fargate task

04

When to stop using a cascade

Cascade vs speech-to-speech in 2026





COMMUNITY DAY

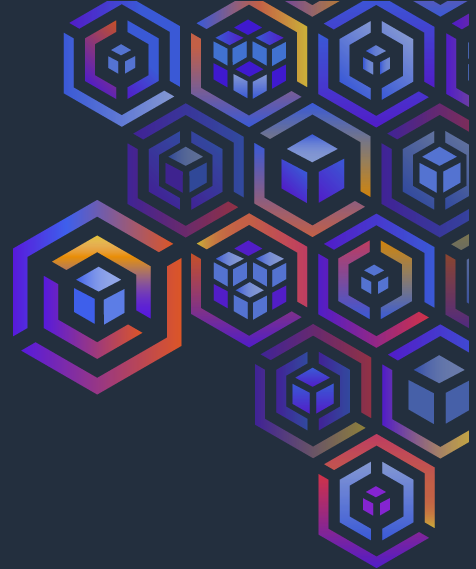
AIRLINE REBOOKING

The scenario

ONE CALLER · ONE RECORDING · AIRPORT AMBIENCE

- › “My flight to Baltimore just got cancelled.
The six-forty.”
- › “Confirmation K as in kilo, seven, R, two, nine, X.”
- › “And I had a checked bag on the original.”

Four slots. One of them is hard.

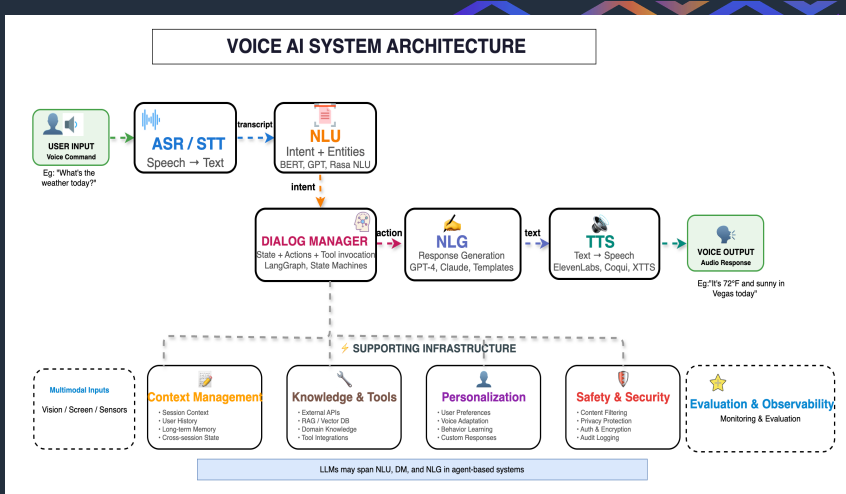




COMMUNITY DAY

THE WORD THIS WHOLE TALK TURNS ON

Two ways to build a voice agent



CASCADE — three models in a relay. This talk lives here.



Transcribe

SEAM

"text"

words

Bedrock

SEAM

"text"

words

Polly



The representation changes at every boundary. Each one is a place you can **measure** — and a place you **pay**.

SPEECH-TO-SPEECH — one model. Audio in, audio out. No text, ever.



Nova 2 Sonic



Nothing to hand off, so nothing to instrument — and **no seam to stand in**.



COMMUNITY DAY

EVERY NUMBER MEASURED, NOT ESTIMATED

One WebSocket. One Fargate task.

CONTROL PLANE · HTTP · BORING

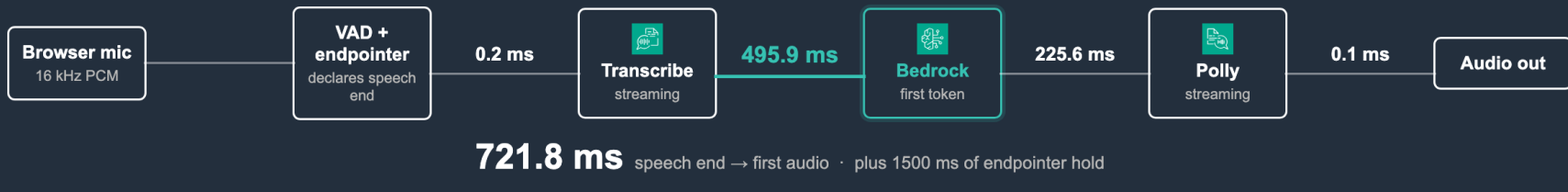
session start

session end

cleanup

people expect the complexity to live here — it does not

DATA PLANE · ONE WEBSOCKET · ONE FARGATE TASK



Lambda

1804 ms · off the critical path



DynamoDB

one item, per turn



S3

audio · trace, after the turn

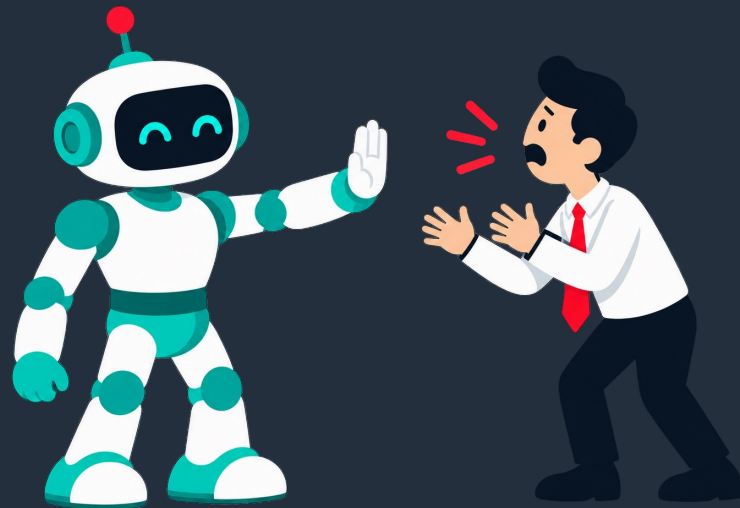


CloudWatch EMF

every number on the panel

Watch it fail

Three runs. One commit. Four booleans.





COMMUNITY DAY

RUN A

naive

700 ms endpoint · no echo guard · no gate

Same audio.
Same commit.

Run A — naive

cuts off · guesses · answers
itself

Run B — shipped

fixed · but cannot be interrupted

Run D — 2026

correct · 736 ms faster



COMMUNITY DAY

THE CONFIDENCE GATE

Ask, or guess?

WHAT TRANSCRIBE RETURNED ON ANOTHER TAKE

“...confirmation is K 7 or 29 X.”

0.33

GATE OFF

Writes K7OR29X.

No flag. No warning.

The wrong passenger flies.

GATE ON · 0.75

Refuses to write it.

Records it as flagged.

“K-7-O-R-2-9-X, correct?”





COMMUNITY DAY

THE ECHO GUARD

The fix that shipped a walkie-talkie

THE CALLER SAID FIVE THINGS. RUN B HEARD TWO.

- 1 "...my flight to Baltimore just got cancelled."
- 2 "I need to get there tonight, or tomorrow morning."
- 3 "No, wait — tonight, if there's anything at all."
- 4 "Confirmation K as in kilo, seven, R, two, nine, X."
- 5 "And I had a checked bag on the original."

The guard stopped the agent answering itself.

It also stopped the caller being heard for 7.3 s a turn.

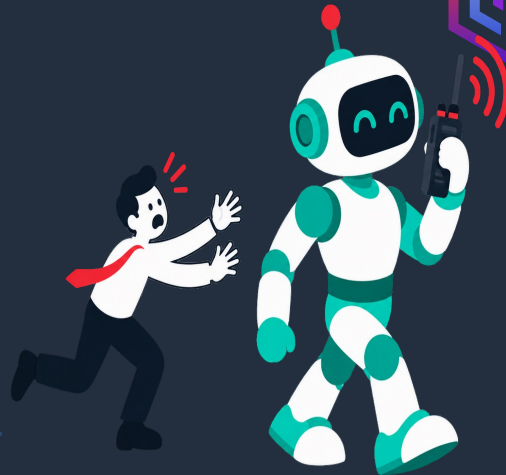
heard

DEAF

DEAF

heard

DEAF





COMMUNITY DAY

EVERY NUMBER HAS A CLOUDWATCH METRIC NAME

If you cannot see it, you cannot defend it

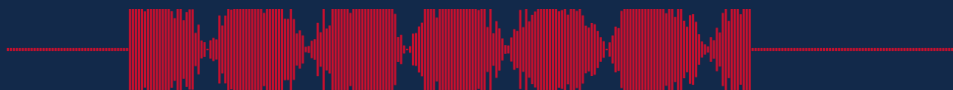


AWS CASCADE VOICE RIG

run B

Start call

SPEAKING DEAF - INBOUND DROPPED



TRANSCRIPT

```
caller My flight to Baltimore tonight got canceled. The confirmation is K 7 or 29 X. [0.36]
>
GATE_BLOCKED confirmation=K7OR29X at 0.36
DEAF inbound audio dropped
TOOL_CALL flight_availability(destination: Baltimore, when: tonight)
agent > Let me check what's open tonight.
agent > Please confirm your booking reference is K-7-0-R-2-9-X.
agent > I have found a nonstop flight to Baltimore tonight at 21:50 with 4 seats available.
```

SESSION ITEM - DYNAMODB

```
{
  "session_id": "fe063bd6",
  "turn": 1,
  "flags": {
    "run": "B",
    "echo_guard": true,
    "endpoint": "fixed 1500ms",
    "confidence_gate": 0.75,
    "barge_in": false
  },
  "slots": {
    "destination": {
      "value": "BWI",
      "confidence": 1
    },
    "flight": {
      "value": 640,
      "confidence": 0.99
    }
  }
},
```

TIMING - CLOUDWATCH EMF

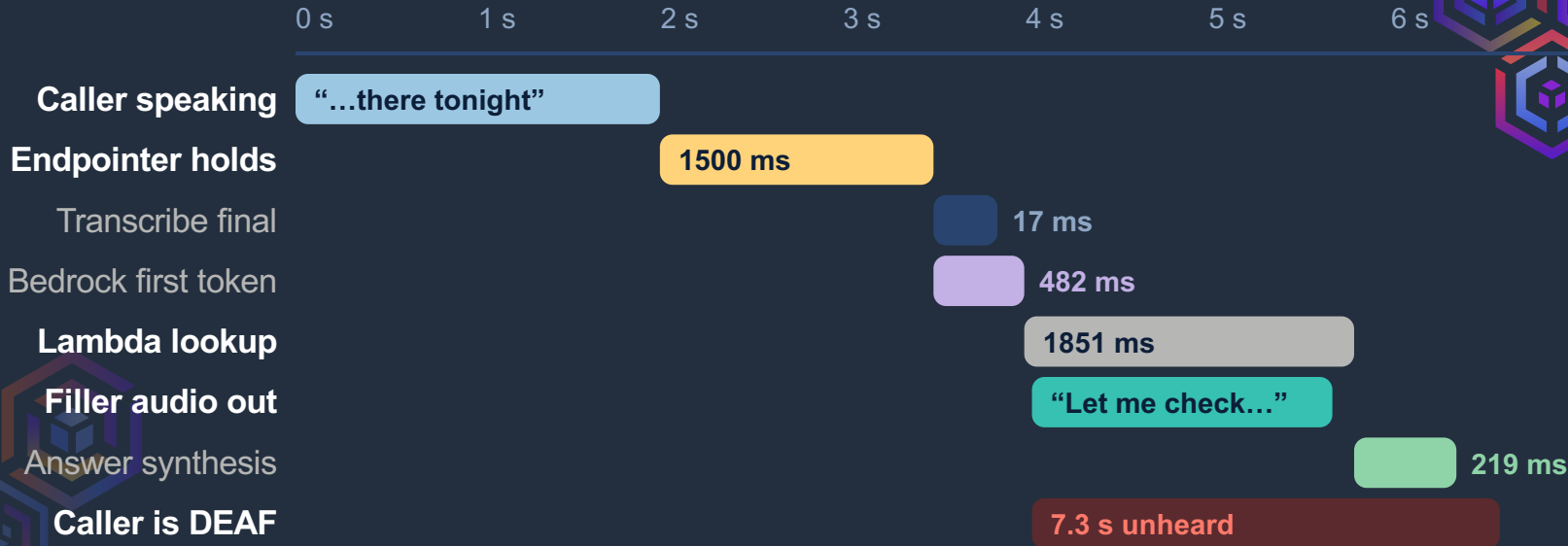
speech end - STT	1ms
STT - first token	464ms
token - Polly	177ms
Polly - audio out	40ms
tool duration	2359ms
filler lead	2258ms
TOTAL	828ms



COMMUNITY DAY

WHERE THE SECONDS ACTUALLY GO

One turn, in time



The caller hears something at 1088 ms. The lookup finishes 1.6 s later.



COMMUNITY DAY

Every fix costs something

Endpointer 700 → 1500 ms

stopped cutting the caller off

+800 ms

every utterance

Confidence gate

stopped a wrong booking reference

+1 turn

when it fires

Echo guard

stopped the agent answering itself

7250 ms

cannot interrupt

Filler at tool decision

caller hears something in 1088 ms

-1620 ms

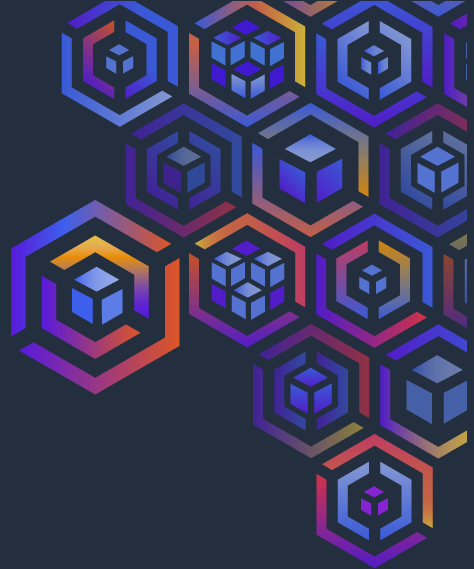
a fix that pays

Semantic endpointing

correct and fast

-736 ms

vs the threshold





COMMUNITY DAY

CASCADE VS SPEECH-TO-SPEECH

Seams, or speed

CASCADE · ~90% STILL SHIP THIS



Every seam is a place to measure —
and a place to pay.

NOVA 2 SONIC

one bidirectional stream

Faster.

The 2026 answer is hybrid.

Speech-to-speech where latency dominates. Cascade where you need tools.



COMMUNITY DAY

Key takeaways

01

Every fix costs something

The gate adds a turn. The guard costs the ability to interrupt.

02

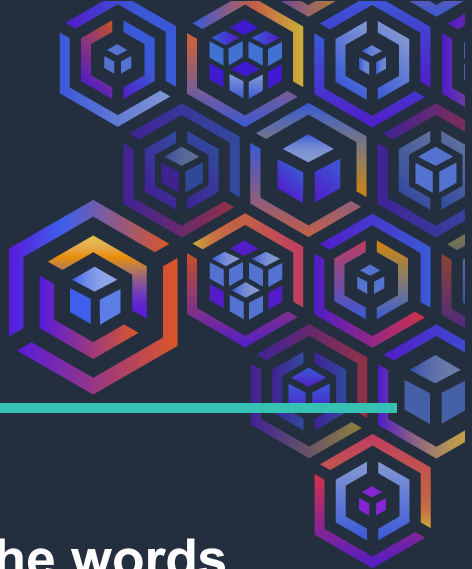
Measure the seam, not the guess

Everyone blames text-to-speech. 59% was the model.

03

Read the words, not the silence

A pause mid-sentence and a pause at the end are identical.





COMMUNITY DAY

Resources

TAKE THIS WITH YOU

- The rig, end to end
Transcribe → Bedrock → Polly on Fargate
- The measurement harness
Scripts that reproduce every number
- Notes on what broke
Including the fix that only looked fixed

WORTH YOUR TIME NEXT

- Amazon Nova 2 Sonic
Bedrock bidirectional streaming
- Transcribe item confidence
What the clarification beat needs
- CloudWatch EMF
Metrics from stdout, no PutMetricData



Build and Host Your Voice Agent on
AWS | AWS Show and Tell

<https://youtu.be/gPHfuGNDIDY>

Thank you

Questions.