STRIDE AI VOICE ENGINE

Technical Benchmark Report

Our first internal benchmark: a modular pipeline compared with Moshi, an end-to-end open-source speech model.

DATE October 2026 | PREPARED FOR Prospective design partners | WEBSITE heystride.co

Executive Summary

Stride AI is building a low-latency voice engine for B2B software vendors in regulated industries (legal, healthcare, corporate operations). This report shares our first internal benchmark, comparing our modular pipeline against Moshi, a well-known end-to-end open-source speech model.

WHAT WE ARE BUILDING

A modular voice engine, delivered through an API, that a software vendor can embed in its own product. Speech recognition, the language model and speech synthesis are separate components, so a customer can choose its models, run the engine in its own environment, and connect the agent to internal systems through tool calls. The focus is on the moments that make voice agents feel unnatural: awkward pauses, slow reactions to interruptions, and audio glitches.

0.20 s
TIME TO STOP
When a caller interrupts (Moshi: 1.21 s)

23 / 23
STOPPED < 1 SEC
Mid-sentence interrupts (Moshi: 6 of 18)

4.16
SPEECH QUALITY
UTMOS scale 1–5 (Moshi: 3.59)

0 / 60
AUDIO GLITCHES
Runaway audio tails in QA sessions

Graceful turn-taking. Stride prioritizes graceful conversational turn-taking. When the caller starts speaking mid-sentence, Stride detects the speech in 128 ms and drops its audio below an audible level (-40 dB) in 0.20 s. In 60 live question-and-answer sessions under the production voice profile, we recorded 0 runaway audio tails; the untuned first setup had 3 in 60 runs.

Response start trade-off. The main trade-off is answer start time. Our median time to start answering is 0.94 s across 60 trials, compared with Moshi's 0.14 s. This comes from a deliberate 0.40 s silence buffer that confirms the caller has finished.

Caveat: All figures come from a single-GPU test environment with a small sample. Network delay is not included. See the limitations on page 5.

Comparative Evaluation

Below is the side-by-side performance comparison between Moshi (end-to-end) and Stride AI (modular). Interruption sessions: Stride 24 (8 scripted scenarios × 3 repetitions; agent mid-sentence in 23, counted for stop time) and Moshi 18. Single-GPU test environment, small sample.

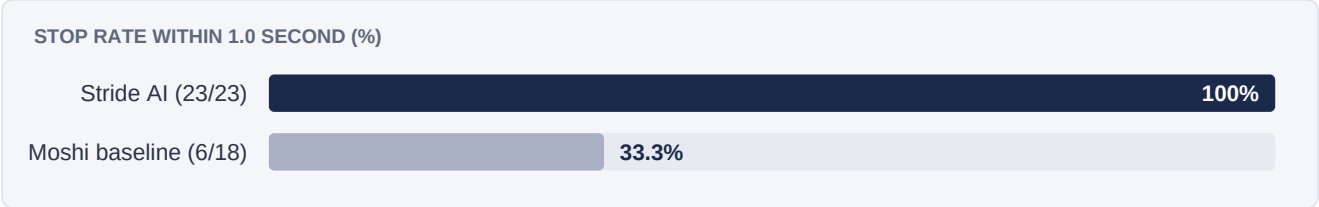
Performance metric	Moshi (end-to-end)	Stride AI (modular)
Wait after caller stops speaking	~0.0 s	0.40 s (0.10–0.20 s in aggressive mode)
Time to start answering, median (60 trials)	0.14 s	0.94 s (95% CI 0.93–0.95)
Slowest start observed (60 trials)	5.6 s	1.3 s
Full answer ready, median / p90	2.5 s / 6.3 s	2.1 s / 3.1 s
Time to stop when interrupted	**1.21 s**	**0.20 s**
Interruptions stopped within 1 s (agent mid-sentence)	**33% (6 of 18)**	**100% (23 of 23)**
Overlapping speech during an interruption	1.20 s	0.20 s
Speech quality, 60 live Q&A turns (UTMOS, 1–5)	3.59	4.16
Speech quality, formal voice profile, 11-sentence test	Not run	4.31 (worst case 3.77)
Audio glitches (runaway audio tails), 60 QA sessions	Not measured	0 of 60

Summary of Benchmark Findings

- **Interruptions are handled reliably.** When a caller cuts in, our agent falls silent in 0.20 s (measured on the caller's clock, until our audio drops below an audible level) and did so within 1 second in 23 of 23 test interruptions. Moshi took 1.21 s on average and met the 1-second bar in 6 of 18.
- **Speech quality is higher.** On the same 60 live question-and-answer turns, predicted speech quality (UTMOS) was 4.16 for our engine vs 3.59 for Moshi. In a separate 11-sentence test, our formal voice profile reached 4.31 (worst case 3.77).
- **Response times are consistent.** Across 60 test questions, our slowest start was 1.3 s, vs 5.6 s for Moshi. Full answers were ready in a median of 2.1 s (p90: 3.1 s).
- **Trade-off.** Moshi begins answering sooner on a typical turn (0.14 s vs our 0.94 s median, or about 0.70 s in our aggressive mode). We chose a deliberate 0.40 s pause before replying so the agent does not talk over a caller who is still finishing a sentence. This is configurable.

Barge-In & Interruption Handling

Of 23 mid-sentence caller barge-ins, Stride stopped speech below 1.0 s in all 23 (100%), with a 0.20 s time to stop and 0.20 s overlap. Moshi stopped below 1.0 s in 6 of 18 (33%), with a 1.21 s time to stop and 1.20 s overlap.



HOW THE 0.20 s IS MEASURED

Measured on the caller's own clock, from the moment the interrupting speech starts until our audio falls below an audible level (-40 dB). Detecting sustained speech takes 128 ms and the stop signal reaches the player in about 15 ms. The 80 ms fade-out, which prevents clicks, is already below the audible level about 55 ms in. Total: about 0.20 s.

REVISABLE TURNS & RESPONSE AFTER INTERRUPTION

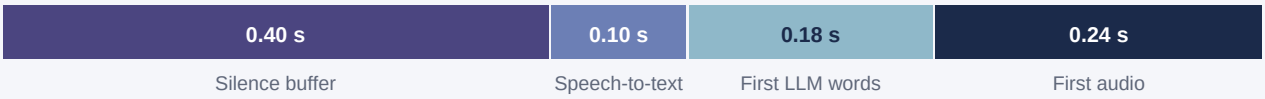
If the caller pauses mid-sentence and resumes within 0.6 s, the system merges both parts and answers once, instead of answering a half-finished sentence.

After the interruption, the agent answered the caller's new question with a median start time of 0.93 s (24 sessions).

Response Latency: A Deliberate 0.40 s Pause

Time to start answering (TTFA) is the time between the caller finishing a sentence and the agent's first audio. Our median is 0.94 s (p90 1.03 s) over 60 trials (20 factual questions, 3 repetitions each), with a 95% confidence interval of 0.93–0.95 s from 1,000 bootstrap resamples. Moshi's median is 0.14 s.
The 0.40 s is a deliberate silence buffer that confirms the caller has really finished.

WHERE THE 0.94 s GOES



Aggressive mode cuts the buffer to 0.10–0.20 s, bringing TTFA to about 0.70 s at the cost of occasionally replying too early. The right setting depends on the use case: a fast intake line or a careful legal consultation. *Stage times are approximate.*

Acoustic Evaluation: Speech Quality

UTMOS is an automated model that predicts how natural speech sounds on a 1–5 scale. Compare live results with live results; the controlled voice-profile score is a separate measurement.

LIVE Q&A BENCHMARK

4.16 vs 3.59

Calculated across 60 live conversational turns with identical queries. Moshi has no separate voice interface, so it could only be scored on live conversations.

FORMAL VOICE PROFILE

4.31 UTMOS

3.77 worst case. Assessed in a separate 11-sentence test (22 samples) of our formal legal voice profile. Controlled score, not a direct live comparison.

TUNING MITIGATION

3 → 0 glitches

Untuned setup scored 3.69 (worst 2.03) with 3 runaway audio glitches in 60 runs. Production profile fixed both (0 glitches in 60 runs).

Automated scoring methodology & listener validity. UTMOS tends to favor flat, clean delivery, which is what the formal profile is tuned for. Read it as a measure of stability rather than warmth. UTMOS is not a substitute for listening tests with human raters, which we plan to run. Quality in this benchmark is automated, not human-rated.

Experimental Setup & Architecture

Environment & timing. All runs used one machine: 1 NVIDIA A100 80 GB GPU (RunPod Secure Cloud). Connection is local WebSocket streaming on the same machine, so network delay is not included. All timings were taken on the caller's clock, not from server logs.

Sample sizes

- **Q&A:** 60 turns (20 questions × 3 repetitions) for response times and live speech quality; same 60 sessions for glitch check.
- **Interruption:** 24 Stride sessions (8 scripted scenarios × 3 repetitions; agent mid-sentence in 23, counted for stop time) and 18 Moshi sessions.
- **Voice profile:** 22 samples (11 sentences) per setting.

Component	Configuration
Speech recognition	faster-whisper large-v3-turbo (GPU, float16)
Language model	IBM Granite 4.2 8B served with vLLM (OpenAI-compatible server; bf16, max 256 tokens, temp 0.3). Tool calls (<|tool_call|>) are parsed on the client side.
Voice activity detection	Silero VAD (threshold 0.5, 128 ms start trigger, 400 ms end-of-speech silence)
Speech synthesis	Qwen3-TTS 12Hz 1.7B CustomVoice, speaker "Ryan", CUDA graphs (~31 ms per frame)
Voice streaming	Clause-by-clause: first 3 frames, then 4 frames per emit (12.5 Hz)
Voice controls	Repetition penalty 1.0 (removes held-vowel stutter); formal, objective legal prompt with no emotional inflection
Precision	bfloat16 for the language and voice models; float16 for speech recognition

Engineering Rigor: Boundaries & Priorities

We would rather be upfront about the boundaries of this benchmark.

LIMITATIONS

- **Network delay is not included.** Everything ran on one machine. In production, add round-trip time between caller, telephony provider, and server.
- **Small samples.** 18–60 runs per test (often the same scripted scenarios repeated three times) is enough to show a clear difference, not to claim production-grade statistics.
- **Scripted test conversations.** Accuracy on real call recordings, accented speech, or noisy audio not yet measured.
- **Quality is automated, not human-rated.** See the UTMOS note.
- **Tool calling not benchmarked here.** Supported, but latency with database lookups or retrieval in the loop is the next test.
- **Not yet certified.** No compliance certification (e.g., HIPAA, SOC 2). Designed for self-hosted or private-cloud deployment so customer data can stay in the customer's environment.

NEXT VALIDATION PRIORITIES

- A larger interruption test set (50–100 trials planned)
- Latency with database lookups or retrieval (RAG) in the loop
- Accuracy on real call recordings, accented speech, and noisy audio
- Human listening tests with raters, alongside automated UTMOS scores
- Compliance requirements, worked through together with design partners

Design Partner Program

We are opening a small design-partner cohort for teams building voice features in legal, healthcare, or operations software. Partners get early access, direct input on the roadmap, and a say in which benchmarks we run next.

HOW TO GET STARTED

1. Review the results and tell us which metrics matter most for your use case.
2. Send us your thoughts in writing or, if you prefer, join a short call.
3. If there is mutual interest, we can sign a short non-binding letter of intent (carries no financial commitment or legal obligation).

CONTACT

Aysima Çelik
Founder

Om Chimurkar
Research Advisor

hello@heystride.co
heystride.co