What Voice AI Needs: Trust or Fluency?

What determines whether your voice channel will be used, more than once

Authors: Panos Konstantinidis, PhD, and Jenny Tantidou, MBA, co-founders of Evenly

Executive Summary

Speech fluency makes a system feel responsive, but reliability determines whether it can be used for meaningful conversations, repeatedly. Trust is earned when the system does not falsify data from systems of record, and when the organisation can improve behaviour over time. This is true for any voice AI that aims to earn customer trust, not only for regulated deployments. A loss of trust in the voice channel can erode the customer's trust in the entire organisation. Customers care whether the system misled them; auditors care whether the failure can be evidenced after the fact. The case becomes unavoidable wherever speech affects money, health, identity, rights, access, safety, or institutional accountability.

Recently, Google released Gemini 3.1 Flash Live and OpenAI released GPT-Realtime-2, GPT-Realtime-Translate, and GPT-Realtime-Whisper. Both vendors describe a fundamental move away from the multi-stage pipeline that has defined conversational AI for a decade. Single-model architectures generate spoken output directly from spoken input, with an impressive conversational responsiveness and speech fluency. They also create a moment of confusion in regulated markets, where buyers reasonably ask whether the orchestrated pipelines that power their banking and accessibility services are now legacy infrastructure. If a bank can deploy GPT-Realtime-2 or Gemini 3.1 Flash Live in its call centre with the responsiveness that either vendor has demonstrated recently, why commission an orchestrated pipeline at all?

This paper makes four arguments. First, the choice between architectures is a question of category, not of vintage. Pipeline systems are modular by construction; native speech-to-speech systems are monolithic. Second, monolithic architectures are unfit for any workload where the spoken output must include a value coming from a system of record, or where regulatory accountability attaches to the spoken content. A monolithic speech-to-speech system cannot provide the control required for accountable speech, because accountable speech requires an enforceable pre-speech representation, and the monolithic architecture has none by construction. This is an inherent architectural property; the next model release will not solve it. Third, the modular architecture is the only configuration that has the flexibility to mix and match modules to consistently deliver state-of-the-art (SoTA) performance at every layer of the stack. No single vendor can lead the frontier in speech recognition, reasoning, and speech synthesis simultaneously and forever. Fourth, in DORA, the EU AI Act, and the European Accessibility Act, the regulatory frameworks now in force make modularity and AI control materially easier to evidence as a compliance posture.

A fifth argument overarches the first four. Trust in customer-facing voice channels is binary in operational reality and contagious in social propagation. A customer who has any reason to suspect the system might quote a wrong balance, an invented dosage, or an incorrect deadline does not tolerate the possibility of inaccuracy; they abandon the channel and tell others. One sufficiently shared incident can render the channel commercially unusable for the institution that experienced it and for peer institutions in the same sector. Consumer protection authorities receive complaints they cannot dismiss, data protection authorities apply GDPR accuracy obligations the monolithic

architecture cannot meet, and sectoral inquiries follow. The architectural decision is therefore beyond discussions about error rates or enforcement of regulatory requirements. It is a question every organisation striving for consumer loyalty or public trust must answer, before the channel is deployed, rather than after the first incident.

Native speech-to-speech is best suited to use cases that do not require authoritative reproduction of system-of-record data. For any organisation deploying voice AI that propagates information from systems of record, or is deployed in regulated sectors, the architectural decision is a one-way street. In such cases, any voice AI must be judged by reliability and trust, not by speech fluency.

1. The architectural fork: modular versus monolithic

The modular pipeline architecture decomposes spoken interaction into discrete components: speech recognition, user intent understanding, function calls to systems of record, response creation, and speech synthesis. Each component runs as a separate service, with auditable text passing between them. Any component can be replaced with a stronger alternative at any time, with no change to the application layer above. The pipeline is, in regulatory language, substitutable at every layer.

The monolithic native speech-to-speech architecture compresses the entire interaction into a single multimodal model that produces audio tokens (output) directly from audio tokens (input). There is no canonical text representation between input and output, no inverse text normalisation step that converts spoken digit sequences into canonical numbers, and no deterministic pronunciation layer. The category includes OpenAI's GPT-Realtime-2, Google's Gemini 3.1 Flash Live, and Kyutai's Moshi. The single defining property that separates the two families is the presence or absence of an inspectable text intermediary. Everything else in this paper follows from that distinction.

Both architectures can call a tool to retrieve a value from a system of record. They differ in what happens to that value next. In the modular pipeline, the retrieved value is not regenerated by the model after retrieval. It is bound to a canonical pre-speech representation, validated, and rendered through a constrained, testable speech layer. By construction, the source-of-record value is protected from model-level falsification before it reaches the customer’s ear. In the monolithic model, the retrieved value enters a generative audio-token process and is statistically reconstructed into speech rather than reproduced from an enforceable intermediate representation. Whether the user hears “fifty,” “fifteen,” or something else is determined by the model’s generation at the moment of speaking, with no architectural mechanism to enforce the original value. A monolithic vendor can in principle achieve auditability through external transcription and comparison after the conversation, but auditability is not reliability. Auditability records what the system said after the customer has heard it. Reliability prevents the customer from hearing it in the first place.

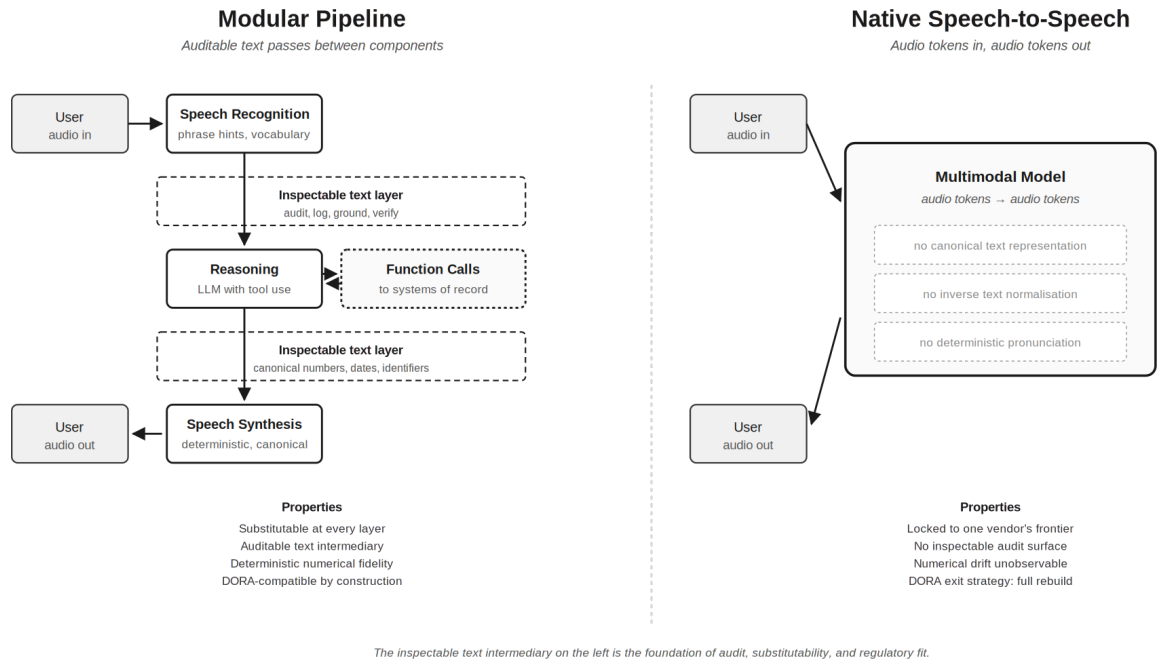

Figure 1. The two architectural families. The inspectable text intermediary of the modular pipeline is the foundation of audit, substitutability, and regulatory fit. The monolithic model has no equivalent control surface.

2. The accuracy problem is structural

The most common objection to this paper's argument is that monolithic systems will gradually get more accurate over time. This objection misunderstands the source of the gap. Researchers from OpenAI itself, in a paper published in September 2025, argued that hallucinations are statistical errors arising from training and evaluation incentives that reward confident guessing over admitting uncertainty, and that they persist even in state-of-the-art systems [1]. The empirical record is consistent across vendors. Koenecke and colleagues studied OpenAI's Whisper in 2024 and found that 38 percent of identified hallucinations contained explicit harms, including fabricated medical content, false associations, and invented authority [2]. The Associated Press subsequently documented that Whisper had been deployed across more than 30,000 clinicians and 40 health systems, with one vendor erasing source audio after transcription, leaving hallucinations undetectable after the fact [3]. Google's own engineers have documented the same architectural failure in their issue tracker. Bug report #1894 on the official googleapis/python-genai repository records that Gemini 2.5 Flash Native Audio, the predecessor to Gemini 3.1 Flash Live, falsifies values retrieved from system-of-record tool calls: the model generates confident factual content from its training data before the tool result returns, asserting values that differ from what the system holds. The reported examples include the model saying “$2,300” when the actual tool result was “$4,420–$4,530,” and “Biden” when the actual answer was “Trump.” Every mitigation the architecture offers (system instructions, function descriptions, interrupt scheduling) failed to prevent the behaviour [11]. The architectural class produces these failures across vendors, regardless of vintage or marketing claims.

For the specific failure mode that matters when a voice agent draws values from a system of record, the SHALLOW benchmark is the most direct evidence. It documents that decoder-based speech models, the architectural family that includes native speech-to-speech, systematically prioritise linguistic fluency over acoustic fidelity, producing phonetically plausible substitutions that pass

conventional word-error-rate thresholds while distorting meaning [4]. Phonetic redundancy is low: "fifteen" and "fifty" differ by a single fricative. Without an inverse text normalisation step, a digit drifted in the audio-token sequence is dropped silently with no text-layer check. OpenAI itself acknowledges the limitation: the Whisper model card states that the company recommends against use in "high-risk domains like decision-making contexts, where flaws in accuracy can lead to pronounced flaws in outcomes" [5]. A subsequent release may improve fluency, latency, and language coverage. It cannot, by construction, add an inspectable text layer where none exists.

The accuracy gap is architectural. The next model release will not solve it.

---

3. Why modular delivers state-of-the-art over time

The most persuasive case for modular is not defensive. It is the simple observation that no single vendor leads the frontier across every layer of the voice stack at the same time and forever, and none ever will. Speech recognition has frontier vendors that include Whisper, Deepgram, AssemblyAI, and others. Reasoning has GPT, Claude, Gemini, and increasingly capable open-weight models. Speech synthesis has ElevenLabs, Cartesia, Azure, Google, and others. The probability that any one speech-to-speech provider sits on the frontier across all three layers simultaneously, and for a sustained period, is in practice zero.

A modular system embeds the strongest available component on every layer, on a rolling basis. When a speech-to-text (STT) module with higher transcription accuracy appears, the pipeline switches that single component. When a text-to-speech (TTS) module with more natural prosody appears, the pipeline switches the synthesis layer. The same applies to the LLM layer when a stronger reasoner is released. Each change is local. A monolithic deployment cannot do this; it delivers one vendor's best across all layers, including the layers where their best is mediocre, with no remediation possible until the vendor ships an updated end-to-end model. State-of-the-art compounds across layers: a modular pipeline running near frontier on each of three layers delivers a near-frontier experience overall, while a monolithic system at frontier on one layer and meaningfully behind on another is dragged down by its weakest link.

4. Why modular is required for any organisation accountable to consumer rights

The European regulatory framework now in force makes modularity materially easier to evidence as a compliance posture. The Digital Operational Resilience Act (DORA), Regulation (EU) 2022/2554, in force since 17 January 2025, requires financial entities to manage ICT concentration risk, maintain exit strategies under Article 28(8) that allow migration without disrupting service, and grant audit rights over providers in the data path of customer-facing services. The EU AI Act, Regulation (EU) 2024/1689, layers transparency, oversight, and accuracy obligations under Articles 13 to 15. The European Accessibility Act, Directive (EU) 2019/882, strengthens the case for verifiable voice outputs in covered accessibility contexts, especially where the user cannot cross-check what the system tells them. Where a voice agent communicates personal data, including balances, appointments, identifiers, account status, eligibility, or health information, GDPR Article 5(1)(d) turns accuracy into an evidentiary obligation, not merely a product-quality preference, extending the architectural argument to any organisation accountable to consumer rights.

A native speech-to-speech architecture sits awkwardly inside this framework. The architecture is, by construction, concentration on a single vendor's audio-token model. For accountable-speech workloads, there is no like-for-like exit strategy short of reimplementation. A modular pipeline maps cleanly onto every requirement: components are substitutable at every layer, exit strategies exist by default, and the text intermediary provides an audit surface that satisfies traceability obligations on the modules where they apply.

A worked example. A retail banking customer asks for the current balance on their savings account. The native speech-to-speech system responds quickly and confidently, drawing the answer through audio tokens with no inverse text normalisation step, and says “fifty euros” when the system of record holds fifteen. Ten minutes later, a direct debit is rejected for insufficient funds. The customer files a complaint with the Bank of Greece, the country's banking regulator, and a separate complaint with the Hellenic Data Protection Authority on the basis that the bank failed to provide accurate personal data under GDPR Article 5(1)(d). The bank's regulatory response requires evidence of what the system said. The audio-token model has no such evidence. The modular pipeline has the text intermediary, the function call to the system of record holding the authoritative balance, and an auditable synthesis log linking text to audio. Multiply by call volume and the cost of the architectural choice becomes a P&L item, not a footnote. Worse still, because the system is a black box, the only path to remediation is full replacement.

There is no exit strategy short of full reimplementation.

---

5. Where each architecture fits

If a workload requires equality with a system of record, regulatory auditability, or low recovery cost on error, it requires the modular pipeline. If none of those three properties bind, the monolithic architecture is appropriate and excellent for the workloads it is designed for. Most production deployments at scale combine both, with native speech-to-speech components used inside an orchestrated pipeline for the parts of the workload where their properties fit.

Native speech-to-speech appropriate for:	Modular pipelines required for:
Consumer voice assistants for entertainment and casual conversation without deterministic data or systems of record Language learning and pronunciation practice Companionship and creative-writing applications Internal productivity tools, where the user reviews output for correctness before acting Low-stakes triage that routes the user to a verified channel for the actual transaction	Banking and financial services: balances, IBANs, transfer amounts, instalment dates, interest rates Healthcare: medication names and dosages, appointment times, clinical instructions, patient identifiers Public administration and tax: identifiers, deadlines, amounts owed, entitlement values Telecommunications and utilities: billing amounts, plan terms, outage information, service identifiers Accessibility outputs for users who cannot cross-check what the system tells them

6. A decision framework

A buyer can classify any voice workload in three questions. A use case that passes all three toward native speech-to-speech can use it. A use case that fails any one of the below requires the modular pipeline.

1. Source of truth. Does the spoken output need to include a value held in a system of record? If yes, the modular pipeline is required.

2. Auditability. Will the system be subject to regulatory, clinical, or legal audit on what it said? If yes, the modular pipeline is required.

3. Recovery cost. If the system says the wrong thing, what is the cost of recovery? If "negligible, the user will ask again," native speech-to-speech is appropriate. If "regulatory incident, clinical risk, or financial loss," the modular pipeline is required.

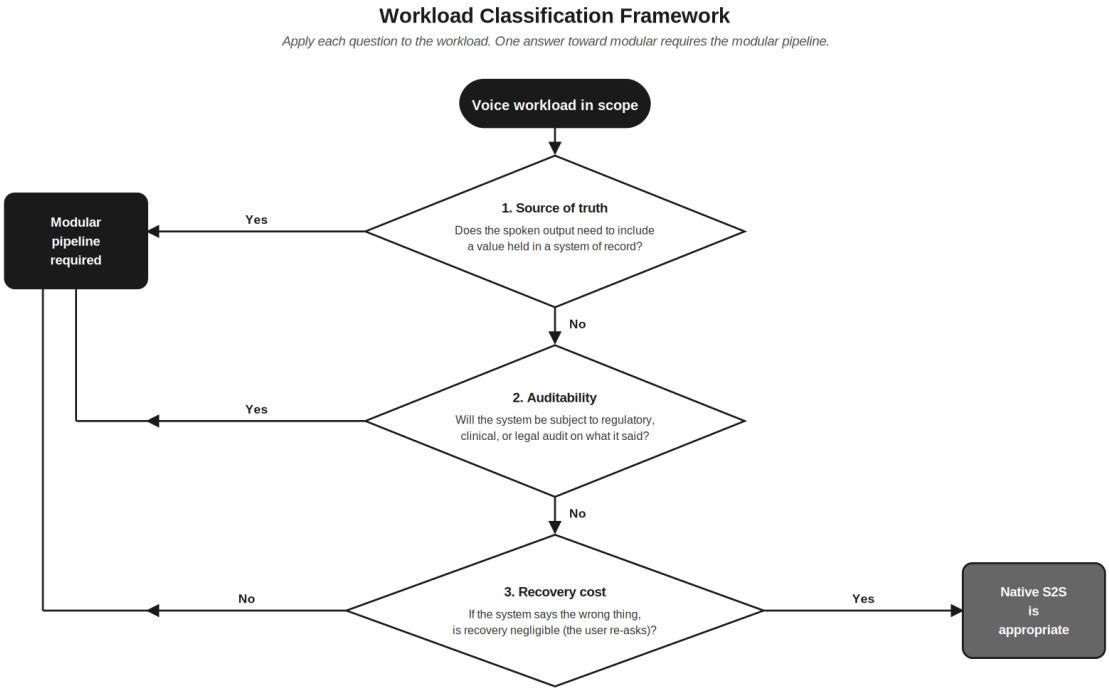

Figure 2. Workload classification framework. Apply to every voice workload in scope.

7. The modular pipeline in production

Evenly focuses on making voice AI accessible, reliable, trusted, and accountable in production. It does this by controlling each layer of the voice stack (speech recognition, reasoning, system-of-record access, and speech synthesis), with validation, observability, and failover engineered between the modules and layers of the architecture.

Evenly is built as a modular pipeline architecture by design, with orchestration across the strongest available component at every layer of the stack. Evenly incorporates best-of-breed speech-to-text (STT), large language models (LLM), and text-to-speech (TTS) modules, with pre-selected substitutes per module evaluated against the same acceptance suite, including latency, language coverage, terminology, numerical rendering, data residency, and known edge cases. Every module

is substitutable through the orchestration layer, which is how the pipeline retains state-of-the-art performance over time: as the frontier moves at any layer, Evenly switches that single component. The pipeline is inherently multilingual: the STT detects language automatically, the LLM reasons across languages without intermediate translation, and the TTS speaks the output language on demand. Evenly STT supports configurable phrase hints for sector-specific vocabulary and terminology, the LLM layer executes function calls to systems of record for any value that must be authoritative, and the TTS layer renders validated text through constrained, testable synthesis.

For any organisation with a voice workload in scope, Evenly will benchmark its modular pipeline against any monolithic alternative on the organisation's own workload, with the organisation's own data and acceptance criteria. The output is an architectural decision document the organisation can defend internally, regardless of which provider it chooses.

References

[1] Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. (2025). *Why Language Models Hallucinate*. arXiv:2509.04664. https://arxiv.org/abs/2509.04664

[2] Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., and Sloane, M. (2024). *Careless Whisper: Speech-to-Text Hallucination Harms*. ACM FAccT '24. https://doi.org/10.1145/3630106.3658996

[3] Burke, G. and Schellmann, H. (October 2024). *Researchers say AI transcription tool used in hospitals invents things no one ever said*. The Associated Press.

[4] Koudounas, A., et al. (2025). *SHALLOW: A Hallucination Benchmark for Speech Foundation Models*. arXiv:2510.16567. https://arxiv.org/abs/2510.16567

[5] OpenAI. *Whisper Model Card*. https://github.com/openai/whisper/blob/main/model-card.md

[6] OpenAI (7 May 2026). *Advancing voice intelligence with new models in the API*.

[7] Google (26 March 2026). *Gemini 3.1 Flash Live: Making audio AI more natural and reliable*.

[8] Défossez, A. et al. (2024). *Moshi: a speech-text foundation model for real-time dialogue*. Kyutai. arXiv:2410.00037.

[9] Regulation (EU) 2022/2554 (DORA). Regulation (EU) 2024/1689 (AI Act). Directive (EU) 2019/882 (EAA), transposed into Greek law as N. 4994/2022.

[10] European Union Agency for Cybersecurity (ENISA). *Guidelines on third-party ICT risk management under DORA*.

[11] Google. Bug Report: Gemini 2.5 Flash Native Audio Preview 12-2025 — Model Hallucinates Before NON_BLOCKING Tool Results Return. googleapis/python-genai issue #1894, filed 29 December 2025. https://github.com/googleapis/python-genai/issues/1894

This paper presents the architectural distinction between modular and monolithic conversational AI systems independent of any specific vendor implementation. All references are independently verifiable.

About Evenly

Evenly designs and operates pipeline-architecture conversational AI systems for accessibility and regulated workloads. The company is headquartered in Athens. Clients include Eurobank, Credia Bank, Athens International Airport, AADE, PPC (ΔΕΗ), Protergia, Affidea, and the Delphi Economic Forum. The Evenly product suite includes Evenly Connect for real-time interpretation, Evenly Events for conference and live event translation, Evenly Dialog for voice-first conversational AI in 120 languages, and Evenly Comply for accessibility scanning and certification.