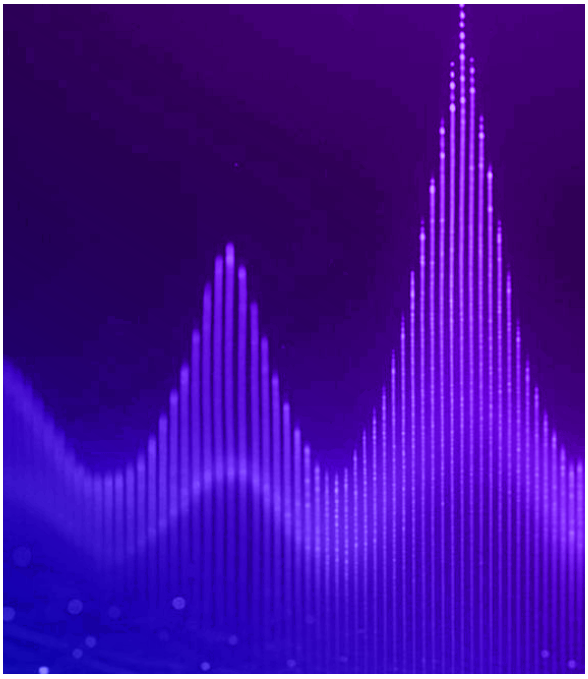
Fortanix | Deepgram

CASE STUDY

Enabling Secure Enterprise Voice AI While Protecting Proprietary Models and Sensitive Data

About the Customer



Deepgram is a leading voice AI foundation-model company trusted by enterprises and developers worldwide for speech-to-text, text-to-speech, and voice-agent applications. Its models, including Nova-3 for real-time speech recognition and Flux for voice agent conversations, deliver sub-300ms latency and a 54% word error rate improvement over prior generations. They are built on a foundation of significant proprietary research, training and optimization, making them valuable intellectual property and central to the company's market leadership.

Driven by data residency, sovereignty, and compliance requirements, as well as the need for low-latency performance that delivers the best user experience, enterprise demand for self-hosted AI is accelerating. Deepgram increasingly works with organizations looking to deploy voice AI models within their own environments.

Challenges

Protecting Deepgram's Model IP

Deepgram's models, including weights, parameters and architecture, represent years of innovation and substantial investment. Deploying frontier models into third-party environments, whether on premises, in AI Factories, or across NeoCloud providers, introduces significant business risk. Model owners must consider that:

- Model weights could be copied or extracted
- Proprietary logic could be reverse-engineered
- AI assets could be reused without authorization

Unlike traditional software, the value of exposed model weights could be permanently diminished once compromised. With this being the case, protecting model IP is about more than cybersecurity. It's about protecting the core business.

Securing Highly Sensitive Voice Data Across Enterprises

Voice AI often operates on some of the most sensitive information an organization handles. Depending on the industry, this could include:

- Patient histories and clinical conversations
- Banking calls and financial details
- Customer support interactions
- Internal meetings and confidential discussions
- Government or security-related communications

Organizations understandably want to unlock the value of their data with AI while keeping it inside controlled environments. That means the model must often be deployed where the data already resides. This creates a dual mandate: protecting the enterprise's sensitive data while also securing the model owner's IP.

The Runtime Exposure Problem

Generally speaking, traditional security controls were designed to protect data at rest or in transit. They certainly weren't built to fully protect AI workloads while they're actively running.

During inference, model weights must be decrypted in memory. Voice data must also be processed in memory in order to generate transcripts, responses, or analytics. But without runtime protections, anyone with sufficient access, from privileged administrators to malicious actors exploiting infrastructure, may be able to see sensitive assets.

For highly regulated organizations, that risk can delay or prevent AI adoption altogether.

The Fortanix Solution

Deepgram addresses these challenges with Fortanix Confidential AI, enabling the secure deployment of proprietary voice AI models across enterprise environments without exposing model IP or sensitive customer data. Fortanix Confidential AI combines confidential computing, attestation, and secure key management to protect both models and data while they are in use.

At a Glance: Fortanix Confidential AI	
Capability	**Business Outcome**
In-Memory Encryption	Models and voice data remain protected during execution
Cryptographic Attestation	Only verified, trusted environments can run workloads
Secure Key Release	Model weights decrypt only under approved conditions
Zero Trust for Admins	Even privileged users cannot access protected workload data

Confidential Execution Environment

Fortanix enables AI workloads to run inside hardware-isolated confidential computing environments, often referred to as trusted execution environments (TEEs). Within these environments:

- ☑ Model weights remain protected in memory
- ☑ Sensitive voice data is securely processed
- ☑ Workloads are isolated from host systems and administrators

This extends protection beyond storage and network encryption into active runtime execution.

Attestation-Based Trust

Before any organization deploys a Deepgram model, Fortanix verifies that the environment meets the desired security requirements. This could include validating the hardware configuration, system integrity, the approved runtime environment, or the trusted workload state. Only verified environments are allowed to run the model, without exceptions.

Secure Key Release

Policy-driven key management means that Deepgram models remain encrypted until execution time. Decryption keys are released only after the environment is fully verified, and they're never exposed outside the protected boundary. This ensures proprietary models are only decrypted inside trusted runtimes.

Secure Collaboration Between Model Owners and Enterprises

With Fortanix Confidential AI, organizations can deploy Deepgram's proprietary models in enterprise environments while the model owner maintains control over its intellectual property. At the same time, enterprises can:

- ☑ Run voice AI on sensitive internal data
- ☑ Keep data within their own environments
- ☑ Support sovereignty and compliance mandates
- ☑ Accelerate AI adoption with confidence

With end-to-end verification that environments are secure before anything runs, Fortanix removes the need for either side to rely on trust alone.

Benefits

Protection of Deepgram's High-Value AI Assets

Deepgram's model weights remain protected during execution, helping prevent theft, replication, reverse engineering and unauthorized access

Market Potential

With Fortanix, Deepgram can support self-hosted, on-premises and controlled enterprise AI deployments that would have previously been too risky from an IP perspective. This allows for new revenue opportunities, particularly considering Deepgram's models are trained on huge domain-specific audio datasets. Protecting them prevents competitors from replicating its advantages, strengthening its market position considerably.

Secure Voice AI in Regulated Industries

With proprietary models and internal data secured, advantages abound: healthcare organizations can securely transcribe patient interactions, financial institutions can responsibly analyze customer calls, and government agencies can deploy voice AI while maintaining operational control.

Faster Enterprise AI Adoption

By reducing security concerns around runtime exposure, organizations can launch their voice AI initiatives to production faster and with far less friction.

Verified Security, Not Assumed Trust

Fortanix replaces implicit trust with cryptographic verification. That means:

- Environments must prove their integrity before execution
- Keys are released only under approved conditions
- Runtime exposure is minimized
- Security controls are measurable and enforceable

As voice becomes one of the most important interfaces in enterprise AI, organizations must balance the dueling priorities of protecting proprietary AI models and safeguarding highly sensitive data. Fortanix Confidential AI enables both.

By combining confidential computing, attestation and secure key management, Fortanix provides a trusted foundation for deploying voice AI securely across enterprise environments without compromising model IP or data privacy.

Ultimately, this allows innovators like **Deepgram** to expand securely into new markets while enabling enterprises to adopt AI with confidence.

"Voice contains the enterprise's most sensitive data, patient conversations, financial transactions, classified briefings. As real-time voice becomes the primary interface of the enterprise, organizations need to deploy voice AI in the environments their security, confidentiality, and regulatory requirements demand. Together with Fortanix and NVIDIA, we've made it possible to run voice AI without compromise. Your data stays protected, your models stay protected, and you still get real-time performance."

Scott Stephenson,
CEO, Deepgram