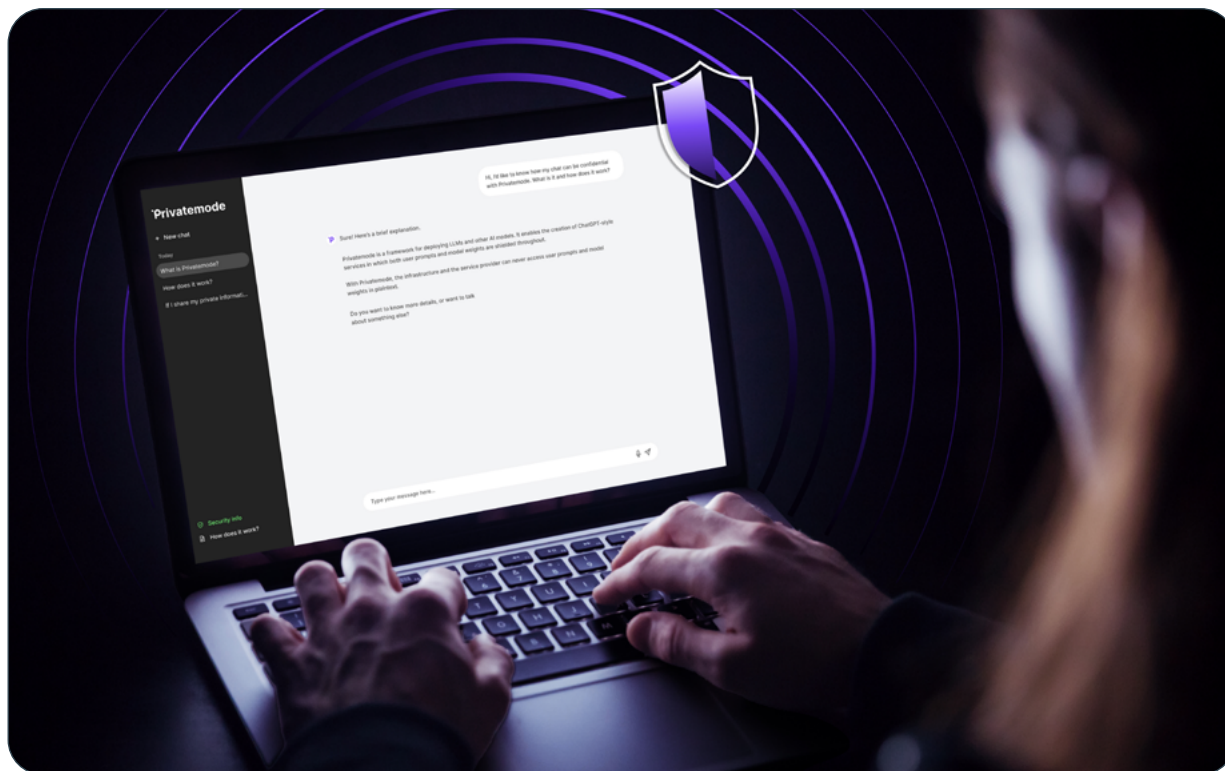


Removing blockers to using generative AI in industries with sensitive data

A Llama-powered LLM service employs confidential computing to bring airtight data security to cloud-based AI applications



At a glance

Edgeless Systems used Llama to create the first cloud-based large language model (LLM) service that uses confidential computing to keep data encrypted at all times. Businesses with sensitive data in tightly governed and high-security industries can now quickly develop and safely use generative AI applications to improve their operations and efficiency.



Industry

Technology



Use case

Developing secure, compliant AI applications



Goal

Enable businesses across sectors to use cloud-based generative AI



Llama version

Llama 3.1 70B Instruct AWQ INT4, Llama 3.3 70B Instruct AWQ INT4



Deployment

Bare metal hosted cloud via Scaleway

Results*

6 months

faster time to launch vs.
on-premises AI projects

50% savings

compared to on-premises
AI deployments

1,000s

of employees onboarded,
and it's just the beginning

The challenge

Protecting sensitive data while capitalizing on generative AI

Businesses in heavily regulated industries like healthcare and banking faced obstacles in adopting generative AI. Many cloud-based solutions on the market didn't meet their strict requirements for data privacy, confidentiality, zero trust and compliance. Yet turning to on-premises AI deployments was a complex and costly investment that many couldn't justify.

Integrating generative AI into business practices with confidence

Edgeless Systems saw the problem unfolding and knew that these businesses needed a solution for AI-powered workplace productivity that wouldn't compromise data privacy and security. The company set out to create a cloud-based service that could fortify chatbots and other AI tools using confidential computing — a technology that isolates sensitive data and computations in hardware-based trusted execution environments (TEEs) to ensure data is encrypted and protected even during processing.



Edgeless Systems is a German-based cybersecurity company that makes the public cloud the safest place for sensitive data. The company builds world-leading, open-source solutions for confidential computing, taking data security for cloud and AI applications to an unprecedented level, enabling encrypted data processing at scale.

- **Industry:**
Technology
- **Company size:**
30 employees and growing

“We wanted to make AI safe for sensitive data so businesses across sectors could confidently use AI to improve their operations.”

Thomas Strottner

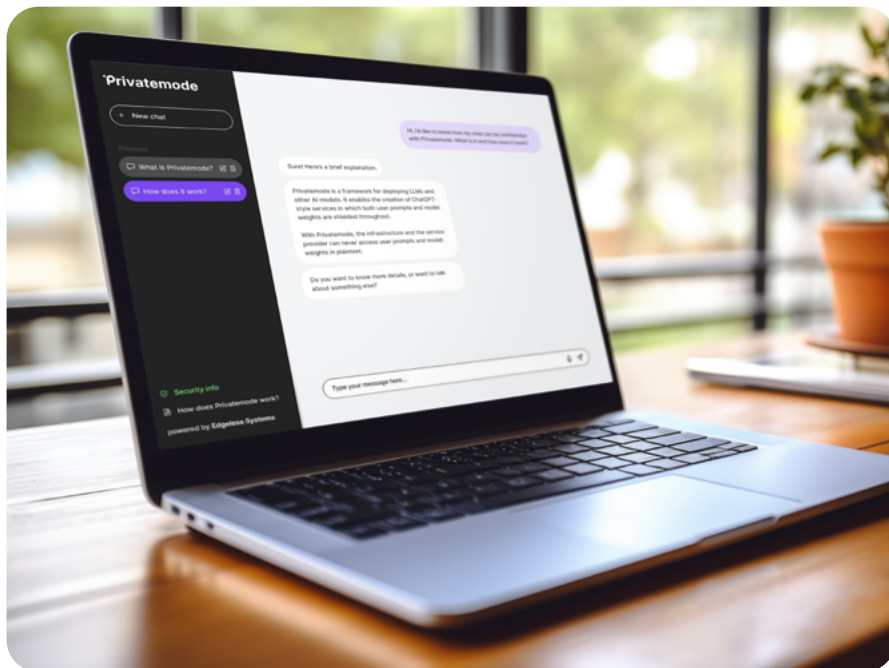
VP of Business Development, Edgeless Systems

The solution

A state-of-the-art LLM service that keeps data safe and secure

Edgeless Systems created Privatemode AI using Llama 3.1 70B Instruct AWQ INT4 and later upgraded to Llama 3.3 70B Instruct AWQ INT4. Privatemode AI is the first end-to-end, cloud-based LLM service that keeps prompts and responses encrypted at all times. “Using open-source components such as Llama is essential for our security model and to ensure transparency from infrastructure to application,” says Moritz Eckert, Vice President of Product and Technology at Edgeless Systems.

The service uses confidential computing, combining confidential virtual machines with NVIDIA H100 GPUs to run AI models in an isolated environment. The approach protects data from the service provider, the infrastructure and underlying cloud platforms, the model and Edgeless Systems, meeting strict regulations for end-to-end data encryption and mitigating risks such as prompt injection attacks or data breaches. In other words, Privatemode AI frees businesses to create AI applications while protecting sensitive information.



Privatemode AI provides end-to-end encrypted AI services for heavily regulated industries.

“Before Privatemode AI, tight security required extremely costly, on-prem deployments. Now companies have a secure option that uses confidential computing to isolate prompts and data from the underlying infrastructure and AI models. AI workloads can access powerful Llama models in the cloud with the same level of security and privacy.”

Moritz Eckert
VP of Product and Technology,
Edgeless Systems

Balancing efficiency and performance with an open, quantized Llama model

In developing the service, Edgeless Systems needed an open-source, self-hostable LLM they could easily download and run on their platform. Proprietary models were off the table. They also needed a model that could perform in multilingual environments for their European — specifically German — customers.

But the real challenge was finding a model that met performance thresholds while running on a single NVIDIA H100 GPU, a requirement that was necessary to deliver on the security benefits of confidential computing.

To identify contenders that were proficient in German and English, Edgeless Systems reviewed existing benchmarks using LMSYS Org’s Chatbot Arena. Then, they shortlisted models based on Massive Multitask Language Understanding (MMLU) 5-shot scores.

Several Llama models made the shortlist, including Llama 3.1 70B. Though they began with the base model, Edgeless Systems soon turned to a quantized version: Llama 3.1 70B Instruct AWQ INT4. Quantized models trade precision for smaller footprints that need less computing power. By using INT4 precisions, Edgeless Systems could benefit from an advanced model with 70 billion parameters without needing multiple GPUs.

From there, Edgeless Systems created their own German test prompts covering client-specific tasks such as translation, summarization and legal text explanation. The company’s scoring scheme rated responses on content accuracy, linguistic clarity and adaptation to context on a scale of one to five. They performed multiple runs and checked answers to validate scoring.

The results were clear. Llama 3.1 70B Instruct AWQ INT4 was superior to other options for speed, language accuracy and multilingual capabilities while providing the right level of efficiency.

Model	Evals		Testing	
	MMLU 5-shot	MMLU 5-shot (German)	Edgeless Systems test score (1-5)	Weighted German grammar mistakes per answer (#mistakes/word count)
gpt-4o (MMLU 0-shot CoT)	88.7		4.99	0.00020
meta-lama/Meta-Llama-3.1-70B-Instruct	83.6	79.27	4.87	0.00052
hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4	81.42		4.75	0.00046
google/gemma-2-27b-it	75.2		4.64	0.00095
VAGOolutions/SauerkrautLM-Nemo-12b-Instruct	69.12	61.49	4.55	0.00136
mistralai/Mistral-Nemo-Instruct-2407	68	62.7	4.49	0.00166
google/gemma-2-9b-it	71.3		4.32	0.00243
meta-llama/Meta-Llama-3.1-8B-Instruct	69.4	60.59	4.30	0.00285
TheBloke/Mixtral-8×7B-Instruct-vo.1-GPTQ			4.22	0.00453
VAGOolutions/Llama-3.1-SauerkrautLM-8b-Instruct	66.55		4.14	0.00414
cortecs/Llama-3-SauerkrautLM-70b-Instruct-GPTQ	79.15	72.7	4.01	0.00089

The quantized Llama model outperformed other open-source options on critical benchmarks.

Using confidential computing for AI applications

Let's take a closer look at how Privatemode AI protects sensitive data. Using confidential computing technology, the platform creates a secure environment that separates the infrastructure and service provider from the data and models.

Client side

The client consists of a lightweight proxy that exposes an **OpenAI-compatible API**, making it easy to plug Privatemode AI directly into existing applications and codebases without changes. It verifies the server's attestation, encrypts prompts and forwards inference requests. Responses are decrypted locally — ensuring end-to-end confidentiality.

For users who prefer a user interface, they also provide a **web app front end** that connects to the same secure back end.

Server side

The server securely processes inference requests using multiple components running inside the confidential container.

Contrast coordinator

Acts as the control plane for attestation and policy enforcement

- Performs remote attestation of AI worker and secret service instances
- Enforces runtime integrity policies
- Issues attestation-bound credentials to bootstrap secure key exchange

Secret service

Handles prompt encryption key management

- Authenticates to clients using coordinator-issued credentials
- Accepts and securely stores client-provided encryption keys
- Releases keys only to attested and verified AI workers

AI workers

Run the open-source vLLM library

- Serve inference using the Llama 3.3 70B model
- Handle prompts securely
- Never expose prompts and responses in plaintext to the infrastructure

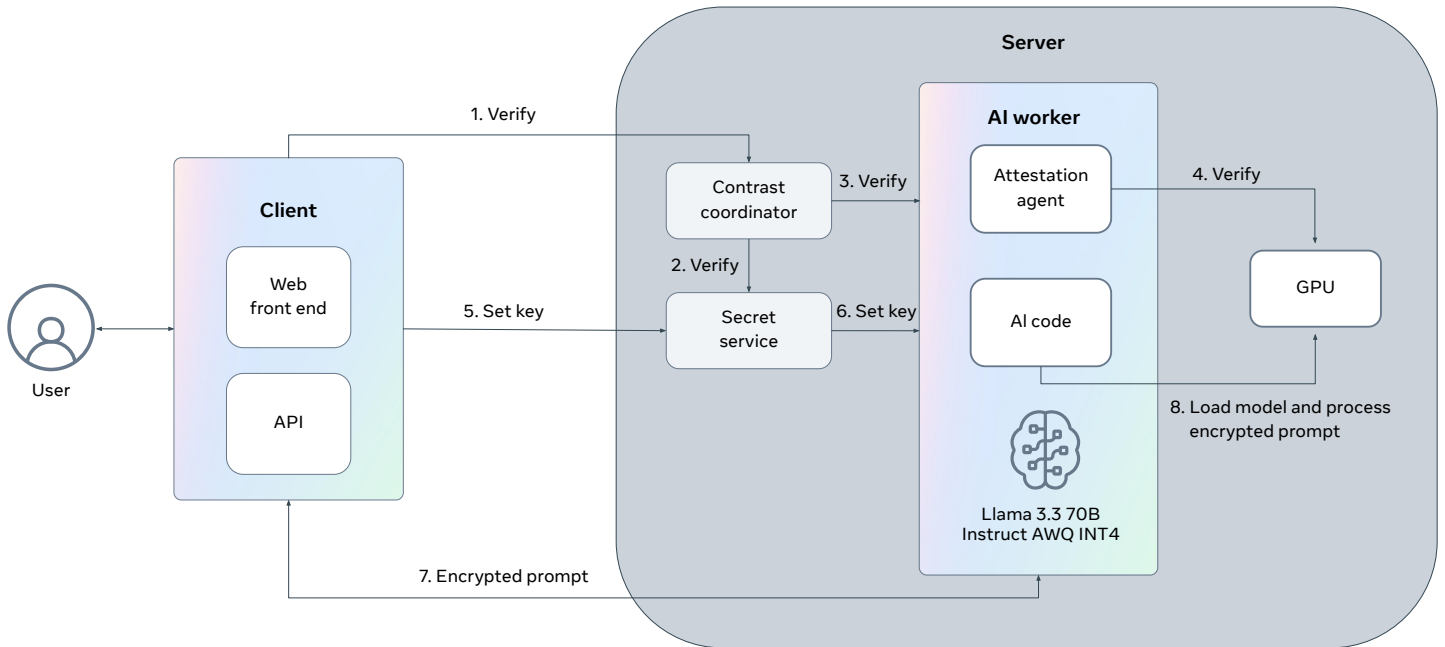
AI code

Is deployed inside the workers and structured in three functions

- **Encryption proxy:** Decrypts requests and encrypts responses — separate from the underlying confidential runtime encryption
- **Disk mounter:** Mounts model weights as read-only volumes with integrity checks, ensuring the Llama model can't be tampered with during the workers lifecycle
- **vLLM inference engine:** Generates responses from prompts using the Llama model

Attestation agent

Authenticates with the secret service using credentials provided by the coordinator and verifies the GPU



Privatemode AI verifies the integrity of each step of the confidential computing service.

Privatemode AI is deployed on Edgeless Systems' Contrast product with confidential containers on a bare metal-hosted cloud using Scaleway's Kubernetes service. The Llama model is hosted within confidential containers and served via the open-source vLLM library.



The outcome

Companies with sensitive data can safely use generative AI in the cloud

Privatemode AI has opened up a world of opportunity for compliance-driven businesses. Now, these organizations have a path to develop secure AI applications to boost employee productivity, improve business operations and save time and money. Healthcare companies, law firms, banks and more are free to spin up confidential AI assistants for all sorts of tasks like quick and secure document review and analysis.

Instead of missing out on AI's potential or only pursuing small, cost-conscious on-premises deployments, these businesses can confidently tap into powerful Llama models in the cloud at scale. Better yet, they can create AI applications fast while keeping expenses in check. Organizations that use Privatemode AI with Llama typically launch their AI projects six months faster and save approximately 50% compared to the complexity of pursuing on-premises deployments.

6 months

faster time to launch vs.
on-premises AI projects

50% savings

compared to on-premises
AI deployments

1000s

of employees onboarded,
and it's just the beginning

“Organizations across sectors like healthcare, finance and government can now use Privatemode AI and powerful Llama models in the cloud without compromising on security or privacy.”

Martin Paloncy

Partnerships Manager, Edgeless Systems

Conclusion

Bringing advanced features to Privatemode AI

Looking ahead, Edgeless Systems plans to add features such as retrieval augmented generation (RAG), tool calling, reasoning and multi-modal capabilities to its LLM service, giving customers the ability to build more sophisticated AI applications. Using the new NVIDIA GPU generation with Grace Blackwell architecture, Edgeless Systems will be able to deploy larger, non-quantized Llama models and extend the security of confidential computing to massive GPU arrays in the cloud.



How can Llama help your business?

See how open-source Llama brings unmatched control, customization and flexibility to generative AI application development and deployment.

[Learn More](#)[Related Stories ▶](#)