Solutions Brief

Edgeless Systems

Advancing Security for Large Language Models with NVIDIA GPUs and Edgeless Systems

Introduction

AI and large language models (LLMs) like ChatGPT are transforming how organizations operate, while driving unprecedented levels of productivity and innovation. However, AI adoption can often be impeded by concerns surrounding data privacy and security. This is particularly true for highly regulated industries like financial services, healthcare, and the public sector.

Applications

Edgeless Systems introduced Continuum AI, the first generative AI framework that keeps prompts encrypted at all times with confidential computing by combining confidential VMs with NVIDIA H100 GPUs and secure sandboxing.

The launch of this platform underscores a new era in AI deployment, where the benefits of powerful LLMs can be realized without compromising data privacy and security. Edgeless Systems, a Germany-based cybersecurity company that develops open-source software for confidential computing, is collaborating with NVIDIA to empower businesses across sectors to confidently integrate AI into their operations.

The confidential LLM platform isn't just a technological advancement—it's a pivotal step towards a future where organizations can securely utilize AI, even for the most sensitive data.

Use Cases

The Continuum technology has two main security goals. It first protects the user data and also protects AI model weights against the infrastructure, the service provider, and others. Infrastructure includes the basic hardware and software stack that the given AI app runs on. This includes all of the underlying cloud platforms, as well. In the case of ChatGPT, this would be Microsoft Azure. The service provider is the entity that provides and controls the actual AI app. In the case of ChatGPT, this would be OpenAI.

Data, Algorithm, and Model Safety

Continuum relies on two core mechanisms: confidential computing and advanced sandboxing. Confidential computing is a hardware-based technology that keeps data encrypted even during processing. Further, confidential computing makes it possible to verify the integrity of workloads.

Confidential computing, powered by NVIDIA H100 Tensor Core GPUs and advanced sandboxing technology, enables customers to protect user data and AI models. It does this by creating a secure environment that separates the infrastructure and service provider from the data and models. This technology also includes popular AI inference services, like NVIDIA Triton Inference Server™.

For more details, check out **Continuum** to stay ahead in the realm of enterprise-grade confidential AI.

Key Points

- 74% of data breaches involved a human element[1]
- $4.5M average cost of data breach[2]
- 90% of executives indicate the potential value of gen AI is moderate or extensive[3]
- Estimated 17% of all cyberattack/data leaks will involve gen AI by 2027[4]

> "Using confidential computing to address a key concern that security-conscious companies and individuals have with existing gen AI services. If applied correctly, confidential computing comprehensively prevents data leaks in AI services—including the involuntary use of user data for AI model (re-)training."
>
> **Felix Schuster,**
> CEO Edgeless Systems