Controlled Intelligence

Paths for the Secure Use of Generative AI in Business

An architecture proposal for the secure local operation of generative AI models in a zero-trust network with active traffic monitoring and multi-layer perimeter protection through solutions from genua.

Artificial intelligence is increasingly becoming part of everyday working life – for example in text creation, data analysis or process automation. But there are specific security requirements, especially in the public sector: The use of cloud-based AI solutions such as ChatGPT or Gemini is hardly acceptable in security-critical environments. The reason for this is not only the possible leakage of sensitive data into external networks but also the inability to control the models and their communication.

1. Challenges for the Secure Use of Generative AI in Business

With the rise of generative AI models such as ChatGPT, Claude or Gemini, companies are experiencing a new wave of digital possibilities – from automated text generation to code completion to intelligent knowledge management. At the same time, a key question arises: How can these tools be integrated securely, in accordance with data protection laws, and in a controlled manner in sensitive business environments?

Especially in regulated sectors – such as the public sector, energy supply or the health industry – classic cloud-based models are often not permitted. The reason: Prompts and usage data are transferred via the internet to externally hosted servers, usually in third countries. This contradicts not only internal compliance guidelines but also legal regulations such as the GDPR or industry-specific security standards.

But even outside of these high-security areas, there is growing awareness that control of data streams, model behaviors, and access obligations is essential – whether within the scope of internal company secrets or the protection of customer data. Companies are therefore increasingly looking for solutions that combine modern AI functionality with local control and technical security.

This white paper proposes one such solution – an architecture proposal of how generative AI can be operated securely, locally and scalably: on-premises, API-compatibly, and embedded in a zero-trust framework.

2. Local Language Model with vLLM: OpenAI-Compatible AI-Infrastructure in a Company‘s Own Data Center

For companies that want to use generative AI securely and in a controlled manner, the combination of a high-performance local language model and an efficient interference infrastructure is decisive. One especially promising solution is offered by the integration of GPT-OSS – a modern, open-source language model from OpenAI – with the vLLM (virtual Large Language Model) interference framework.

vLLM was developed to make large language models available especially efficiently. It offers an API that is compatible with the OpenAI interface which allows existing tools, plugins, and internal software solutions to continue to be used locally – without extensive modifications. This reduces not only the implementation effort and complexity but also opens the possibility to make existing automations and processes AI-capable with minimal integration effort.

An Overview of the Advantages

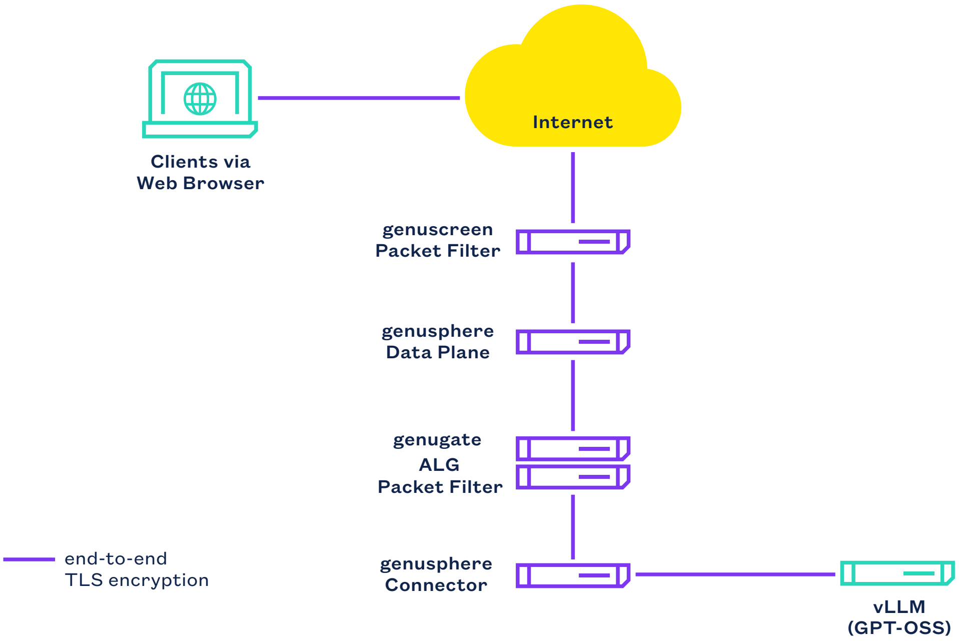
Local operation of GPT-OSS:
No data stream into the Internet, full control of model accesses and resources

High-performance interference through vLLM:
Scalable and memory-efficient, suitable for production environments

Flexibility and security:
Can be combined with internal security layers such as network segmentation, firewalling, and active traffic analysis

This architecture enables sovereign use of generative AI by companies without sacrificing central control, data protection or compliance.

3. Architecture Overview: Secure AI Operation in the Company Network



Construction of a holistically protected infrastructure for the secure use of generative AI

The key to the secure use of generative AI lies not solely in the choice of the model but rather in the construction of a holistically protected infrastructure. The combination of local LLM operation and a multi-layer zero-trust approach forms the basis for a controlled, GDPR-compliant and proprietary AI strategy.

The architecture shows how a model such as GPT-OSS, made available via vLLM, can be integrated in a controlled network scenario. Various security layers are used here that secure the use of AI – from the external border of the network to the active monitoring of the data traffic on the inside.

genuscreen – the Proven Perimeter Protection

Serving as the first line of defense is genuscreen, the high-performance firewall and VPN appliance from genua. It reliably protects the company network against attacks from the Internet and enables fine-grained control of the network traffic. This ensures already at the perimeter that only defined traffic reaches – and leaves – the internal network.

genugate – High-Resistance Firewall with Application Control

Within the architecture, genugate serves as an especially robust application layer firewall. Through the principle of zone- and media data separation, access to sensitive infrastructure – such as the server on which the vLLM and the language model run – is strictly regulated. Only exactly defined communication is allowed. Content can optionally even be passed on via manually examined releases, e.g., in the case of file transfers. This ensures that the organization retains control over data steams at all times.

genusphere – Zero Trust Application Access for Internal AI Services

While the internal operation of generative AI already satisfies many security requirements, there is often an additional requirement in practice: controlled access of AI applications from the outside – for example, by remote workers, external partners or distributed organization units. Here, genua uses genusphere as a solution for Zero Trust Application Access (ZTAA) that makes internal applications available securely without granting network access.

Unlike conventional VPNs or reverse proxies, genusphere enables a segmented provision of applications – completely separated from the internal network. Access rights are assigned in a context-dependent, role-based, and device-specific manner. This means: Only users who are authorized under the defined conditions can access certain functions of an application.

In the context of a locally operated AI architecture, this results in the following advantages:

- AI applications such as vLLM can be made accessible selectively to the outside – for example, as API, web interface or internal knowledge portal.
- Access is authenticated and context sensitive, e.g., using Device Trust, Identity Federation, and granular policies.
- The application remains internally isolated – genusphere only communicates the application layer, not the underlying infrastructure.

genusphere thereby offers a scalable, secure, and GDPR-compliant way to make generative AI usable even across location borders – without compromises in network- or data security.

Offline Operation with Update Window
To ensure maximum security, the AI server can be operated completely offline. Within clearly defined update cycles, model weights, security rules or system components are updated – via authorized and controlled processes. This variant is especially well suited for use in high-security environments.

4. Sovereign AI Use Needs Secure Architectures

Generative AI has the potential to fundamentally change the way in which companies work – more efficiently, more intelligently, and more creatively. But with this potential come new risks as well: data protection, model control, and the integrity of sensitive information are at stake.

This white paper shows that it is already possible today to operate high-performance AI systems locally and securely – on the basis of GPT-OSS, embedded in a zero trust architecture with multi-layer network safeguarding, intelligent traffic monitoring, and the controlled provision of interfaces. The combination of vLLM and security solutions from genua, such as genusphere, enable sovereign use of AI, independent of the cloud and in accordance with compliance and regulatory provisions.

Companies that take AI seriously should take their architecture just as seriously. Not only to avoid risks but also to retain control of their own digital future.

Further Information:

genua.eu/it-security-solutions

About genua

genua GmbH secures sensitive IT networks in the public and enterprise sectors, at critical infrastructure organizations, and in the classified industry with highly secure and scalable cyber security solutions. The company focuses on comprehensive network protection and internal network security for IT and OT. The range of solutions includes firewalls and gateways, VPNs, remote maintenance systems, internal network security, and cloud security through to remote access solutions for mobile working.

genua GmbH is a company of the Bundesdruckerei Group. With more than 400 employees, it develops and produces IT security solutions exclusively in Germany. Since the founding of the company in 1992, regular certifications and approvals from the German Federal Office for Information Security (BSI) provide proof of the high security and quality standards of the products. Customers include, among others, BMW, the German Armed Services, THW as well as the Würth Group.

Part of the
Bundesdruckerei
Group

bdr.

genua GmbH

Domagkstrasse 7 | 85551 Kirchheim, Germany
+49 89 991950-0 | info@genua.eu | www.genua.eu