



Kontrollierte Intelligenz

Wege zum sicheren Einsatz generativer KI in Unternehmen

Ein Architekturvorschlag zum sicheren lokalen Betrieb generativer KI-Modelle in einem Zero-Trust-Netzwerk mit aktiver Traffic-Überwachung und mehrschichtiger Perimeterabsicherung durch Lösungen von genua.



Künstliche Intelligenz hält zunehmend Einzug in den Berufsalltag – von der Texterstellung über die Datenanalyse bis hin zur Prozessautomatisierung. Doch gerade im öffentlichen Sektor stellen sich spezifische Sicherheitsanforderungen: Der Einsatz von Cloud-basierten KI-Lösungen wie ChatGPT oder Gemini ist in sicherheitskritischen Umgebungen kaum tragbar. Grund hierfür ist nicht nur der mögliche Abfluss sensibler Daten in externe Netze, sondern auch die fehlende Kontrollierbarkeit der Modelle und ihrer Kommunikation.

1. Herausforderungen beim sicheren Einsatz generativer KI in Unternehmen

Mit dem Aufstieg generativer KI-Modelle wie GPT-4, Claude oder Gemini erleben Unternehmen eine neue Welle digitaler Möglichkeiten – von automatisierter Textgenerierung über Code-Vervollständigung bis hin zu intelligentem Wissensmanagement. Gleichzeitig stellt sich jedoch eine zentrale Frage: Wie lassen sich diese Tools sicher, datenschutzkonform und kontrollierbar in sensible Unternehmensumgebungen integrieren?

Besonders in regulierten Branchen – etwa dem öffentlichen Sektor, der Energieversorgung oder der Gesundheitsbranche – sind klassische Cloud-basierte Modelle oft nicht zulässig. Der Grund: Prompts und Nutzungsdaten werden über das Internet an fremdgehostete Server übertragen, meist in Drittländer. Dies widerspricht nicht nur internen Compliance-Vorgaben, sondern auch gesetzlichen Regelungen wie der DSGVO oder branchenspezifischen Sicherheitsstandards.

Doch auch außerhalb dieser Hochsicherheitsbereiche wächst das Bewusstsein dafür, dass Kontrolle über Datenströme, Modellverhalten und Zugriffspflichten essenziell ist – sei es im Rahmen interner Betriebsgeheimnisse oder beim Schutz von Kundendaten. Unternehmen suchen daher zunehmend nach Lösungen, die moderne KI-Funktionalität mit lokaler Kontrolle und technischer Absicherung kombinieren.

Dieses Whitepaper stellt eine solche Lösung vor – ein Architekturvorschlag, wie generative KI sicher, lokal und skalierbar betrieben werden kann: On-Premises, API-kompatibel und eingebettet in ein Zero-Trust-Framework.



2. Lokale Sprachmodelle mit vLLM: OpenAI-kompatible KI-Infrastruktur im eigenen Rechenzentrum

Für Unternehmen, die generative KI sicher und kontrolliert einsetzen wollen, ist die Kombination aus einem leistungsstarken lokalen Sprachmodell und einer effizienten Inferenz-Infrastruktur entscheidend. Eine besonders vielversprechende Lösung bietet die Integration von Llama 3.3 – einem modernen Open-Source-Sprachmodell von Meta – mit dem Inferenz-Framework vLLM (Virtual Large Language Model).

vLLM wurde entwickelt, um große Sprachmodelle besonders effizient bereitzustellen. Es bietet eine kompatible API zur OpenAI-Schnittstelle, die es ermöglicht, bestehende Tools, Plugins und interne Softwarelösungen lokal weiterhin zu nutzen – ohne umfangreiche Anpassungen. Dies reduziert nicht nur Implementierungsaufwand und Komplexität, sondern schafft auch die Möglichkeit, bestehende Automatisierungen und Abläufe mit minimalem Integrationsaufwand KI-fähig zu machen.

Die Vorteile im Überblick

Lokaler Betrieb von Llama 3.3:

Kein Datenabfluss ins Internet, volle Kontrolle über Modellzugriffe und Ressourcen

Kompatibilität mit OpenAI-API:

Reibungslose Integration in bestehende Systeme und Anwendungen

Hochperformante Inferenz durch vLLM:

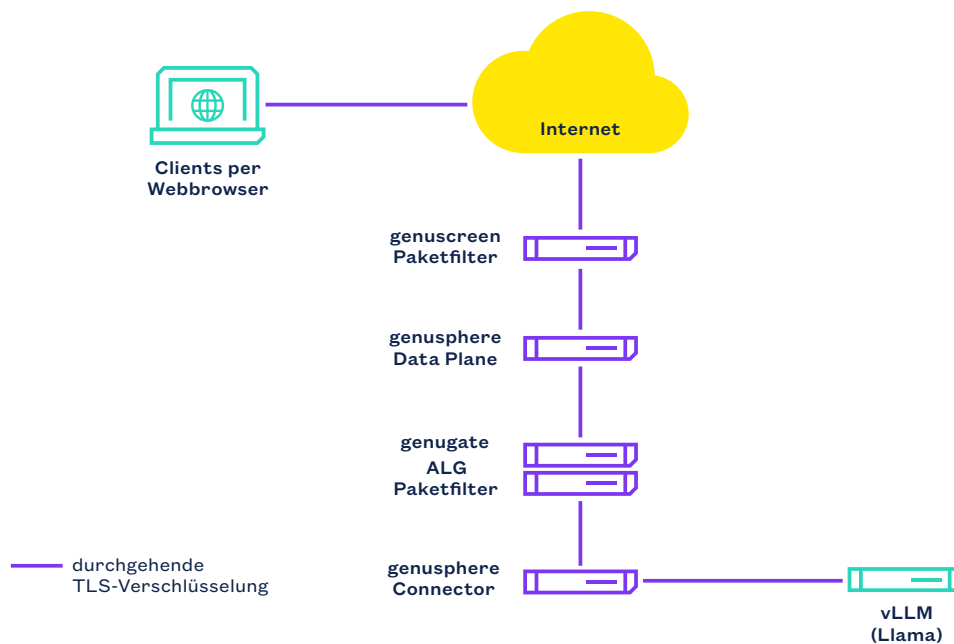
Skalierbar und speichereffizient, für Produktionsumgebungen geeignet

Flexibilität und Sicherheit:

Kombinierbar mit internen Sicherheitsschichten wie Netzwerksegmentierung, Firewalling und aktiver Verkehrsanalyse

Diese Architektur ermöglicht Unternehmen eine souveräne Nutzung generativer KI, ohne auf zentrale Kontrolle, Datenschutz oder Compliance zu verzichten.

3. Architekturüberblick: Sicherer KI-Betrieb im Unternehmensnetz



Aufbau einer ganzheitlich geschützten Infrastruktur
zum sicheren Einsatz generativer KI

Der Schlüssel zu einem sicheren Einsatz generativer KI liegt nicht allein in der Auswahl des Modells, sondern im Aufbau einer ganzheitlich geschützten Infrastruktur. Die Kombination aus lokalem LLM-Betrieb und einem mehrschichtigen Zero-Trust-Ansatz bildet die Grundlage für eine kontrollierte, datenschutzkonforme und unternehmenseigene KI-Strategie.

Die Architektur zeigt, wie sich ein Modell wie Llama 3.3, bereitgestellt über vLLM, in ein kontrolliertes Netzwerkszenario integrieren lässt. Dabei kommen verschiedene Sicherheitsschichten zum Einsatz, die den KI-Betrieb absichern – von der Außengrenze des Netzwerks bis zur aktiven Überwachung des Datenverkehrs im Inneren.

genuscreen – der bewährte Perimeterschutz

Als erste Schutzschicht dient genuscreen, die leistungsfähige Firewall und VPN-Appliance von genua. Sie schützt das Unternehmensnetz zuverlässig vor Angriffen aus dem Internet und ermöglicht eine feingranulare Steuerung des Netzwerkverkehrs. Auch verschlüsselte Verbindungen lassen sich damit analysieren und kontrollieren. So wird bereits am Perimeter sichergestellt, dass nur definierter Traffic in das interne Netz gelangt – und von dort heraus.

genugate – High Resistance Firewall mit Anwendungskontrolle

Im Inneren der Architektur agiert genugate als besonders robuste Application Layer Firewall. Durch das Prinzip der Zonen- und Mediendatentrennung wird der Zugriff auf sensible Infrastruktur – etwa den Server, auf dem vLLM und das Sprachmodell laufen – strikt reguliert. Nur exakt definierte Kommunikation ist erlaubt. Optional lassen sich Inhalte sogar über manuell geprüfte Freigaben weitergeben, etwa bei Dateiübertragungen. So bleibt die Kontrolle über Datenströme jederzeit in der Hand der Organisation.

Zugriff mit genusphere – Zero Trust Application Access für interne KI-Dienste

Während der interne Betrieb generativer KI bereits viele Sicherheitsanforderungen erfüllt, ergibt sich in der Praxis häufig ein zusätzlicher Bedarf: kontrollierter Zugriff auf KI-Anwendungen von außen – etwa durch mobile Mitarbeitende, externe Partner oder verteilte Organisationseinheiten. Genau hier setzt genusphere an als Lösung für Zero Trust Application Access (ZTAA), die interne Anwendungen sicher verfügbar macht, ohne Netzwerkzugriffe zu gewähren.

Im Unterschied zu klassischen VPNs oder Reverse-Proxys ermöglicht genusphere eine segmentierte Bereitstellung von Applikationen – vollständig entkoppelt vom internen Netzwerk. Zugriffsrechte werden kontextabhängig, rollenbasiert und gerätespezifisch vergeben. Das bedeutet: Nur wer unter den definierten Bedingungen berechtigt ist, kann auf bestimmte Funktionen einer Anwendung zugreifen.

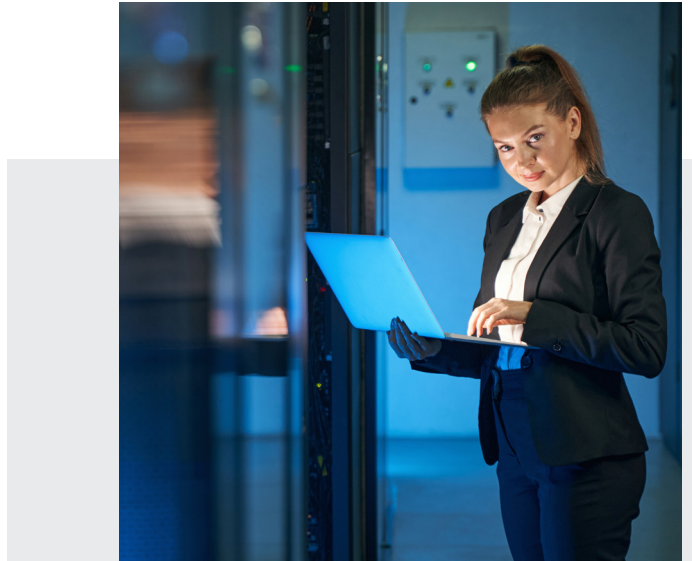
Im Kontext einer lokal betriebenen KI-Architektur ergeben sich daraus folgende Vorteile:

- KI-Anwendungen wie vLLM können selektiv nach außen bereitgestellt werden – etwa als API, Web-Oberfläche oder internes Wissensportal.
- Zugriffe erfolgen authentifiziert und kontextsensitiv, etwa mit Device Trust, Identity Federation und granularen Policies.
- Die Anwendung bleibt intern isoliert – genusphere vermittelt lediglich die Anwendungsebene, nicht die darunterliegende Infrastruktur.

Damit bietet genusphere eine skalierbare, sichere und DSGVO-konforme Möglichkeit, generative KI auch über Standortgrenzen hinweg nutzbar zu machen – ganz ohne Kompromisse bei Netzwerk- oder Datensicherheit.

Update-Fenster und Air-Gap-Szenarien

Um maximale Sicherheit zu gewährleisten, kann der KI-Server vollständig offline betrieben werden. Innerhalb klar definierter Update-Zyklen werden Modellgewichte, Sicherheitsregeln oder Systemkomponenten aktualisiert – über autorisierte, kontrollierte Prozesse. Alternativ lässt sich ein Air-Gap-Betrieb realisieren, bei dem keinerlei direkte Kommunikation nach außen möglich ist. Diese Variante eignet sich insbesondere für den Einsatz in Hochsicherheitsumgebungen.



4. Fazit: Souveräne KI-Nutzung braucht sichere Architekturen

Generative KI hat das Potenzial, die Art und Weise, wie Unternehmen arbeiten, grundlegend zu verändern – effizienter, intelligenter und kreativer. Doch mit diesem Potenzial gehen auch neue Risiken einher: Datenschutz, Modellkontrolle und die Integrität sensibler Informationen stehen auf dem Spiel.

Dieses Whitepaper zeigt, dass es bereits heute möglich ist, leistungsstarke KI-Systeme lokal und sicher zu betreiben – auf Basis von Llama 3.3, eingebettet in eine Zero-Trust-Architektur mit mehrschichtiger Netzwerkabsicherung, intelli-

gener Verkehrsüberwachung und kontrollierbarer Schnittstellenbereitstellung. Die Kombination aus vLLM und Sicherheitslösungen von genua wie genusphere erlaubt einen souveränen Umgang mit KI – unabhängig von der Cloud, im Einklang mit Compliance und regulatorischen Vorgaben.

Unternehmen, die KI ernst nehmen, sollten ihre Architektur ebenso ernst nehmen. Nicht nur, um Risiken zu vermeiden – sondern auch, um die Kontrolle über ihre eigene digitale Zukunft zu behalten.

Weitere Informationen:

genua.de/it-sicherheitsloesungen



0525-01-DE

Über genua

Die genua GmbH mit Sitz in Kirchheim bei München sichert sensitive IT-Netzwerke im Public- und im Enterprise-Sektor, bei KRITIS-Organisationen und in der geheimschutzbetreuten Industrie mit hochsicheren und skalierbaren Cyber-Security-Lösungen. Dabei fokussiert sich das Unternehmen auf den umfassenden Schutz von Netzwerken, Kommunikation und interne Netzwerksicherheit für IT und OT. Das Lösungsspektrum umfasst Firewalls und Gateways, Virtual Private Networks, Fernwartungssysteme, interne Netzwerksicherheit und Cloud Security bis hin zu Remote-Access-Lösungen für mobiles Arbeiten und Homeoffice.

Die genua GmbH ist ein Unternehmen der Bundesdruckerei-Gruppe. Mit mehr als 400 Mitarbeitern entwickelt und produziert sie IT-Security-Lösungen ausschließlich in Deutschland. Seit der Unternehmensgründung 1992 belegen regelmäßige Zertifizierungen und Zulassungen durch das Bundesamt für Sicherheit in der Informationstechnik (BSI) den hohen Sicherheits- und Qualitätsanspruch der Produkte. Zu den Kunden zählen u. a. Arvato Systems, BMW, die Bundeswehr, das THW sowie die Würth-Gruppe.



genua GmbH

Domagkstraße 7 | 85551 Kirchheim bei München
+49 89 991950-0 | info@genua.de | www.genua.de

