Inspiration durch Wissen

Einsatz Künstlicher Intelligenz im Umfeld von Verschlusssachen

Toni Schnell, inxire GmbH

OMNISECURE
Berlin, 21.01.2026

Agenda

- Einleitung
- Architektur
- Datenfluss
- Maßnahmen zur Sicherung der KI
- KI-Verfahren für VS
- Finetuning und RAG
- Live-Demo
- Fragen und Antworten

KI im Umfeld von Verschlusssachen

Einleitung

- Ausgangslage
  - Verschlusssachen unterliegen höchsten rechtlichen, organisatorischen und technischen Schutzanforderungen
  - Gleichzeitig steigt der Bedarf an schneller, fundierter Entscheidungsunterstützung
  - Künstliche Intelligenz kann diesen Bedarf adressieren – wenn sie kontrollierbar ist

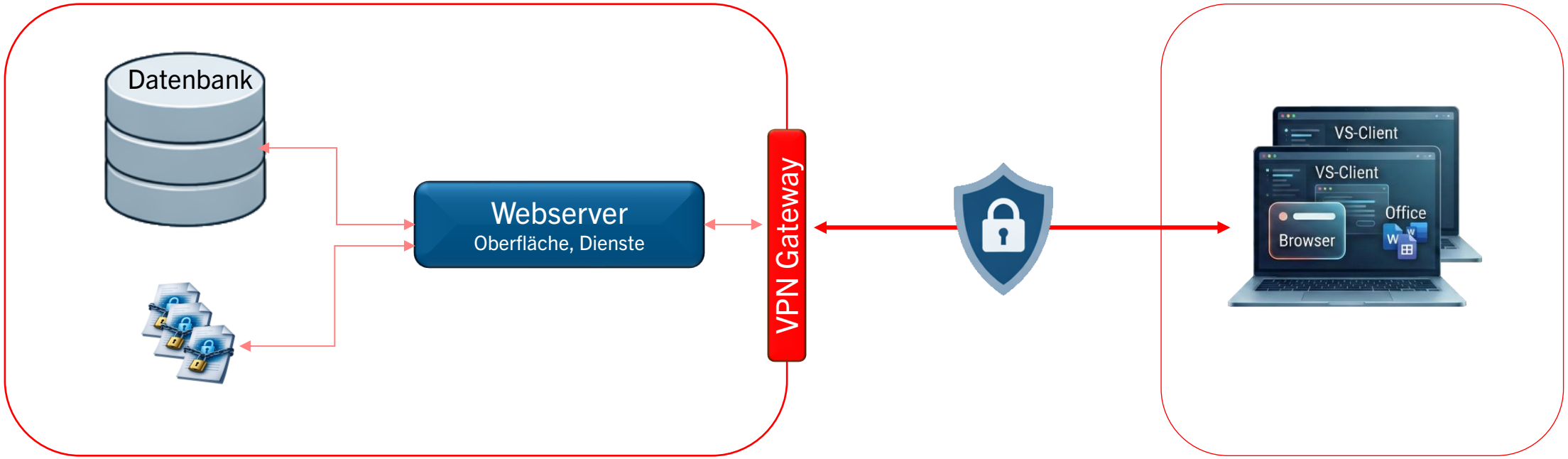
- Ziele
  - Digitale Souveränität
  - Vertrauenswürdige KI-Nutzung
  - Auch bei VS-GEHEIM

KI im Umfeld von Verschlusssachen

ChatGPT im Tresor ?

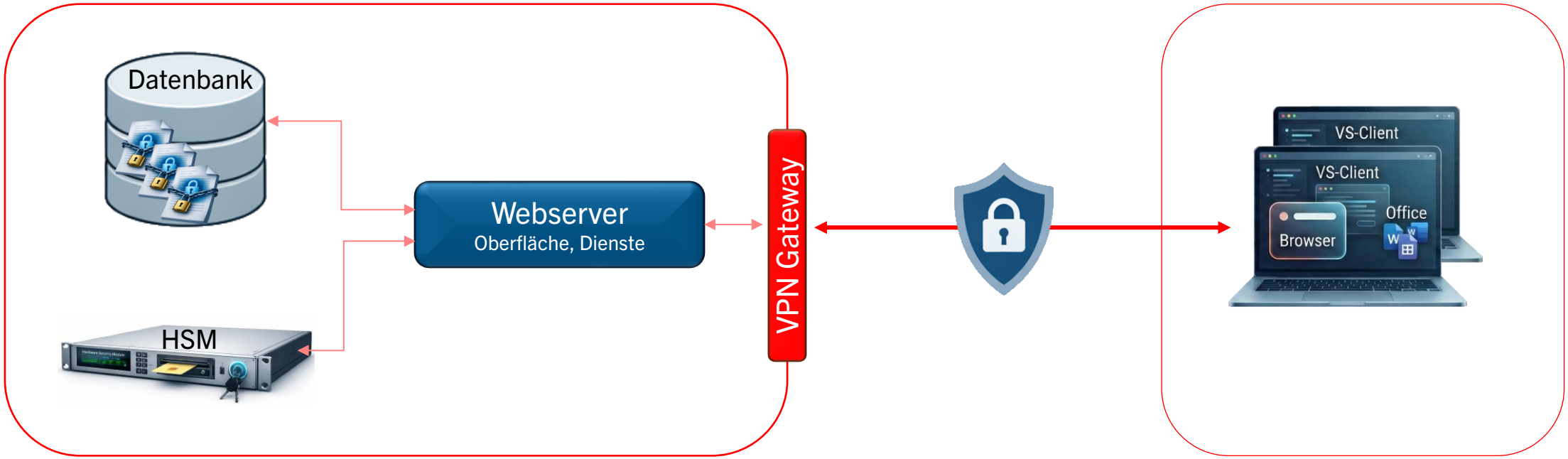
Architektur

Verschlusssachen mit klassischer Webarchitektur



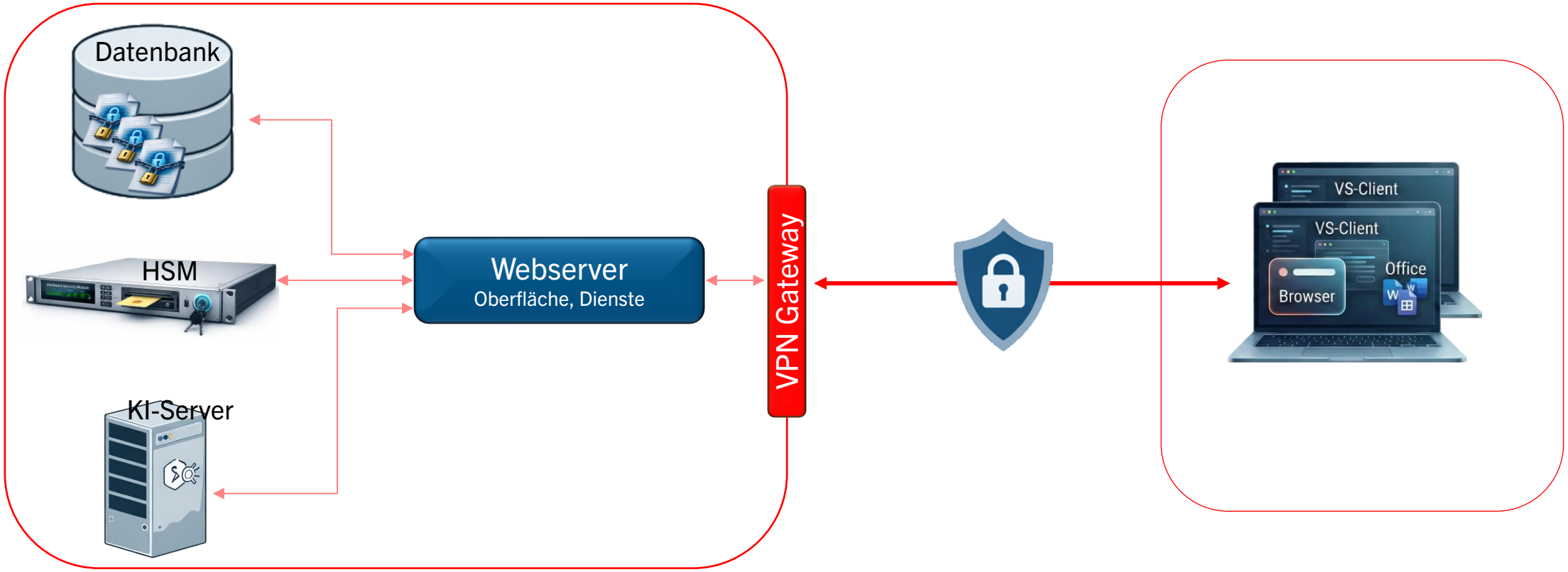
Architektur

Verschlusssachen mit klassischer Webarchitektur und HSM



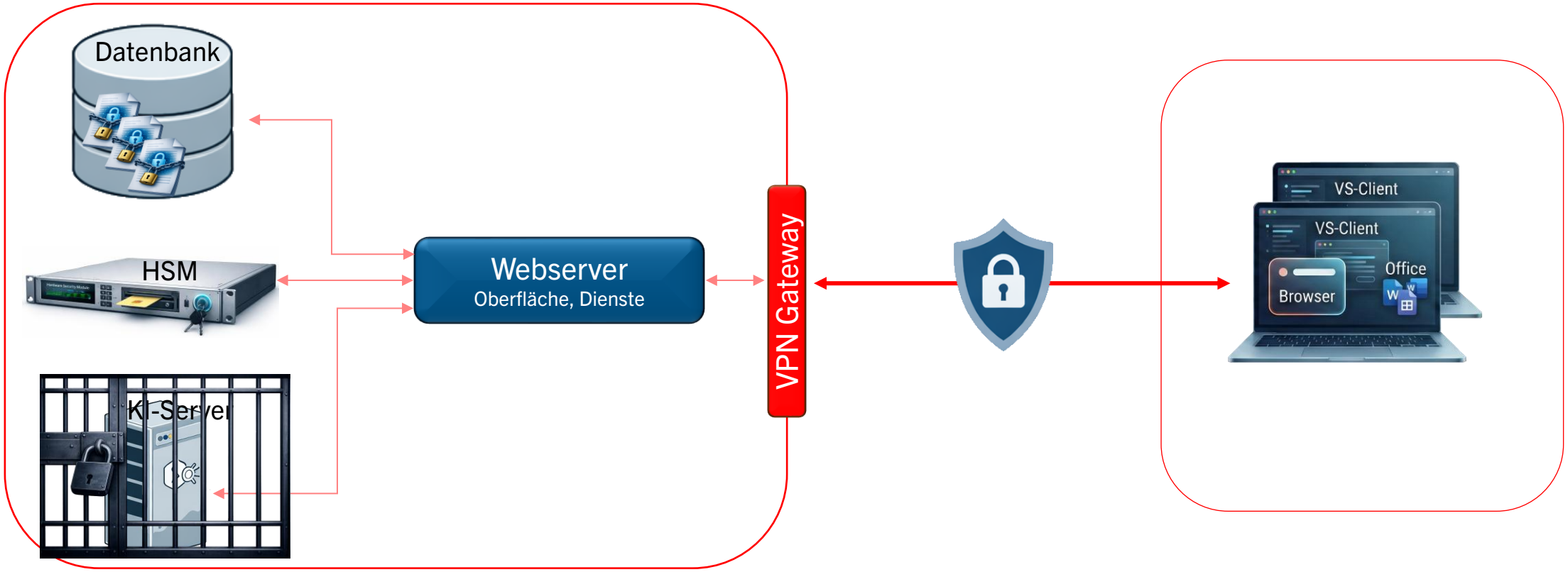
Architektur

Verschlusssachen mit klassischer Webarchitektur, HSM und KI-Server



Architektur

Verschlusssachen mit klassischer Webarchitektur, HSM und KI-Server

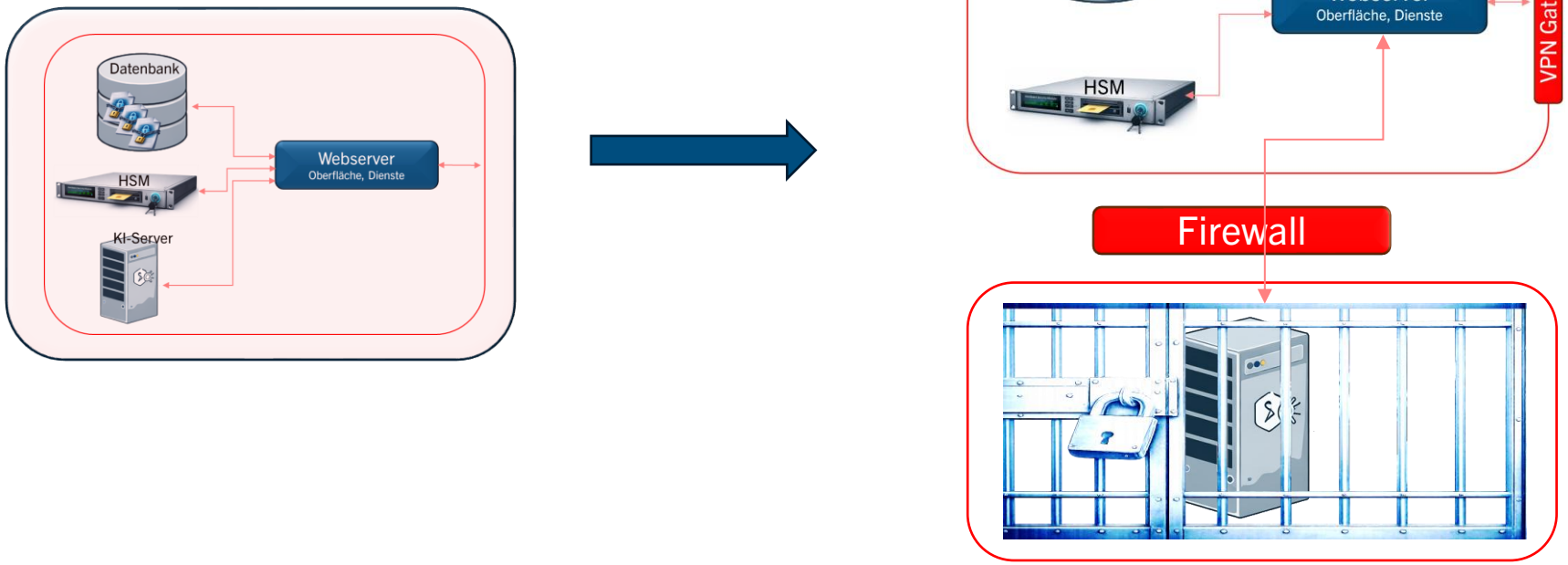


KI-Server ist kein vertrauenswürdiger Teil der VS-Domäne

Isolierte Ausführungsdomäne

Getrennte Hardware

- KI-Server auf eigene Hardware auslagern
- Mit Firewall sichern



Isolierte Ausführungsdomäne (VM)

Durchsetzung der Isolation durch Separation Kernel

- L4Re Separation Kernel als Sicherheitsanker
- Minimale Trusted Computing Base (TCB)
- Strikte Isolation von Speicher, CPU und I/O
- Alle sicherheitsrelevanten Komponenten als getrennte VMs
- Kein implizites Vertrauen zwischen Domänen

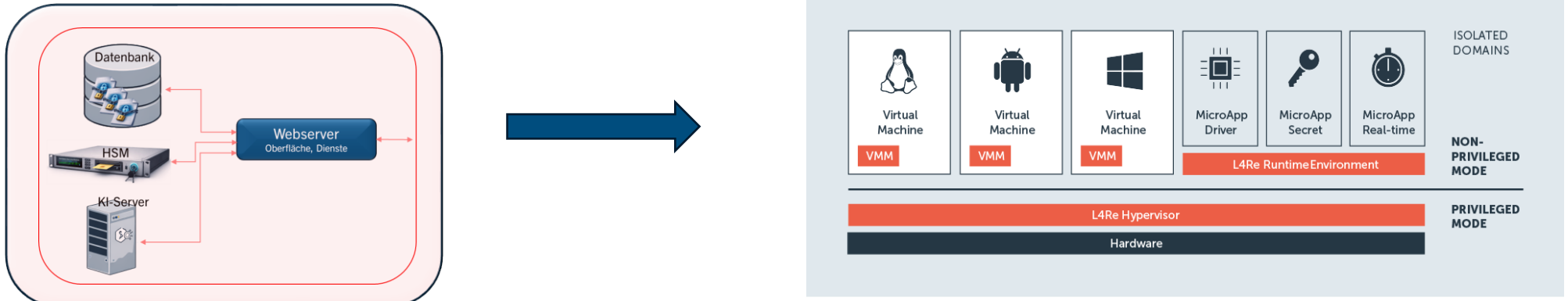
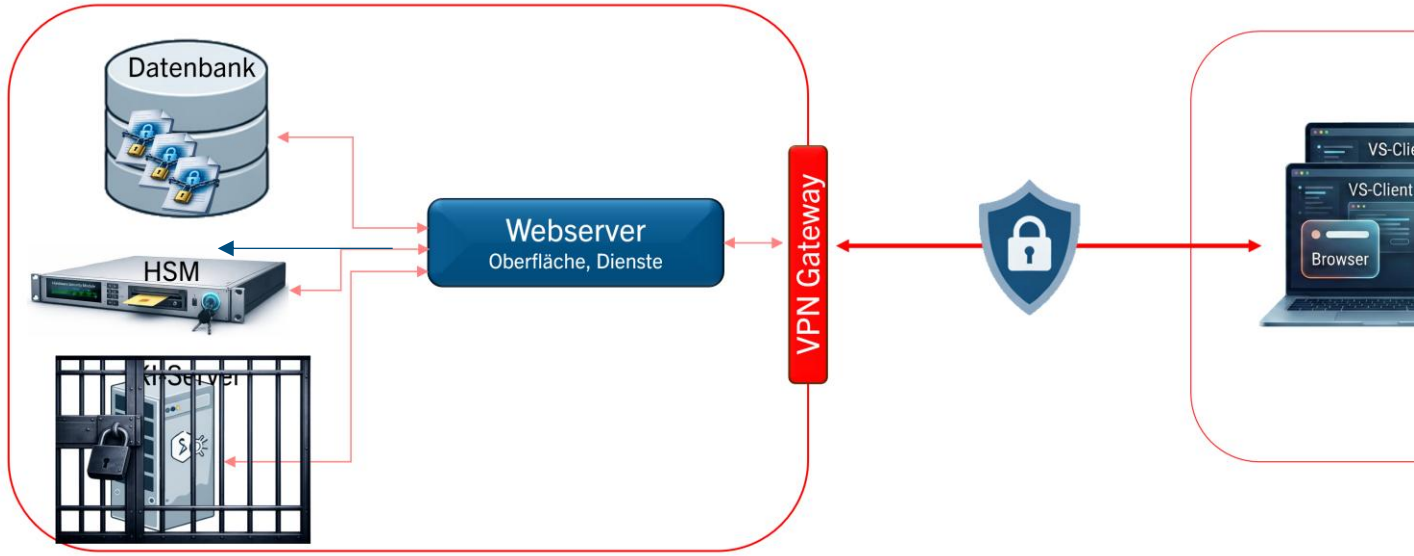


Bild Quelle: l4re.org

Datenfluss

Technische Vorbedingungen

- Benutzer wird eindeutig identifiziert
- Benutzerbezogene Freigabe der Schlüsselverwendung (z. B. PIN) für kryptographische Operationen im HSM
- PQC-fähige Verfahren für die Verschlüsselung
- Inhalte liegen als Cyphertext mit Wrapped eDEK in der DB

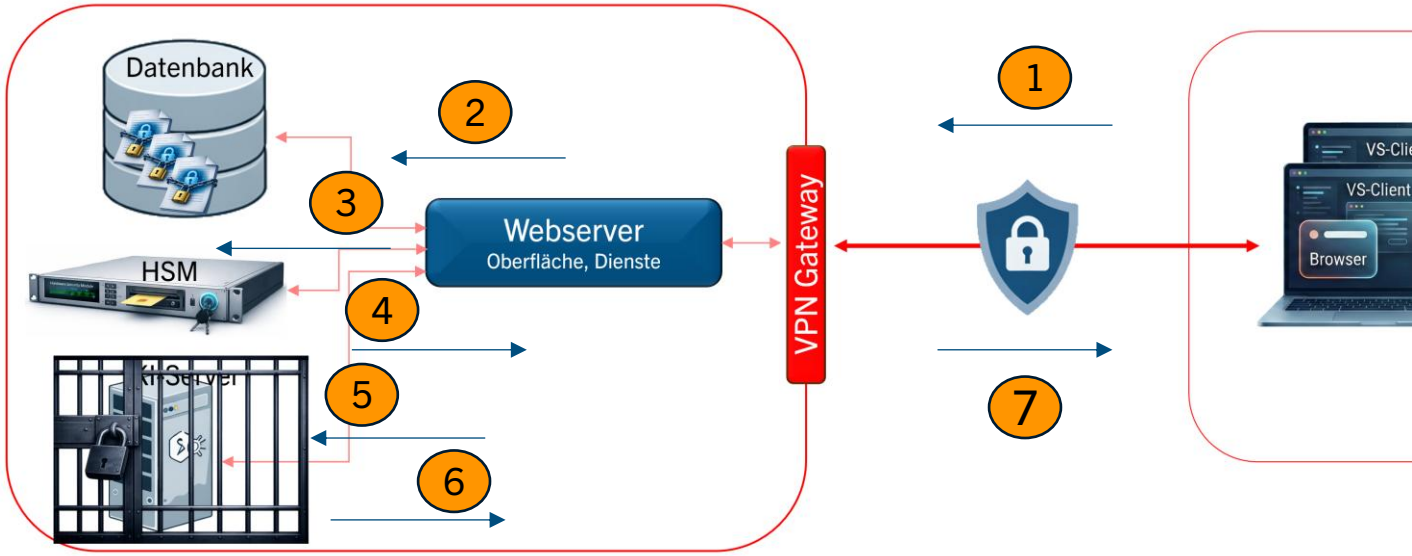


- Nur HSM kennt privaten Schlüssel und verlässt diesen nicht
- Alle Netzübergänge mit mTLS gesichert

Datenfluss

Laufzeit der VS-Verarbeitung

1. Anfrage „Dokument analysieren“
2. Prüfung Need-to-know
3. Dokumentenschlüssel (wrapped eDEK) wird im HSM entschlüsselt
4. Ciphertext aus der Datenbank wird mit eDEK im isolierten Arbeitsspeicher entschlüsselt

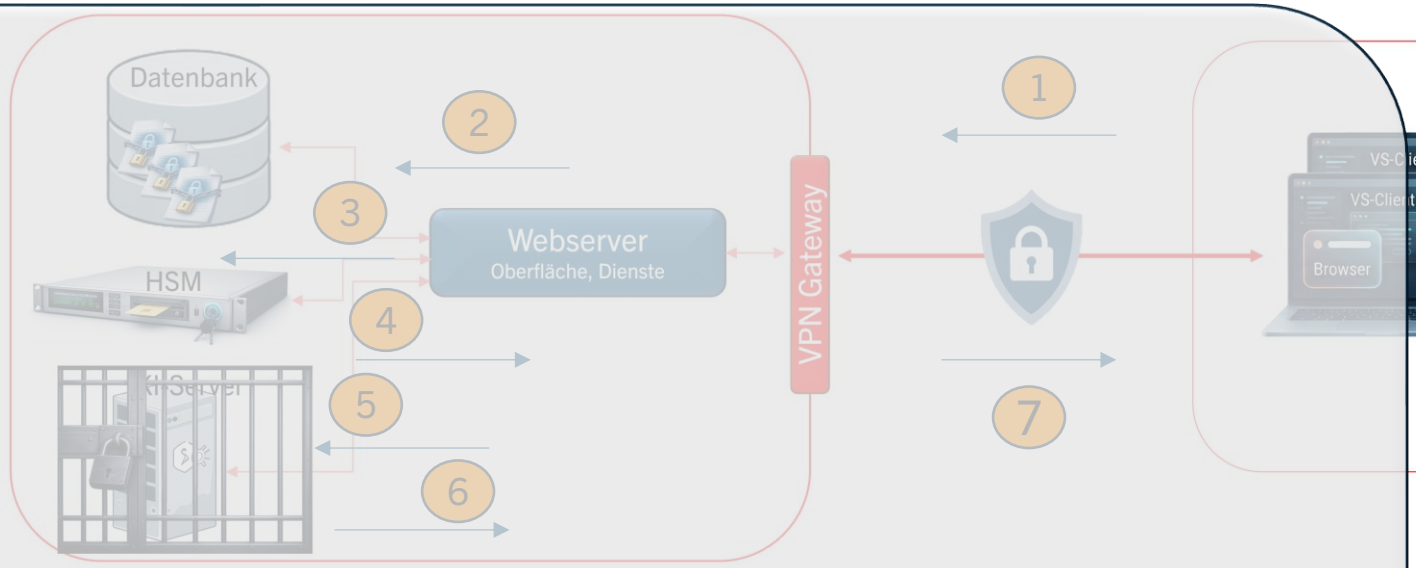


5. Klartext wird ausschließlich transient, zweckgebunden und nicht persistent an den KI-Server übergeben
6. Ergebnis von KI-Server wird kontrolliert an Webserver übergeben
7. Webserver validiert, protokolliert und weiterverarbeitet

Datenfluss

Sicherheitsprinzip (Zusammenfassung)

1. Anfrage „Dokument analysieren“
2. Prüfung Need-to-know
3. Dokumentenschlüssel (wrapped eDEK) wird im HSM entschlüsselt
4. Ciphertext aus der Datenbank wird mit eDEK im isolierten Arbeitsspeicher entschlüsselt
5. Klartext wird ausschließlich transient, zweckgebunden und nicht persistent an den KI-Server übergeben
6. Ergebnis des KI-Servers wird durch den Webserver validiert, protokolliert und weiterverarbeitet
7. Ergebnis des KI-Servers wird durch den Webserver validiert, protokolliert und weiterverarbeitet



Die KI verarbeitet VS-Inhalte ausschließlich transient, isoliert und zweckgebunden; Schlüsselmaterial und Freigabeentscheidungen verbleiben vollständig außerhalb der KI.

Technische Maßnahmen zur Sicherung des KI-Servers

- Gesicherte Kommunikation ausschließlich über den Webserver
- Strikte Trennung von Steuerung und Inhalt
- Kein direkter Zugriff auf Datenbank oder HSM
- Auswirkungen einer kompromittierten KI sind auf den isolierten KI-Server begrenzt
- Prompt Injection hat keine sicherheitsrelevanten Auswirkungen außerhalb der isolierten VM

Wir verhindern nicht jedes Fehlverhalten der KI – wir stellen sicher, dass es keine sicherheitsrelevanten Konsequenzen haben kann.

Organisatorische Maßnahmen zur Sicherung des KI-Servers

- KI-Governance
- Klare Rollen- und Verantwortlichkeitstrennung
- Festgelegter Einsatzrahmen
- Betriebsregeln für Modell- und Systemänderungen
- Regelmäßige Überprüfung & Re-Evaluierung
- Incident- und Abschaltkonzept

Die technische Isolation begrenzt den Schaden - die organisatorischen Maßnahmen verhindern Fehlanwendung

Nutzung der KI für VS

Prozesse bei VS

KI-Verfahren für VS
Entstehung

- Sprach- und Stilunterstützung
- Strukturierungs- und Gliederungshilfen
- Zusammenfassung
- Übersetzung
- Metadaten-Extraktion
- OCR

KI-Verfahren für VS

Vereinnahmung und Klassifizierung

- Vorschlag für Klassifizierungsmerkmale*
- Erkennung sensibler Inhalte (z.B. Personen)
- Formelle Dokumentenprüfung

* Bedingt in der Regel eine trainierte Fachdomäne

KI-Verfahren für VS

Verarbeitung

- Semantische Suche und kontextbasierte Recherche
- Entitäten- und Beziehungsanalyse des VS-Bestands
- Strukturierung und thematische Clustering-Verfahren
- Frage-Antwort-Unterstützung auf Basis vorhandener VS-Inhalte
- Analyse und Auswertung (z.B. Zeitreihenanalyse)
- Klassifizierung
- Mustererkennung

KI-Verfahren für VS

Weitergabe

- Weitergabe als Need-to-know
  - Hinweise auf potenziell berechtigte Empfänger
- Weitergabe in andere Sicherheitsdomäne
  - Regel- und kontextbasierte Vorschläge
  - Erkennung sensibler Inhalte

KI-Verfahren für VS

Zusammenfassung

Künstliche Intelligenz steigert Effizienz im VS-Umfeld – Verantwortung, Kontrolle und Freigabe verbleiben jederzeit beim Menschen

KI-Verfahren für VS

Feintuning von KI-Modellen

- Spezifische Analysen und Auswertungen erfordern Fachdomänenwissen
- Feintuning dient der Anpassung eines bestehenden Modells an definierte fachliche Kontexte
- Feintuning erfolgt ausschließlich auf freigegebenen Trainingsdaten
- Nachvollziehbarkeit durch Versionierung, Trainingsdokumentation und Freigabeprozess
- Trennung von Modelltraining und Produktivbetrieb

KI-Verfahren für VS

Feintuning von KI-Modellen – Besonderheiten im VS-Umfeld

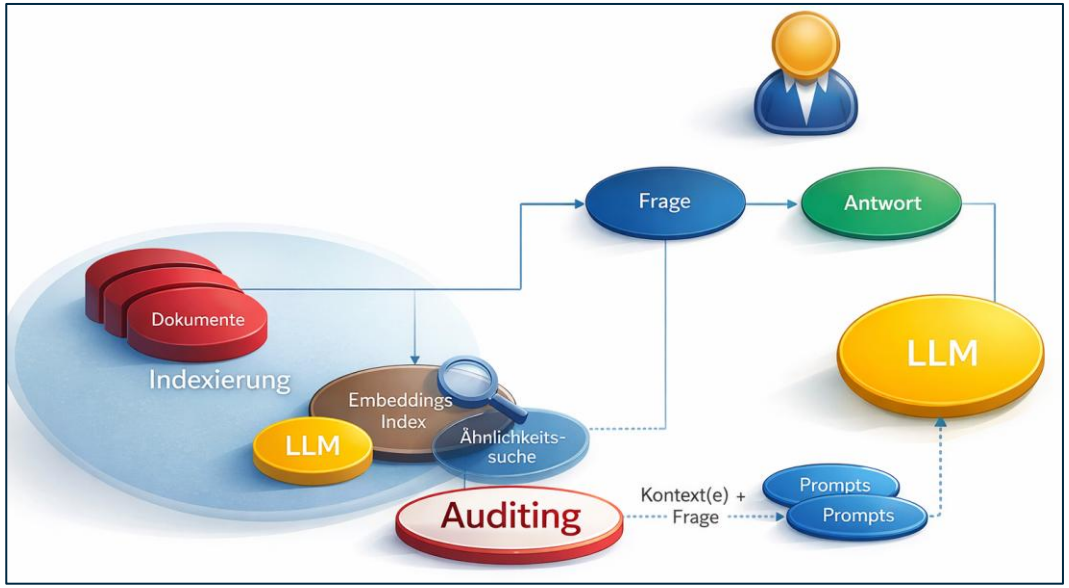
- Ein feingetuntes Modell ist mindestens mit der höchsten Einstufung der verwendeten Trainingsdaten zu behandeln
- Beim Feintuning werden Inhalte der Trainingsdaten dauerhaft im Modell verankert, sodass Rückschlüsse auf diese Daten prinzipiell nicht ausgeschlossen werden können.
- KI-Modelle erzeugen probabilistische Ergebnisse und können halluzinieren
- Feingetunte KI-Modelle mit VS-Bezug erfordern besonderen Schutz
- Einsatz nur für klar abgegrenzte Aufgaben
  - Klassifikation ✅
  - Textgenerierung ❌

Feintuning verlagert Informationen dauerhaft ins Modell – und ist damit eine sicherheitsrelevante Architekturentscheidung mit besonderem Verantwortungsbedarf.

RAG (Retrieval-Augmented Generation)

Datenfluss und VS-Kontext

- Dokumente werden indexiert und als Embeddings gespeichert
- Nutzerfrage wird mit dem Embeddings-Index abgeglichen, um relevante Dokumente zu finden
- Ähnlichkeitssuche liefert kontextrelevante Inhalte
- Auditing stellt Nachvollziehbarkeit sicher
- LLM generiert die Antwort auf Basis von Frage, Kontext und Prompts.



- Vektordaten sind mindestens gemäß der VS-Einstufung der Quelldaten zu behandeln
- Inhaltliche Rückschlüsse aus Vektordaten können grundsätzlich nicht ausgeschlossen werden

KI-Verfahren für VS

RAG vs. Feintuning im VS-Umfeld

Kriterium	RAG (Retrieval-Augmented Generation)	Feintuning
Lage des Wissens	Außerhalb des Modells (Dokumente, Vektordatenbanken)	Im Modell selbst
VS-Risiko	Kontrollierbar, da Wissen außerhalb des Modells gehalten und separat gesteuert wird	Erhöht, da Wissen nicht mehr vom Modell trenn- oder selektiv entfernbar ist
Einstufung	Dokumente & Vektordaten separat einstufbar	Modell mindestens mit der höchsten Einstufung aller verwendeten Trainingsdaten zu behandeln
Änderbarkeit / Widerruf	Hoch (Dokumente entfernbar, Indizes neu erzeugbar)	Niedrig bis nicht möglich
Auditierbarkeit	Hoch (Datenquellen, Zugriffe, Datenfluss nachvollziehbar)	Eingeschränkt (Modellparameter sind nur mathematisch, nicht semantisch auditierbar)
Typische VS-Use-Cases	Suche, Q&A, Kontextbereitstellung	Klassifikation, Erkennung klar definierter Muster
Empfehlung im VS-Umfeld	Bevorzugter Standardansatz ✅	Nur in Ausnahmefällen mit gesonderter Freigabe ⚠️

RAG ermöglicht die Trennung von Wissen und Modell –
Feintuning integriert Wissen in das Modell.

KI-Verfahren für VS

Auswertung von strukturierten Daten

- Strukturierte Daten (z. B. relationale Tabellen) unterliegen ebenfalls der VS-Einstufung
- Eine VS-Einstufung auf Datensatzebene ist praktisch nicht umsetzbar
- Daten werden zu logisch abgegrenzten Datensätzen (z. B. Exporte, CSV-Dateien) zusammengefasst
- KI-gestützte Auswertung erfolgt ausschließlich innerhalb des definierten VS-Datenflusses

Strukturierte VS-Daten sind für KI geeignet, wenn sie als eingestufte Datenpakete verarbeitet werden – nicht als offener Datenstrom.

KI-Verfahren für VS

Live-Demo

- Bewertungs-Analyse
- Trend-Analyse
- Frage an das Dokument

KI-Verfahren für VS

Zusammenfassung

- KI wird ausschließlich unterstützend eingesetzt und trifft keine fachlichen, klassifizierungs- oder freigaberelevanten Entscheidungen
- Abgeleitete Daten (z. B. Übersetzungen, Analyseergebnisse, Vektoren) sind mindestens mit der Einstufung der zugrunde liegenden VS-Inhalte zu behandeln
- KI-Verarbeitung erfolgt nachvollziehbar, protokolliert und innerhalb definierter Sicherheitsdomänen

KI unterstützt Prozesse – Verantwortung und Entscheidungen verbleiben beim Menschen; der Umgang mit Wissen hängt vom gewählten Verfahren ab.

KI im Umfeld von Verschlusssachen

Fragen und Antworten

Vielen Dank für Ihr Interesse

www.inxire.com