bedag

KI-Sandbox | ChatBE/BE-AI

BFH eGov Lunch 2026/4 vom 16.11.2026

Dominic Salzmann, Bedag

Agenda

1. KI-Sandbox
   - → Was ist die KI-Sandbox mit PoC LLM?
   - → Dedicated Inferencing
   - → Die KI-Sandbox der Staatskanzlei (STA)
2. KIGV@BE
   - → Wie soll es im Kanton Weitergehen
   - → Architekturübersicht
   - → KI-Infrastruktur im ISDS Kontext
3. Fragen & Diskussion

KI-Sandbox mit PoC LLM
Die Infrastruktur

bedag

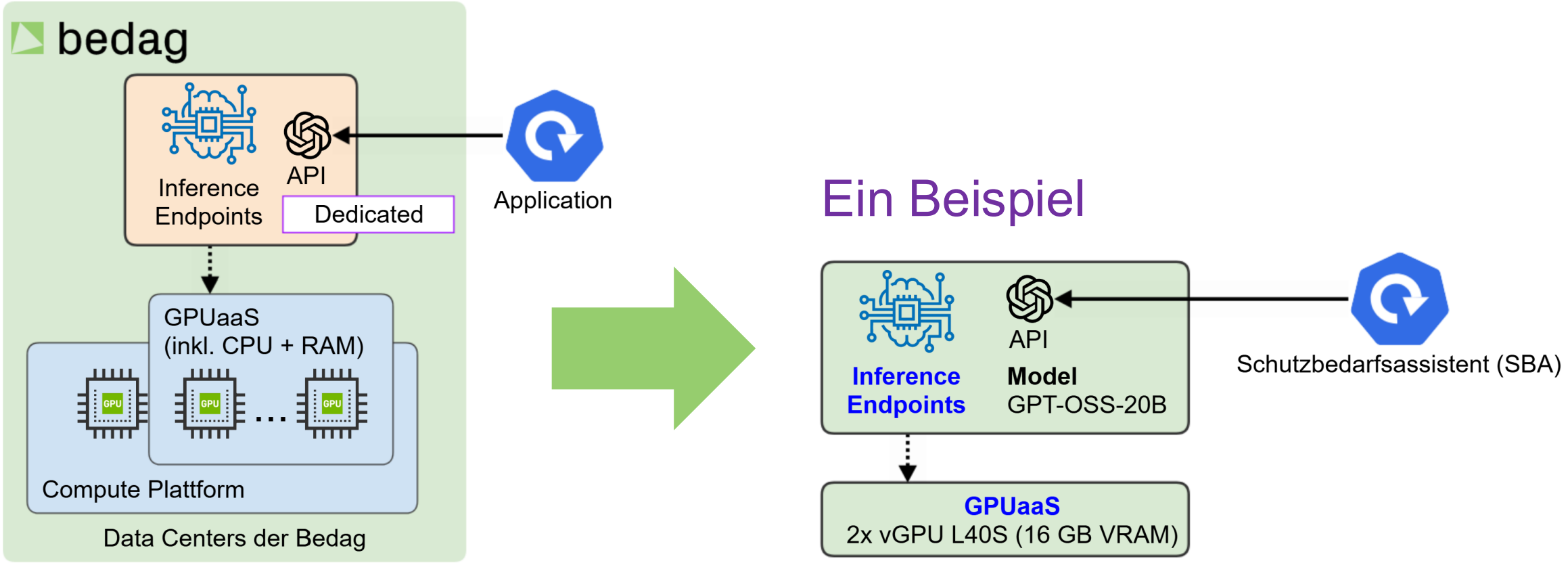
Was ist die KI-Sandbox mit PoC LLM?

- Die Sandbox ist eine dedizierte OnPremise-Infrastruktur für den Betrieb und das Testen von Large Language Models (LLMs)
  - Server-Hardware: Supermicro-Server
  - Grafikkarten fürs Inferencing: NVIDIA L40S
  - Betriebs-Plattform: Kubernetes (K8S)
  - Inferenz-Engine: Virtual Large Language Mode (vLLM)
- Leistungsumfang
  1. Beinhaltet ein fixes kontingent an Inferenz-Leistung
     → Fixe Bundles mit CPU, RAM und GPU
  2. Die Aufwände für den Modellbetrieb
  3. Optionaler Engineering-Support für Projekte im Bereich KI

bedag

Dedicated Inferencing

- Basiert auf unserem Service «Dedicated Inferencing (LLM)»



bedag

Die KI-Sandbox der Staatskanzlei (STA)

- Die Staatskanzlei nutzt aktuell folgende Mengen an Ressourcen

Anzahl GPU (N)	GPU-VRAM (GB)	CPU-Cores (N)	RAM (GB)
3x NVIDIA L40S	144	36 vCores	600

- Im Rahmen ihres Sandbox-Projekts nutzen sie die Sandbox für verschiedene Proof of Concepts im Kanton Bern
  1. Schutzbedarfsassistent
  2. Datenschutzservice (DSS)
  3. Protokoll Composer
  4. ChatBE/BE-AI
  5. …

bedag

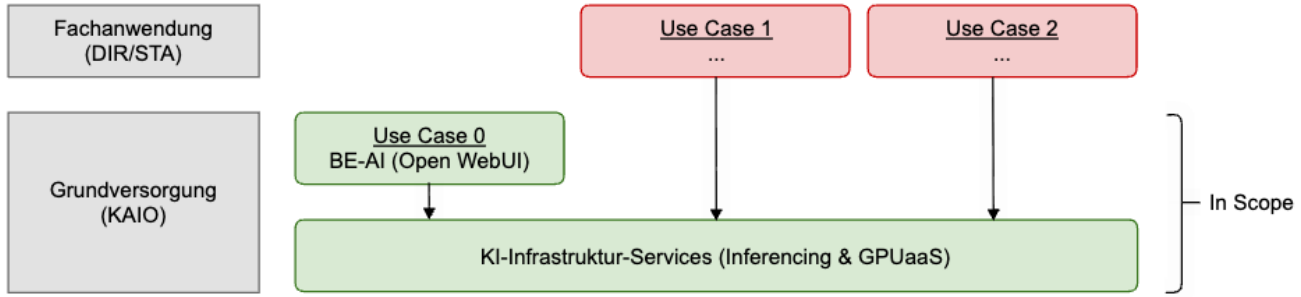
KIGV@BE
«KI-Sandbox vNext»

bedag

Ein paar Hintergrundinformationen zu KIGV@BE

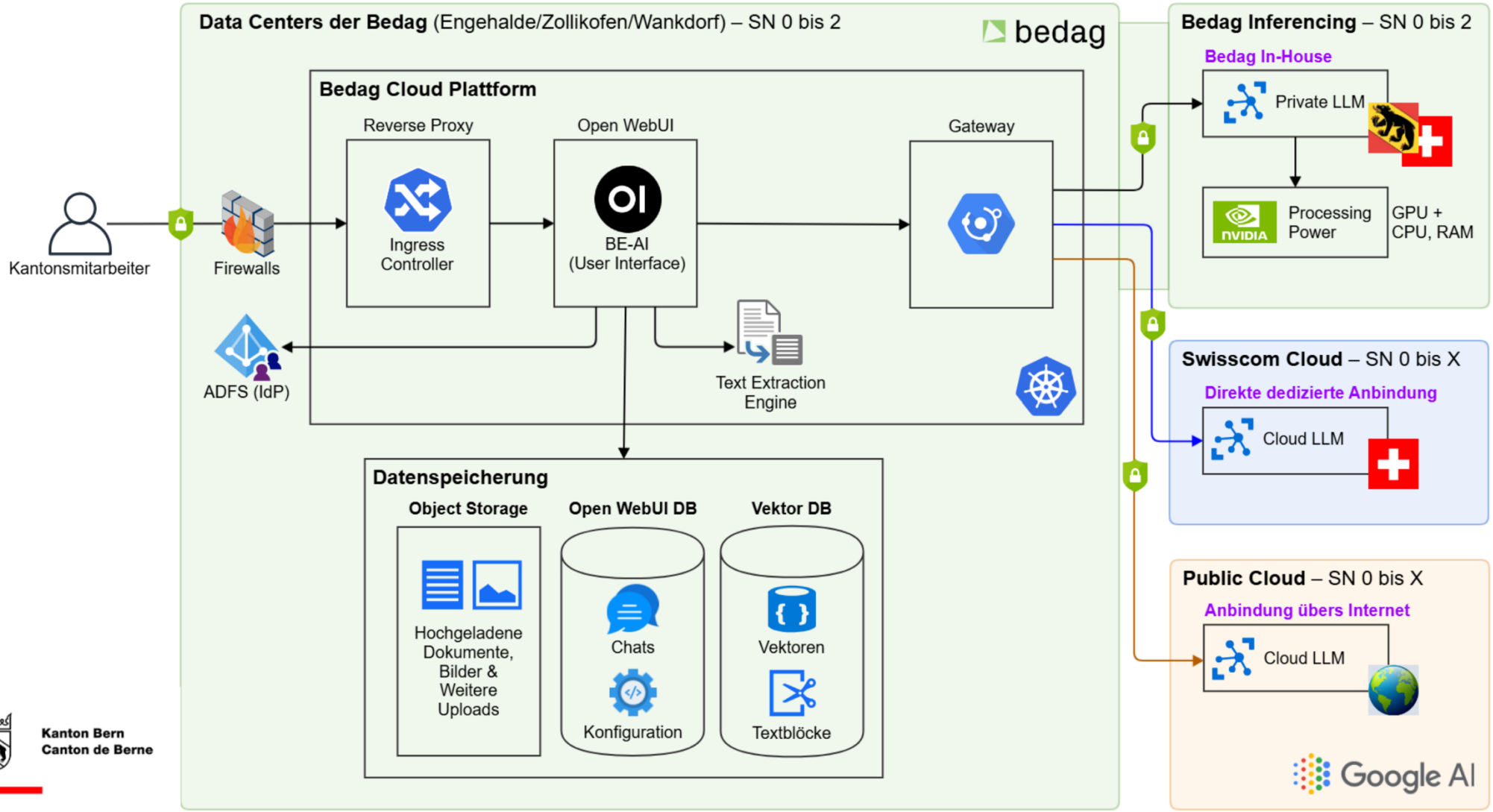
- Den Mitarbeitenden der kantonalen Verwaltung soll generative KI (GenKI) als Teil der Grundversorgung zur Verfügung stehen.
- Damit dies kontrolliert, sicher und für die ganze Verwaltung wiederverwendbar geschieht, wird eine zentral betriebene Plattform aufgebaut.
- Das Vorhaben KIGV@BE baut auf der KI-Sandbox auf, wobei ChatBE als «UseCase 0», mit Priorität umgesetzt wird.

➢ Bis Ende 2026 soll bei der Bedag eine Infrastruktur bereit stehen, über welche Sprachmodelle für Daten der Schutzniveaus 0 bis 2 genutzt werden können.
Ab Januar 2027 wird sie in einem Pilotbetrieb eingesetzt.

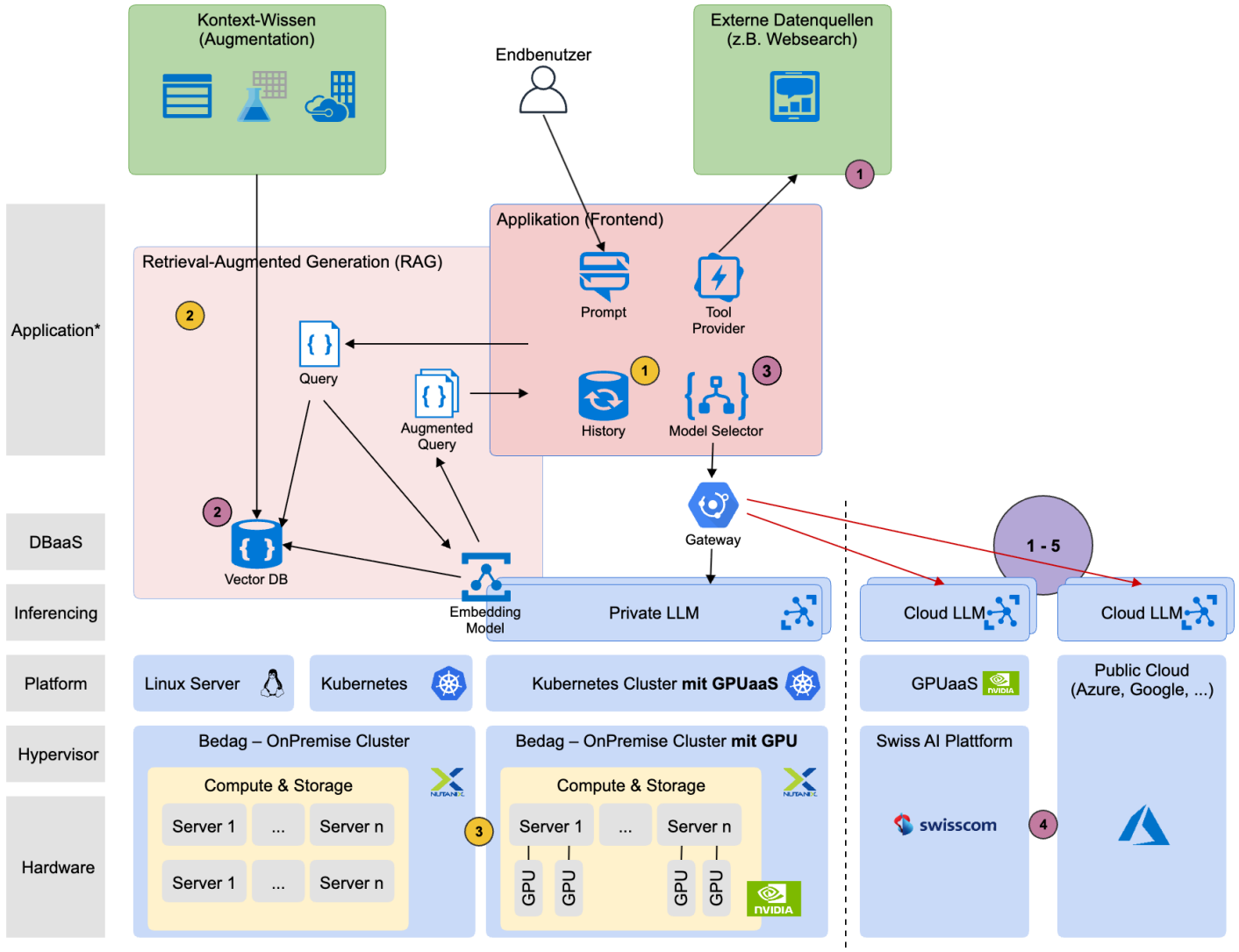


bedag

Architekturübersicht



KI-Infrastruktur im Kontext ISDS



Information Security

Allgemein gelten hier die ordentlichen Service der Bedag gemäss Vertrag Applikationsplattformen (APF) inkl. Grundschutz bis und mit Schutzstufe 2.

1. Persistierte Daten bleiben On-Premise
2. Sekundärdaten ohne spezifischen Anspruch an RTO/RPO
3. Identische Basisinfrastruktur (Grundschutz, APF)
4. Vertäge mit passenden SLA bzgl. IS

Data Security

Betreffend Data Security definieren die Daten den Schutzbedarf und die Verantwortung liegt beim Data Owner.

Beachtensewert:
1. Tool Provider arbeiten in der Regel nicht mit den Rechten des Endbenutzers (Autorisierung)
2. Daten in der Vektor-Datenbank sind Endbenutzern oftmals indirekt zugänglich (Autorisierung)
3. Hybride Modellverfügbarkeit möglich. Welche Daten dürfen wo bearbeitet werden?
4. Welche Rahmenbedingungen gelten für welche Provider?

Souveränität & Datenhoheit

1. Örtliche Datenschutzbestimmungen CH / EU / US
2. Fehlender Zugriff für Analyse und Betrieb
3. Keine (wenig) Kontrolle über Datenverarbeitung
4. Kein direkter Zugriff (fehlende Datenhoheit)
5. Ungeügende Datenschutzgesetzte (z.B. US)

bedag

Fragen & Diskussion



bedag

bedag

Herzlichen Dank!