

KI-Infrastruktur für Unternehmen

Welches Modell, welche Infrastruktur – und was das wirklich kostet.

KI-Strategie Modellvergleich Digitale Souveränität Stand August 2026

Die Architektur

LAYER 1	Interface	Das täglich genutzte Tool. Stabil und einheitlich.	
LAYER 2	Modelle	Claude, Mistral, Apertus, Open-Weight – austauschbar, pro Use Case gewählt.	
LAYER 3	Konnektoren	Web, Office, CRM, APIs – die KI verbindet sich mit dem Unternehmen.	
LAYER 4	Skills & Plugins	Kodifiziertes Firmenwissen. Heben Junioren auf Senior-Niveau.	+ Multiplikator
LAYER 5	Wissensdatenbank	Company Memory – strukturiert abgelegt, über RAG zugänglich.	

Modellvergleich

FRONTIER US	Claude (Anthropic)	Stärkste Modell-Palette, MCP-Ökosystem. US CLOUD Act.
M365-INTEGRIERT	Microsoft 365 Copilot	Natürlich in Office. Schwächer bei mehrstufigen Tasks. US CLOUD Act.
EUROPÄISCH	Mistral (FR)	Kein US CLOUD Act. Günstigste Cloud-Option. On-prem möglich.
SCHWEIZERISCH	Apertus 1.5 (CH)	Open-weight, Schweizer Recht, Schweizerdeutsch. Eigene Infra nötig.
CHINESISCH / OPEN-WEIGHT	Kimi K3, DeepSeek V4, Qwen3, GLM-5	Frontier-Niveau. Governance- und Supply-Chain-Risiken beachten.

Digitale Souveränität

Datenresidenz	Wo liegen die Daten? Lösbar – via 3P (Frankfurt), Mistral (FR) oder Apertus on-prem (CH).
Jurisdiktion	Wer kann Zugriff erzwingen? US CLOUD Act gilt für alle US-Unternehmen, unabhängig vom Server-Standort.

Vorgehen

- Projektverantwortliche bestimmen
Führungsaufgabe, nicht IT.
- Use Cases priorisieren
2–3 konkrete Use Cases mit messbarem Mehrwert.
- Modelle im eigenen Kontext testen
Realität schlägt Benchmark.
- Skills entwickeln & Teams schulen
AI Champions benennen und befähigen.

5 PRINZIPIEN FÜR EINE ROBUSTE KI-STRATEGIE
01 KI ist ein Stack, kein Produkt. Wer die Schichten trennt, bleibt frei.
02 Wer auf einen einzigen Vendor setzt, bindet seine Strategie an dessen Roadmap. Modellverfügbarkeit ist ein geopolitisches Risiko.
03 Der Weg zum Modell entscheidet die Qualität: Direkte Interfaces schlagen Middleware bei komplexen Aufgaben.
04 Souveränität auf zwei Achsen denken: Residenz ist lösbar, Jurisdiktion ist der härtere Teil.
05 Die billige Abo-Ära endet. Self-hosted Open-Weight für Volumen, Frontier-Cloud für Komplexität.