WHITEPAPER

Der souveräne KI-Stack

Generative KI in Organisationen einsetzen, ohne neue Abhängigkeiten zu schaffen

Ein Orientierungsrahmen für IT-Verantwortliche im öffentlichen Sektor und im Mittelstand

WE SUCCESS Consulting GmbH
Digitale Transformation · Strategieberatung · Stuttgart
Version 1.0 · 2026

Inhalt

1 Management Summary 3
2 Das Souveränitätsdilemma 3
3 Was „souverän“ wirklich bedeutet 4
4 Referenzarchitektur eines souveränen KI-Stacks 4
5 Die Bausteine im Einzelnen 5
5.1 Modelle: groß genug, nicht größer 5
5.2 Inferenz: die unterschätzte Schicht 5
5.3 Wissensschicht (RAG): hier entsteht der Wert 6
5.4 Orchestrierung und Oberfläche 6
6 Build, Buy oder Hybrid – die Entscheidungslogik 6
7 Ein Stufenplan für den Aufbau 7
8 Wirtschaftlichkeit ohne Illusionen 7
9 Typische Fallstricke 8
10 Fazit und nächste Schritte 8

1 Management Summary

Generative KI ist in den meisten Organisationen längst kein Pilotthema mehr, sondern eine Erwartung – von Mitarbeitenden, von Führungskräften, zunehmend auch von Bürgerinnen und Bürgern. Gleichzeitig entsteht eine unbequeme Spannung: Wer den schnellsten Weg zum Nutzen wählt, schiebt seine sensibelsten Daten und einen Teil seiner Handlungsfähigkeit zu einer Handvoll außereuropäischer Plattformanbieter. Was als Produktivitätsgewinn beginnt, kann sich als neue, schwer reversible Abhängigkeit verfestigen.

Dieses Whitepaper beschreibt einen dritten Weg zwischen Verzicht und blinder Übernahme: den souveränen KI-Stack. Gemeint ist eine Architektur, die den produktiven Nutzen generativer KI erschließt, dabei aber Kontrolle über Daten, Betrieb und technologische Optionen in der eigenen Organisation hält. Souveränität wird hier nicht als ideologische Maximalforderung verstanden, sondern als nüchterne Risikosteuerung: die Fähigkeit, Anbieter zu wechseln, Daten zu schützen und Entscheidungen umzukehren.

Die Kernaussagen vorweg:

- Souveränität ist kein binärer Zustand, sondern ein Spektrum aus mehreren Ebenen – Daten, Betrieb, Technologie und Recht –, die sich getrennt steuern lassen.
- Open-Weight-Modelle haben die Lücke zur Spitze so weit geschlossen, dass für die überwiegende Mehrheit realer Anwendungsfälle keine proprietäre Cloud-API mehr zwingend nötig ist.
- Der größte Hebel liegt nicht im Modell, sondern in der Wissensschicht: gut kuratierte, sauber angebundene eigene Dokumente schlagen ein größeres Modell mit schlechtem Kontext fast immer.
- Der wirtschaftlich tragfähige Einstieg ist fast nie „alles selbst hosten“, sondern eine bewusst gewählte Hybrid-Architektur mit klaren Abbruchkanten.

Der Rest dieses Dokuments liefert das Begriffsgerüst, eine Referenzarchitektur, ein Reifegradmodell für den schrittweisen Aufbau sowie die typischen Fallstricke, die Projekte in der Praxis ausbremsen.

2 Das Souveränitätsdilemma

Die Anziehungskraft der großen KI-Plattformen ist rational. Sie bieten Spitzenmodelle, sofort verfügbar, ohne Investition in Hardware oder Betrieb, abgerechnet pro Nutzung. Für ein erstes Experiment ist das die offensichtlich richtige Wahl. Das Problem entsteht erst dort, wo aus dem Experiment ein Geschäftsprozess wird.

Drei Mechanismen verwandeln Bequemlichkeit in Abhängigkeit. Erstens der Datenabfluss: Sobald Fachverfahren, Vertragstexte, Personaldaten oder Bürgerinformationen durch eine externe API laufen, verlässt die Organisation den Bereich, in dem sie technische und rechtliche Kontrolle ausübt – unabhängig von Zusicherungen in den Geschäftsbedingungen. Zweitens der Lock-in: Prompts, Workflows und Integrationen werden auf die Eigenheiten eines Anbieters zugeschnitten und lassen sich später nur mit Aufwand lösen. Drittens das politische Risiko: Preise, Konditionen, Verfügbarkeit und sogar der Marktzugang können sich aus geopolitischen oder regulatorischen Gründen kurzfristig ändern – eine Variable, die außerhalb des eigenen Einflusses liegt.

Für europäische Organisationen verschärfen zwei Rahmenbedingungen das Bild. Der öffentliche Sektor unterliegt einem Anspruch auf digitale Souveränität, der inzwischen politisch und teils regulatorisch unterlegt ist. Und alle Organisationen bewegen sich im Spannungsfeld von DSGVO und dem in Stufen wirksam werdenden EU AI Act, die beide Nachvollziehbarkeit darüber verlangen, wo und wie personenbezogene Daten verarbeitet werden. Die naheliegende Antwort – „dann nutzen wir eben keine KI“ – ist keine: Sie tauscht ein steuerbares Risiko gegen einen sicheren Wettbewerbsnachteil.

3 Was „souverän“ wirklich bedeutet

Der Begriff Souveränität wird in der Debatte oft als Alles-oder-Nichts verwendet – und verfehlt damit die Praxis. Nützlicher ist ein Ebenenmodell, in dem sich vier Dimensionen unabhängig voneinander einstellen lassen. Eine Organisation kann auf einer Ebene hohe Souveränität verlangen und kann auf einer anderen Ebene bewusst Kompromisse eingehen.

Ebene	Leitfrage	Typischer Hebel
Daten	Verlassen sensible Daten je den eigenen Kontrollbereich?	On-Premise- oder EU-Hosting, Anonymisierung
Betrieb	Wer betreibt das System und kann es abschalten oder ändern?	Eigener Betrieb vs. verwalteter Dienst
Technologie	Könnten wir den Anbieter mit vertretbarem Aufwand wechseln?	Offene Modelle, offene Schnittstellen
Recht	Welcher Rechtsraum und welche Behörde haben Zugriff?	Vertragsgestaltung, Gerichtsstand, Zertifizierung

Die wichtigste Unterscheidung in der technologischen Ebene betrifft den Begriff „offen“. Open-Weight-Modelle stellen die trainierten Gewichte zum Herunterladen und lokalen Betrieb bereit – das genügt, um sie unabhängig von einem Anbieter auszuführen. Das ist nicht dasselbe wie quelloffen im engeren Sinn, bei dem auch Trainingsdaten und Trainingscode offenliegen. Für den praktischen Souveränitätsgewinn – die Fähigkeit, ein Modell selbst zu betreiben und den Anbieter zu wechseln – ist die Verfügbarkeit der Gewichte der entscheidende Faktor, nicht die Reinheit der Lizenz. Diese Differenzierung lohnt sich, weil sie verhindert, dass Projekte an einer akademischen Definition scheitern, während der pragmatische Nutzen längst erreichbar wäre.

4 Referenzarchitektur eines souveränen KI-Stacks

Ein souveräner KI-Stack ist modular. Jede Schicht erfüllt eine klar abgegrenzte Aufgabe und lässt sich einzeln austauschen – genau das ist die technische Grundlage von Souveränität. Von unten nach oben:

Schicht	Aufgabe
Infrastruktur	Rechenkapazität für Inferenz – GPU-Server im eigenen Rechenzentrum, bei einem EU-Anbieter oder in einer souveränen Cloud.
Modell	Das oder die Sprachmodelle, die Texte verstehen und erzeugen. In der Regel mehrere Modelle unterschiedlicher Größe für unterschiedliche Aufgaben.
Inferenz	Die Laufzeitumgebung, die Modelle effizient ausführt und über eine standardisierte Schnittstelle bereitstellt.
Wissen (RAG)	Die Anbindung eigener Dokumente, damit das Modell auf organisationsspezifischem Wissen statt nur auf Trainingswissen antwortet.
Orchestrierung	Die Logik, die Anfragen an das passende Modell und die passende Wissensquelle leitet, Rechte prüft und Abläufe steuert.
Oberfläche	Der Zugang für Nutzende – Chat-Oberfläche, Integration in Fachanwendungen oder Programmierschnittstelle.

Zwischen den Schichten stehen offene, standardisierte Schnittstellen. Wenn die Inferenzschicht eine etablierte API-Konvention bedient, kann das darüberliegende System ohne Umbau zwischen einem selbst gehosteten Modell und – falls nötig – einem externen Dienst wechseln. Diese Austauschbarkeit ist der eigentliche Wert der Architektur. Sie macht Souveränität von einer einmaligen Entscheidung zu einer dauerhaften Option.

5 Die Bausteine im Einzelnen

5.1 Modelle: groß genug, nicht größer

Die verbreitetste teure Fehlannahme lautet, man brauche stets das größte verfügbare Modell. In der Praxis sind kleinere, lokal betreibbare Open-Weight-Modelle für klar umrissene Aufgaben – Zusammenfassen, Klassifizieren, Beantworten von Fragen zu eigenen Dokumenten – oft vollkommen ausreichend, deutlich günstiger und schneller. Sinnvoll ist meist ein Portfolio: ein leistungsfähigeres Modell für anspruchsvolle Aufgaben, ein schlankes für Routine. Welche konkreten Modelle gerade führend sind, ändert sich im Quartalsrhythmus; die Architektur sollte deshalb so gebaut sein, dass ein Modellwechsel eine Konfigurations- und keine Projektfrage ist.

5.2 Inferenz: die unterschätzte Schicht

Die Inferenzschicht entscheidet darüber, ob ein lokal betriebenes Modell wirtschaftlich läuft. Für den Einzelplatz- oder Testbetrieb genügen schlanke Werkzeuge, die ein Modell mit minimalem Aufwand bereitstellen. Für den produktiven Mehrbenutzerbetrieb braucht es eine Laufzeitumgebung, die parallele Anfragen effizient bündelt und die teure GPU-Kapazität auslastet. Die Wahl zwischen beiden ist eine der folgenreichsten – und am häufigsten übersprungenen – Entscheidungen im Aufbau.

5.3 Wissensschicht (RAG): hier entsteht der Wert

Retrieval-Augmented Generation verbindet das Sprachmodell mit den eigenen Dokumenten: Zu jeder Frage werden die passenden Stellen aus dem eigenen Bestand gesucht und dem Modell als Kontext mitgegeben. Damit antwortet das System auf Basis des aktuellen, organisationsspezifischen Wissens statt nur aus dem allgemeinen Training. Der Aufwand verschiebt sich damit von der Modellfrage hin zur Datenqualität: saubere Dokumentenaufbereitung, sinnvolle Segmentierung, korrekte Behandlung von Sonderfällen wie zusammengesetzten Dateinamen oder Umlauten, und – sicherheitskritisch – eine Rechteprüfung, die sicherstellt, dass niemand über den Umweg der KI Zugriff auf Dokumente erhält, die er regulär nicht sehen dürfte. Erfahrungsgemäß steckt hier der Großteil der eigentlichen Projektarbeit.

5.4 Orchestrierung und Oberfläche

Die Orchestrierungsschicht ist die Stelle, an der Governance technisch greifbar wird: Routing zwischen Modellen, Durchsetzung von Zugriffsrechten, Protokollierung für Nachvollziehbarkeit, Trennung von Arbeitsbereichen nach Mandant oder Fachbereich. Die Oberfläche schließlich entscheidet über die Akzeptanz. Eine vertraute Chat-Oberfläche senkt die Einstiegshürde; der höhere Nutzen entsteht jedoch dort, wo KI nicht als separates Werkzeug erscheint, sondern direkt in die bestehenden Fachanwendungen eingebettet ist.

6 Build, Buy oder Hybrid – die Entscheidungslogik

Die Frage „selbst bauen oder einkaufen“ wird zu oft als Glaubensfrage geführt. Tatsächlich ist sie pro Anwendungsfall zu beantworten, entlang zweier Achsen: der Sensibilität der Daten und der Geschäftskritikalität des Prozesses.

Konstellation	Empfehlung	Begründung
Unkritische Daten, Experiment	Externer Dienst (Buy)	Schnell, günstig im Test, kein Souveränitätsrisiko bei unsensiblen Daten.
Sensible Daten, hohe Kritikalität	Eigener Betrieb (Build)	Volle Kontrolle über Daten und Verfügbarkeit rechtfertigt den Aufwand.
Gemischt / im Wandel	Hybrid	Souveräner Kern für Sensibles, externe Spitzenleistung für den Rest – mit klaren Grenzen.

Für die meisten Organisationen ist der realistische Einstieg ein Hybrid mit eindeutig definierten Abbruchkanten: Sensible Verarbeitung läuft auf dem souveränen Kern, unkritische oder besonders anspruchsvolle Aufgaben dürfen einen externen Dienst nutzen – aber die Regel, welche Daten welche Grenze überschreiten dürfen, ist explizit, technisch durchgesetzt und protokolliert. Souveränität bedeutet hier nicht, auf externe Dienste vollständig zu verzichten, sondern jederzeit zu wissen und zu steuern, welche Daten welchen Weg nehmen – und den Schalter umlegen zu können.

7 Ein Stufenplan für den Aufbau

Souveräne KI lässt sich nicht in einem Sprung einführen, ohne Budget und Akzeptanz zu überfordern. Bewährt hat sich ein Aufbau in vier Stufen, bei dem jede Stufe für sich genommen bereits Nutzen stiftet.

1. **Orientierung.** Anwendungsfälle erheben und nach Datensensibilität und Nutzen priorisieren. Ergebnis ist eine kurze Liste, nicht eine Plattform. Diese Stufe kostet vor allem Denkarbeit.
2. **Kontrollierter Pilot.** Ein einzelner, klar umrissener Anwendungsfall mit echten, aber überschaubaren Daten – betrieben auf vorhandener oder klein dimensionierter Hardware. Ziel ist Lernen über Datenqualität und Akzeptanz, nicht Skalierung.
3. **Souveräner Kern.** Aufbau der dauerhaften Architektur für sensible Anwendungsfälle – mit produktiver Inferenz, sauberer Wissensanbindung, Rechte- und Protokollierungskonzept. Ab hier wird Betrieb zum Thema.
4. **Hybride Skalierung.** Erweiterung um weitere Anwendungsfälle und gezielte Einbindung externer Dienste für unkritische Aufgaben – innerhalb der unter Punkt 6 definierten Grenzen und Governance.

Der häufigste Fehler besteht darin, mit Stufe drei zu beginnen: erst die Plattform zu bauen und dann nach Anwendungsfällen zu suchen. Das führt zu teuren Systemen ohne Nutzer. Die umgekehrte Reihenfolge – erst der nachgewiesene Bedarf, dann die Infrastruktur – ist sowohl wirtschaftlich als auch politisch tragfähiger.

8 Wirtschaftlichkeit ohne Illusionen

Die Frage nach den Gesamtkosten wird oft auf den Vergleich „Cloud-API pro Anfrage gegen einmaliger GPU-Kauf“ verkürzt. Das greift zu kurz. Eine ehrliche Betrachtung umfasst neben der Hardware auch Strom und Kühlung, Betrieb und Wartung, Personalkompetenz sowie die Kosten der Datenaufbereitung – die in fast allen Projekten den größten und am häufigsten unterschätzten Posten bilden.

Der wirtschaftliche Vorteil des eigenen Betriebs tritt erst ab einem gewissen, konstanten Nutzungsvolumen ein: Solange Anfragen selten und unregelmäßig sind, ist die nutzungsbasierte Abrechnung externer Dienste günstiger; bei hohem, planbarem Dauerbetrieb kehrt sich das Verhältnis um. Der Souveränitätswert – Datenschutz, Unabhängigkeit, Reversibilität – ist dabei kein Posten in der Tabelle, sondern eine Risikoabsicherung, deren Wert sich erst im Ernstfall zeigt. Wer souveräne KI allein über die Kosten pro Anfrage rechtfertigt, argumentiert auf dem falschen Feld; der eigentliche Gegenwert ist die vermiedene Abhängigkeit.

Diese Einschätzungen ersetzen keine individuelle Wirtschaftlichkeitsrechnung. Sie hängt von Volumen, Datenarten, vorhandener Infrastruktur und regulatorischen Anforderungen ab und sollte für die konkrete Organisation aufgestellt werden.

9 Typische Fallstricke

- **Modell-Fixierung.** Energie fließt in die Wahl des größten Modells, während die Wissensschicht vernachlässigt wird – die über die Qualität tatsächlich entscheidet.
- **Plattform vor Anwendungsfall.** Eine große Infrastruktur wird gebaut, bevor ein konkreter, nachgefragter Nutzen feststeht.
- **Rechte als Nachgedanke.** Zugriffsrechte werden erst spät bedacht – mit dem Risiko, dass die KI Dokumente zugänglich macht, die ein Nutzer regulär nicht sehen dürfte.
- **Datenqualität unterschätzt.** Der Aufwand für Aufbereitung, Segmentierung und Pflege der Dokumente wird zu niedrig angesetzt.
- **Betrieb vergessen.** Ein Pilot läuft, aber niemand hat geklärt, wer das System dauerhaft betreibt, aktualisiert und überwacht.
- **Scheinsouveränität.** Ein „europäisches“ Label wird akzeptiert, ohne zu prüfen, welcher Rechtsraum und welche Behörde tatsächlich Zugriff auf die Daten haben.

10 Fazit und nächste Schritte

Souveräne KI ist eine Frage bewusster Gestaltung: den Nutzen generativer KI erschließen und zugleich die Kontrolle über Daten, Betrieb und Optionen behalten. Der Weg dorthin führt nicht über die größte Plattform, sondern über eine modulare, austauschbare Architektur, eine sorgfältig gepflegte Wissensschicht und einen Aufbau in Stufen, bei dem der nachgewiesene Bedarf der Infrastruktur vorausgeht.

Der sinnvollste erste Schritt ist klein: ein halbtägiger Standortbestimmungs-Workshop, der die zwei bis drei Anwendungsfälle mit dem besten Verhältnis aus Nutzen und Datensensibilität identifiziert – und damit die Grundlage für jede weitere Entscheidung legt.

Über WE SUCCESS Consulting GmbH

WE SUCCESS Consulting begleitet Organisationen im öffentlichen Sektor und im Mittelstand bei digitaler Transformation, digitaler Souveränität und dem praxistauglichen Einsatz von KI – von der Strategie bis zur konkreten Architektur. Für einen Standortbestimmungs-Workshop oder ein unverbindliches Gespräch erreichen Sie uns über die im Briefkopf angegebenen Kontaktdaten.

WE SUCCESS Consulting GmbH
Geschäftsführer Andreas Pelzner
Wachaustr. 11
70469 Stuttgart

📱 +49 170 6368503
✉️ andreas@we-success.de
🔗 www.we-success.de
🔗 www.linkedin.com/in/andreas-pelzner