IT-PLANUNGSRAT

Berlin
04.02.2026

Nachnutzung generativer KI-Systeme

Vorstudie zu Nachnutzungspotentialen bestehender generativer KI-Systeme der öffentlichen Verwaltung

Version	Datum	Autor:in	Aktion
1.0	19.12.2025	Jens Tiemann, Dorian Wachsmann Kompetenzzentrum Öffentliche IT (ÖFIT) am Fraunhofer FOKUS https://www.oeffentliche-it.de/	Fertigstellung
1.0	04.02.2026	IT-Planungsrat	Veröffentlichung

Kompetenzzentrum Öffentliche IT

Gefördert durch:

Bundesministerium des Innern

aufgrund eines Beschlusses des Deutschen Bundestages

Inhaltsverzeichnis

Das Wichtigste in Kürze

In dieser Vorstudie geht es um die Möglichkeiten der Nachnutzung und Vereinheitlichung von KI-Systemen. Dafür werden sieben Systeme betrachtet, die Funktionen großer Sprachmodelle (LLMs) für verschiedene Bereiche der Verwaltung in Bund, Ländern und Kommunen verfügbar machen.

Die Systeme ähneln sich weitgehend in ihren Anwendungsfällen und weisen viele technische Gemeinsamkeiten auf. Dazu zählen die konsequente Umsetzung von Microservice-Architekturen sowie die Nutzung ähnlicher Open-Source-Frameworks.
Die Hebung von Nachnutzungspotenzialen und der Aufbau eines gemeinsamen KI-Ökosystems verspricht der Verwaltung Vorteile. Eine „einheitliche KI-Plattform" für alle, die von der Mehrheit der Akteure getragen wird, erscheint derzeit jedoch unrealistisch. Das liegt daran, dass

- die Dynamik im Bereich KI Parallelentwicklungen fast unvermeidbar macht,
- die Einzelprojekte häufig weit fortgeschritten sind und bereits genutzt werden und
- die Lösungen unterschiedliche Bedarfe und Nutzergruppen ansprechen.

Die Etablierung einer Referenzarchitektur ist jedoch ein realistisches Ziel, für das es auch schon Ansätze gibt. Prioritär sollte angestrebt werden, Microservice-Komponenten austauschbar zu machen, sodass:

- unterschiedliche Projekte spezifische Anwendungsfälle entwickeln können und
- diese mit geringem Aufwand in andere Lösungen integriert werden können.

Dafür sind ein einheitliches Systemverständnis und standardisierte Schnittstellen notwendig. Darauf folgt die gemeinsame Weiterentwicklung genutzter Open Source Software und eine regelmäßige Sicherheitsüberprüfung kritischer Komponenten.
Somit sehen wir die größte Chance in verschiedenen, auf spezifische Bedarfe zugeschnittene Lösungen, die aber technisch ähnlich und vergleichbar organisiert sind. Diese sollten in produktivem Austausch stehen und so gemeinsam voneinander profitieren.

Zum Aufbau eines KI-Ökosystems wurden in dieser Vorstudie die folgenden Handlungsempfehlungen entwickelt, die sowohl an der Anbieter- als auch an der Nutzerperspektive ansetzen:

- Begriffe und Systemstrukturen schärfen
- Konvergente Architekturen vorantreiben
- Nutzergruppen und deren Bedarfe systematisch erfassen
- Compliance als Querschnittsthema operationalisieren

Die Autoren danken dem Kompetenzteam „Künstliche Intelligenz“ im Schwerpunktthema Datennutzung des IT-Planungsrats, den Ansprechpartnern der genannten GenKI-Lösungen und den Vertretern von KIVA.arc für ihre Unterstützung bei dieser Untersuchung.
Die Eigenschaften der GenKI-Lösungen wurden bis Oktober 2025 erfasst. Aktuell verfügbare und geplante Eigenschaften einzelner Lösungen klären Sie am besten direkt mit den jeweiligen Anbietern.

1 Ziele der Vorstudie

Die Nutzung von KI-Systemen in der Verwaltung hat durch die rasante Entwicklung insbesondere generativer KI in den letzten Jahren stark zugenommen. Dabei lässt sich beobachten, dass in verschiedenen Teilen der öffentlichen Verwaltung auch immer mehr eigene Systeme entwickelt, pilotiert und eingesetzt werden, um den eigenen Bedarfen gerecht zu werden.
Bei der Nutzung von Anwendungen generativer KI in der Verwaltung besteht das Potenzial, ein qualitativ hochwertiges KI-Ökosystem zu entwickeln, bevor zu viele Parallelentwicklungen unnötige Ressourcen verbrauchen und Doppelstrukturen hervorrufen. Vorteile wie Skaleneffekte und Kosteneffizienz durch nachnutzbare Software sprechen dafür, dass die Verwaltung auf kommunaler, Landes- und Bundesebene die Prinzipien Nachnutzung und "Einer-für-Alle" (EfA) auch für KI-Anwendungen verfolgen sollte. So könnten hohe Fixkosten und knappe Ressourcen durch eine gemeinsame Nutzung von KI-Lösungen und Infrastruktur verringert werden. Gleichzeitig wird die drohende Abhängigkeit von ausländischen Akteuren minimiert, was die Handlungsfähigkeit und Sicherheit der öffentlichen Verwaltung stärkt.

1.1 Fragestellung

In Kooperation mit dem Kompetenzzentrum Künstliche Intelligenz im Schwerpunktthema Datennutzung des IT-Planungsrats hat das Kompetenzzentrum Öffentliche IT (ÖFIT) in einer Vorstudie untersucht, inwieweit bestehende GenKI-Lösungen kompatible KI-Architekturen und interoperable IT-Infrastrukturen nutzen und welche Empfehlungen zum weiteren Vorgehen für eine nachfolgende Studie gegeben werden können.
Das Ziel dieser Vorstudie ist eine Bestandsaufnahme von ausgewählten Systemen mit Bezug zu generativer künstlicher Intelligenz, die sich im gesamten föderalen Kontext in Betrieb bzw. in Entwicklung befinden. Diese Bestandsaufnahme soll einen Vergleich in Bezug auf technische, rechtliche, organisatorische und prozessuale Umsetzung zwischen den Lösungen ermöglichen. Dazu werden relevante Eigenschaften der KI-Systeme in Steckbriefen festgehalten und in einer Zusammenfassung wesentliche Gemeinsamkeiten und Unterschiede der betrachteten KI-Systeme hervorgehoben. Das kann als erster Schritt dienen, um perspektivisch ein gemeinsames KI-Ökosystem zu entwickeln, das auf einem oder einer Kombination mehrerer bestehender Systeme basiert.

Die Fragestellungen für den Vergleich sind nach den folgenden Aspekten gruppiert:

- Modell- und Anbieter-Unabhängigkeit
- Lizenzen und Open Source
- Vertrauenswürdigkeit & Ethik
- Modularität
- Daten- und Geheimnisschutz
- Mandantentrennung und Skalierbarkeit

Die insgesamt 33 Fragestellungen aus der Projektskizze finden sich in Anhang "Fragestellungen zur Untersuchung der GenKI-Lösungen" und wurden im Rahmen dieser Vorstudie vereinheitlicht und die Steckbriefstruktur überführt.

Die Vorstudie betrachtet die folgenden, vom Kompetenzteam KI benannten Systeme:

- **KIPITZ** - KI-Plattform des ITZBund
- **PLAIN** - DevOps-Plattform der Auslands-IT des Auswärtigen Amts (Umsetzung durch Bundesdruckerei)
- **LLMoin** - KI-System der Hansestadt Hamburg
- **NRW.Genius** - KI-Plattform des Landes Nordrhein-Westfalen
- **F13** - Open Source KI-System des Landes Baden-Württemberg
- **AIGude** - KI-System des Landes Hessen
- **MUCGPT** - KI-System der Stadt München

1.2 Methodik

Es wurde ein mehrstufiges, flexibles Vorgehen entwickelt, um über qualitativen Interviews mit Projektverantwortlichen möglichst passgenaue Informationen zu gewinnen:

1. Zuerst wurde eine Internet- und Literaturrecherche zu den genannten sieben Projekten durchgeführt, um bereits öffentlich verfügbare Informationen zu nutzen und eine erste Einordnung vorzunehmen.

2. Parallel dazu wurde eine initiale Referenzarchitektur entworfen, sowie ein gemeinsames Begriffsverständnis etabliert, um möglichst genaue Informationen auf dem gleichen Abstraktionsniveau erheben zu können.
3. Auf Basis der Fragestellungen des Kompetenzteams und den entwickelten Begriffsverständnis wurde die Steckbriefstruktur mit den vier Abschnitten „Allgemeines, „Nichtfunktionales", „Funktionales" und „Technik-Stack" entwickelt und mit den bereits gesammelten öffentlich verfügbaren Informationen ausgefüllt.
4. Die initiale Referenzarchitektur aus Schritt 2 sowie die teilweise ausgefüllten Steckbriefe aus Schritt 3 dienten bei der Kontaktaufnahme mit den Projektverantwortlichen dazu, möglichst präzise und effizient das noch fehlende Wissen zu erfragen und die Steckbriefe zu komplettieren.
5. Je nach Bedarf wurden ein oder mehrere qualitative Interviews geführt, welche zur Abstimmung und Finalisierung der Steckbriefe dienten.
6. Zusätzlich wurde eine vereinfachte Systemübersicht entwickelt, welche als Einstieg zu Vergleichen und in die technischen Details der jeweiligen Systeme dienen.

Die Analyse konzentriert sich auf die Identifikation von Gemeinsamkeiten und charakteristischen Unterschieden zwischen den verschiedenen GenKI-Lösungen. Dabei werden insbesondere die Art des Angebots und die technischen Ansätze beleuchtet, um ein fundiertes Verständnis der Vielfalt in diesem Bereich zu schaffen. Das Ergebnis eines solchen Vergleichs ist, dass Unterschiede und Gemeinsamkeiten ersichtlich werden, Zielgruppen und Bedarfe erkannt und somit ein erster Schritt Richtung einer möglichen Konvergenz der Systeme gegangen werden kann.

Die in den Steckbriefen dargestellten Informationen spiegeln naturgemäß einen zeitlich begrenzten Stand wider, der sich aufgrund der schnellen Entwicklungszyklen kontinuierlich wandelt. Zukünftige Planungen werden nur dann einbezogen, wenn sie konkrete Erkenntnisse für die Zielsetzung dieser Vorstudie liefern und zum besseren Verständnis von Gestaltungsoptionen oder Entwicklungsrichtungen beitragen. Aufgrund des Themenumfangs kann sich nur eine Auswahl für diese Vorstudie besonders relevanter Informationen in den Steckbriefen finden. Anhand der Steckbriefe werden besondere Ausrichtungen der GenKI-Lösungen deutlich. Zur Systemauswahl sind die Steckbriefe weder vorgesehen noch geeignet, auch weil es noch weitere GenKI-Lösungen zum Einsatz in der öffentlichen Verwaltung gibt.

Die Nutzung einer GenKI-Anwendung in der öffentlichen Verwaltung ist voraussetzungsreich: Es braucht aus technischer Sicht wesentliche Komponenten wie die Anwendung selbst, ein oder mehrere Sprachmodelle sowie die IT-Infrastruktur. Aus organisatorischer Sicht müssen ggf. externe Dienste eingebunden werden. Auch muss vor dem Wirkbetrieb die notwendige Dokumentation insbesondere zu Fragen des Datenschutzes, der IT-Sicherheit und ggf. zu den Anforderungen der KI-Verordnung vorhanden sein. Die in dieser Vorstudie genannten GenKI-Systeme unterscheiden sich teilweise stark in ihrem Angebot und damit auch in dem Umfang, in dem die hier genannten Aspekte von ihnen abgedeckt werden.

2 Inhaltliche Grundlagen

Dieses Kapitel beschreibt inhaltliche Rahmenbedingungen der Vorstudie und führt die wichtigsten Begrifflichkeiten ein, die, nach unserer Beobachtung, regelmäßig verschiedentlich genutzt werden und so potenziell zu Missverständnissen führen können. Wann immer möglich, haben wir uns an der Referenzarchitektur KI-Plattform für die Öffentliche Verwaltung (KIVA.arc) und deren Verwendung von Begriffen orientiert [KIVA.arc].
Einige der Länder entwickeln und nutzen (geteilte) KI-Infrastrukturen, auf denen die zu betrachtenden (Fach-)Anwendungen aufbauen. Uns geht es jedoch ausschließlich um die Nachnutzungsmöglichkeiten der Anwendung. Ob KI-Infrastrukturen ebenfalls geteilt und länderübergreifend bereitgestellt werden könnten, ist nicht Teil der vorliegenden Untersuchung.

2.1 Begriffsverständnis zu GenKI-Systemen

In dieser Vorstudie werden Systeme verglichen, welche die Nutzung generativer Künstlicher Intelligenz (GenKI) unter Verwendung großer Sprachmodelle (LLMs) für die Verwaltung ermöglichen. LLMs erstellen kontextbezogen neue Texte, indem sie Muster reproduzieren, die zuvor aus großen Datenmengen „erlernt" wurden. Auch wenn die Nutzung anderer GenKI-Anwendungsfälle (beispielsweise Generierung von Bildern) bei einigen der betrachteten Systeme zumindest geplant ist, liegt der Fokus der Vorstudie auf Anwendungen, die LLMs nutzen.
Abbildung 1 zeigt auf, wo sich GenKI und LLMs verorten lassen. Künstliche Intelligenz als ein Oberbegriff vereint eine ganze Reihe von Methoden und Ansätzen mit dem gemeinsamen Ziel, intelligentes Verhalten nachzubilden. Maschinelles Lernen ist dabei eine Teilmenge, die sich besonders in den letzten Jahren durch die Fortschritte durch Neuronale Netzwerke als besonders erfolgreich herausgestellt hat. Deep Learning beschreibt dabei den Ansatz, tiefe neuronale Netze mit mehreren Schichten zu konstruieren. Die Abbildung verdeutlicht, dass GenKI lediglich ein Ansatz im großen Feld der KI ist, der derzeit allerdings als sehr erfolgreich wahrgenommen wird.

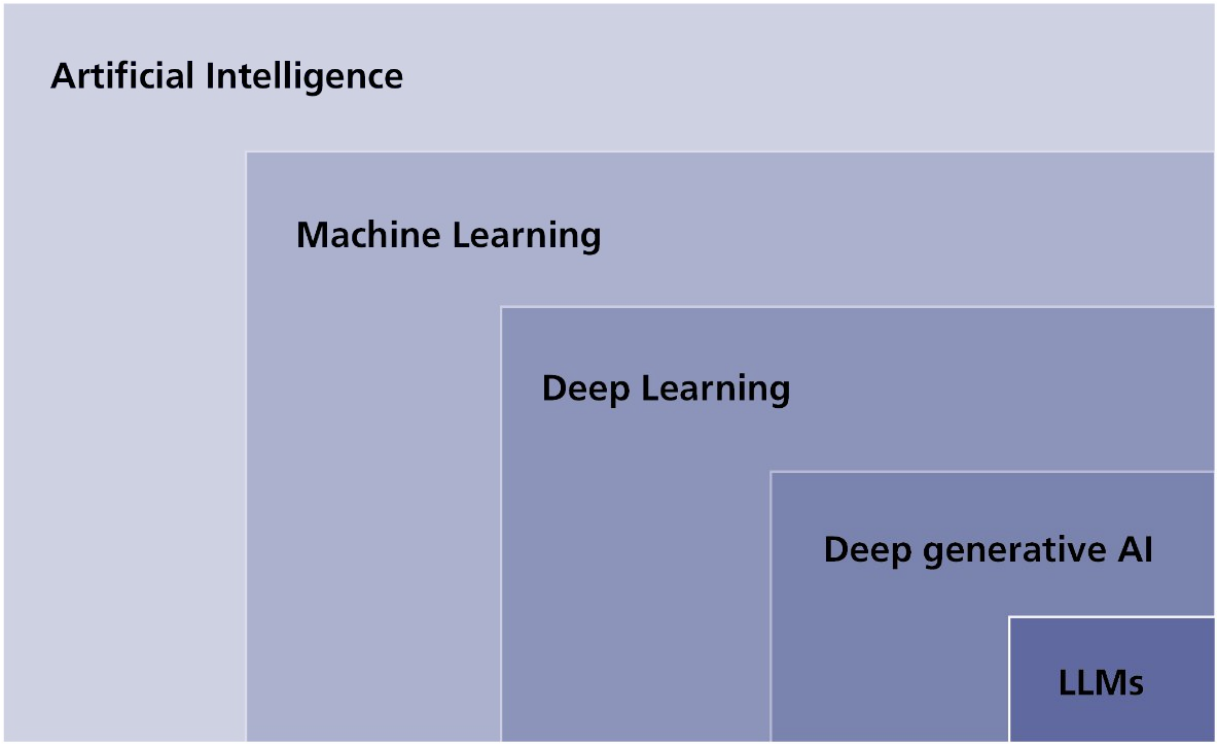


Abbildung 1: Fokus der Vorstudie

Im Folgenden sprechen wir von einem vollständigen GenKI-System, wenn es (mindestens) aus den drei Layern **Frontend**, **Services** und **Inferenz** (siehe Abbildung 2) besteht.
Unter **Frontend** verstehen wir die Bedienoberflächen für Nutzer:innen und Administrator:innen, um auf die implementierten KI-Services oder auch eine Kombination aus Services zuzugreifen. Der **Inferenz**-Layer ist für die Verwaltung und Ausführung der LLMs verantwortlich.
Services können sein:

- Retrieval Augmented Generation (RAG): LLMs werden um einen Dokumentenkorpus (in Form einer vektorbasierten Datenbasis) ergänzt, um die Antwortqualität für bestimmte Kontexte stark zu verbessern.
- MCP-Server (Model Context Protocol) bieten einen standardisierten Zugang zu externen Datenquellen oder softwarebasierten Tools.
- Tools: Softwarebasierte Werkzeuge, mit welchen innerhalb einer Anwendung die LLM-Funktionalität ergänzt werden kann. Typischerweise für Aufgaben gedacht, für die ein LLM nicht optimiert ist (beispielsweise Taschenrechner).

- Task based Services: Microservices, welche spezielle Funktionalitäten auf Basis von LLMs bereitstellen, zum Beispiel freier Chat, Zusammenfassungen, Übersetzungen.
- Agents: Agenten erfüllen (semi-)autonom komplexere Aufgaben und nutzen dafür ggf. auch andere Services.

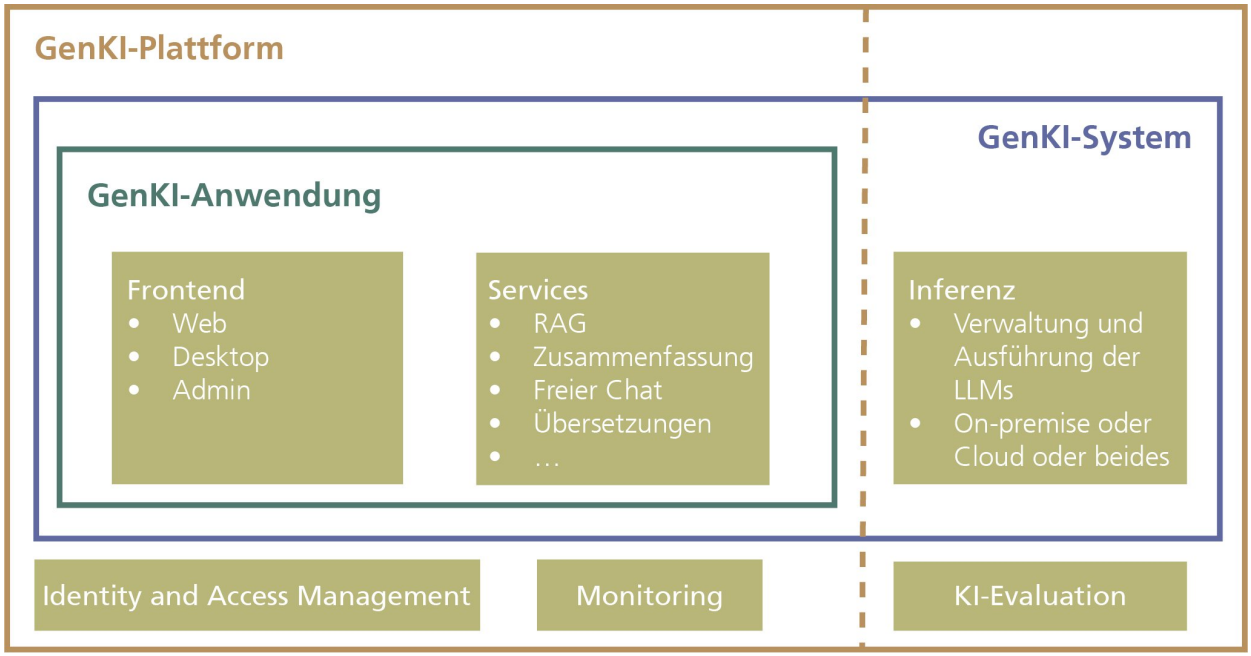


Abbildung 2: Begriffsklärung: GenKI-Plattform, -System, -Anwendung

Eine GenKI-Anwendung ist eine Teilmenge des GenKI-Systems, bestehend aus der Kombination von einem oder mehrerer Services in Verbindung mit einem Frontend. Ein Beispiel für eine GenKI-Anwendung ist eine Chatanwendung realisiert als Webanwendung, mit implementierter Schnittstelle zum Inferenz-Layer. Die GenKI-Anwendung benötigt nach diesem Verständnis eine – wie auch immer geartete – KI-Infrastruktur, welche für die Inferenz verantwortlich ist. Die Kombination von GenKI-Anwendung und Inferenz komplettiert das GenKI-System. Sowohl Anwendung als auch KI-Infrastruktur können on-premise oder cloudbasiert gehostet werden. Die KI-Infrastruktur stellt ein oder mehrere LLMs bereit und kann dabei getrennt von der GenKI-Anwendung (weiter-)entwickelt und betrieben oder bei Bedarf ausgetauscht werden.

Eine Plattform verstehen wir als eine „digitale Grundlage“, auf der entwickelt, betrieben oder interagiert werden kann. Somit kann eine GenKI-Anwendung (auch neben anderen Anwen-

dungen) Teil einer Plattform sein und kann auf dieser genutzt werden. Die Plattform stellt weitere Funktionalitäten zum praktischen Betrieb zur Verfügung, wie die Verwaltung von Zugriffsrechten (Identity and Access Management) und Werkzeuge zum Monitoring. Sie kann ebenfalls den Inferenz-Layer implementieren und LLMs bereitstellen oder diese über eine externe KI-Infrastruktur beziehen.

Eine GenKI-Lösung ist ein konkretes Produkt, entweder eine Anwendung, System oder Plattform. Die meisten der betrachteten GenKI-Lösungen verstehen sich selbst als Plattform, wobei die Breite der bereitgestellten Funktionalitäten variiert. Dazu braucht es immer die Plattform-Software, die beispielsweise als Open Source bereitgestellt werden kann und Hardware, also die Infrastruktur, auf der die Plattform-Software betrieben wird. [ÖFIT]

2.2 Perspektiven auf Vergleichbarkeit

Aus einer technischen Perspektive lässt sich eine Vergleichbarkeit der Lösungen durch die Betrachtung der Architektur, des Technologie-Stacks und den verwendeten Schnittstellen herstellen. Für den Technik-Stack haben wir uns für ein Fünf-Schichten-Modell entschieden. Anders als bei manchen Schichtmodellen (wie zum Beispiel dem TCP/IP-Stack) bauen die Schichten in diesem Modell nicht hierarchisch aufeinander auf. Das Modell hilft jedoch, eine strukturierte Übersicht über die verwendeten Technologien zu gewinnen und dabei insbesondere zwischen KI, Daten und „klassischen" Softwaremodulen zu unterscheiden:

- **Applikationsschicht**: User Interface, Anwendungsentwicklung
- **Technologieschicht**: Monitoring, Container Orchestrierung und Management, Authentifizierung, Schnittstellen und APIs zur Kommunikation zwischen Microservices
- **KI-Schicht**: LLMs, Gateway, Evaluation, KI-Performanz
- **Datenschicht**: Datenbanken und Datenprozesse
- **Compute-Schicht**: Server und Rechenkapazitäten für LLMs und Backend

Dabei ist zu beachten, dass die Compute-Schicht nicht weiter beachtet wurde, wenn das Nachnutzungsangebot eine „reine" Softwarelösung ist.

Die primäre Zielgruppe der betrachteten Systeme ist die deutsche Kernverwaltung des Bundes, der Länder und der Kommunen. Der Fokus auf die verwaltungsinterne Arbeit bringt verwaltungsspezifische Anforderungen an die Systeme mit sich. Diese sind technischer, organisationaler und rechtlicher Natur. Verwaltungsspezifische Compliance bezüglich der Datenver-

arbeitung, dem Datenschutz, der IT-Sicherheit müssen eingehalten werden, hinzu kommen Regelungen aus der KI-Verordnung (KI-VO) [Bitkom].

Alle GenKI-Lösungen werden unter Berücksichtigung einer spezifischen Zielgruppe sowie der geplanten Nutzungsart entwickelt und bereitgestellt. Bei der Bereitstellung unterscheiden wir zwischen zwei verschiedenen Betriebsmodellen, der Mitnutzung und der Nachnutzung.

- **Mitnutzung** bedeutet, es gibt einen zentralen Akteur, welcher die GenKI-Lösung als Software-as-a-Service (Saas) bereitstellt und betreibt. Andere, beispielsweise nachgeordnete Behörden oder Kommunen, unter Umständen auch andere Länder, nutzen gemeinsam aber getrennt nach Mandanten die bereitgestellte Lösung.
- Im Unterschied dazu gibt es bei der **Nachnutzung** keinen zentralen Akteur, der die Lösung bereitstellt. Stattdessen wird die Lösung als Softwarepaket bereitgestellt und Nachnutzende betreiben diese selbst beispielsweise beim eigenen Dienstleister. Die Bereitstellung der KI-Anwendung oder Plattform-Software erfolgt idealerweise als Open-Source.

Eine weitere, stärker formalisierte Möglichkeit ist die **Nutzung nach dem EfA-Prinzip** („Einer-für-Alle"). Es ist ein Prinzip der interföderalen Zusammenarbeit und insbesondere zentraler Baustein bei der Umsetzung des Online-Zugangsgesetz (OZG). Das EfA-Prinzip soll zu einer effizienteren Digitalisierung beitragen, indem Parallelentwicklungen vermieden und die Verbreitung in der Fläche gefördert wird. Das EfA-Prinzip dient dabei als ein organisatorischer Rahmen zur Bereitstellung von Softwarelösungen zwischen den Ländern und definiert Kriterien und Mindestanforderungen. Das Konzept von Marktplätzen unterstützt den Austausch als Ansatzpunkt für technische und rechtliche Anforderungen. [FOKUS]
Während bei den stark mit dem EfA-Prinzip verknüpften Online-Diensten auch durch das Onlinezugangsgesetz eine starke Governance vorhanden ist, ist die Ausgangslage bei der Nutzung von Anwendungen generativer KI in der Verwaltung derzeit anders: Die Nutzung von generativer KI in der Verwaltung ist ein relativ neues Gebiet, das in dieser Form keine Vorläufer hat und über das zunächst Erfahrungen gesammelt werden. Auch handelt es sich bei den Anwendungen der in dieser Vorstudie untersuchten Lösungen um freie Assistenzfunktionen, die noch nicht in formalisierte Prozesse eingebunden (im Gegensatz beispielsweise zur Funktion einer automatisierten Entscheidung, die in einem Verwaltungsprozess eingebunden ist).

Trotz der unterschiedlichen Ausgangslage ist es sinnvoll, sich im Sinne der Unterstützung der Digitalisierung auf die Nutzung des EfA-Prinzips vorzubereiten.
Dabei kann jedes Land in der Betriebsphase sowohl betreibendes als auch mitnutzendes Land sein. Mindestanforderungen gibt es sowohl für die Dienste als auch für Rollen und Verantwortlichkeiten in der Betriebsphase. Ersteres umfasst Anforderungen an Oberflächengestaltung und Design, Fachlogik, Nutzerkonto, Payment, Datenaustauschstandard, Routing, Organisation, IT-Sicherheit; Letzteres regelt Rollen und Verantwortlichkeiten bezüglich Betriebsverantwortlichkeit, Steuerungskreis, Support und weiteres. [BMDS][ITPLR]

3 Vorstellung der GenKI-Lösungen

In diesem Kapitel werden die betrachteten GenKI-Lösungen in Kürze vorgestellt. Die ausführlichen Steckbriefe finden sich in einem externen Dokument. Dieses Kapitel dient somit als Einführung, die wesentliche Aspekte der Lösungen zusammenfasst.

Die Steckbriefe sind auf die Erfassung einer Momentaufnahme ausgelegt. Geplante Weiterentwicklungen der Lösungen wurden nur dann in die Steckbriefe aufgenommen, wenn diese für den Vergleich der GenKI-Lösungen einen Mehrwert bieten, beispielsweise im Hinblick auf eine mögliche Konvergenz oder – noch interessanter – grundsätzliche Unterschiede aufzeigen.

In den Steckbriefen findet sich neben einer kurzen (Selbst-)Beschreibung allgemeine Informationen zu den Verantwortlichen der Projekte, sowie Zielgruppe, Umsetzungsstand und den Nutzungsmöglichkeiten der Lösung. Außerdem wurde eine Übersicht über die Lösung in Form einer Abbildung erstellt.

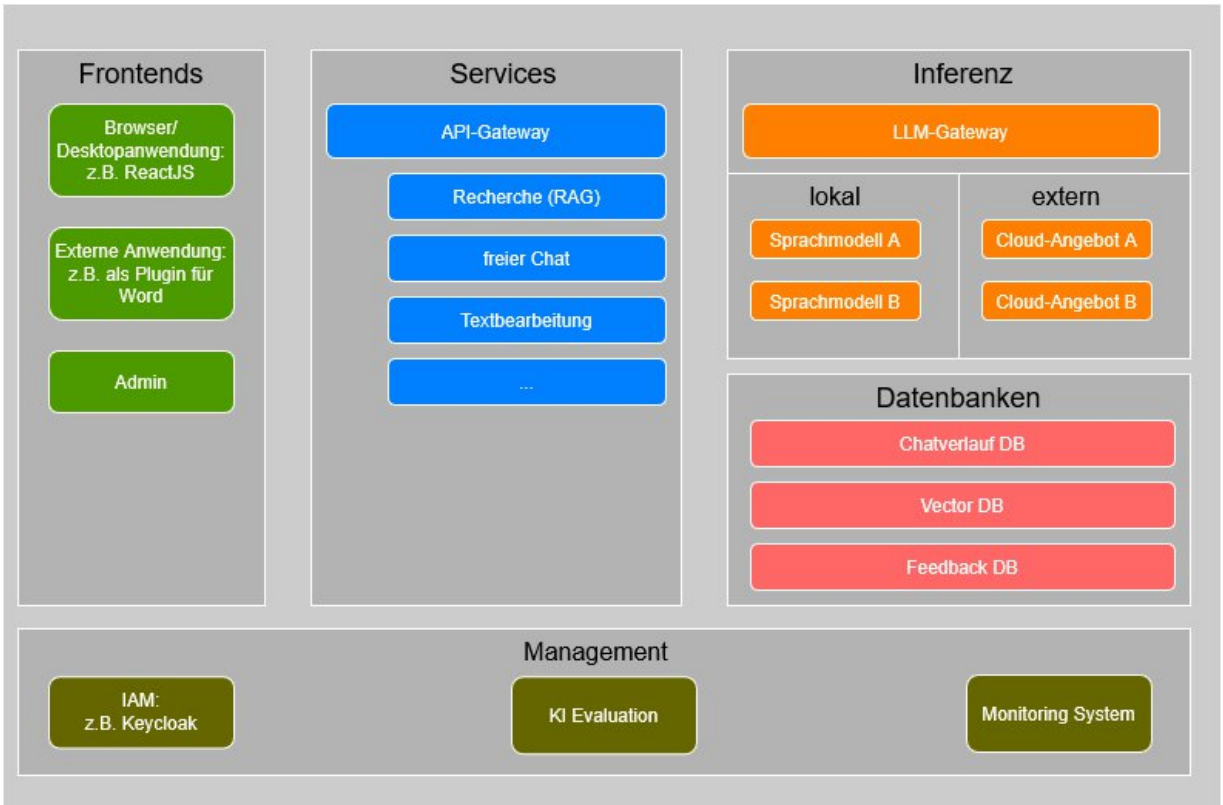


Abbildung 3: Vorlage für die Systemübersicht

Als Vorlage dient eine Systemübersicht einer GenKI-Plattform (siehe Abbildung 3), die auf dem obigen Begriffsverständnis aus Abbildung 2 aufbaut und sich zudem an die KIVA.arc-Referenzarchitektur anlehnt:

- Im Frontend-Layer werden die von der Anwendung bereitgestellten Benutzerschnittstellen gelistet. Das ist mindestens eine Browser- oder Desktopanwendung für Endnutzende.
- Im Service-Layer werden wesentliche Services gelistet, die in der Anwendung implementiert sind, Hilfsfunktionen werden nicht aufgeführt. Hinzu kommt das API-Gateway - inklusive Framework, falls vorhanden. Jeder blaue Kasten entspricht dabei einem eigenen Microservice.
- Der Inferenz-Layer ist nur dann Teil der Übersicht, wenn es sich um ein vollständiges GenKI-System handelt, andernfalls werden die implementierten Schnittstellen zur Inferenz gelistet. Unter Datenbanken werden drei verschiedene in das System integrierte Datenspeicher gelistet, für den Chatverlauf, die Vektordaten sowie das Nutzer:innen-Feedback.
- Schließlich gibt es den für eine Plattform nötigen Management-Layer. In diesem werden die Werkzeuge für Identity and Access Management (IAM), KI-Evaluation und allgemeines Monitoring gelistet.

Je nach Einordnung und Umfang der Lösung wird die Vorlage angepasst und mit den konkret von der Lösung implementierten Komponenten und Services ergänzt.

3.1 KIPITZ

KIPITZ versteht sich als eine KI-**Plattform** („Portal") zur gemeinsamen, behördenübergreifenden Nutzung von GenKI-Anwendungen in der Bundesverwaltung. Bereitgestellt werden Softwarelösungen sowie die Möglichkeit der Entwicklung eigener Apps, bzw. der Beauftragung des ITZBund zur Entwicklung spezieller Lösungen. Es wird somit ein vollständiges **GenKI-System zur Mitnutzung** angeboten. Damit hat KIPITZ den Anspruch, als zentrale GenKI-Lösung für alle Bundesbehörden den Betrieb zu zentralisieren, aber gleichzeitig Raum für spezifische Bedarfe und Eigenentwicklung zu geben, unter Beibehaltung der Garantien zu Fragen der Compliance. Derzeit können Bundesbehörden KIPITZ nutzen, perspektivisch soll die Plattform über den noch im Aufbau befindlichen Deutschland-Stack föderal den Ländern und Kommunen angeboten werden.

Verantwortliche	ITZBund
Homepage	https://www.itzbund.de/SharedDocs/Videos/DE/Mediathek/20250117-Gehres-Kipitz.html https://maki.beki.bund.de/a/bmi-makimo-app/steckbriefe/BD20254689782
Entwicklungsstand	Im produktiven Einsatz VS-NfD freigegebene Instanz verfügbar (BMF)
Zielgruppe	Alle Bundesbehörden Endanwender aus der Verwaltung sollen das KI-Portal unkompliziert nutzen können
(Nach)Nutzungsszenarien	Mitnutzung der Plattform inklusive aller verfügbaren Apps Nutzung auch von anderen Behörden entwickelter bzw. beauftragter Apps

Tabelle 1: Kurzübersicht KIPITZ

Zu den bereits entwickelten und zur Mitnutzung stehenden GenKI-Anwendungen zählen u. a. Transkription, Lösungen zur Barrierefreiheit und eine RAG-Implementierung. Außerdem wird ein Software Development Kit (SDK) angeboten, durch das eigene Anwendungen entwickelt werden können. Auch kann das ITZBund beauftragt werden, spezifische Lösungen zu entwickeln oder eigene Lösungen auch anderen Organisationen zur Verfügung zu stellen.

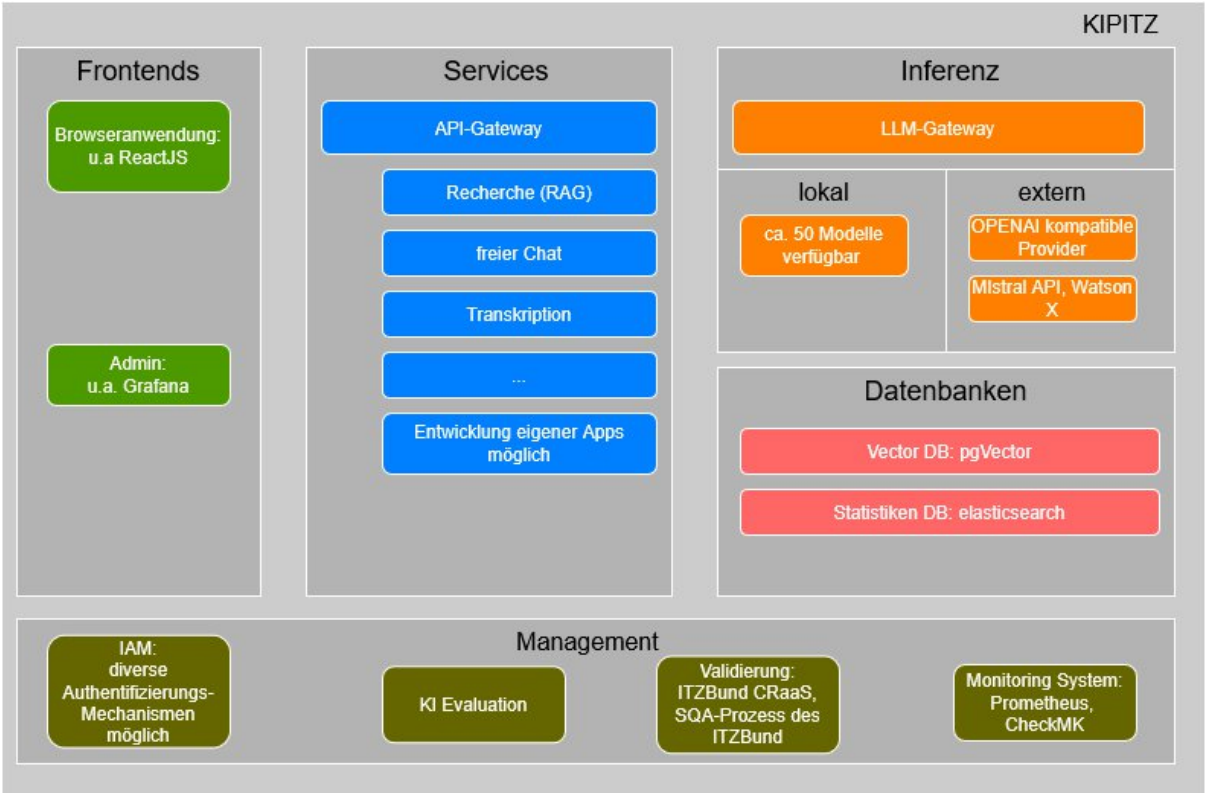


Abbildung 4: Systemübersicht KIPITZ

GenKI-Anwendungen auf der KIPITZ Plattform sind weitgehend modellagnostisch, für den lokalen Betrieb im ITZBund sind derzeit rund 50 Modelle verfügbar. Behörden können entscheiden, welche LLMs sie nutzen wollen. Die Anbindung von Fremdmodellen ist ebenfalls möglich, sowohl extern in der Cloud betrieben als auch on-premise auf den Servern des ITZBund. Eine Instanz ist VS-NfD freigegeben.

Gemeinsam mit an dem Angebot interessierten Behörden erfolgt eine Bewertung, welche GPU-Ressourcen benötigt werden. Anschließend werden die Hardwareinvestitionen von den Behörden finanziert. Kosten für die Nutzung werden in Rahmenverträgen festgelegt. Der Betrieb erfolgt durch das ITZBund. Parallel dazu können Sprachmodelle aus der Public Cloud genutzt werden, beispielsweise zur Einführung von KI und Sammlung erster Erfahrungen sowie für unkritische Anwendungen.

3.2 PLAIN

PLAIN (Platform Analysis and Information Systems) ist eine Cloud-**Plattform**-Lösung **zur Mitnutzung** rund um Daten, Künstliche Intelligenz (KI) und Machine Learning. Das Angebot richtet sich damit vor allem an Datenlabore und Fachabteilungen der Bundesverwaltung für die Entwicklung von Daten- und KI-Anwendungen, sowie Prototypen. Die Cloud ist als DevOps-Plattform im Sinne der Modernisierungsagenda der BReg ausgeprägt.

Verantwortliche	Auslands-IT des Auswärtiges Amt (AA) gemeinsam mit der Bundesdruckerei
Homepage	www.plain.diplo.de/ https://maki.beki.bund.de/a/bmi-makimo-app/steckbriefe/BD20242162685
Entwicklungsstand	Steht zur Nutzung bereit, wird im Dialog mit den Kunden lfd. weiterentwickelt
Zielgruppe	Steht für alle Bundesbehörden zur Verfügung, um ein permanentes Prototyping zu betreiben; Analysen zu fahren sowie eigene und fremde Software zu deployen
(Nach)Nutzungsszenarien	Mitnutzung Bundesebene mittels Verwaltungsvereinbarung zwischen Behörde und AA Nachnutzung durch Länder prinzipiell denkbar

Tabelle 2: Kurzübersicht PLAIN

Aufgrund der Tatsache, dass PLAIN kein GenKI-System im Verständnis dieser Vorstudie ist, wird auf die Abbildung zur Systemübersicht verzichtet.

PLAIN ist eine Kombination aus Plattform-as-a-Service Angeboten, die zentrale Software-as-a-Service-Angebote bereitstellt und vollständig mandantenfähig ist. Das bedeutet, dass jede Behörde eigenständig in ihrem Cluster arbeiten kann, Software-as-a-Service konsumieren oder selbst bereitstellen kann. Besonderer Wert wird auf Kollaboration beispielsweise mittels gitlab gelegt. Die Plattform ist nach BSI-Grundschutz aufgebaut wird in einem Rechenzentrum in Deutschland auf Hardware der Auslands-IT betrieben. Bei Bedarf können Sprachmodelle gehostet und auch bereitgestellt werden. Aktuell genutzt werden bspw. Gemma und Mistral. Grundsätzlich können beliebige GenKI-Anwendungen und GenKI-Systeme unter Mitnutzung von Managementsystemen (beispielsweise IAM) auf PLAIN betrieben werden, beispielsweise derzeit eine F13 Instanz im produktiven Betrieb.

3.3 LLMoin

LLMoin versteht sich als ein innovativer KI-Assistent, der die Verwaltung dabei unterstützt, Arbeitsprozesse zu optimieren, den Zugang zu Informationen zu verbessern und Mitarbeitende gezielt bei anspruchsvollen Text- und Rechercheaufgaben zu entlasten. LLMoin wurde von der Freien und Hansestadt Hamburg gemeinsam mit IT-Dienstleister Dataport entwickelt und wird als Webanwendung von Dataport **zur Mitnutzung** auch für andere Länder bereitgestellt. Das Angebot ist somit ein vollständiges **plattformbasiertes GenKI-System**.

Verantwortliche	Senatskanzlei Hamburg, Amt für IT und Digitalisierung
Homepage	https://www.hamburg.de/politik-und-verwaltung/behoerden/senatskanzlei/aktuelles/pressemeldungen/hamburg-etabliert-ki-assistenten-llmoin-als-bundesweit-nutzbares-erfolgsmodell-1127988 https://www.dataport.de/nachricht/kuenstliche-intelligenz-hamburg-startet-mit-ki-assistent-llmoin-in-den-regelbetrieb https://maki.beki.bund.de/a/bmi-makimo-app/steckbriefe/HH20258004063
Entwicklungsstand	Die Lösung ist seit 2024 in Hamburg im Einsatz
Zielgruppe	Angebot für Behörden, Ämter und Betriebe der Trägerländer, Kommunen, Städte, öffentliche Einrichtungen und Unternehmen
(Nach)Nutzungsszenarien	Mitnutzung über den IT-Dienstleister Dataport gewünscht

Tabelle 3: Kurzübersicht LLMoin

LLMoin soll die tägliche Arbeit der Verwaltung durch KI-basierte Werkzeuge unterstützen. Zu den Funktionen gehören Textzusammenfassung und -generierung, Recherche, ein freier Chat und Prompt-Bibliotheken (vordefinierte Vorlagen für wiederkehrende Aufgaben). LLMoin wird bereits von mehreren Ländern und Kommunen mitgenutzt.

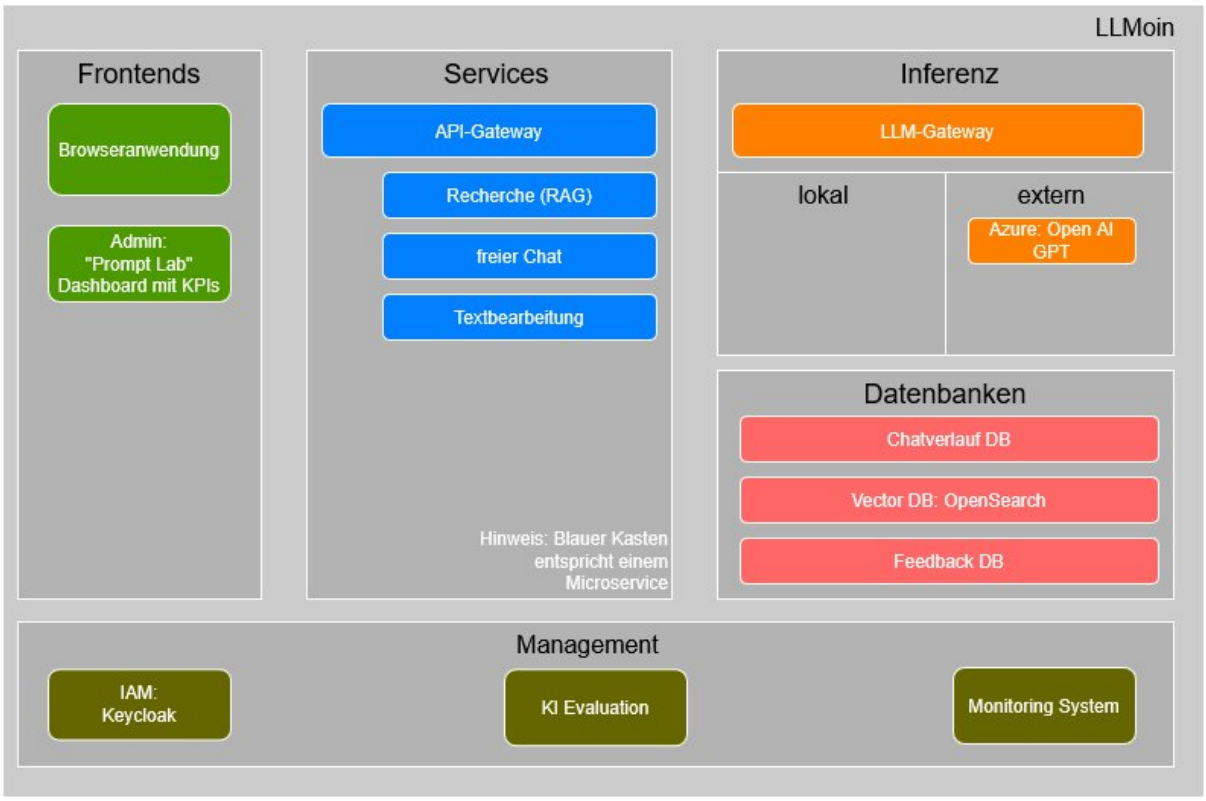


Abbildung 5: Systemübersicht LLMoin

Abbildung 5 zeigt die Systemübersicht über LLMoin, die auf den uns verfügbaren Informationen basiert, im Layer Management und Datenbanken unvollständig ist. Die Anwendung wird als Browseranwendung den Endnutzenden bereitgestellt. Die Services werden über eine API-Gateway angesteuert. Zum Zeitpunkt der Datenerhebung verfügte der Inferenz-Layer über eine externe Anbindung an die Open AI GPT-Modelle über die Azure Cloud; On-premise Lösungen sind geplant.

3.4 NRW.Genius

NRW.Genius versteht sich als zentrale **Plattform** (Genius-Plattform) für die Integration von KI in Fachverfahren. NRW.Genius basiert auf einer hybriden skalierbaren Infrastruktur und stellt somit ein vollständiges **GenKI-System zur Mitnutzung** innerhalb NRWs zur Verfügung. Es soll dazu beitragen, Innovation und Effizienz in der öffentlichen Verwaltung voranzutreiben. Als Webanwendung soll sie die tägliche Arbeit in der Verwaltung durch KI-basierte Werkzeuge unterstützen.

Verantwortliche	Ministerium für Heimat, Kommunales, Bau und Digitalisierung in Zusammenarbeit mit Landesbetrieb IT.NRW
Homepage	https://www.it.nrw/informationstechnik/digitale-innovation/ki-labor
Entwicklungsstand	NRW.Genius steht technisch zur Nutzung bereit Derzeit kontinuierliche Skalierung innerhalb der gesamten Landesverwaltung Nordrhein-Westfalen Kontinuierliche Weiterentwicklung von Produkt und Plattform
Zielgruppe	Mitarbeitende der gesamten Landesverwaltung NRW, perspektivisch auch bundesweiter Einsatz denkbar
(Nach)Nutzungsszenarien	Mitnutzung in der gesamten Landesverwaltung Nordrhein-Westfalen, perspektivisch EfA-Nachnutzung möglich, da für den flächendeckenden Einsatz konzipiert

Tabelle 4: Kurzübersicht NRW.Genius

Zu den Services, die NRW.Genius anbietet, gehören Textzusammenfassung und -generierung, Recherche, Chat sowie "Chat with your documents". Diese finden sich in Abbildung 6 aufgelistet unter Services. Jeder Service ist containerisiert und entspricht einem Microservice.

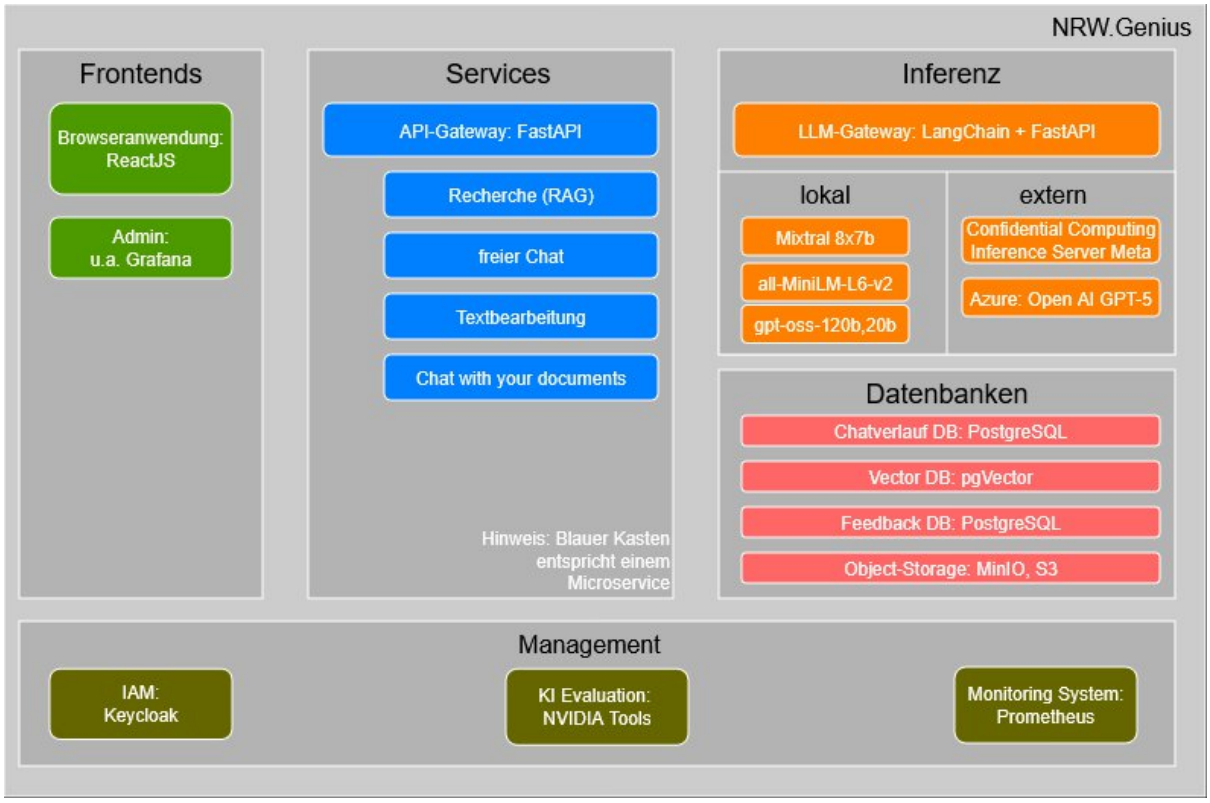


Abbildung 6: Systemübersicht NRW.Genius

NRW.Genius implementiert ein sogenanntes „LLM-Hub", über das sowohl on-premise als auch cloudbasierte LLMs angebunden sind. Je nach Schutzbedarf der Daten und Präferenz der Kunden können entweder on-premise Modelle angesteuert werden oder, bei geringem Schutzbedarf, auch externe Cloudangebote. Eine VS-NfD Freigabe ist bisher nicht erfolgt.

F13

3.5 F13

F13 ist das vollständig **plattformbasierte GenKI-System** des Landes Baden-Württemberg. Die **Plattform-Software** steht als „F13-OS" Open Source zur **Nachnutzung** zur Verfügung. F13-OS versteht sich als Grundlage für souveräne und modellagnostische KI-Systeme, die – da Open Source – auf eigener Infrastruktur betrieben und gemeinschaftlich weiterentwickelt werden können. F13-OS ist modular, stellt eine Reihe von GenKI-Funktionalitäten bereit und kann mit anderen Anwendungen kombiniert werden. Damit soll das Projekt einen Beitrag zu einer modernen, transparenten und kooperativen Verwaltungsdigitalisierung leisten.

Verantwortliche	Land Baden-Württemberg InnoLab_bw
Homepage	https://f13-os.de/ https://maki.beki.bund.de/a/bmi-makimo-app/steckbriefe/BW20252926711
Entwicklungsstand	Seit Juli 2025 als quelloffene Software zur Nachnutzung und Weiterentwicklung auf openCode bereitgestellt Weiterentwicklungen geplant
Zielgruppe	frei verfügbar auf openCode, gesamte Verwaltung als potenzielle Nutzerin
(Nach)Nutzungsszenarien	Mitnutzung im Land Baden-Württemberg SaaS-Nachnutzung geplant Nachnutzung der Open Source Software

Tabelle 5: Kurzübersicht F13

F13 soll die tägliche Arbeit in der Verwaltung durch KI-basierte Werkzeuge unterstützen. Zu den Funktionen gehören ein KI-Chatbot mit „Chat with your Document" Option, Textzusammenfassung und -generierung sowie Recherche in eigenen Wissensdatenbanken. Das Land Baden-Württemberg betreibt eine F13 Open Source Community, die zum einen bei technischen, organisationalen oder rechtlichen Fragen unterstützen soll, als auch die Weiterentwicklung von GenKI-Funktionalitäten gemeinschaftlich organisiert. Mehrere Bundesländer und Bundesbehörden sind in der Community aktiv.

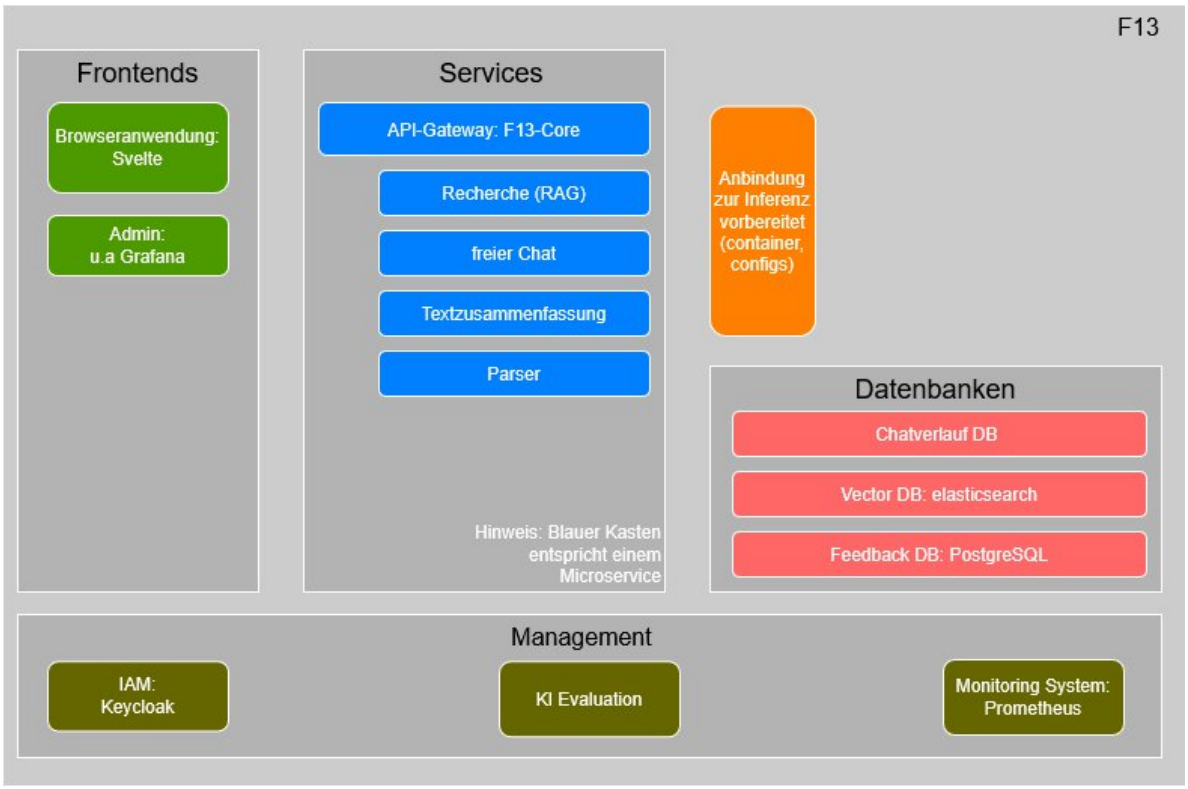


Abbildung 7: Systemübersicht F13

F13 verfolgt konsequent eine Microservice-Architektur. Jeder Service ist in einem eigenen Container gekapselt und die Anbindung zur Inferenz ist in jedem Service in Form von Containern bzw. config-Dateien angelegt. Die Kommunikation zwischen den Containern ist erfolgt über REST-APIs, also zwischen Frontend und F13-Core, zwischen Core und Services sowie eine OPEN-AI konforme REST-API.

3.6 AIGude

AIGude ist ein vollständiges **plattformbasiertes GenKI-System**, das speziell für die hessische Landesverwaltung entwickelt wurde. Das System soll die Recherche innerhalb öffentlicher sowie verwaltungsinterner Quellen schneller, genauer, sicherer und effektiver gestalten und die Arbeit mit Texten effizient unterstützen. Die **Mitnutzung** durch andere Landesbehörden ist in Prüfung, die **Nachnutzung** für Kommunen über einen IT-Dienstleister ebenfalls.

Verantwortliche	Hessische Ministerium für Wirtschaft, Energie, Verkehr, Wohnen und ländlichen Raum Referat Z 5 Digitalisierung, IKT gemeinsam und in Abstimmung mit dem Hessischen Ministerium für Digitalisierung und Innovation
Homepage	noch keine öffentliche Seite, da noch kein Marketing
Entwicklungsstand	Pilot auf souveräner Cloud-Infrastruktur des Landes verfügbar derzeit MVP - Initial Nutzende der Landesverwaltung, wird erweitert
Zielgruppe	Mitarbeitende der Landesverwaltung von Hessen, perspektivisch evtl. kommunale Nutzende
(Nach)Nutzungsszenarien	Mitnutzung durch andere Landesbehörden in Prüfung, Nutzung für Kommunen über IT-Dienstleister ekom21 in Prüfung Teile als Open Source zur Nachnutzung geplant

Tabelle 6: Kurzübersicht AIGude

Zu den Funktionen gehören Assistenz bei der Textbearbeitung, Recherchen in landeseigenen Daten sowie ein freier Chat mit dem geschützten System.

Das Hosting erfolgt in einer proprietären Cloud-Umgebung, betrieben durch die HZD, den IT-Dienstleister der Landesverwaltung. Eine Anbindung von externen Sprachmodellen ist derzeit nicht geplant.

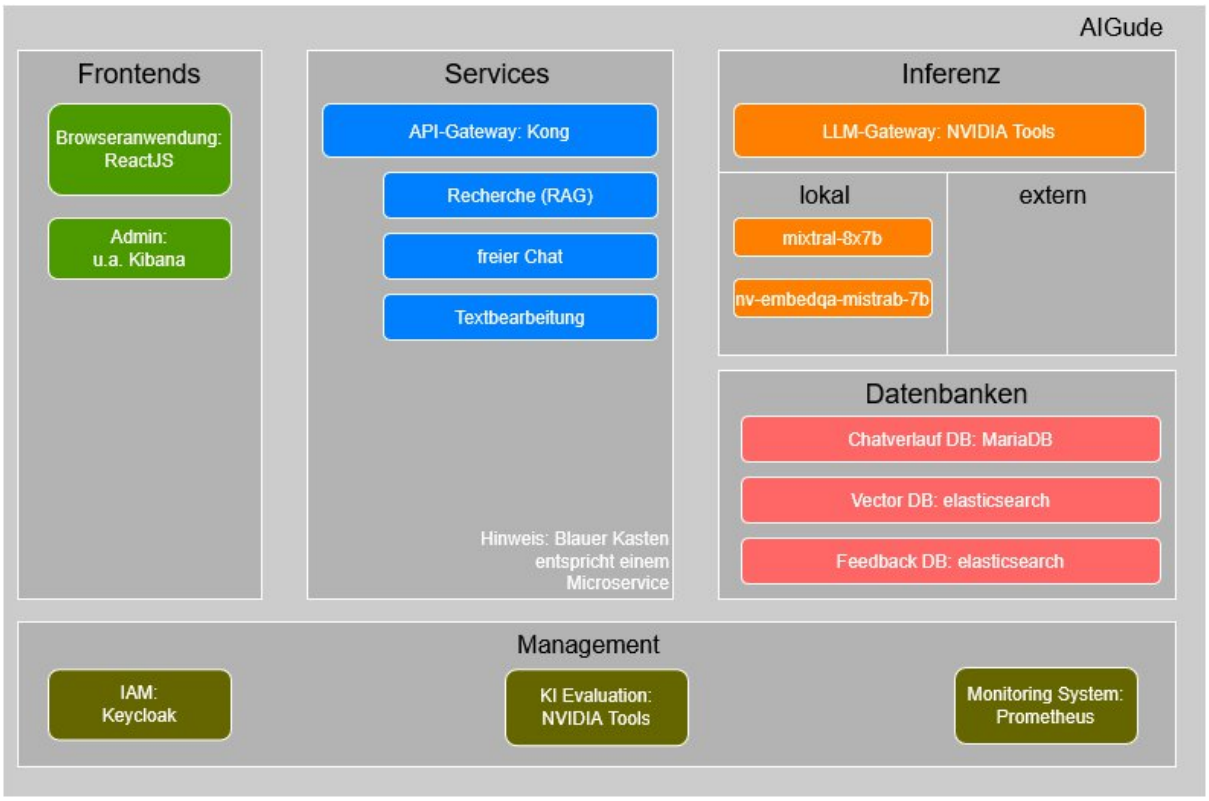


Abbildung 8: Systemübersicht AIGude

Wie die meisten anderen Lösungen verwendet AIGude als API-Gateway und zur Verwaltung der implementierten Microservices FastAPI, als eine Besonderheit jedoch in Verbindung mit Kong. Derzeit wird ausschließlich ein lokales LLM (mixtral-8x7b) auf 5 dedizierten GPUs betrieben. Außerdem gibt es eine weitere GPU für das Embedding bei RAG. Es existiert eine ETL-Pipeline, welche – genauso wie die Microservices – in Python implementiert wurde.

3.7 MUCGPT

MUCGPT ist eine **plattformbasierte GenKI-Anwendung**, um GPT-Modelle für die Verwaltung nutzbar zu machen. Damit soll die tägliche Arbeit durch KI-basierte Werkzeuge unterstützt werden. Die Anwendung verbindet sich mit einem oder mehreren OpenAI-kompatiblen Endpunkten.

MUCGPT ist eine Anwendung neben anderen, die auf der KI-Infrastruktur von München entwickelt und bereitgestellt wird. Die KI-Infrastruktur basiert auf Open Source Software und beinhaltet Basiskomponenten, wie ein LLM-Gateway (Litellm), KI-Observability (Langfuse) und eine Websuche-Komponente (SearXNG). Die Anwendung wird als Open Source auf GitHub frei zur **Nachnutzung** angeboten.

Verantwortliche	KI Competence Center, IT-Referat der Landeshauptstadt München
Homepage	https://ki.muenchen.de/ki-systeme/mucgpt https://ki.muenchen.de/blog/2025-07-23-oss-genai-stack
Entwicklungsstand	Seit Februar 2023 produktiv in München in Einsatz Nachnutzung in Stuttgart
Zielgruppe	frei verfügbar auf GitHub, gesamte Verwaltung als potenzielle Nutzerin
(Nach)Nutzungsszenarien	Nachnutzung Open Source Software-Plattform, Frontend-Demo öffentlich verfügbar

Tabelle 7: Kurzübersicht MUCGPT

Neben dem Chat liegt die Kernfunktionalität von MUCGPT im Teilen und Erstellen von vorkonfigurierten Assistenten, die eine bestimmte Aufgabe erledigen. Außerdem verfügt MUCGPT über Funktionen, um auf externe Werkzeuge und Schnittstellen zuzugreifen.

Beispiele hierfür sind die Übersetzung in einfache Sprache oder das Brainstorming Werkzeug. Beim Brainstorming wird eine Mindmap erstellt, die in weiteren Anwendungen weiterverwendet werden kann. Zukünftig ist die standardisierte Anbindung von Tools über MCP möglich.

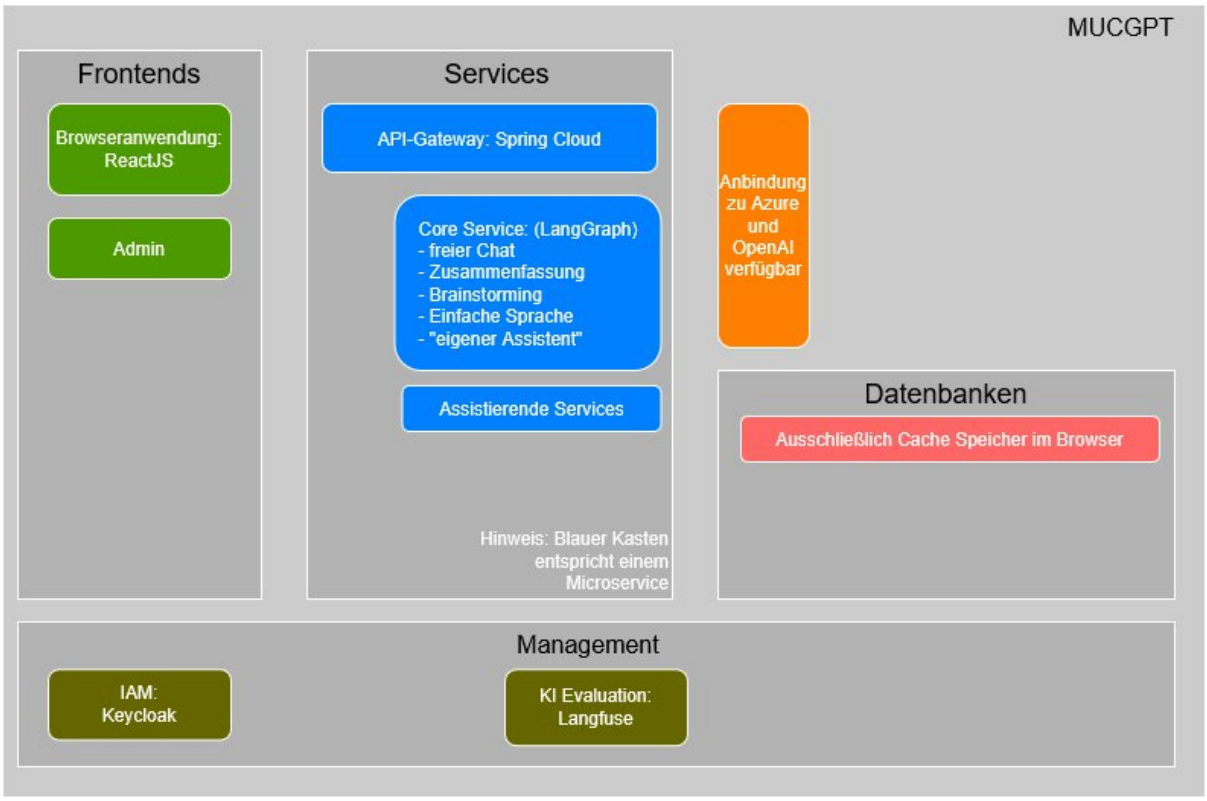


Abbildung 9: Systemübersicht MUCGPT

Wie die meisten Lösungen wurde die Anwendung mit LangChain/LangGraph entwickelt. Als eine Besonderheit nutzt MUCGPT derzeit keine Datenbanken. Stattdessen wird der Chatverlauf ausschließlich im Browser der Nutzer:innen gespeichert.

MUCGPT kann per Konfiguration an OpenAI kompatible Sprachmodelle angeschlossen werden. Im Münchner Betrieb erfolgt die Anbindung über ein LLM-Gateway an Azure (gpt4.1).

4 Vergleich der Lösungen

Der Vergleich der Lösungen erfolgt mit Fokus auf die konkreten Nach- und Mitnutzungsangebote beziehungsweise -möglichkeiten. Es wird explizit auf die Trennung zur primären Nutzung verwiesen. Damit ist gemeint, dass die Lösungen in den meisten Fällen für einen Einsatzzweck, beispielsweise zur Nutzung in einer Landesverwaltung, entwickelt wurden und dort auch im Einsatz sind.
Angedachte oder bereits praktizierte Nach- und Mitnutzungsangebote:

- **Software-as-a-Service**: LLMoin (für Länder), KIPITZ (für Bundesbehörden), F13 (geplant)
- **Plattform**-as-a-Service: PLAIN (für Bundesbehörden)
- **Bereitstellung des Gesamtsystems**: NRW.Genius (geplant), AIGude (geplant), PLAIN (für Länder, geplant)
- **Open Source Software:** MUCGPT, F13-OS

Zu sehen ist, dass sich die Angebote sowohl in Ausrichtung als auch im Umfang unterscheiden, visualisiert in Abbildung 10. Die Abbildung zeigt eine etwas andere Perspektive als eingeführt in Kapitel 3. Die Anwendung, bestehend aus Services und Frontend bildet hier die oberste Schicht. Externe Funktionsblöcke – dazu zählen insbesondere die LLMs, die extra gelistet sind – bilden eine zweite Schicht. Die unterste Schicht bildet die IT-Infrastruktur auf der die Anwendung, die Funktionsblöcke und LLMs laufen.

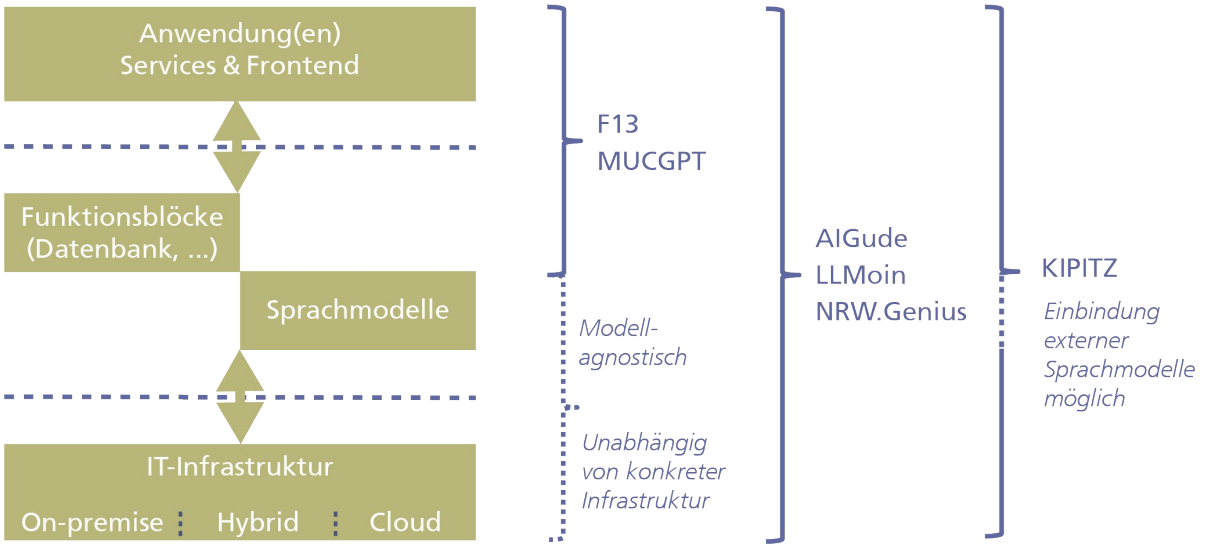


Abbildung 10: Lösungen aus Nachnutzungsperspektive

Die Open Source Software Angebote F13 und MUCGPT werden auf der IT-Infrastruktur einer betreibenden Organisation oder in der Cloud gehostet und sind (weitgehend) modellagnostisch, können also durch weitere Schnittstellenentwicklung ergänzt werden. Aus Nachnutzungssicht wird somit eine GenKI-Anwendung zur Nutzung und Weiterentwicklung angeboten.

AIGude, LLMoin und NRW.Genius ähneln sich bezüglich des Nachnutzungsangebots, unterscheiden sich jedoch beim Stand der Entwicklung, des Vertriebes und der Zielgruppe. In allen drei Fällen besteht das Angebot aus einem vollständigen GenKI-System, bereitgestellt über SaaS.

Die KIPITZ Plattformlösung hat als Zielgruppe die gesamte Bundesverwaltung. Das Angebot ist ebenfalls SaaS, also eine gemeinsame Nutzung der Infrastruktur, bereitgestellt durch das ITZBund. Es können sowohl eigene Anwendungen entwickelt als auch bereits bestehende genutzt werden. Je nach Bedarf können unterschiedliche LLMs entweder on-premise deployed oder, falls unbedenkliche Use-Cases, auch auf Cloudlösungen zurückgegriffen werden.

Außerhalb des Vergleichs der Abbildung verortet sich PLAIN und sollte aus Nachnutzungsperspektive getrennt zwischen Bundesbehörden und (perspektivisch) Ländern betrachtet werden. PLAIN – als PaaS Lösung – wird von der Bundesdruckerei betrieben und ermöglicht die gemeinsame Nutzung auf Servern der Auslands-IT. Behörden können die bereitgestellte Infrastruktur und Software für die Entwicklung eigener Anwendungen nutzen und diese anderen über die Plattform bereitstellen. Das heißt auch, dass je nach Bedarf unterschiedliche Modelle auf den Servern gehostet werden können. Gleichzeitig gibt es Überlegungen, die gesamte Plattform zur Nachnutzung an Länder zu geben, sodass diese sie auf ihrer eigenen Infrastruktur hosten können.

Nach der Kurzvorstellung in Kapitel 3 werden die GenKI-Lösungen im Detail miteinander verglichen. Für weitere Details zu den GenKI-Lösungen wird auf die detaillierten Steckbriefe verwiesen.

4.1 Modell- und Anbieter-Unabhängigkeit

Als zentrales funktionales Element nimmt das Sprachmodell im Inferenz-Layer wesentlich Einfluss auf die Eigenschaften des GenKI-Systems, nicht nur auf konkrete Leistungsfähigkeit, sondern auch auf strategische Ziele wie souveräner Betrieb von KI. Details zu den verwendeten

Sprachmodellen finden sich in den Steckbriefen auf dem Blatt „Technik-Stack“ in dem „KI-Schicht“ sowie in den Systemübersichten in Kapitel 3.
Über alle Projekte hinweg zeigt sich ein hohes Bewusstsein für die Vorteile von Modell- und Anbieter-Unabhängigkeit. Die Umsetzungsstände sind jedoch unterschiedlich weit fortgeschritten. Manche Pläne zur Anbindung weiterer Modelle anderer Anbieter befinden sich noch in der Planungsphase, wie es beispielsweise bei LLMoin der Fall ist. NRW.Genius bietet bereits verschiedene interne sowie cloudbasierte Anbindungen an. Flexible, individuell planbare Lösungen für LLMs finden sich bei KIPITZ. Komplett modellagnostisch, selbstverantwortlich zu organisierende Lösungen finden sich bei Open Source Softwarelösungen wie F13. Bei diesen sind Container und Schnittstellen für die Anbindung vorbereitet. Die unterschiedlichen Strukturen der einzelnen GenKI-Lösungen führen zu unterschiedlichen Konsequenzen in Bezug auf die Modell- und Anbieterunabhängigkeit sowie inwieweit das als Aufgabe des Anbieters der GenKI-Lösung gesehen wird.
Prinzipiell lässt sich zwischen zwei Betriebsarten unterscheiden: On-premise, also auf „eigener“ Hardware, und (meist subscription-based) Cloudlösungen. Zum Zeitpunkt der Datenerhebung (ohne Berücksichtigung der Roadmaps) nutzen die Lösungen folgende Betriebsarten für die LLMs:

- Ausschließlich on-premise: AIGude
- Ausschließlich cloudbasiert: LLMoin
- Sowohl on-premise als auch cloudbasiert: KIPITZ, NRW.Genius
- Je nach deployment, abhängig vom Betreiber: MUCGPT, F13-OS

PLAIN konstituiert für sich genommen eine Cloud (Community Cloud), die grundsätzlich in beliebigen physikalischen Umgebungen betreibbar ist.
Die Kostenstruktur ist abhängig von den Betriebsarten. Je nach Betriebsart kommt es zu regelmäßigen Kosten für subscription-based Angebote, Investitionen für die Beschaffung eigener Hardware oder laufende Kosten aufgrund des Betriebs. KIPITZ verfolgt hier einen interessanten Ansatz: Behörden, welche den Dienst nutzen wollen, finanzieren die Hardware, welche vom ITZBund im Anschluss betrieben wird. Als Grundlage für das Kalkulationsmodell wird die Hardware in Token umgerechnet, die der Behörde zur Nutzung bereitstehen. Auf diese Weise wird Hardware zentralisiert, was zu höher Auslastung und damit Effizienz führt.
Fast alle Lösungen lassen sich auch mit europäischen Modellen nutzen. Das gilt vor allem für Modelle von Mistral, die beispielsweise von NRW.Genius, AIGude, F13 (in den Umsetzungen

des Landes Baden-Württemberg) und der MUCGPT-Umsetzung für München betrieben werden. Bei KIPITZ und PLAIN können sie bei Bedarf ebenfalls on-premise betrieben werden. Auf PLAIN können grundsätzlich beliebige Modelle betrieben werden.

4.2 Lizenzen und Open-Source

Die Nutzung von Open Source Software für die öffentliche Verwaltung ist aus strategischen Gründen wichtig, u. a. verringert sie die Abhängigkeit von einzelnen Anbietern und erlaubt einen leichteren Austausch zwischen verschiedenen Systemen. Die zur Realisierung der GenKI-Lösungen verwendeten Software-Komponenten finden sich in den Steckbriefen auf dem Blatt „Technik-Stack", der Status der GenKI-Lösung in Bezug auf Open Source ist auf dem Blatt „Allgemeines" unter dem Punkt „Angebot" zu finden.

Alle Lösungen setzen zumindest in Teilen auf Open Source Software, was bei aktuellen IT-Projekten Best Practice ist. Das zieht sich durch den gesamten Technik-Stack. Für die Frontend-Entwicklung wird zumeist auf React.js gesetzt, eine offene Javascript Library. Alle Lösungen setzen für IAM auf die Open Source Anwendung Keycloak. Für das Monitoring und die Überwachung sowie Dashboard-Visualisierung wird oftmals auf Prometheus und Grafana zurückgegriffen.

Als das am häufigsten verwendete Framework zur Anwendungsentwicklung wird LangChain genutzt. Auch auf Ebene der Datenbanken wird viel Open Source Software genutzt. Vor allem PostgreSQL findet Einsatz, in Verbindung mit pgVector für Vektordatenbanken für RAGs. Open Source LLMs werden hauptsächlich Metas LLama Modelle sowie Mistrals mixtral-Modelle verwendet.

Vollständig Open Source sind F13 und MUCGPT. AIGude plant, mindestens einzelne Microservices als Open Source zu veröffentlichen und so mit anderen Projekten zu teilen, sowie ebenfalls von den bereits Open Source zugänglichen Microservices (beispielsweise von F13) zu profitieren. Somit sind hier Nachnutzungspläne zwischen den Lösungen auf Ebene der Microservices zu erkennen. Darüber hinaus prüft AIGude die Möglichkeit einer umfassenderen Open Source Stellung des KI-Systems.

4.3 Vertrauenswürdigkeit und Ethik

Eine Herausforderung beim Einsatz generativer KI ist die schwierige Nachvollziehbarkeit, wie Output und Ergebnisse zustande kommen. Daher muss die Qualität der Ergebnisse einer Anwendung (laufend) überprüft werden und Nutzer:innen müssen die Grenzen der Technologie

kennen. Details zu diesem Punkt finden sich in den Steckbriefen auf dem Blatt „Nicht-funktionales“ unter „Funktionen zur sachgerechten Nutzung“.
Generell zeigt sich, dass alle Lösungen einige Funktionen zur sachgerechten Nutzung implementieren. Besonders häufig werden Prompt-Bibliotheken sowie Schablonen genannt. Durch die Bereitstellung einer Auswahl vorgefertigter Prompts können nicht nur Anwendungsszenarien veranschaulicht, sondern auch Kompetenzanforderungen an die Nutzenden reduziert werden.
MUCGPT bietet die Erstellung eigener Assistenten an. Das beinhaltet die Möglichkeit der Anpassung des Systemprompts, der erlaubten Tools, Startbeispiele, vorgeschlagene Prompts und der Beschreibung.
Besonders hervorzuheben ist, dass alle Projekte ihrer Verantwortung aus der KI-Verordnung nachkommen, den Kompetenzaufbau der Nutzer:innen zu ermöglichen. Dazu werden unterschiedliche Lernmaterialen in Text und Videoform bereitgestellt. Auch werden ganze Lehrangebote und Consulting zu Prozessanpassung angeboten, wie beispielsweise bei LLMoin.
Inwieweit die Lösungen Methoden implementieren, welche die Akkuratheit des Outputs der Modelle verifizieren, konnte nicht abschließend erhoben werden. Hier ist auch auf den Stand der Technik zu achten. Es handelt sich um ein aktuelles Forschungsbiet, aktuell gibt es nur wenig praktische Ansätze. Am besten lassen sich Ergebnisse überprüfen, wenn eine RAG-Pipeline implementiert ist und der Output auf Basis hinterlegter Dokumente mit entsprechenden Verweisen zur Nachprüfung erstellt wird. RAG-Services werden bereits von den meisten Lösungen implementiert.

4.4 Modularität

Als moderne IT-Anwendungen nach Stand der Technik verwenden alle betrachteten GenKI-Lösungen eine Schichtenarchitektur und Container-Virtualisierung, wodurch sich Ansatzpunkte und Schnittstellen für den Austausch von Modulen zwischen GenKI-Anwendungen bieten. Konkrete Angaben finden sich in den Steckbriefen auf dem Blatt „Technik-Stack“ unter „Applikationsschicht“ und „Technologieschicht“.
Alle Lösungen verfolgen eine Microservice-Architektur mit Kubernetes und Docker. Auch die Kommunikation zwischen den Komponenten erfolgt zumeist über REST-API, beispielsweise zwischen Frontend und API-Gateway, sowie zwischen dem Gateway und den Microservices. Als Framework wird dafür mehrfach FastAPI genannt. Der Vorteil einer gemeinsamen Referenzarchitektur mit einheitlichen Schnittstellen zeigt sich am Beispiel von AIGude und F13.

AIGude nutzt einzelne Microservices von F13 als Grundlage für die Weiterentwicklung eigener Microservices. Auf diese Weise können Teile einer Anwendung geteilt und gemeinsam weiterentwickelt werden. Da viele der GenKI-Lösungen ähnliche Anwendungsfälle abdecken, findet sich an dieser Stelle und auf diesem Level viel Potenzial zum Austausch und gemeinsamer Entwicklung.
Beachtet werden sollte dabei die teilweise unterschiedliche Granularität der Microservices. MUCGPT implementiert einen Core-Service als einen Microservice, der alle Services Chat, Summarize etc. in einem Container realisiert. Im Unterschied dazu legt beispielsweise F13 Wert auf eine strikte Trennung zwischen den Services.
Organisationswissen lässt sich bei den meisten Lösungen einbinden, wobei sich die Art und Weise sowie das Sicherheitslevel unterscheiden. Vor allem zwei Varianten sind implementiert. Das ist (a) eine Möglichkeit für Nutzende, eigene Dokumente „ad hoc" hochzuladen und Wissen zu extrahieren oder mit den Dokumenten zu chatten („Frag mein PDF"). Die meisten Lösungen haben außerdem (b) einen RAG-Services implementiert, wobei für die Vektordatenbank zumeist auf Open-Source Lösungen gesetzt wird. An dieser Stelle spielt auch die Mandantentrennung sowie Berechtigungskontrolle eine große Rolle, da sichergestellt werden muss, dass nur berechtigte Personen auf bestimmte Datensätze Zugriff erhalten.

4.5 Daten- und Geheimnisschutz

Datensicherheit ist ein übergreifendes Thema, das schon in der IT-Architektur verankert werden muss. Zunächst technische Aspekte, wie der Autorisierung von Nutzer:innen, der Wahl eines Sprachmodell-Angebots sowie einer Ausführungsumgebung haben unmittelbar auch rechtliche Konsequenzen. In den Steckbriefen finden sich Angaben zu „Datenverarbeitung / Datenschutz" auf dem Blatt „nicht-funktionales" sowie Angaben zu Authentifizierung und Verarbeitung auf dem Blatt „Technik-Stack" unter „Technologieschicht" und „Compute-Schicht".
Daten- und Geheimnisschutz wird von hoher Bedeutung, sobald ein System Daten mit Schutzbedarf verarbeiten können soll. In diesem Kontext sind vor allem die Datenschutzgrundverordnung sowie die europäische KI-Verordnung einzuhalten. NRW.Genius hat dafür drei verschiedene Schutzbedarfe festgelegt, für welche auf Mandantenebene der Ort der Datenverarbeitung ausgewählt werden kann. Bei hohem Schutzbedarf (sensible oder personenbezogene Daten), werden Anfragen derzeit durch on-premise LLMs verarbeitet. Bei normalen Schutzbedarf können freigegebene Cloud Modelle zum Einsatz kommen. Bei Daten ohne

Schutzbedarf wird auf Modelle zurückgegriffen, welche über die Azure Cloud angebunden sind. Eine Verarbeitung von Verschlusssachen nach VSA ist nicht vorgesehen.
NRW. Genius setzt auf Confidential Computing, eine Sicherheitstechnik, die Daten während der Verarbeitung schützt. Dabei laufen Berechnungen in einem hardwarebasierten Trusted Execution Environment (TEE) ab. Erst nachdem deren Vertrauenswürdigkeit geprüft ist, werden vertrauliche Daten verarbeitet. Bei der Verarbeitung in externen Rechenzentren bzw. bei cloud-basierten Diensten ergänzt diese Technologie die Verschlüsselung von Daten im Speicher und bei der Übertragung.
KIPITZ verfolgt eine ähnliche Trennung der Verarbeitung von Daten wie NRW.Genius. Bei unkritischen Anwendungsfällen und Daten können externe Modelle angebunden werden. Für schutzbedürftige Anfragen wird auf die Hardware des ITZBund zugegriffen, die außerdem als einziges der betrachteten Projekte VS-NfD freigegeben ist.
F13 gibt allgemeine Hinweise zu Datenschutz, bspw. zur Datensparsamkeit und Zweckgebundenheit. Außerdem wird auf die Community verwiesen, die bei Fragen unterstützen kann.
MUCGPT verbietet unter Berücksichtigung der aktuell genutzten Modelle die Nutzung personenbezogener Daten.

4.6 Compliance-Unterstützung

Der Einsatz von GenKI im Verwaltungskontext erfordert die Berücksichtigung von umfangreichen rechtlichen Fragenstellungen, daher sind Unterstützungsangebote bei Compliance Fragen ein wichtiges Thema für die Nachnutzung von GenKI-Lösungen. In den Steckbriefen findet sich die „Compliance-Unterstützung“ auf dem Blatt „nicht-funktionales“.
Welche Compliance-Unterstützung notwendig ist, hängt wesentlich von der Art des Nutzungsangebotes ab, also Mitnutzung durch SaaS oder Nachnutzung durch Open Source. Im ersten Fall liegt die Verantwortung für Datenschutz, IT-Sicherheit und regulatorische Vorgaben beim Anbieter der SaaS Lösung. Mitnutzende benötigen Teile der Dokumentation des Angebots als Informationsquelle für die eigene Dokumentation und Vorlage zu den Themen wie Datenschutz, Einführungskonzepten, Barrierefreiheit und Schulungen. Diese können auch von anderen Mitnutzenden bereitgestellt werden, wie es beispielsweise Hamburg bei LLMoin anbietet. Diese Dokumente müssen auf die entsprechenden rechtlichen und organisatorischen Gegebenheiten angepasst werden, liefern jedoch eine gute Grundlage für eine möglichst niedrigschwellige Umsetzung der Compliance.

Etwas schwieriger gestaltet sich das bei Open Source Lösungen. F13 bietet eine Beschreibung der nötigen Dokumente auf der Website und weist außerdem auf die Community-Unterstützung hin. Zu diesem Zweck wurden eigene Kanäle eingerichtet, zum Beispiel zu Fragen der Nutzung oder des Deployments. Außerdem ist die Bereitstellung von Grundlagen und Best-Practices in Form von White-Label-Dokumenten geplant.
MUCGPT stellt keine Dokumente zur Nachnutzung bereit, es wird darauf verwiesen, dass die Bedingungen bei der Nachnutzung sehr individuell sein können.
PLAIN als PaaS-Lösung verweist auf eine vertikal geschnittene Verantwortungsteilung zwischen Plattformbesitzer, Plattformbetreiber und Mandanten. Ein Datenschutzkonzept deckt den Plattformbereich ab und dient gleichermaßen als Blaupause für die Mandantenbereiche.

4.7 Mandantentrennung und Skalierbarkeit

Cloudbasierte Systeme und Plattformen erlauben durch die gemeinsame Nutzung von Ressourcen eine weite Skalierbarkeit, nicht nur in Bezug auf Verarbeitungskapazitäten der Anwendungen, sondern auch im Verbund von Nutzern, was Konzepte der Mandantentrennung erfordert. In den Steckbriefen finden sich Angaben zur Skalierbarkeit auf dem Blatt „Technik-Stack" im „Compute-Schicht", die „Mandantentrennung" wird auf dem Blatt „Funktionales" behandelt.
Insbesondere für die Mitnutzung von Software ist es unerlässlich, Mandantentrennung auf der Plattform zu implementieren und Berechtigungskonzepte festlegen zu können.
Beispielsweise NRW.Genius implementiert die Mandantentrennung auf Datenbankebene. Daten werden mandantenspezifisch isoliert und durch gruppenbasierte Zugriffskontrollen geschützt. Die Mandantenstruktur basiert auf Active Directory Gruppen, sodass jede Organisation ihre Nutzenden eigenständig verwalten kann. LLMoin bietet hier zwei verschiedene Modelle an: entweder die private-Lösung, bei der jeder Mandant über einen eigenen Infrastruktur-Namespace verfügt oder die shared-Lösung, wo die Mandantentrennung direkt auf der Anwendungsebene durch geteilten Namespace erreicht wird. KIPITZ und PLAIN sind durchgängig auf Mandantentrennung ausgelegt. Auf PLAIN können LLMs zwischen Mandanten (Behörden) geteilt werden. F13 hat Mandantentrennung in ihrer Roadmap als Feature gelistet.

5 Ergebnisse und Empfehlungen

Abschließend werden die wesentlichen Erkenntnisse aus dem Vergleich der GenKI-Lösungen und der Arbeit an dieser Vorstudie kurz zusammengefasst und Handlungsempfehlungen abgeleitet.
Insgesamt sehen wir die parallelen Entwicklungen von GenKI-Systemen aufgrund der Dynamik des Themenfeldes als unvermeidbar und nicht unbedingt als nachteilig an. Stattdessen sollte die derzeitige Ausgangslage als Chance gesehen werden, die unterschiedlichen verwaltungsseitigen Bedarfe so digital souverän wie möglich zu erfüllen, auch indem durch das Lösungsportfolio auf außereuropäische Anbieter weitgehend verzichtet werden kann.

Nach dem detaillierten Vergleich der GenKI-Lösungen lassen sich folgende Erkenntnisse zusammenfassen:

GenKI-Lösungen bieten ähnliche Services auf unterschiedlichen Wegen an

Die betrachteten GenKI-Lösungen lassen sich in einem Angebotsspektrum verorten – von verhältnismäßig einfach nutzbaren SaaS-Diensten bis zu hochflexibler Open-Source-Software. Entlang eines Kontinuums stehen sich einfache Nutzung mit klaren Betriebs- und Compliance-Garantien und maximale Offenheit/Anpassbarkeit gegenüber.

- SaaS-Angebote ermöglichen schnelle, unkomplizierte Nutzung, sind jedoch nur begrenzt anpassbar; Garantien zu Betrieb, Weiterentwicklung und Compliance müssen vertraglich geklärt werden.
- SaaS mit Weiterentwicklungsoptionen bzw. austauschbaren Services kombiniert Einfachheit mit selektiver Flexibilität (beispielsweise Eigenentwicklung via SDK, modulare Service-Austausche).
- Open-Source-Lösungen bieten volle Flexibilität und Wechselfähigkeit, erfordern aber hohe interne Kompetenzen und eigenständige Compliance-Umsetzung; Unterstützung erfolgt primär über Communities.

Eine GenKI-Lösung kann auch sowohl als SaaS angeboten, als auch als Open-Source-Software bereitgestellt werden. Dagegen sprechen teilweise Fragen der Geheimhaltung (beispielsweise aufgrund traditioneller Sicherheitsphilosophie oder der Gefahr der Aufdeckung von Interna).

Nachnutzung wird mitgedacht, explizite EfA-Umsetzung ist selten

Obwohl die Nachnutzung bei allen GenKI-Lösungen mitgedacht wird – von hoher Konfigurierbarkeit über den geplanten Austausch von Fachanwendungen bis hin zum schon realisierten Einsatz in anderen Verwaltungseinheiten oder der Bereitstellung als Open Source Software – wird die explizite EfA-Nachnutzung bislang nur ansatzweise thematisiert. So versteht sich KI-PITZ im Sinne des EfA-Prinzips als die zentrale Plattform für die Bundesverwaltung. Ansonsten dominiert derzeit eher der konkrete Aufbau von KI-Infrastrukturen und die Einführung neuer Anwendungen in die Verwaltungspraxis. Das liegt daran, dass die formalen EfA-Anforderungen zusätzlich die Komplexität im Entwicklungsprozess sowie bei Abstimmungen erhöhen und neue rechtliche Fragenstellungen aufwerfen. Die Komplexität kann begrenzt werden, wenn klar abgegrenzte Teile von GenKI-Systemen nachgenutzt werden, eingebettet in einem gemeinsamen Verständnis einer technischen und organisatorischen Struktur von GenKI-Systemen.

Ein Hauptaufwand der Nachnutzung von KI-Anwendungen bleiben Compliance-Themen. Communities können Einführungen, Erfahrungsaustausch und Compliance-Umsetzung erleichtern; daraus ergeben sich auch Beratungs- und Hosting-Geschäftsmodelle.

Eine weitergehende Unterstützung bieten vorbereitete Checklisten, Muster-Dokumentation oder sogar bedingte Garantien für Open-Source-Software. Wichtig ist, dass die Dokumentations- und Prozesskomplexität tatsächlich reduziert wird.

Organisatorische Schnittstellen zwischen zentralen Bausteinen des GenKI-Systems erkennbar

Einige der GenKI-Lösungen trennen konsequent – technisch und organisatorisch - zwischen GenKI-Anwendung und KI-Infrastruktur. Das kann bedeuten, dass ein Team für die Entwicklung und Bereitstellung der KI-Infrastruktur verantwortlich ist, während ein anderes Team die GenKI-Anwendung entwickelt. Anschließend wird für den Betrieb der GenKI-Anwendung auf die KI-Infrastruktur zurückgegriffen. Diese Trennung ergibt auch aus Kompetenzperspektive Sinn, da KI-Kompetenz gebündelt werden kann. Das versetzt Teams auch ohne spezifische KI-Kompetenz in die Lage, GenKI-Anwendungen zu bauen.

Weiterhin spricht für eine solche Trennung, dass die Fachanwendungen von der dynamischen Entwicklung bei großen Sprachmodelle entkoppelt werden können: Sprachmodelle nutzen Spezialhardware, zukünftig sind leistungsfähigere Sprachmodelle denkbar, wie auch spezialisierte Sprachmodelle aus vertrauenswürdigen Cloud-Strukturen. Beachtet werden müssen in

jedem Fall die organisatorischen Implikationen. Durch unterschiedliche Betriebsmodelle ändern sich Zuständigkeiten und sowohl die IT-Strategie als auch die Beschaffung muss die aus dieser Schnittstelle folgenden Optionen berücksichtigen.

Aus dem Stand der analysierten GenKI-Lösungen und den übergreifenden Betrachtungen ergeben sich die folgenden **Handlungsempfehlungen**:

1) Zielgruppen und deren Bedarfe systematisch erfassen

Es sollte ein kompaktes Szenarien-Portfolio erstellt werden, das derzeit typische Anwendungsfälle (beispielsweise Recherche, Chat-Assistenz, Dokumentenaufbereitung, usw.) umfasst und diesen konkreten Gruppen von Behörden und Organisationen zuordnet. Aus diesen Szenarien sind zentrale Bedarfe abzuleiten, etwa nach Anpassbarkeit und Integrationsgrad oder nach Datenschutzanforderungen bzw. Compliance-Unterstützung. Parallel dazu braucht es die unterschiedlichen Kompetenzprofile – fachlich, technisch und organisatorisch – um diese mit den passenden Angebotsformen (SaaS, SaaS mit Möglichkeiten zu Eigen- oder Weiterentwicklungen, Open-Source-Software) zu verknüpfen, um sowohl eine koordinierte Angebotspalette ableiten zu können als auch neue notwendige Kompetenzen bei Einführung und Betrieb von GenKI-Anwendungen in der öffentlichen Verwaltung aufbauen zu können.

2) Begriffe und Systemstrukturen schärfen

Zentrale Begriffe wie Plattform, Anwendung oder KI-Infrastruktur sind konsistent zu definieren, um ein gemeinsames Verständnis zu fördern und Vergleichbarkeit zu ermöglichen. Eine technische Referenzarchitektur dient dabei als Basiskonzept, braucht aber eine Ergänzung aus der organisatorischen Perspektive. Während die technische Referenzarchitektur schrittweise immer detaillierter die technische Realisierung adressiert, steht bei der Strukturierung nach organisatorischen Gesichtspunkten die Umsetzung beziehungsweise Umsetzbarkeit der Einführung in einer Behörde im Mittelpunkt.

Somit umfasst das organisatorische Referenzmodell nicht nur eine technische Abgrenzung zwischen Anwendung, Inferenz und Datenspeicherung, sondern auch (Vorschläge für die) organisatorische Schnittstellen im Hinblick auf Aspekte wie Betriebsmodelle, Zuständigkeiten, Compliance und Beschaffungsprozesse. Abgestimmt auf die oben genannten bedarfsgerech-

ten Anwendungsszenarien sowie der dafür notwendigen Kompetenzen erleichtert die Schärfung von Begriffen und Systemstrukturen die Konvergenz von GenKI-Lösungen sowie Vergleich und Kombinierbarkeit für die nachnutzenden Organisationen.

3) Compliance als Querschnittsthema bearbeiten

Bei der Nutzung von GenKI-Lösungen helfen praxisnahe Checklisten und vorbereitete Dokumente als Teil der Dokumentation bei Einführung und Betrieb. Die unterschiedlichen Angebotsformen von GenKI-Lösungen können mit abgestuften Compliance-Support-Stufen verknüpft werden, um beispielsweise Checklisten zwischen Lösungen verschiedener Anbieter auszutauschen sowie die nutzende Organisation möglichst weitgehend zu unterstützen.
Die Unterstützung bei Compliance-Anforderungen kann, im Idealfall auf Basis standardisierter Angebotsformen, auch unabhängig von konkreten GenKI-Lösungen angeboten werden. Ein Schritt dahin kann die Etablierung einer anbieterneutralen, Community-getriebenen Wissensbasis und Austauschplattform (Marktplätze) sein, um von Dokumentation über Erfahrungen bis zu Anwendungsbeispielen breit verfügbar zu machen. Ein Startpunkt könnte der Marktplatz der KI-Möglichkeiten (https://maki.beki.bund.de) sein.

4) Konvergente Architekturen und Open-Source vorantreiben und parallel Nachnutzungspotenziale realisieren

Alle betrachteten GenKI-Lösungen haben einen nachvollziehbaren Zweck und erzielen konkreten Nutzen. Die Analyse der Interviews zeigt außerdem, dass eine Einigung auf eine einzelne Lösung derzeit nicht realisierbar scheint. Zudem ist unklar, welche zusätzlichen technische, organisatorische und rechtliche Aufwände zu leisten wären.
Realistischer ist eine (weitere) technische sowie organisatorische und lizenzrechtliche Annäherung, von der alle Seiten profitieren. Eine bereits erkennbare Möglichkeit wäre, Microservices konsequent nach gleichen Standards zu entwickeln und Open Source zu stellen. So könnte ein produktiver Austausch entstehen, ohne bestehende IT-Strukturen oder Geschäftsmodelle aufzugeben.
Es sollte angemerkt werden, dass es inzwischen eine ganze Reihe verschiedener Lösungen gibt, die ein breites Spektrum an Anwendungsfällen abdecken. Daher ist es sinnvoll, vor einer Neuentwicklung auf kommunaler oder Landesebene auf das Nachnutzungspotenzial bereits bestehender Lösungen zurückzugreifen, falls dies organisatorisch und technisch möglich ist.

5) Positionierung zum Deutschland-Stack klären und aktiv gestalten

Der Deutschland-Stack befindet sich noch in einer frühen Entwicklungsphase, in der parallel zur Nennung erster konkreter Standards und Technologien eine systematische Klärung der übergeordneten Zielsetzungen erforderlich ist. Typische Software-Komponenten wie Datenbanken sowie eine eigene Gruppe „KI" in der Plattform-Schicht werden Teil des D-Stack werden und bieten damit einen geeigneten Bezugsrahmen für die Nutzung und die Entwicklung von GenKI-Anwendungen in der öffentlichen Verwaltung.

Entwickler- und Nutzergruppen aus dem GenKI-Umfeld sollten diese frühe Phase aktiv nutzen, um durch einen strukturierten Konsensbildungsprozess zentrale Bedarfe und Anforderungen an den Deutschland-Stack zu formulieren. Soweit bereits vorhanden, können Ergebnisse aus der Arbeit an einem KI-Ökosystem als Grundlage dienen, um spezifische Anforderungen aus dem GenKI-Kontext in die Weiterentwicklung des D-Stack einzubringen. Dies stärkt nicht nur die Relevanz des D-Stack für GenKI-Anwendungen, sondern sichert auch die Berücksichtigung praxisorientierter Bedarfe.

Fragestellungen für die Hauptstudie

Entsprechend dem Zielbild dieser Vorstudie und unter Berücksichtigung der vorgeschlagenen Handlungsempfehlungen ergeben sich folgende Fragestellungen für die Hauptstudie:

- Welche Nutzergruppen gibt es und was sind deren spezifische Bedarfe?
- Wie kann ein technisch praktikabler und rechtlich sicherer Austausch von Service-Komponenten zwischen den GenKI-Lösungen verbessert werden?
- Inwieweit kann das EfA-Prinzip auf diesen klar definierten Bereich angewendet werden? Welche weiteren Bereiche gibt es, auch perspektivisch, in Bezug auf verwaltungsspezifische Sprachmodelle?
- Wie kann Compliance als Querschnittsthema das KI-Ökosystem stärken und die Nachfrage nach konformen GenKI-Lösungen erleichtern?

Abbildungsverzeichnis

Tabellenverzeichnis

Quellenverzeichnis

Die Informationen zu den benannten GenKI-Lösungen in dieser Vorstudie finden sich unter dem Einstiegslink, der bei der Kurzvorstellung der GenKI-Lösungen in Kapitel 3 angegeben ist oder beruhen auf internen Dokumenten und Interviews.
Hier werden nur zusätzlich verwendete Dokumente aufgeführt. Das letzte Abrufdatum der Onlinequellen ist der 16.12.2025.

[Bitkom] Bitkom e.V.: Umsetzungsleitfaden zur KI-Verordnung (EU) 2024/1689, https://www.bitkom.org/Klick-Tool-Umsetzungsleitfaden-KI-Verordnung

[BMDS] Mindestanforderungen an „Einer für Alle"-Services *(insb. Technik),* https://www.digitale-verwaltung.de/SharedDocs/downloads/Webs/DV/DE/EfA/efa-mindestanforderungen.pdf?__blob=publicationFile&v=2

[FOKUS] Fraunhofer FOKUS: EfA im Fokus - Studie zur Organisation der kommunalen Nachnutzung von EfA-Leistungen und der Nutzung des EfA-Marktplatzes, https://www.fokus.fraunhofer.de/de/dps/projekte/EfA-Studie.html

[ITPLR] Mindestanforderungen an den Betrieb von „Einer für Alle"-Services *(insb. Rollen)*, https://www.it-planungsrat.de/fileadmin/beschluesse/2023/Beschluss2023-07_EfA_Mindestanforderungen.pdf

[KIVA.arc] Innenministerium Baden-Württemberg: Referenzarchitektur KI-Plattform für die Öffentliche Verwaltung, https://baden-wuerttemberg.usercontent.opencode.de/innenministerium/kiva/docs/architecture/high_level_architecture/

[ÖFIT] Kompetenzzentrum Öffentliche IT: Kompetenzen für den Einsatz generativer Künstlicher Intelligenz in der Verwaltung, https://www.oeffentliche-it.de/publikationen/kompetenzen-fuer-den-einsatz-generativer-kuenstlicher-intelligenz/

Anhang: Fragestellungen zur Untersuchung der GenKI-Lösungen

Das KI-Kompetenzteam hat folgende Fragestellungen zur Untersuchung der GenKI-Lösungen zusammengestellt. Dazu soll die Vorstudie die KI-Systeme und die verwendeten Schnittstellen-, Daten- und Programmierungsstandards anhand der folgenden technischen, rechtlichen und prozessualen Rahmenbedingung untersuchen:

Modell- und Anbieter-Unabhängigkeit:

- Lassen sich verschiedene Sprachmodelle im System betreiben und welche Inferenz-Lösungen sind im Einsatz?
- Welche Kosten sind damit verbunden?
- Wie weit wird auf europäische souveräne Sprachmodelle gesetzt?
- Gibt es eine Roadmap für zukünftige Entwicklungen und Erweiterungen der Plattformen?

Lizenzen und Open Source:

- Vor- und Nachteile von Open-Source sollten mitgedacht oder beschrieben werden.
- KI-Anwendungen sollten mit niedrigem Aufwand geteilt, weiterentwickelt und wiederverwendet werden können. Sprachmodelle sollten mit niedrigem Aufwand austauschbar sein.
- Zu welchem Grad sind Systeme (Anwendungen und Sprachmodelle) Open Source oder proprietär?

Vertrauenswürdigkeit & Ethik:

- Wie können Ergebnisse geprüft werden?
- Wie kann die Qualität der Ergebnisse gemessen und evaluiert werden?
- Werden Anwendende auf Grenzen der Technologie hingewiesen?
- Werden allgemeine Regeln zu Barrierefreiheit eingehalten? (z. B. Kompatibilität mit gängigen Screenreadern, hier gerne Vergleich von Dokumentation der Anbieter und keine Testverfahren.)

Modularität:

- Folgen die Systeme einer Microservice-Architektur und verwenden gängige Schnittstellenstandards?
- Wie sind Datenbanken angebunden und welche Services sind im Einsatz (Load-Balancing, Parser, Multi-RAG, etc.)?
- Wie wird Organisationswissen (geschütztes und ungeschütztes) in den Systemen eingebunden (z. B. RAG-Lösungen)?

Daten- und Geheimnisschutz:

- Wie sind Datenintegrität und Informations-Sicherheit umgesetzt?
- Können Daten je nach Anwendungsfall on-premise, in privater Cloud oder in public Cloud verarbeitet werden?
- Werden die Vorgaben der Europäischen KI-Verordnung und der Datenschutz-Grundverordnung eingehalten? (Erwähnung der relevanten Vorgaben und Vergleich)
- Wie werden Authentifizierung und Autorisierung umgesetzt (z. B. Anbindung an behördliche Single Sign-on-Lösungen)?
- Wird eine BSI-Zulassung angestrebt?

Mandantentrennung und Skalierbarkeit:

- Welches Berechtigungskonzept wird bei unterschiedlichen Nutzergruppen eingesetzt?
- Mandantentrennung: Inwiefern wird die Trennung der Daten und Wissensdatenbanken von unterschiedlichen Nutzern gewährleistet?
- Inwiefern wird die vorhandene KI-Infrastruktur durch mehrere Behörden genutzt, um Leistungsspitzen zu decken und Ressourcen zu schonen? (z. B. Infrastruktur-Ressourcen, wie Grafikkarten.)
- Sind die Systeme für verschiedene Nutzungsszenarien skalierbar (z. B. kleine Kommunen vs. große Bundesbehörden)?

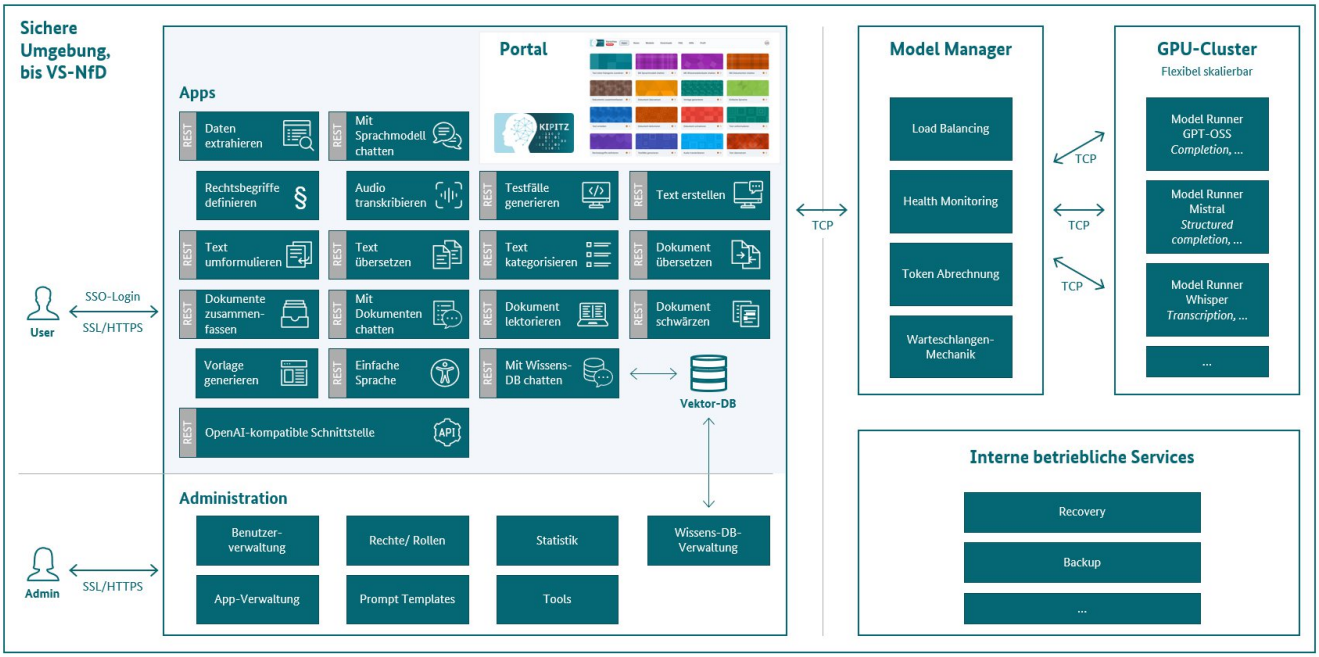
Bei der technischen Orientierung an generischen KI-Architekturen und Best Practices sollen beispielsweise folgende Funktionen betrachtet werden:

- Multi-LLM / Modellagnostik
- Möglichkeit zum Erstellen eigener Assistenten
- Einrichten von Prompt-Bibliotheken
- APIs für Funktionen und Modelle (zur Integration in Digitalisierungsvorhaben)
- Berechtigungsverwaltung (Abbildungsorganisation und Rollen)
- Betrieb Cloud und on-Premise
- Kollaborationsmöglichkeiten (z. B. Bewertung und Teilen von Assistenten)
- Einbinden eigener Modelle
- Websuche
- Anzeigen von Quellen

Anhang: Architekturbilder der Projekte

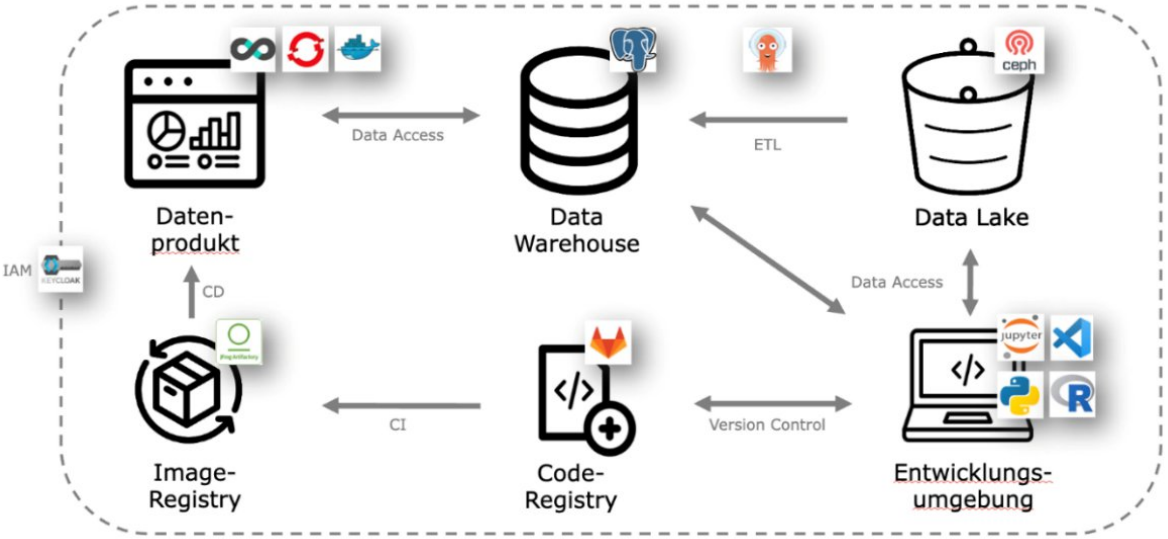
In diesem Anhang finden sich Übersichts-Architekturbilder der einzelnen GenKI-Systeme als Überblick und zum Vergleich. Es wurden Darstellungen ausgewählt, die möglichst über dargestellte Datenpfade einen ersten Einblick in die Funktionsweise als auch in wichtige Komponenten des Systems bieten. Bei öffentlich verfügbaren Grafiken ist die Quelle angegeben.

KIPITZ



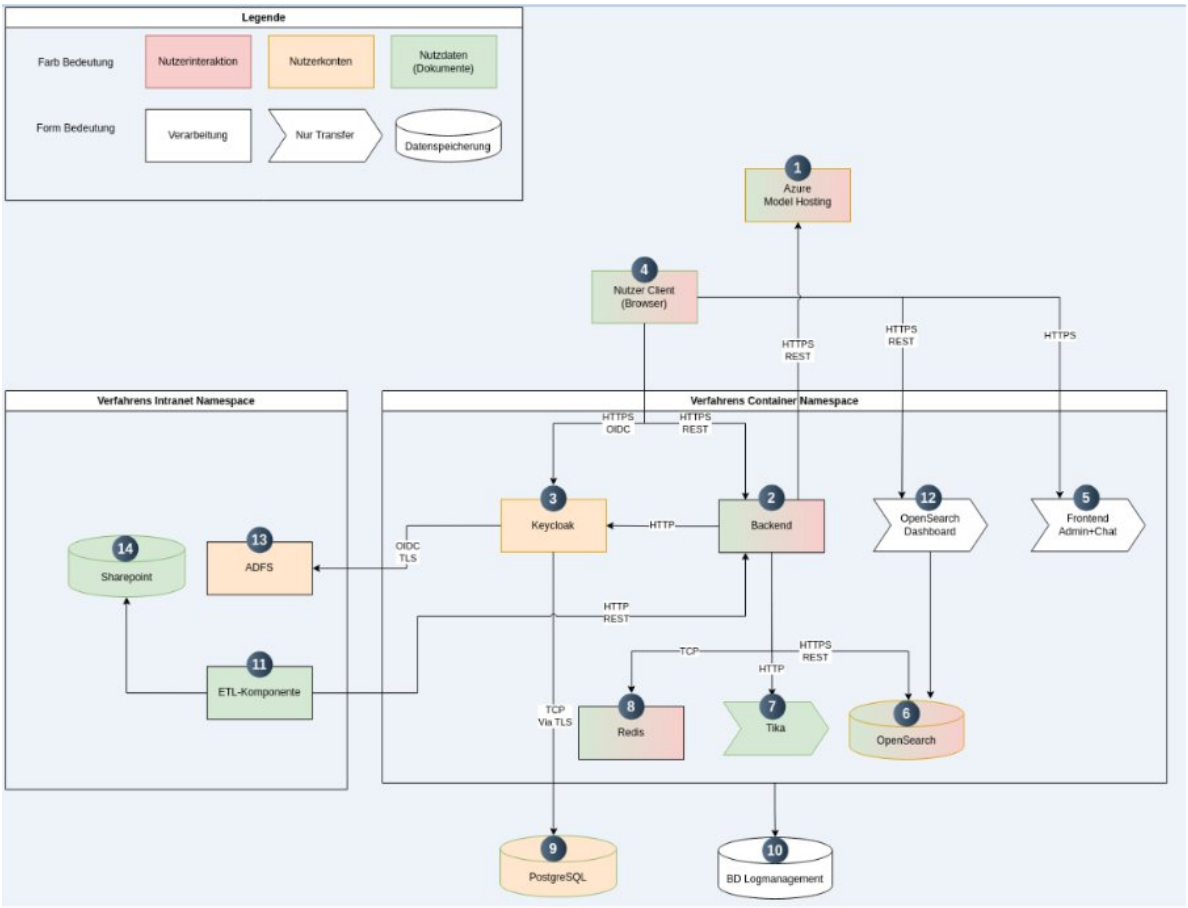
Technisches Architekturbild KIPITZ

PLAIN



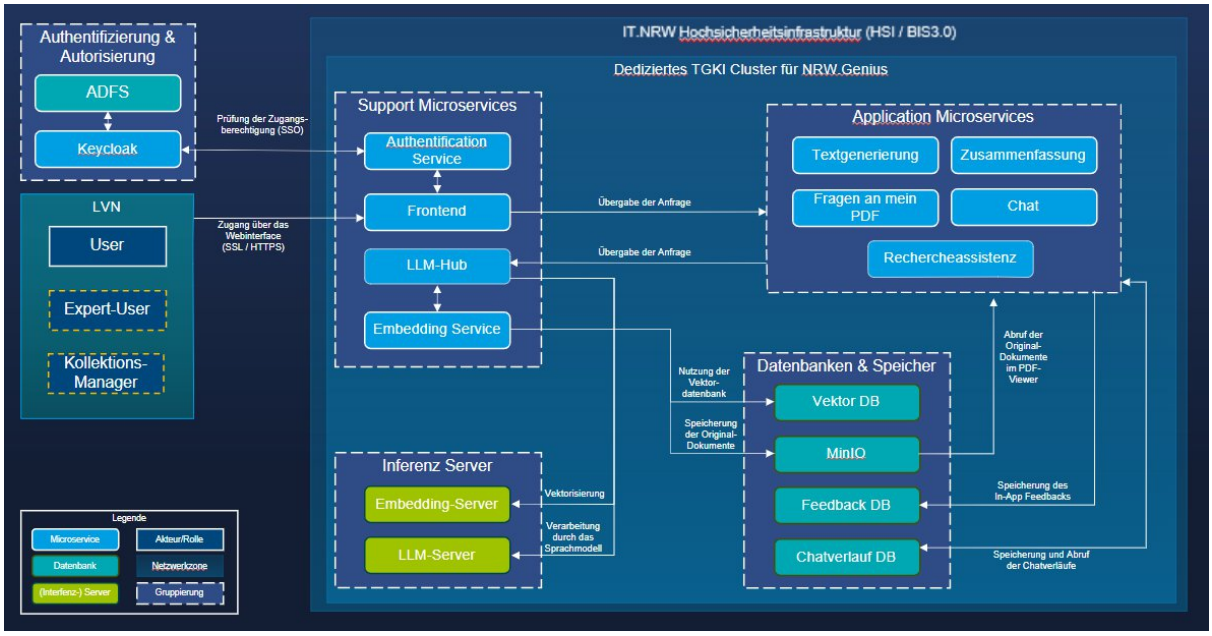
Funktionsweise und Plattform Services PLAIN

LLMoin



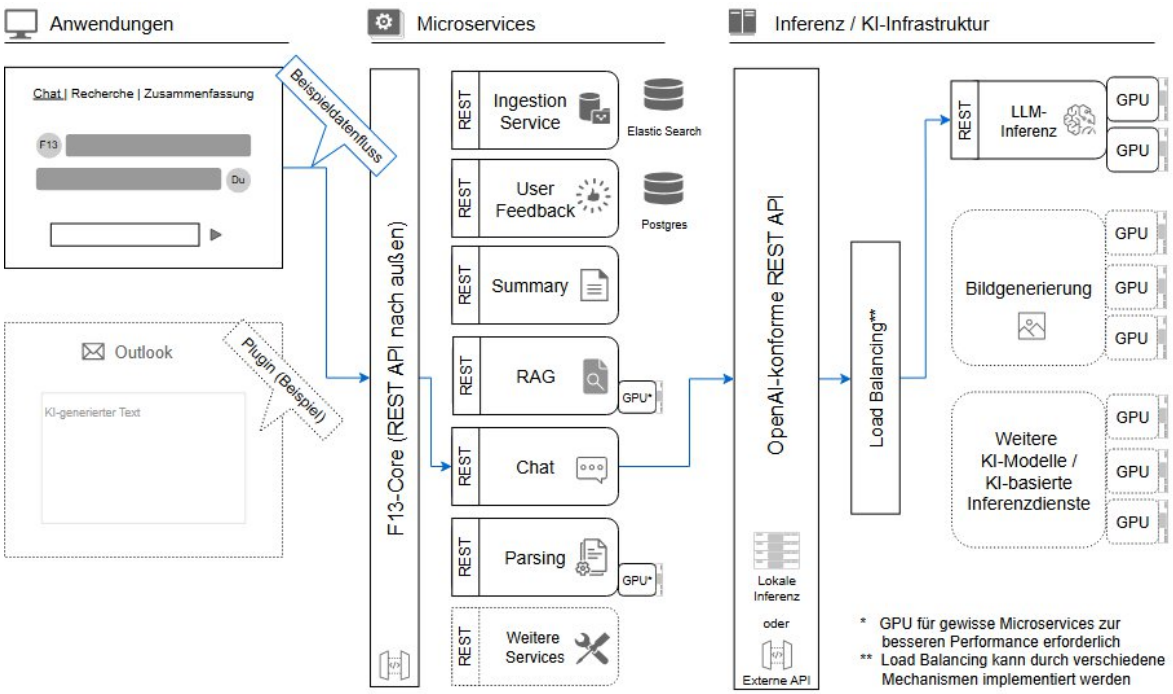
Technisches Architekturbild

NRW.Geni-us



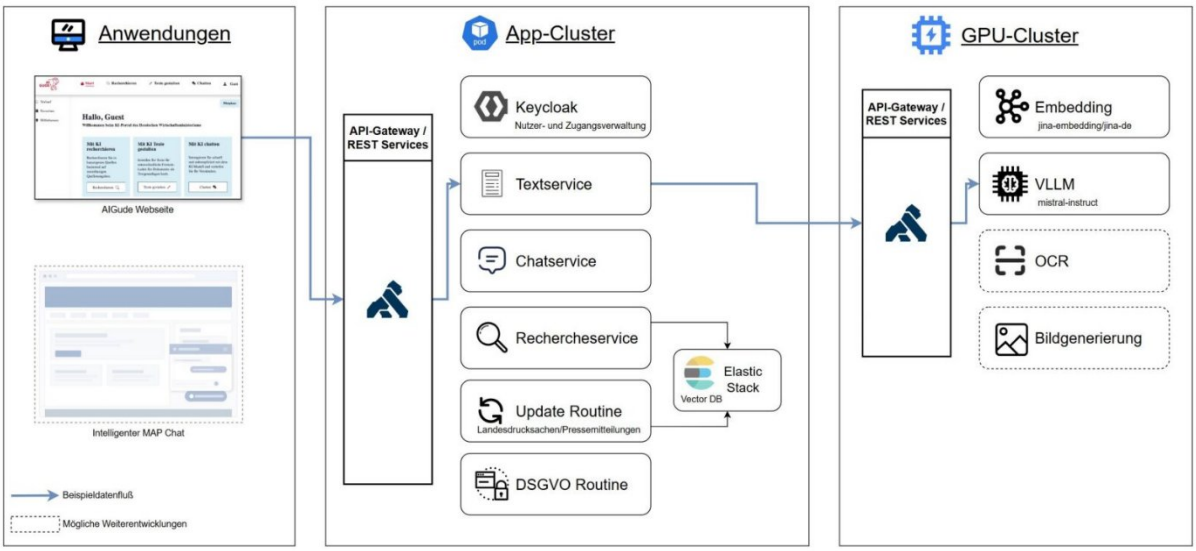
NRW.Genius On-Prem Architektur (Release 1.0)

F13



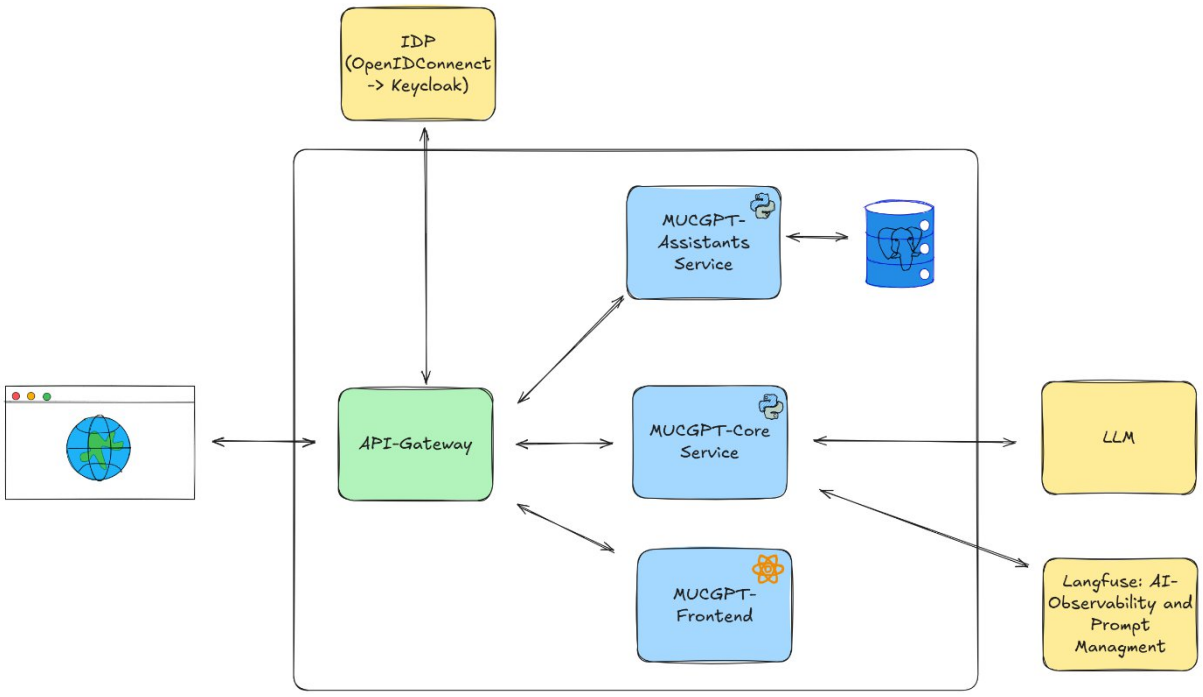
F13 Softwarearchitektur (Quelle: https://f13-os.de/kennenlernen/softwarearchitektur)

AI-Gude



AIGude Architekturbild

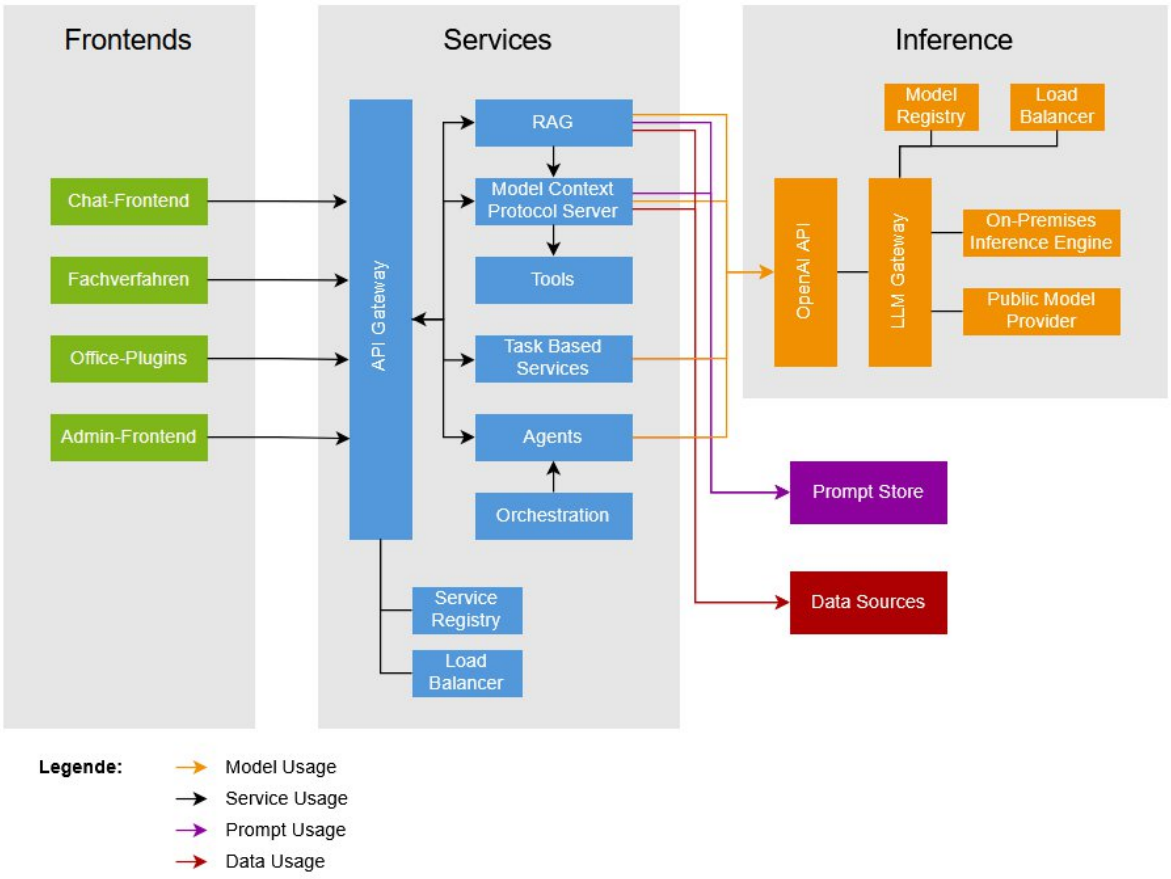
MUCGPT



MUCGPT Architecture Overview

(Quelle:https://github.com/it-at-m/mucgpt/blob/main/docs/architecture.png)

zum Vergleich: KIVA.arc High-Level-Architektur



(Quelle: https://baden-wuerttemberg.usercontent.opencode.de/innenministerium/ki-va/docs/architecture/high_level_architecture/)