Infrastruktur und Standards für Generative KI in der Öffentlichen Verwaltung

Zwischenbericht mit Empfehlungen zu Design und Umsetzung

eines nachhaltigen und souveränen KI-Ökosystems

Eine Ausarbeitung im Rahmen des Kompetenzteams Künstliche Intelligenz im Schwerpunktthema Datennutzung des IT-Planungsrates

Kompetenzteam Künstliche Intelligenz
Maßnahmenteam 2
Version: 0.9
Datum: 18.11.2024

Inhalt

Inhalt

1 Zusammenfassung

In diesem Zwischenbericht geben wir, das Maßnahmenteam „Infrastruktur und Standards“ des KI-Kompetenzteams, einen Überblick über den aktuellen Implementierungsstand einiger KI-Vorhaben in der deutschen Verwaltung sowie bestimmter Erfahrungen und Empfehlungen aus unseren Projekten. Dabei fokussiert sich dieser Bericht auf die Nutzung Generativer KI (GenAI).

Aus dieser Analyse leiten wir **zwei Empfehlungen** ab:

1. Mittel- bis langfristig ist es notwendig, eine interföderale Plattform zu schaffen zur behördenübergreifenden Entwicklung, Beschaffung sowie Nutzung von GenIA-Anwendungen. Dafür soll **eine Machbarkeitsstudie** zeitnah Voraussetzungen, Rahmenbedingungen und Umsetzungsoptionen identifizieren.
2. Um in der Zwischenzeit der hohen Dynamik der Thematik gerecht zu werden, wird **kurzfristig der Aufbau eines Kompetenznetzwerks KI Infrastruktur** vorgeschlagen, um gemeinsame Standards und Architekturvorlagen zu erarbeiten.

Die öffentliche Verwaltung ist bei der Nutzung von KI und insbesondere GenAI vergleichsweise schnell und erfolgreich vorangegangen. Der „Marktplatz der KI-Möglichkeiten“ des BMI, der ebenfalls im Rahmen des Kompetenzteams KI des IT-Planungsrates zur interföderalen Nachnutzung aufbereitet wird, ermöglicht einen ersten Überblick über bestehende Anwendungen auf den verschiedenen Verwaltungsebenen. So können potenziell **nachnutzbare KI-Systeme** schnell und unbürokratisch identifiziert werden. Um die überall **knappen Ressourcen** möglichst effizient nutzen zu können, sind nachnutzbare KI-Anwendungen von elementarer Wichtigkeit. Damit diese Nachnutzung tatsächlich gelingt, sind **kompatible Architekturen** und **interoperable Infrastrukturen** unabdingbar. Um diese sicherzustellen, benötigen wir gemeinsame offene Standards und Vorlagen, an denen wir uns als öffentliche Verwaltung wie auch Dienstleister bei der Entwicklung und/oder Beschaffung von GenAI-Anwendungen orientieren können. Nicht zuletzt zahlen diese Standards und Vorlagen auch auf das Ziel ein, die **digitale Souveränität** des Staates bei der Nutzung von KI-Anwendungen mit sicherzustellen.

Der vorliegende Zwischenbericht beginnt in Kapitel 2 mit der Darstellung der Ziele, des Adressatenkreis sowie des thematischen Fokus. In Kapitel 3 werden verschiedene Nutzungspotenziale von Generativer KI in der Verwaltung illustriert.

Kapitel 4 beschreibt mögliche KI-Architekturen, die diese Nutzungspotenziale ermöglichen, bevor Kapitel 5 die bereits bekannten föderalen Ansätze zum Aufbau von KI-Plattformen beleuchtet:

- Das EfA-Prinzip als Arbeitsrahmen für KI-Anwendungen,
- Das Projekt KIPITZ zur Entwicklung einer KI-Plattform des Bundes,
- Die Bereitstellung von Entwicklungsumgebungen für KI durch GovDigital,
- Das Projekt PLAIN zur Entwicklung einer KI- und Datenanalyse Platform der Bundesdruckerei und des Auswärtigen Amtes und schließlich
- Die Plattform des GovTech Campus

Darüber hinaus wird im Anhang des Berichts eine Auswahl an bestehenden GenAI-Anwendungen bei den im Maßnahmenteam beteiligten Ländern vorgestellt. Hier wird die Vielfältigkeit von KI und seinen potenziellen Anwendungen, aber auch die Gemeinsamkeiten der Systeme wie **F13** in Baden-Württemberg, **LLMoin** in Hamburg, **AIGude** in Hessen oder **NRW.Genius** in Nordrhein-Westfalen deutlich.

Kapitel 6 analysiert eine Auswahl der Gemeinsamkeiten dieser Vorhaben. Dabei werden beispielsweise die Aspekte OpenSource und Multi-LLM-Ansätze (Large Language Model) beleuchtet.

Der Bericht endet mit der Ableitung von zwei fachlichen Empfehlungen an den IT-Planungsrat.

2 Einleitung

2.1 Ziele des Dokuments

KI trifft auf eine öffentliche Verwaltung, die mit umfassenden Herausforderungen kämpft:

Auf der einen Seite wächst sowohl die Anzahl der inhaltlichen Aufgaben und Dienstleistungen, die auf den verschiedenen föderalen Ebenen erfüllt werden müssen, als auch die Erwartung der Bürgerinnen und Bürger an die Effizienz, Nutzerfreundlichkeit, Verfügbarkeit und Professionalität dieser Dienstleistungen. Allein die Digitalisierung bestehender Verwaltungsprozesse in Zeiten knapper Haushaltsmittel fordert enorme Kraftaufwände über Landes- und Kommunalgrenzen hinweg.

Auf der anderen Seite trifft dieser wachsende Aufgabenumfang und Aufholbedarf auf einen immer akuter werdenden Fachkräftemangel: So prognostiziert beispielsweise die Bundesagentur für Arbeit, dass bis 2032 bis zu 35 Prozent der Mitarbeitenden in Rente gehen wird[1]. Rein zahlenmäßig kann dieser Verlust an Arbeitskraft von den nachfolgenden Generationen nicht abgedeckt werden – selbst wenn sich mehr Berufseinsteiger für eine Laufbahn im öffentlichen Dienst entscheiden würden und die klassischen Lebensläufe mit jahrzehntelanger Betriebszugehörigkeit noch zeitgemäß wären. Da beides nicht der Fall ist, entsteht also sowohl am oberen als auch am unteren Ende der Altersverteilung unter den Mitarbeitenden des öffentlichen Dienstes eine klaffende Lücke.

Diesen Herausforderungen gegenüber steht das immense Potenzial Künstlicher Intelligenz:

Klassifizierende KI-Systeme können schon seit längerem zur Kategorisierung, Automatisierung und Prozessierung von großen Datenmengen genutzt werden, sei es bei der Bilderkennung bspw. im Rahmen von Geodaten oder bei der richtigen Zuordnung von Akten. Neu hinzugekommen ist seit zwei Jahren die Generative KI, die nicht nur bestehende Daten erkennt und klassifiziert, sondern komplett neue Ergebnisse erstellen kann, die qualitativ bereits einen hohen Standard erreichen.

Das Ergebnis dieses Aufeinandertreffens wird in diversen Studien, Whitepapern und anderen Veröffentlichungen beschrieben: So schätzt ein kürzlich publiziertes McKinsey-Papier, dass der Fachkräftemangel in der öffentlichen Verwaltung durch den flächendeckenden GenAI-Einsatz um ein Drittel verringert werden könnte[2]. Gleichzeitig nennt eine andere Studie aus demselben Haus die zentrale Bereitstellung und technische Skalierung der Infrastruktur als eine von zwei Grundvoraussetzungen für diesen Effekt[3].

Diesen Grundvoraussetzungen widmen wir uns im vorliegenden Bericht, indem wir einige der derzeitigen Entwicklungen auf Länder und Bundesebene beschreiben. Wir sehen ein großes Risiko, dass die Potentiale der KI-Technologie in der deutschen Verwaltung nicht ausgeschöpft werden

[1] https://www.handelsblatt.com/politik/deutschland/digitalisierung-arbeitsagentur-will-bis-zu-19-millionen-euro-fuer-ki-zahlen/100079417.html, abgerufen am 18.11.2024

[2] https://www.mckinsey.com/de/~/media/mckinsey/locations/europe%20and%20middle%20east/deutschland/news/presse/2024/2024-07-15%20genai%20and%20talent%20in%20public%20sector/mckinseymit%20mut%20und%20augenmass%20bitte.pdf, S. 6, abgerufen am 29.10.2024

[3] https://www.mckinsey.de/~/media/mckinsey/locations/europe%20and%20middle%20east/deutschland/publikationen/2024-09-03%20generative%20kuenstliche%20intelligenz%20in%20der%20oeffentlichen%20verwaltung/mckinsey-artikel_generative%20knstliche%20intelligenz%20in%20der%20ffentlichen%20verwaltung.pdf, S. 2, abgerufen am 29.10.2024

können, da eine Zersplitterung der kommunalen, Landes- und Bundesinitiativen droht. **Gleichzeitig ist zu beobachten, dass - mangels strategischer, deutschlandweiter Kooperationskultur, fehlender gemeinsamer Infrastrukturen und einem stark ausgeprägtem Föderalismus - die gleichen oder ähnlichen Vorhaben mehrfach und teils unkoordiniert umgesetzt werden**. Aufgrund dieser Situation entstehen Anwendungen und Infrastrukturen teilweise mehrfach, ohne untereinander standardisiert bzw. harmonisiert zu sein. Um diesem Risiko vorzubeugen, sowie redundante Entwicklungen und Fehlinvestitionen zu vermeiden, skizzieren wir darüber hinaus einen Handlungsrahmen. Dieser **leitet sich aus den praktischen Erfahrungen** von Pilotprojekten auf Bundes- und Länderebene ab. Er bündelt somit konkretes Verwaltungswissen, um es interföderal zur Verfügung zu stellen und daraus entscheidungsrelevante Empfehlungen zu geben.

2.2 Adressatenkreis

Wir richten uns mit diesem Zwischenbericht **primär an den IT-Planungsrat**. Für Bundes- und Landesverwaltungen sowie Kommunen, die vor der Herausforderung stehen, für den Einsatz und Betrieb von KI-Anwendungen innerhalb der Verwaltung eine entsprechende Infrastruktur (neu) aufzubauen oder zu beauftragen, kann dieser Bericht wertvolle Hinweise und Denkanstöße enthalten.

2.3 Thematischer Fokus und definitorische Grundlagen

„Künstliche Intelligenz“ ist ein vager Begriff, hinter dem sich eine ganze Reihe technologischer Verfahren verbirgt. Aufgrund der hohen Heterogenität dieser Technologien und der auf ihnen fußenden Anwendungen fokussieren wir uns auf Informationen zur KI-Infrastruktur für den Einsatz von **generativen KI-Anwendungen (GenAI) in der allgemeinen Verwaltung.** Darüber hinaus verstehen wir in diesem Zwischenbericht den Begriff der „**KI-Infrastruktur**“ als die Summe aller technischer Elemente, die zur erfolgreichen Umsetzung einer GenAI-Anwendung nötig sind. Eine „**Plattform**“ dagegen ist die Summe aus Infrastruktur, konkreten Anwendungen und Begleitprozessen wie beispielsweise Identity Access Management (IAM) und Dokumentationspflichten. Die „**Architektur**“ einer KI-Anwendung beschreibt zuletzt das Zusammenspiel aller technischen Elemente und Prozesse, von Schnittstellen und Interaktionen über Softwaredesignmuster und -elemente bis hin zu Datenflüssen.

Die KI-Infrastruktur kann mehrere Komponenten enthalten:

- **Rechenleistung**: KI-Modelle, insbesondere sog. Große Sprachmodelle (LLMs), benötigen enorme Rechenkapazitäten. Diese werden oft durch Hochleistungsrechner (High-Performance Computing, HPC), spezialisierte Hardware wie GPUs (Graphics Processing Units) oder TPUs (Tensor Processing Units) sowie Cloud-Computing-Ressourcen bereitgestellt.
- **Datenspeicher und -management**: KI-Systeme brauchen eine leistungsfähige Dateninfrastruktur, die Speicherlösungen umfasst, welche sowohl große Datenmengen als auch hohe Übertragungsgeschwindigkeiten unterstützen. Datenbanken, Data Lakes und Data Warehouses spielen hier eine zentrale Rolle. Insbesondere für LLMs sind Vektordatenbanken erforderlich.

- **Netzwerkinfrastruktur**: KI-Anwendungen benötigen eine schnelle, sichere und zuverlässige Netzwerkinfrastruktur, insbesondere wenn sensible Daten genutzt werden und/oder Rechenkapazitäten über mehrere Standorte oder Cloud-Umgebungen verteilt sind.
- **Sicherheit und Datenschutz**: Insbesondere im fachlichen Verwaltungsbereich sind eine ganze Reihe von Sicherheitsmaßnahmen erforderlich, um sensible Daten zu schützen und den rechtlichen Anforderungen zu entsprechen, die sich u.a. aus der DSGVO und der KI-Verordnung, aber auch aus Fach-, Landes- und Urheberrecht ergeben.
- **Orchestrierung und Monitoring**: Der Betrieb einer KI-Infrastruktur erfordert Systeme und Prozesse, die Workflows orchestrieren, Automatisierung unterstützen und das Ressourcenmanagement übernehmen können. Diese Komponente umfasst auch Monitoring- und Logging-Systeme, um die Performance und den Zustand der KI-Modelle und -Systeme zu überwachen.

Die KI-Architektur beschreibt also, wie alle Elemente der KI-Infrastruktur zusammenwirken, die im Rahmen einer KI-Plattform von mehreren Nutzenden in Anspruch genommen werden können.

3 Nutzungspotenziale generativer KI in der Verwaltung

Im Bereich generativer AI sind folgende beispielhafte Anwendungsfälle denkbar und bei Behörden in Deutschland teilweise bereits umgesetzt. Umsetzungsbeispiele aus einigen Bundesländern finden sich im Anhang (**F13** in Baden-Württemberg, **LLMoin** in Hamburg, **AIGude** in Hessen und **NRW.Genius** in Nordrhein-Westfalen).

3.1 Text

Beispiele für die Nutzung von GenAI für Textaufgaben sind unter anderem:

- Automatisierte Beantwortung von Bürger- und Terminanfragen
- Erstellung von Berichten und Zusammenfassungen
- Unterstützung bei internen und externen Übersetzungen
- Automatisierte Erstellung von Schreiben, Bescheiden und Dokumenten
- Entwurf von Verordnungen und Richtlinien
- Automatische Protokollerstellung bei Sitzungen
- Erstellung von Erklärungen und Präsentationen für Bürger
- Automatisierte Analyse von Feedback

3.2 Medien

Sprachmodelle können auch für die Erstellung von Inhalten in anderen Medienarten genutzt werden, z.B.:

- Automatisierte Erkennung und Analyse von Bilddaten
- Erstellung von Audio-Lerneinheiten
- Generierung von Videosequenzen zur Reduzierung der Barrieren in Webseiten

3.3 Technische Inhalte

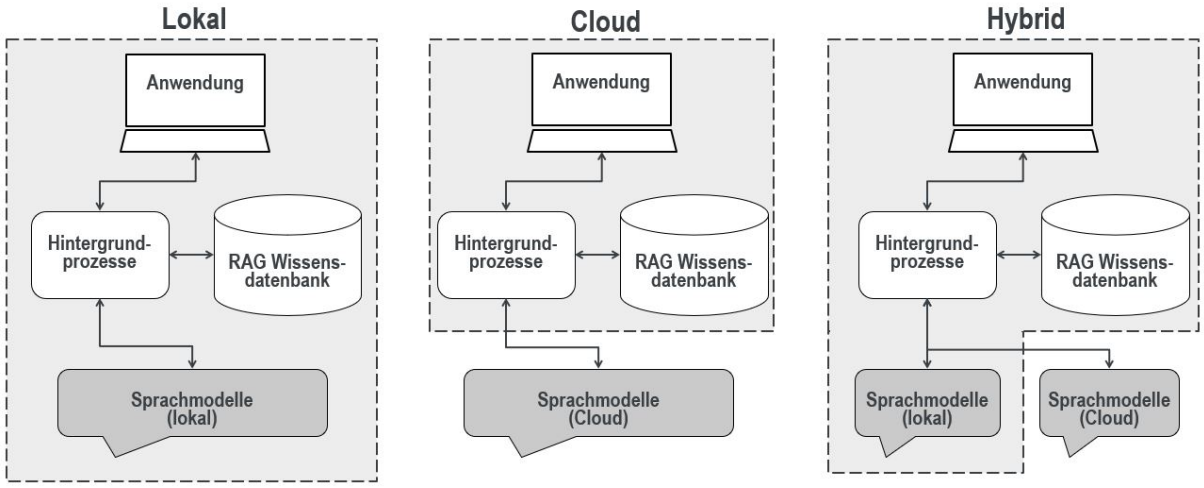
Zur Beschleunigung der Software-Entwicklung, kann GenAI Inhalte generieren, wie zum Beispiel:

- Erstellung von Code-Snippets
- Erstellung von technischen Notationen (z.B. BPMN, DMN)

4 Typische KI-Architekturen

KI-Anwendungen können und werden aufgrund der unterschiedlichen Kritikalitätsstufen der zu verarbeitenden Daten mit unterschiedlichen Architekturen umgesetzt. Die drei grundlegenden Betriebsarten für KI-Anwendungen im öffentlichen Sektor sind **Lokal (On-Premises), Cloud** und **Hybrid**.

4.1 Betriebsmöglichkeiten



Bei **lokalem Betrieb** wird das gesamte KI-System auf lokaler Infrastruktur der jeweiligen Behörde betrieben. Das gilt für die gesamte Anwendung, Hintergrundprozesse, und verwendete KI-Sprachmodelle. Dies bietet maximale Kontrolle und Datenschutz, was besonders für sensible, also personenbezogene und -beziehbare und/oder als vertraulich eingestufte Daten im öffentlichen Sektor vorteilhaft ist. Allerdings bringt es hohe Kosten und Wartungsaufwand mit sich, und die Skalierbarkeit ist begrenzt, da die lokale Hardware bei steigendem Bedarf aufgerüstet werden muss.

Im **Cloud-Betrieb** werden einige bis alle Prozesse in der Cloud betrieben, was geringere Initialkosten, schnellere Verfügbarkeit sowie eine hohe Skalierbarkeit der jeweiligen Anwendung ermöglicht. In Verbindung mit Service-Unterstützungsangeboten können auch Updates und Wartungen automatisch erfolgen, was wiederum den in der Organisation selbst anfallenden Aufwand reduziert. Allerdings entstehen Datenschutzrisiken, da Daten extern verarbeitet werden, was im öffentlichen Sektor besonders kritisch sein kann. Ebenso bestehen evtl. Risiken für die Informationssicherheit durch die Kontrollabgabe an externe Dienstleister. Zudem besteht eine Abhängigkeit vom Cloud-Anbieter.

Bei **hybridem Betrieb** werden die beiden Ansätze Lokal und Cloud kombiniert. Sensible Daten und Anwendungen können lokal betrieben werden, während weniger kritische Prozesse in der Cloud laufen. Dies ermöglicht Flexibilität und optimierte Ressourcennutzung. Allerdings ist die Umsetzung komplexer, da beide Infrastrukturen synchronisiert werden müssen.

Zusammenfassend bietet das lokale Modell höchste Datensicherheit, ist jedoch am kostenintensivsten. Die Cloud ist kostengünstiger und flexibel, birgt jedoch Datenschutz- und evtl. Informationssicherheitsrisiken.

Die Entscheidung für das jeweilige Betriebsmodell muss von Anwendung zu Anwendung individuell getroffen werden. Bestimmte Plattformen bieten alle drei Ansätze, andere wiederum sind nur für einen zugeschnitten. Unabhängig vom gewählten Betriebsansatz sollten alle in Kapitel 2.3 genannten Infrastrukturkomponenten im jeweiligen Betriebsmodell verankert sein.

5 Ansätze zum Aufbau von gemeinsam nutzbaren KI-Plattformen

Die Einrichtung einer KI-Infrastruktur ist mit sehr hohen Initial- und Betriebskosten verbunden. In der öffentlichen Verwaltung gibt es eine zunehmende Anzahl gemeinsam nutzbarer oder bereits gemeinsam genutzter KI-Plattformen, die im Folgenden vorgestellt werden.

5.1 Arten von KI-Plattformen

Grundsätzlich ist zwischen drei Arten von Plattformen zu unterscheiden:

- **Entwicklungsplattformen** ermöglichen den Aufbau und Test neuer Ideen für potenzielle KI-Anwendungen (auch Piloten, Prototypen oder PoCs [Proof of Concept] genannt). Diese Plattformen sind für eine schlanke und schnelle Entwicklung ausgelegt und enthalten viele der dafür notwendigen Instrumente, eignen sich in der Regel aber weder rechtlich noch technisch für den dauerhaften Betrieb der jeweiligen Anwendung. Sie enthalten unterschiedliche Modelle und Hilfswerkzeuge wie beispielsweise Coding-Unterstützungstools, UI/UX-Designelemente etc., häufig jedoch keine Möglichkeit zum dauerhaften Monitoring oder prozessualer Dokumentation.
- **Betriebsplattformen** wiederum haben das Ziel, eine verlässliche und skalierbare Basis für den dauerhaften und sicheren Betrieb von KI-Anwendungen über mehrere Behörden hinweg zu ermöglichen. Sie enthalten häufig höhere Rechenkapazitäten und Datenspeicher, eine engere Auswahl an Modellen und bieten Dokumentations- und Monitoringmöglichkeiten.
- **Reallabore** oder **Sandboxes** sind in Abgrenzung dazu besonders abgesicherte Plattformen, auf denen kritische Systeme verprobt und optimiert werden können, die aus verschiedenen Gründen (noch) nicht auf den beiden erstgenannten Ebenen umgesetzt werden können. Solche Reallabore sind im Rahmen der KI-Verordnung vorgesehen, aktuell im deutschsprachigen Raum aber noch eher selten.

Viele der KI-Aktivitäten der öffentlichen Verwaltung laufen aktuell auf Entwicklungsplattformen, da es sich um Prototypen und Piloten handelt. Für den dauerhaften Einsatz von KI-Anwendungen ist jedoch der Aufbau von Betriebsplattformen sowie deren nahtlose Verzahnung mit den unterschiedlichen Entwicklungsplattformen von elementarer Bedeutung. Während Betriebsplattformen häufig Aufgabe der öffentlichen IT-Dienstleister sind, werden Entwicklungsplattformen mittlerweile von fast allen privaten IT-Dienstleistern angeboten. **Für einen erfolgreichen Transfer von Pilot- in Betriebsanwendungen ist der Aufbau von offenen Standards, Vorlagen und offenen Schnittstellen unabdingbar, damit die Migration über Plattformen hinweg möglichst reibungslos funktioniert.**

5.2 Eine für Alle (EfA-Ansatz)

Im Rahmen der OZG-Umsetzung wurde das "Eine für Alle" (EfA)-Prinzip konzipiert. Länder und Kommunen sollten demnach nicht jedes digitale Verwaltungsangebot eigenständig neu entwickeln, sondern von den Digitalisierungsvorhaben anderer Länder profitieren. Das EfA-Prinzip ließe sich auch auf den Bereich der generativen KI und damit notwendigen Infrastrukturen, Standards und Schnittstellen übertragen, um so frühzeitig redundante Entwicklung gleichartiger KI-Anwendungen zu vermeiden und bestehende KI-Anwendungen nachnutzen zu können.

5.3 Aktuelle Beispiele

Nachfolgend werden einige Beispiel-Vorhaben genannt, die eine übergreifende KI-Plattform konzipieren oder pilotieren.

5.3.1 Projekt KIPITZ (Bund / ITZBund)

Das sich im Aufbau befindende KI-Portal „**KIPITZ**“ des ITZBund kann im strukturellen Aufbau als Vorbild für eine flexible KI-Infrastruktur für große Sprachmodelle und entsprechender Anwendungen für die öffentliche Verwaltung und für die potenzielle Umsetzung für Länder und Kommunen angesehen werden. Dabei können verschiedene Modelle angebunden und sowohl in vorgefertigten Anwendungen genutzt als auch in eigene Entwicklungen integriert werden. Der modulare Aufbau der einzelnen Komponenten erlaubt einen schnellen Austausch, z.B. sobald neuere Versionen verfügbar sind, aber auch die Erweiterbarkeit der einzelnen Infrastruktur-Elemente. Über Plug-Angebote kann KIPITZ zudem in andere Anwendungen wie bspw. Outlook etc. integriert werden.

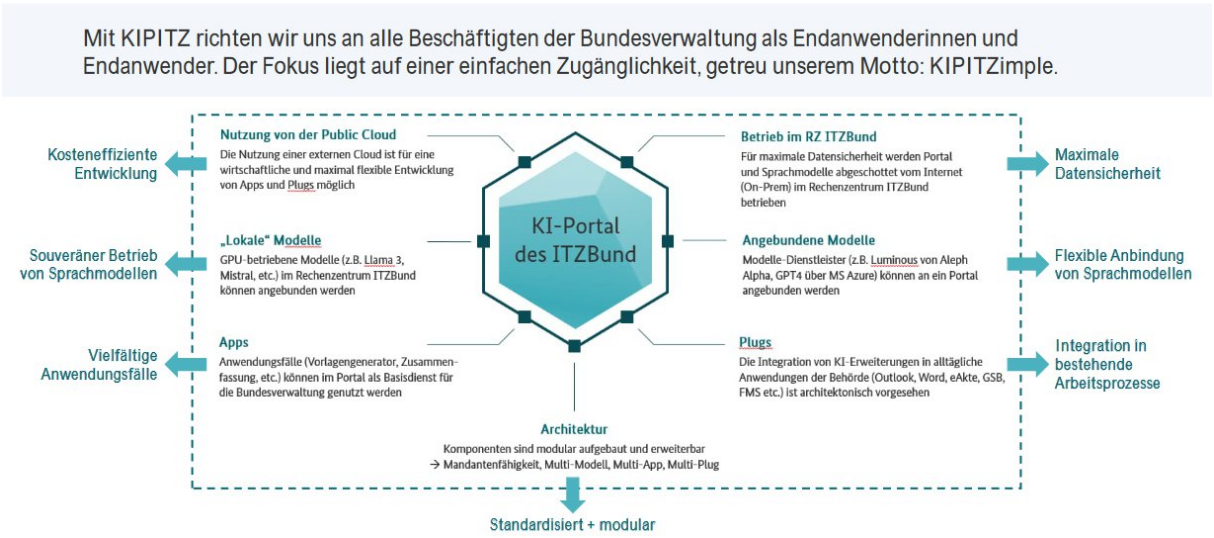


Abbildung: Überblick über KIPITZ[4]

[4] https://egovernmentwettbewerb.de/wp-content/uploads/2024/07/ITZ_Bund_KIPITZ_Kat_1.pdf, S. 3, abgerufen am 29.10.2024

5.3.2 GovDigital / DVC

GovDigital ist die Genossenschaft der öffentlichen IT-Dienstleister. Mit dem Programm gd.KI Ökosystem soll eine schnelle und sichere Implementierung von KI für die öffentliche Verwaltung ermöglicht werden. Ein Bestandteil ist die „KI-Werkstatt“, welche zur proaktiven Einbringung von Projekten und gemeinsamen Entwicklung innerhalb der Genossenschaft einlädt. Darüber hinaus ist eine „LLM-Sandbox“ geplant, welche ermöglichen soll, mit KI-Sprachmodellen in einer gesicherten Entwicklungsumgebung von GovDigital zu arbeiten. Dabei wird auch auf den Aufbau von Technologiekompetenz sowie die Notwendigkeit rechtskonformer Lösungen Bezug genommen. In der „Kollaborationsschicht“ werden schließlich die Bereitstellung sowie der Austausch von LLM-Services orchestriert. KI-Werkstatt und Sandbox sind seit Juni 2024 in Betrieb, die Kollaborationsschicht soll zeitnah folgen.

Ebenfalls von GovDigital umgesetzt werden soll die digitale Verwaltungscloud (DVC), auf der unterschiedliche KI-Anwendungen umsetzbar sein werden. Die ersten Anwendungen wurden im Sommer 2024 migriert und sind für alle Nutzenden der DVC abrufbar.

5.3.3 PLAIN

Die Bundesdruckerei hat zusammen mit der Auslands-IT des Auswärtigen Amtes die Datenanalyse und KI-Plattform PLAIN (Platform Analysis and INformation System)[5] entwickelt. PLAIN ermöglicht die Verarbeitung und Analyse großer Datenmengen aus verschiedenen Quellen in einer hochsicheren Umgebung und soll somit perspektivisch als zentrale Infrastruktur für die ressortübergreifende Zusammenarbeit in der Bundesverwaltung dienen. Die Plattform umfasst dabei Infrastructure as a Service (IaaS), Platform as a Service (PaaS) und Software as a Service (SaaS). Ein weiterer Bestandteil von PLAIN ist der „Community-Mandant,“ der 2024 eingeführt wurde und das kollaborative Arbeiten weiter verstärkt. Dieser gemeinschaftliche Datenraum ermöglicht es, KI-Anwendungen ressortübergreifend zu teilen und Synergien zu schaffen. PLAIN wurde als Beta-Version bereits 2023 veröffentlicht und befindet sich aktuell in der Pilotphase.

5.3.4 GovTech Campus Deutschland

Der **GovTech Campus Deutschland** wird eine flexible Software-as-a-Service Plattform für KI-Anwendungen auf Basis von großen Sprachmodellen anbieten, die für alle Verwaltungsorganisationen zugänglich sein soll. Somit können sich zukünftig Chancen bieten, insbesondere für kleinere Landesverwaltungen und Kommunen, KI-Anwendungen auf Basis von großen Sprachmodellen ohne Aufbau einer eigenen KI-Infrastruktur für die eigene Verwaltungsarbeit bereitzustellen.

[5] https://www.bundesdruckerei.de/de/innovation-hub/plain, abgerufen am 18.11.2024

6 Gemeinsamkeiten bestehender GenAI Anwendungen – eine Auswahl

Im Rahmen der Arbeit des Maßnahmenteams haben wir die folgenden und weitere Gemeinsamkeiten in unseren Pilotprojekten und Vorhaben identifiziert.

6.1 Multi-LLM/ Modellagnostik

Große Sprachmodelle (LLMs, z.B. Lllama 3.1, GPT 4.0) unterscheiden sich in verschiedenen Aspekten wie Kontextgröße, Kosten der Token, Aktualität, Mehrsprachigkeit usw. – zudem werden immer neue Modelle und neue Versionen bekannter Modelle entwickelt, sodass die regelmäßige Evaluation und ggf. der Austausch verwendeter Modelle wichtig ist. Auch mit Blick auf die Nachnutzbarkeit ist die Anbindbarkeit unterschiedlicher Modelle wichtig. Empfohlen wird daher, **Anwendungen modellagnostisch zu konzipieren, d.h. kompatibel mit unterschiedlichen Modellen zu entwickeln**. Welches Modell für welchen Anwendungsfall geeignet ist, kann dabei ebenfalls über jeweils aktuelle Benchmarks recherchiert werden. Wichtige Kriterien können die Abdeckung bestimmter Sprachen sein, aber auch das Verständnis für bestimmte Aufgabentypen und/oder Zusammenhänge, Toxizität und Akkuratesse oder Robustheit und Performanz. Diese Modelle können, je nach Anwendungsfall und damit verbundenen datenschutzrechtlichen- und Informationssicherheitsaspekten, entweder On-Premises oder extern gehostet werden.

6.2 Open Source

Ein robustes KI-Ökosystem in der deutschen Verwaltung wird davon leben, dass Funktionen, Anwendungen und Sprachmodelle mit möglichst wenig Aufwand geteilt, weiterentwickelt und wiederverwendet werden können. Diesbezüglich ist anzustreben, dass KI-Funktionen, KI-Anwendungen und KI-Sprachmodelle für die deutsche Verwaltung als Open-Source-Produkte implementiert werden. Bei Open-Source-Software steht der Quellcode offen zur Verfügung. Das bedeutet, jeder kann den Code einsehen, modifizieren und weiterentwickeln. Dies gewährleistet eine hohe Transparenz, da Nutzer die Funktionsweise der Software verstehen und überprüfen können.

Durch den offenen Zugang zu Software und die Möglichkeit, den Quellcode zu prüfen, fördert Open Source zudem die digitale Souveränität: Institutionen haben die Freiheit, Software nach ihren eigenen Bedürfnissen anzupassen, ohne von proprietären Anbietern abhängig zu sein. Ein weiterer Aspekt der Open-Source-Anwendungen ist ihre besondere Betrachtung in einigen rechtlichen Kontexten, wie beispielsweise der KI-Verordnung.

Bei der Entscheidung für oder gegen Open-Source-Produkte müssen also Transparenz- und Souveränitätsaspekte auf der einen und Wartungs- und Sicherheitsaspekte auf der anderen Seite abgewogen werden. Mit Blick auf die besonderen Bedürfnisse der öffentlichen Verwaltung wird empfohlen, wo möglich auf Open-Source-Produkte zurückzugreifen.

6.3 Kontextwissen durch Retrieval Augmented Generation

Mittlerweile ist bekannt, dass KI-Tools nicht immer das liefern, was man von ihnen erwartet. Während GenAI längere Texte relativ gut zusammenzufassen kann, hapert es beispielsweise bei der Generierung eigener Inhalte: Die KI-Werkzeuge halluzinieren. Halluzinationen bezeichnen dabei die von einem KI-Modell generierten Inhalte, die zwar realistisch erscheinen, aber von den vorgegebenen Quelleninputs abweichen. Man spricht von fehlender Übereinstimmung oder mangelnder faktischer Richtigkeit. Große Sprachmodelle werden mit „Weltwissen“ trainiert. Dennoch liefern sie mitunter ungenügende Ergebnisse, auch wenn die generierten Texte überzeugend klingen. Eine preisgünstige Variante, um LLMs ein gewisses Maß an Wissen zu entlocken, liegt darin, die Modelle mit Hilfe eigener Daten "einzugrenzen" und die Prompts exakter zu gestalten. Eine wesentliche Methode, um das in der Praxis umzusetzen, ist Retrieval Augmented Generation (RAG).

6.4 Modularisierung

Damit Module wiederverwendet werden können, müssen entsprechende Schnittstellen geschaffen und möglichst offen und/oder standardisiert gestaltet werden. Generative KI-Funktionen können dadurch ohne Medienbruch in anderen Arbeitsprogrammen verfügbar gemacht werden (wie beispielsweise bei KIPITZ im Rahmen von Plugs). Aber auch Schnittstellen zu bekannten Datenplattformen oder zur wissenschaftlichen Auswertung bestimmter Daten können Sinn machen. Daher sollte sowohl im Rahmen der architektonischen Entscheidungen einzelner KI-Anwendungen als auch beim Design der KI-Infrastruktur ein Schnittstellenkonzept festlegt werden, welche Daten von wem in welchem Format ein- und ausgespielt werden können.

6.5 Datenschutz und Informationssicherheit

Auch rechtliche Rahmenbedingungen können technische Umsetzungskonsequenzen haben: Für KI-Anwendungen selbst sind dies insbesondere Vorgaben in Bezug auf Erklärbarkeit und Nachvollziehbarkeit, aber auch technische Datenschutzmaßnahmen („privacy by design“) sollten Berücksichtigung in der Architektur der individuellen KI-Anwendung finden. Darüber hinaus müssen KI-Plattformen und -Anwendungen den BSI-Anforderungen entsprechen, wenn sie in entsprechend sensiblen Umgebungen eingesetzt werden sollen.

Infrastruktur und Plattformen wiederum sollten die Kontrolle der Einhaltung solcher Vorgaben erleichtern, beispielsweise indem sie Dokumentations- und MLOps-Elemente enthalten oder Schnittstellen zu den entsprechenden Prüfbehörden ermöglichen.

7 Fazit & Empfehlungen

In diesem Zwischenbericht nebst Anhang haben wir einen Überblick über die Umsetzung von GenAI-Projekten in der öffentlichen Verwaltung gegeben sowie die Vorhaben zum Aufbau einer gemeinsamen Infrastruktur dargestellt. Wir haben die grundlegenden Gemeinsamkeiten unserer bereits bestehenden KI-Anwendungen und -Systeme hervorgehoben. Daraus lassen sich Designgrundsätze für künftige Anwendungen und Systeme ableiten.

Wir haben in einer Vielzahl vorangegangener Bund-Länder-Digitalisierungsvorhaben gesehen, wie fehlende interföderale Abstimmung uns bremsen kann. Dies führt dazu, dass Deutschland insgesamt kostenineffizient und im europäischen Vergleich nur langsam mit der Digitalisierung vorankommt. Dies darf uns als Verwaltung nicht wieder passieren, wenn wir die in Kapitel 3 genannten Potenziale Generativer KI effektiv und effizient zur Bewältigung der aktuellen und anstehenden Herausforderungen nutzen wollen.

Da beim Thema KI noch nicht alle Weichen gestellt sind, besteht die Chance, frühzeitig strategisch zu koordinieren, sinnvolle Standardisierungs- und Harmonisierungsmaßnahmen zu definieren und umzusetzen. So kann eine wirtschaftliche und digital souveräne Grundlage für die verwaltungsübergreifende KI-Umsetzung etabliert werden.

Hierzu schlagen die im Maßnahmenteam 2 beteiligten KI-Expert:innen folgende Maßnahmen vor.

1. **Beauftragung einer Machbarkeitsstudie zur Implementierung einer behördenübergreifenden KI-Plattform**

   Ausgehend von der Prämisse, dass KI-Funktionen kosteneffektiv und qualitativ hochwertig für alle Ebenen der Verwaltung bereitzustellen sind, soll ein Konzept für deren Umsetzung im Rahmen einer Machbarkeitsstudie aufgezeichnet werden. Anhand von definierten Kriterien und Rahmenbedingen soll unter anderem die Machbarkeit der Bereitstellung einer zentralen Entwicklungs-, Experimentier-, und Betriebsplattform gem. Kapitel 5.1 untersucht werden.

   Die Machbarkeitsstudie soll die Nachnutzbarkeit der im Kapitel 5 aufgelisteten, bereits bestehenden Ansätze anhand technischer, rechtlicher und prozessualer Rahmenbedingung prüfen. Durch die Nutzung bereits aufgebauter Infrastrukturen durch Bund, Länder und Kommunen könnten gemeinsame Anwendungen wesentlich schneller nachgenutzt werden, zudem müssten kostspielige Komponenten wie bspw. Rechenkapazität nicht von jeder Behörde einzeln aufgebaut werden. Falls sich keiner der etablierten Ansätze nachnutzen lässt, ist der Aufbau einer neuen KI-Plattform, oder die Harmonisierung mehrerer bestehender Ansätze zu prüfen.

   Wir empfehlen, die Durchführung einer Machbarkeitsstudie in Q2/Q3 2025 zu beauftragen. Die Beauftragung erfolgt direkt durch das Kompetenzteam KI. Zur Erstellung der Studie wird eine Kooperation der FITKO mit der ÖFIT empfohlen. Das Kompetenzteam KI wird die Studie begleiten und als koordinierender Ansprechpartner für die FITKO zur Verfügung stehen. Zudem stellen wir anheim, weitere geeignete Institutionen bei der Durchführung der

Studie einzubinden, die die weitere Erfahrung und das Wissen für das Thema mitbringen. Zur Erstellung der Studie ist es sehr wichtig, dass die bereits beteiligten Akteure bei Bund, Ländern, kommunalen Spitzenverbänden sowie IT-Dienstleistern einbezogen werden, um den pragmatischen Bezug zu bestehenden Vorhaben herzustellen, sowie die Relevanz und Umsetzbarkeit der Empfehlungen zu optimieren.

2. **Aufbau eines KI-Kompetenznetzwerks zu Infrastruktur- und Architekturthemen**

   Auch wenn die Nutzung einer gemeinsamen Infrastruktur aktuell noch nicht möglich ist, sollten die jetzt entwickelten KI-Anwendungen der verschiedenen Verwaltungen mindestens untereinander kompatibel bzw. interoperabel sein und perspektivisch auf eine gemeinsame Plattform migriert werden können. Um diesen sehr akuten und kurzfristigen Handlungsbedarf zu decken, schlagen wir den Aufbau eines Netzwerkes vor, in dem KI-Expert:innen aus Bund, Ländern, Kommunen und IT-Dienstleistern gemeinsame Erfahrungen, Standards und Vorlagen austauschen können und so die Interoperabilität ihrer jeweiligen Systeme sicherstellen.

Bei Nichteinhaltung dieser Empfehlungen gehen wir davon aus, dass spezifische KI-Silos in den Behörden entstehen werden. Diese Silos würden eine behördenübergreifende Nachnutzung und gemeinsame Weiterentwicklung von GenAI-Anwendungen stark erschweren, beziehungsweise unmöglich machen. Darüber hinaus verursachen Parallelentwicklungen **höhere Kosten** für alle Behörden – sowohl finanziell als auch kapazitiv. Und schließlich würde ein vorhersehbares und kostspieliges Zersplittern der deutschen Verwaltung-IT-Landschaft in einem zukunftsweisenden Digitalthema der **Reputation der öffentlichen Verwaltung schaden**.

Angesichts der geschilderten Herausforderungen und des enormen Beitrages, den gerade Generative KI zur Lösung selbiger leisten kann, sind eine gemeinsame Strategie und verbindliche architektonische Standards unerlässlich. Wir sind zuversichtlich, dass der IT-Planungsrat als Koordinierungsgremium diese Bemühungen teilt und unterstützen wird.

8 Kontaktdaten und Mitwirkende

Bei Fragen und Anmerkungen kontaktieren Sie bitte das Kompetenzteam Künstliche Intelligenz:

kompetenzteam.ki@sk.hamburg.de

Verantwortlich

Die Erstellung des Dokuments erfolgte im Rahmen des Schwerpunktthemas Datennutzung des IT-Planungsrats im Kompetenzteam Künstliche Intelligenz unter der Leitung von Dr. Annika Busse (Senatskanzlei der Freien und Hansestadt Hamburg).

Mitwirkende

Alphabetische Aufzählung der im Maßnahmenteam „Infrastruktur und Standards“ mitwirkenden Personen:

Johannes Ast, Staatsministerium Baden-Württemberg

Tobias Heinrich, Sächsische Staatskanzlei

Janina Jäger, Senat der Freien und Hansestadt Hamburg – Senatskanzlei

Alain Knorr, Staatskanzlei des Saarlandes

Erin Polster, Hessisches Ministerium für Wirtschaft, Energie. Verkehr, Wohnen und ländlichen Raum

Nick Schirmer, Staatskanzlei des Landes Nordrhein-Westfalen

Torsten Tuschinski, Staatskanzlei des Landes Nordrhein-Westfalen

Stand: 0.9, 18.11.2024

9 Anhang:

9.1 Baden-Württemberg

Strategische Vorgehensweise

Baden-Württemberg strebt ein technologisch, souveränes und über Verwaltungsebenen und Ländergrenzen hinweg interoperables Ökosystem für KI-gestützte Verwaltungsanwendungen an. Um im technologisch volatilen KI-Sektor frühzeitig Erfahrung zu sammeln und informierte Entscheidungen treffen zu können, hatte die Landesverwaltung im Jahr 2021 die Entwicklung eines KI-Assistenzsystem-Prototyps beschlossen. Der F13-Prototyp wurde 2022 entwickelt und ab 2023 in der gesamten Landesverwaltung getestet.

Auf Grundlage der Erkenntnisse hat das Staatsministerium Baden-Württemberg die Software F13 als KI-Modell-agnostische Anwendungssammlung weiterentwickelt. Über F13 werden verschiedene KI-Anwendungen (Apps) den Mitarbeitenden verfügbar gemacht. Das Ministerium des Innern, für Digitalisierung und Kommunen und die Landesoberbehörde IT Baden-Württemberg (BITBW) entwickeln seit 2024 im Projekt „KI4BW“ eine Integrationsplattform für KI-Anwendungen. Über diese Plattform sollen F13 sowie weitere KI-Grundfunktionen und KI-Anwendungen zentral, sicher und datenschutzkonform bereitgestellt werden. Begleitend werden KI-Bedarfe der Ministerien im KI4BW-Netzwerk ressortübergreifend gebündelt und ausgetauscht.

Wesentliche Anwendungsfälle

F13 ist eine Sammlung an KI-gestützten Anwendungen zur Textverarbeitung. Die Anwendungen wurden in Workshops mit Mitarbeitenden konzipiert und im F13-Prototyp getestet. Seitdem werden die Anwendungen überarbeitet und um neue ergänzt. Stand 18.11.2024 sind folgende Funktionen im Regelbetrieb:

- Chat: Ein KI-Chatbot, mit dem Anwendende frei interagieren und generische Textarbeiten umsetzen können.
- Recherche: Auf Retrieval Augmented Generation basierend, können Anwendende bis zu fünf Dokumente hochladen und Fragen zu ihrem Inhalt stellen. Außerdem können die täglich aktualisierten Drucksachen des Landtags Baden-Württembergs sowie die Pressemitteilungen der Landesregierung als Wissensgrundlage genutzt werden.
- Zusammenfassungsfunktion: Anwendende können Dokumente hochladen und kürzen lassen. Ergebnistexte und Originaltexte können parallel angezeigt und miteinander abgeglichen werden.

Anwendungsarchitektur

Alle nicht-KI Komponenten F13s werden auf Servern der BITBW betrieben. KI-Sprachmodelle werden sowohl auf Grafikprozessoren im Rechenzentrum der BITBW betrieben als auch auf Grafikprozessoren in der STACKIT Cloud. Damit verfolgt F13 ein hybrides Architekturmodell.

9.2 Hamburg

Strategische Vorgehensweise

Die Stadt Hamburg erprobt in verschiedenen Bereichen die Nutzung von künstlicher Intelligenz im Verwaltungskontext. Über einen eigens eingerichteten Fonds – den InnoTecHH Fonds – unterstützt die Senatskanzlei Hamburger Behörden dabei, innovative Ideen für den Einsatz von KI und anderen neuen Technologien zu entwickeln und schnell zu erproben. Ende 2023 wurde im Rahmen des InnoTecHH Fonds das LLM-Pilotprojekt „LLMoin“ ins Leben gerufen. Nach zwei Pilotierungsphasen mit behördenübergreifenden Testgruppen wurde der Mehrwert einer LLM-Anwendung für die Verwaltung bestätigt und Ende 2024 soll sukzessive der Rollout in der Hamburger Verwaltung beginnen. Neben dem LLM-Textassistenten werden weitere LLM-Anwendungen im Rahmen von Pilotprojekten untersucht, unter anderem ein Chatbot für Bürger:innen bei der Antragsunterstützung und intelligente Dokumenten-Suche.

Wesentliche Anwendungsfälle

LLMoin umfasst folgende Funktionen:

- **Zusammenfassung**: Lange Texte und Dokumente werden zusammengefasst. Mögliche Optionen bei der Zusammenfassung sind Länge, Sprachstil, Textart (Stichpunkte oder Fließtext). Es können eigene Texte eingegeben, mehrere Dokumente hochgeladen oder beides miteinander kombiniert werden.
- **Recherche**: Nutzer*innen können Fragen entweder an vorher definierte Datenquellen (Dokumentenarchiv, individuell pro Behörde festgelegt) oder ad-hoc hochgeladene Dokumente stellen. Die zur Frage relevanten Textstellen sowie der gesamte umliegende Text wird den Nutzer:innen angezeigt und erlaubt damit eine Verifizierung der Antwort anhand der Quellen.
- **Textgenerierung**: Aus Stichpunkten und anderen Vorgaben wird ein Text generiert. Es können Dokumente hochgeladen oder per Copy & Paste eingefügt werden, die zur Textgenerierung verwendet werden sollen. Mögliche Optionen sind: Ausgabeformat, Länge, Sprachstil, Textart. LLMoin erstellt bspw. Entwürfe für Vermerke aus unstrukturierten Daten oder gibt Feedback zu Rechtschreibung und Grammatik eines schon existierenden Dokuments oder schreibt den Inhalt in eine bürgernahe Sprache um.
- **Freies Prompting**: Im Gegensatz zu den geführten Prompts der ersten drei Funktionen erlaubt das „freie Prompting“, frei und flexibel Eingaben zu formulieren, um maßgeschneiderte Antworten zu erhalten. Ergänzend können Dokumente zur Verwendung im Chat hochgeladen werden.

Anwendungsarchitektur

LLMoin wird auf Dataports **Plattform für Generative AI** entwickelt und betrieben. Die Gen AI Plattform basiert auf einer RAG-Architektur, die modellagnostisch und skalierbar ist. In der RAG-Architektur können unterschiedliche Sprachmodelle nahtlos integriert werden. Die Sprachmodelle stehen als API-Schnittstelle zur Verfügung (LLM on Cloud). LLMoin verwendet aktuell GPT-Modelle, die auf der Microsoft Cloud (Azure) in Europa betrieben werden.

9.3 Hessen

Strategische Vorgehensweise

In Hessen wurde ein interministerieller Arbeitskreis **KI und Innovation** unter Federführung des Hessischen Ministeriums für Digitalisierung und Innovation gegründet. In diesem Kreis sowie in seinen Unterarbeitsgruppen werden strategische und operative Aspekte der KI-Implementierung im Land ressortübergreifend beraten und zur Entscheidung vorbereitet werden.

In den Justiz-, Finanz- und Innenministerien laufen verschiedene Verbundaktivitäten im Bereich KI.

Wesentliche Anwendungsfälle

Im Hessischen Wirtschaftsministerium werden verschiedene Anwendungsfälle im Bereich GenAI konzipiert und erprobt. Das Wirtschaftsministerium entwickelt unter anderem in einem iterativen Verfahren das KI-System „**AIGude**“ zur KI-gestützten Recherche und Textgestaltung. Das barrierefreie und mit Responsive Design entwickelte System umfasst aktuell folgende Funktionen:

Recherche: Kleine Anfragen aus zwei Legislaturperioden, der Koalitionsvertrag und die KI-Verordnung stehen aktuell für KI-gestützte Recherchen zur Verfügung. Diese werden kurzfristig durch weitere Datenquellen ergänzt. Antworten auf Recherche-Anfragen werden mit Quellen hinterlegt.

Textzusammenfassung: Zwecks schnellerer Inhaltserfassung können längere Texte gekürzt werden. Die Zusammenfassung kann nach Länge sowie nach Stil erstellt werden. Die Übersetzung in einfache Sprache ist mit dieser Funktion möglich.

Fließtextgenerierung: Anhand von Stichpunkten können Texte neu generiert werden. Die gewünschte Textlänge sowie der gewünschte Textstil kann durch die Nutzenden definiert werden.

Textumformulierung: Bestehende Textblöcke können mithilfe des Sprachmodells neu formuliert werden.

Diese Funktionen werden in einem iterativen Entwicklungsverfahren im Laufe des Pilotprojektes ergänzt und nach Nutzerbedarf angepasst. Gleichzeitig werden sie für weitere Nutzergruppen innerhalb der Landesverwaltung als MVP bereitgestellt.

Die Hessische Zentrale für Datenverarbeitung entwickelt darüber hinaus weitere GenAI Anwendungsfälle:

Verwaltungs- und IT-Prozess-Agenten: Die Erstellung von User-Stories, Sequenzdiagrammen, Architekturskizzen und Sitzungsprotokollen („Speech to Summary“) sowie Textumformulierung für sprachliche Verfeinerung wird aktuell erprobt

Programmierassistenz: Für UI, Code Completion und IDE/Entwicklungsumgebung werden Assistenten getestet.

ChatBots: Für Wohngeld wird ein Chatbot mit der Datengrundlage Wohngeldgesetz und eWoG-Handbüchern erprobt. Weiterhin sind Recherche-Chatbots für die KI-Verordnung und interne Protokolle im Test.

Anwendungsarchitektur

Die oben genannten GenAI-Systeme werden im „Forschungslabor Cloud“ der Hessischen Zentrale für Datenverarbeitung pilotiert. Aktuell werden On Premises Installationen der Sprachmodelle Mixtral 8x7b, gemma2 und MiniCPM genutzt. Die Plattform basiert auf NVIDIA AI Enterprise. Zur Unterstützung der Recherche wurde eine Wissensdatenbank auf Basis von Elasticsearch implementiert.

9.4 Nordrhein-Westfalen

Strategische Vorgehensweise

Die Landesregierung Nordrhein-Westfalen will die Schlüsseltechnologie Künstliche Intelligenz zum Vorteil der Bürgerinnen und Bürger sowie der Wirtschaft einsetzen. Zur Nutzung der Potenziale Künstlicher Intelligenz wurde das beim Landesbetrieb Information und Technik NRW (IT.NRW) angesiedelte **KI-Labor** als Ansprechpartner für die gesamte Landesverwaltung etabliert.

Wesentliche Anwendungsfälle

IT.NRW entwickelt aktuell im iterativen Verfahren das speziell auf die Bedürfnisse der öffentlichen Verwaltung zugeschnittene KI-System „**NRW.Genius**“. NRW.Genius enthält eine Reihe von KI-basierten Werkzeugen, um die tägliche Arbeit zu unterstützen und ist Infrastruktur- und Sprachmodell-agnostisch konzipiert.

Dabei unterstützt die Anwendung in Ausbaustufe 1 folgende konkrete Anwendungsfälle:

- **Zusammenfassung:** Diese Funktion ermöglicht, Texte wie PDF-Dokumente oder andere Textdateien zu kürzen und die wesentlichen Inhalte auf einen Blick zusammenzufassen.
- **Textgenerierung:** Mit dieser Funktion kann man automatisch Texte erzeugen lassen, beispielsweise E-Mails oder Berichte.
- **Rechercheassistent:** Diese intelligente Suchfunktion ermöglicht, in umfangreichen Dokumentensammlungen (RAGS) effizient nach relevanten Informationen zu suchen.
- **Chat:** Freie Interaktion mit dem Sprachmodell hinter NRW.Genius durch die Formulierung von Anweisungen oder Fragen („Prompts“).
- **Fragen an mein PDF:** Mit dieser Funktion können Informationen direkt aus PDF-Dateien extrahiert werden. Nutzer können ihre PDF-Dateien hochladen und gezielte Fragen stellen, um spezifische Informationen schnell und präzise zu erhalten.

Anwendungsarchitektur

NRW.Genius verarbeitet sowohl die Eingaben („Prompts") der Beschäftigten als auch die Begleitdokumente, auf die sich diese Prompts beziehen. Da diese Daten unterschiedliche Kritikalitätsstufen in Bezug auf Vertraulichkeit und Datenschutz haben können, ist NRW-Genius Infrastruktur-agnostisch konzipiert und soll drei verschiedene KI-Umgebungen anbieten, welche die Verarbeitung von Daten unterschiedlicher Kritikalitätsstufen ermöglichen. Aus den genannten Vertraulichkeits- und Datenschutzgründen sollen nicht alle Sprachmodelle des NRW.Genius zur Auswertung in allen KI-Umgebungen genutzt werden können. Die konzipierten KI-Umgebungen sind:

- Bei der **On-Premise-Umgebung** wird das KI-System von IT.NRW auf den Servern von IT.NRW betrieben.
- In der **Private Cloud-Umgebung** wird das KI-System von IT.NRW auf Servern vertrauenswürdiger Drittanbieter betrieben.
- Bei der **Public Cloud** wird das KI-System im Auftrag von IT.NRW auf Servern von Drittanbietern betrieben.

Zum Einsatz kommen in Ausbaustufe 1 die Sprachmodelle Mixtral und GPT-4o.

9.5 Saarland

Die saarländische Landesregierung wird in Q1/25 ihre neue Digitalisierungsstrategie verabschieden. Diese umfasst auch das Thema Künstliche Intelligenz und deren Einsatz in der saarländischen Landesverwaltung. Der aktuelle Prozess der Entwicklung und Erarbeitung einer solchen KI-Strategie wird derzeit flankiert durch eine ressortübergreifende Unterarbeitsgruppe des saarländischen Digitalisierungsboards. An dieser AG sind alle Ressorts beteiligt; zudem erfolgt eine Begleitung mit externer Expertise durch das DFKI (Deutsches Forschungsinstitut für Künstliche Intelligenz) sowie PD (Partnerschaften Deutschland).

9.6 Sachsen

Der Freistaat Sachsen hat am 10. September 2021 seine eigene KI-Strategie vorgestellt. Als wichtiges Teilziel der Strategie soll die KI die Sächsische Verwaltung bürgerfreundlicher und effektiver machen. Angesichts dessen wird zunächst mit Hochdruck an einer Richtlinie zum Einsatz von KI in der Sächsischen Staatsverwaltung gearbeitet und ChatGPT in vielen Ressorts getestet oder bereits eingesetzt. In einem zweiten Schritt sollen die Mitarbeiter die Stärken und Schwächen von Generativer KI unter Ausnutzung von Large Language Modellen (LLM) kennenlernen. Hier wird angestrebt insbesondere folgende Anwendungsfälle zu testen:

- Texte zusammenfassen
- Texte generieren
- Texte umformulieren (verbessern)
- In Texten recherchieren (FAQ)
- Anwendungsfälle selbst erstellen (Freies Prompting)

Bei allen Anwendungsfällen soll die Einbindung von eigenen Wissensdatenbanken über RAG unterstützt werden, so dass die Ergebnisse viel präziser werden. Hinsichtlich der zu nutzenden KI-Software/Infrastruktur strebt der Freistaat Sachsen die Nachnutzung von Lösungen aus anderen Bundesländern an. Da momentan konkrete Nachnutzungsmodelle noch nicht zur Verfügung stehen, wird der Test der LLM zunächst mit einer Standardsoftware durchgeführt. Die Standardsoftware soll folgende Anforderungen erfüllen:

- Unterstützung mehrerer kommerzieller und Open-Source-Modelle (beispielsweise OpenAI, Anthropic, Google, Mistral, Meta)
- Einbindung weiterer Modelle über API-Schlüssel
- Assistenten für Erstellung von Prompts
- Chatfunktion
- Prompts über eine Prompt-Bibliothek teilen
- Workflows zur Abarbeitung sich wiederholender gleicher Prompts erstellen
- Eine Standard-API für alle Modelle zur Nutzung von KI-Anwendungsfällen in Digitalisierungsvorhaben
- Einbindung eigener Vektordatenbanken
- Suche in Wissensdatenbanken
- Datenanalyse
- Möglichkeit bei fehlenden Daten das Internet abzufragen
- Umsetzung eines internen Berechtigungskonzeptes (Rollen und Organisation)
- Bei Cloud gehosteten LLM kein Modelltraining zuzulassen (vertragliche Absicherung)
- Cloud gehosteten LLM müssen innerhalb der EU betrieben werden
- Abrechnung soll über eine Nutzerpauschale erfolgen (keine Verbrauchsabrechnung basierend auf einer Tokennutzung)