Bundes-LLM

R&D service: Scientific preparation of a federal large-scale language model

Programm / Ausschreibung	DST 24/26, DST 24/26, AI Ökosysteme 2025: AI for Tech & AI for Green	**Status**	laufend
Projektstart	01.10.2025	**Projektende**	30.04.2026
Zeitraum	2025 - 2026	**Projektlaufzeit**	7 Monate
Keywords	R&D service; requirements for the development of a national LLM; framework conditions (incl. data protection, technical, security, ethical, ecological); assessment of data base for LLM training (general data, federal data, protected data); stakeholde		

Projektbeschreibung

Die Studie befasst sich mit technologischen und gesellschaftlichen Aspekten (rechtlich, ethisch, ökologisch) von Implementierungsvarianten eines Bundes-LLM. Diese sind: (i) Training und Anpassung eines Basis-LLMs, (ii) Feinabstimmung bestehender LLMs (Modelle Dritter oder das eigene Basismodell) auf die Anwendungsbedürfnisse einzelner Stakeholder, (iii) LLM-basierte Plattformen wie RAG-Systeme und deren Anpassung an die Anwendungsbedürfnisse einzelner Stakeholder, (iv) Sovereignty-First-Implementierungen, die vollständig auf einer On-Premises-Infrastruktur basieren, (v) Hybride Flexibilitätsmodelle, bei denen sensible Daten vor Ort verarbeitet werden, während öffentliche Dienste über eine sichere Cloud-Umgebung bereitgestellt werden, (vi) Ökosystempartnerschaften, die auf der Zusammenarbeit mit europäischen KI-Initiativen beruhen, um Ressourcen und Fachwissen gemeinsam zu nutzen, die F&E-Kosten zu senken und den Wissenstransfer zu beschleunigen.

Das Ergebnis der Studie soll den Entscheidungsträgern eine solide, rechtssichere und ethische Grundlage für die Entwicklung und den Einsatz einer Bundes-LLM für die öffentliche Verwaltung bieten, wobei auch das Potenzial für Anwendungen in Industrie und Unternehmen berücksichtigt wird. Die Studie soll Schlüsselfragen klären hinsichtlich (i) des professionellen Requirements Engineering für unterschiedliche Nutzergruppen und Usecases, (ii) Datensicherheit und Technologiesouveränität durch rechtliche, nachhaltige und ethische Integration, (iii) Projektentwicklung, Projektmanagement und Betrieb im öffentlichen Interesse, (iv) Kosteneffizienz durch Open-Source-Komponenten. Es werden Roadmaps für die verschiedenen Anwendungsfälle und Empfehlungen für die politische Umsetzung vorgelegt.

Abstract

The study addresses technological and societal aspects (legal, ethical and ecological) of various cases of the implementation of a Bundes-LLM, including: (i) foundation model training and adaptation, (ii) fine-tuning of existing LLMs (third party models or one's own base model) to the application needs of individual stakeholders, (iii) LLM-based platforms such as RAG systems, and their adaptation to the application needs of individual stakeholders, (iv) Sovereignty-First Implementations entirely

based on-premises infrastructure, with all data processed and stored within national borders, (v) Hybrid Flexibility Models where sensitive data are processed on-premises, while public services are delivered through a secure cloud environment, (vi) Ecosystem Partnerships based on the collaboration with European AI initiatives to share resources and expertise, reducing R&D costs and accelerating capability transfer.

The outcome of the study is designed to provide a robust, lawful and ethical foundation for decision-makers regarding the development and deployment of a Bundes-LLM for public administration, also taking into account the potential for applications in industry and businesses. The study will clarify key questions and inform strategic choices regarding (i) professional requirements engineering for diverse user groups, (ii) ensuring data and technology sovereignty through legal, sustainable and ethical integration, (iii) project development, management, and operation in the public interest, (iv) cost efficiency through opensource components. Roadmaps for the different usecases and policy implementation recommendations are provided.

Endberichtkurzfassung

Basierend auf den Ergebnissen der Auswertung des Rücklaufs der Befragung zum Einsatz von LLM-basierten Systemen in den Ministerien und im BKA und dem State-of-the-Art in der Forschung zu KI-basierten Applikationen, lassen sich folgende drei Kernbotschaften ableiten:

Kernbotschaften

Vielfalt statt Monokultur: Synthese als Leitprinzip für öffentliche LLM-Architekturen:
Die vielfältigen Anforderungen an KI-Technologie in der öffentlichen Verwaltung erfordern den Einsatz unterschiedlicher Large Language Models (LLMs) sowie kleinerer, spezialisierter Modelle (Small Language Models, SLMs). Dazu braucht es, statt einer Monokultur, eine national abgestimmte Referenzarchitektur, die Interoperabilität und Transparenz fördert.
Kontinuierliche Qualitätssicherung im KI-Lebenszyklus: Essenziell für belastbare und adaptierbare KI-Infrastrukturen -- im Allgemeinen sowie in der Verwaltung im Besonderen -- ist ein erweiterbarer, lebenszyklus weiter Test- und Validierungsrahmen, der Data Drift, Modell-Updates, KI-Stack-Änderungen und domänenspezifische Risiken berücksichtigt. Dadurch werden nicht nur die technischen Systeme auf dem aktuellen Stand gehalten bei gleichzeitiger Gewährleistung der bereits erreichten Systemperformanz (Ergebniskorrektheit, Laufzeitverhalten, Betriebskosten udgl.) und erforderlichen Sicherheitslevels, sondern es werden auch ethische Anforderungen, wie Inklusivität, Oversight und Transparenz, im Life-Cycle des jeweiligen Modells sichergestellt.
Digitale Souveränität durch PPP-Strukturen: Ein wesentlicher Schlüssel für digitale Souveränität und Effizienz im öffentlichen Sektor ist eine optimale Bündelung und Nutzung der im Land verfügbaren KI/IT-Kompetenzen, technischen und organisatorischen Infrastrukturen. Insbesondere, da es sich bei der Entwicklung und Ausrollung von KI-basierten Systemen und Applikationen um einen nach wie vor hoch dynamischen IT-Bereich handelt, bei dem domänenspezifisches Know-How, KI-Forschung und praktische Anwendung, forschungsgetriebene Prototypenentwicklung und ein stabiler und sicherer 24/7 Betrieb inklusive Benutzer:innen-Support Hand in Hand gehen. Daneben ist der Austausch und die Verknüpfung mit entsprechenden europäischen Initiativen im Sinne einer allgemeinen Stärkung der Europäischen KI-Landschaft förderlich.

Im Folgenden werden Entwicklungsphasen für eine national abgestimmte KI-Referenzarchitektur skizziert, wobei auf verschiedene technische Umsetzungsvarianten – Einbindung bestehender LLMs in größere Systeme, Modelloptimierung und Basismodelltraining – eingegangen wird. Bei jeder Phase sind Fragen der Umsetzbarkeit, Souveränität, Anforderungen an Humanressourcen und Rechenkapazitäten, Qualitätssicherung, ethische und rechtliche Themen sowie ökologische Auswirkungen mituntersucht und in der Projektdokumentation detailliert.

Phase 1 (KI-Stack -- Basis): Im Sinne einer effizienten Umsetzung KI-basierter Applikationen in der öffentlichen Verwaltung, bei der frühest möglich für die öffentliche Verwaltung nutzbringende Ergebnisse erzielt werden können, empfehlen wir ein phasenorientiertes Vorgehen bei dem zuerst die technischen und prozeduralen Grundlagen geschaffen werden, um LLMs in flexibler, sicherer und rechtskonformer Weise in größere Systemkontexte und Anwendungszusammenhänge einzubauen und einen gesicherten 24/7 Betrieb zu gewährleisten . Anmerkung: Dieser Prozess wurde bereits mit dem vom BKA initiierten GovGPT gestartet, sowie dem Umsetzungsplan für ausgewählte Applikationen in der öffentlichen Verwaltung im laufenden Jahr 2026, inkl. der Entwicklung des „LLM as a service“ durch das BRZ.

Phase 2 (KI-Stack -- Flexible Erweiterbarkeit): Parallel dazu und basierend auf den bereits gemachten Erfahrungen dieser ersten Umsetzung eines gemeinsam nutzbaren KI-Stacks und den technischen Hintergrunddokumenten (techn. Realosierungsvarianten und Ausbaustufen, ethische, rechtliche und ökologische Aspekte) aus der „Bundes-LLM“ Studie schlagen wir eine detaillierte Ausarbeitung eines technischen und prozeduralen Konzepts für die systematische Erweiterung des KI-Stacks für die öffentliche Verwaltung vor. Dazu ist der Austausch bzw. die Zusammenarbeit von Stakeholdern essentiell, wie z.B. den Domänenexpert:innen in den Wirkungsbereichen, dem BRZ, KI-Forschung, Systementwicklung und AI:AT. Diese Arbeit führt direkt in die Umsetzung des erweiterbaren Stacks , auf Basis dessen, was bereits technisch umgesetzt wurde. Benötigte Personalressourcen und -kompetenzen, Hardware- und Softwareanforderungen und damit verbundene Kosten lassen sich aus der „Bundes-LLM“-Studie ableiten.

Während Phase 1 und 2 mit dem Thema gemeinsamer, flexibel erweiterbarer KI-Stack und der Integration bestehender LLMs befasst, dient Phase 3 zur Optimierung von bestehenden Modellen und Phase 4 zur Entwicklung kleinerer, spezialisierter von Grund auf trainierter Basismodelle.

Phase 3 (Modelloptimierung): Dazu gehören die Anpassung bestehender Modelle (i) auf spezifische Anforderungen aus den Wirkungsbereichen, die über die Einbindung bestehender Modelle nicht erreicht werden kann; (ii) lokale, spezialisierte Agenten, die effizient und kostengünstig lokal betrieben werden können. Dabei ist zwischen Finetuning -Ansätzen und Wissensdestillation (Knowledge Distillation) zu unterscheiden. Beim Finetuning geht es um die Anpassung eines vortrainierten LLMs durch Training mit domänenspezifischen Daten (z. B. Verwaltungsanweisungen). Bei der Wissensdestillation werden kleinere, lokal deploybare Modelle erstellt, die Eigenschaften bzw. Kapazitäten von größeren 'Teacher'-Modellen erben.

Für die Modelloptimierung kann je nach Anforderung eine Kombination aus Finetuning und Wissensdestillation sinnvoll sein, z.B.: beginnend mit dem Finetuning eines (kleineren) Teacher Modells mittels Methoden der Low Rank Adaptation (LoRA) auf domänenspezifischen Daten, dann die Verwendung des LoRA-adaptierten Teacher für Wissensdestillation auf ein ein kleineres bzw. effizienteres Student-Modell und eventuell weitere LoRA auf dem Student-Modell für zusätzliches Finetuning. Für die Auswahl geeigneter Methoden und deren optimaler Konfiguration ist das Verständnis der theoretischen Grundlagen

unerlässlich.

Phase 4 (Basismodelltraining): Die Entwicklung eines nationalen Foundation Modells wie z.B. dem schweizer Modell Apertus oder dem niederländischen Modell GPT-NL ist zeit- (minimum 2 Jahre Entwicklungszeit) und ressourcenaufwändig (Daten, Speicherplatz, Rechenleistung, Energiekosten, Personal) und entsprechend kostspielig. Des Weiteren müssen die Modelle gewartet und regelmäßig retrainiert werden, um ihr Wissen auf dem aktuellen Stand zu halten. Soll ein Basismodell kommerzialisiert oder auch nur zentral gehostet werden, muss eine Architektur aufgebaut werden, die Training, Hosting, Skalierung und gegebenenfalls Abrechnung zuverlässig vereint und eine entsprechende Rechenzentrumsinfrastruktur muss aufgebaut, betrieben und gewartet werden.

Im Gegensatz zum Training eines sehr großen Modells (LLM) kann das Training eines Small Language Models (SLM) von Grund auf für sehr spezialisierte Anforderungen wirtschaftlich und technisch sinnvoll sein, besonders dann, wenn (i) bestehende Modelle bestimmte spezifischen Anforderungen prinzipiell nicht erfüllen können, wie z.B. in hochspezialisierten Domänen oder wenn die Daten in einer Fachsprache vorliegen, die in allgemeinen Trainingsdaten kaum bzw. nicht vorkommt; (ii) wenn mit streng vertraulichen Daten gearbeitet werden soll, die niemals mit fremden Basismodellen (auch nicht per Finetuning) in Berührung kommen dürfen; (iii) wenn das Modell direkt auf Kleinstgeräten (Smartphones, Sensoren, Wearables) laufen soll, muss die Architektur von Beginn an auf minimale Parameterzahl optimiert werden. Ebenso wie für Phase 3, erfordert Phase 4 tiefgreifendes Verständnis der theoretischen Grundlagen. Beide Phasen sind eng mit entsprechender Grundlagenforschung verknüpft.

Projektkoordinator

- Österreichische Studiengesellschaft für Kybernetik, abgekürzt "ÖSGK"

Projektpartner

- Software Competence Center Hagenberg GmbH