

Behörde/Institution
Landeshauptstadt München
Verwaltungsebene
Kommune
Website / URL
https://www.muenchen.de/
Einreichungskategorie
Interne Verwaltungsanwendungen
Möchten Sie das Projekt noch in einer zweiten Kategorie einreichen?
Transformation
Projektbeschreibung
Die Landeshauptstadt München (LHM) hat die Bedeutung von Künstlicher Intelligenz (KI), insbesondere generativer KI-Systeme, erkannt und 2024 das KI Competence Center (https://ki.muenchen.de/kicc/) im IT-Referat gegründet. Open Source Software (OSS) spielt dabei eine zentrale Rolle in der KI-Strategie der LHM. KI-Service-Layer auf OSS-Basis: Unser internes KI-Service-Layer ermöglicht den einfachen Einsatz generativer KI (GenAI). Aktuell nutzen wir LiteLLM (https://github.com/BerriAI/litellm) und Langfuse (https://github.com/langfuse/langfuse). LiteLLM fungiert als Proxy-Service, der verschiedene Sprachmodelle, Embeddings und Spezialmodelle wie Reranker über eine einheitliche API zusammenfasst – egal ob Cloud-basiert oder lokal. So erhalten interne Nutzer, eigene Anwendungen und Drittsoftware standardisierten, kontrollierten Zugriff auf KI-Funktionen. Langfuse ist ein Open-Source-Tool zur Überwachung von KI-Systemen, das Tracing, Evaluation und Lastabschätzung über Dashboards ermöglicht. Damit erfüllen wir die Nachvollziehbarkeitsanforderungen des AI-Acts und legen die Basis für stabilen Betrieb. Mittelfristig planen wir, weitere OSS-Komponenten zu integrieren, z.B. Docling (https://github.com/docling-project/docling) für dokumentbasiertes Parsing und SearXNG (https://github.com/searxng/searxng) als Websuche-Instanz. Neben der Nutzung verbessern wir OSS aktiv, etwa durch Penetrationstests für LiteLLM und einen Patch für Langfuse, den wir der Community bereitstellten. Eigene KI-Systeme als OSS: MUCGPT (https://ki.muenchen.de/ki-systeme/mucgpt) ist eine webbasierte ChatGPT-ähnliche Oberfläche, die den Mitarbeitern Münchens als modernes Textwerkzeug dient. Die hinterlegten Modelle sind konfigurier- und austauschbar. Andere Kommunen wie Stuttgart nutzen MUCGPT bereits. Zur Entwicklung verwenden wir OSS-Bibliotheken wie Langchain/Langgraph für Sprachmodelle, FastAPI für Webservices sowie Qdrant, Neo4j und PostgreSQL als Datenbanken.
Beschreiben Sie den technischen Innovationsgrad des Projektes und den Beitrag zur Verwaltungsmodernisierung.
München ist es gelungen, nicht nur KI-Prototypen zu entwickeln, sondern eine nachhaltige Lösung zu etablieren. Der KI-Service-Layer, basierend auf bewährten Open-Source-Komponenten, ermöglicht bestehenden und neuen Anwendungen, KI-Modelle einfach, stabil und nachvollziehbar zu nutzen. Unsere eigenentwickelten KI-Systeme wie MUCGPT werden als zentrale Werkzeuge in der gesamten Stadtverwaltung eingesetzt und entlasten die Mitarbeitenden spürbar. MUCGPT ist eine webbasierte Oberfläche ähnlich ChatGPT, die individuell konfigurierbare Sprachmodelle nutzt. Andere Kommunen, wie Stuttgart, setzen MUCGPT ebenfalls ein. Wir stellen unsere Eigenentwicklungen als Open Source zur Verfügung und laden alle ein, mit uns gemeinsam die Verwaltung fit für KI zu machen. Für die Entwicklung setzen wir Open-Source-Bibliotheken ein: Langchain/Langgraph für Sprachmodelle, FastAPI für Webservices sowie Qdrant, Neo4j und PostgreSQL als Datenbanken.

Welchen ökonomischen Nutzen hat das Projekt?
MUCGPT hat sich mittlerweile als zentrales Werkzeug innerhalb der Stadtverwaltung etabliert. Seit der Produktivsetzung im März 2024 nutzen über 9.000 Beschäftigte MUCGPT mit mehr als 5.000 täglichen Aufrufen. Dabei erstellen die Nutzer eigene Assistenten, die fachspezifische Workflows abbilden. Die Infrastrukturkosten für MUCGPT belaufen sich auf etwa 10 Cent pro Nutzer und Monat. Das ist äußerst kostengünstig im Vergleich zu den Lizenzkosten eines kommerziellen Systems wie Microsoft Copilot: Diese belaufen sich auf ca. 25 Euro pro Nutzer und Monat. Zusätzlich investiert das KICC fünf Tage im Monat in die Weiterentwicklung der OSS.
Beschreiben Sie die Nachhaltigkeit der Lösung.
Mittlerweile werden sieben interne Services bzw. Fachanwendungen über unseren Service-Layer mit GenAI-Funktionalitäten versorgt – und es werden laufend mehr. Über die Administrationsoberfläche können neue Anwender bzw. Services innerhalb weniger Minuten freigeschaltet werden. Dies ermöglicht eine schnelle Verprobung von Prototypen und steigert die Developer Experience (DX).
Wie trägt das Projekt zur Stärkung der Digitalen Souveränität bei?
Dank unseres modularen KI-Service-Layer sind wir unabhängig von spezifischen KI-Modelanbietern. Es ermöglicht es uns, innerhalb kurzer Zeit zwischen Cloud-Providern oder auf eine interne Instanz zu wechseln, ohne dass die Nutzenden (d. h. die Fachanwendungen, die die KI-Modelle beziehen) negative Auswirkungen wahrnehmen oder Umbaumaßnahmen durchführen müssen. Das gesamte KI-Service-Layer sowie alle Eigenentwicklungen können containerisiert betrieben werden und sind somit Plattform-agnostisch. Unabhängigkeit aufgrund offener Schnittstellen und Standards: Wir setzen auf offene Schnittstellen und Standards, um die Austauschbarkeit von Komponenten auch in Zukunft sicherzustellen. Die LiteLLM-Schnittstelle ist mit der Spezifikation von OpenAI kompatibel und setzt damit den aktuellen de-facto Standard um. Sie wird bereits von diversen Fachverfahren als Drop-in-Replacement für Sprachmodelle von OpenAI verwendet. Langfuse unterstützt das OpenTelemetry-Protokoll, und kann somit durch andere Tracing Plattformen ersetzt werden. MUCGPT benutzt das OpenID-Connect-Protokoll zur Authentifizierung von Benutzern, welche momentan mittels Keycloak erfolgt (https://www.keycloak.org/).
Projektwebsite
https://github.com/it-at-m/mucgpt
Umsetzungspartner (Unternehmen, Wissenschaft, sonstige)
KI Competence Center IT@M
Zusätzliche Dokumente
Keine