BAZG LLM-Chat-Assistant

Studiengang: BSc in Informatik
Vertiefung: Distributed Systems and IoT
Betreuer*in: Yannis Valentin Schmutz
Experte: Pierre-Yves Voirol (Abacus Research AG)
Industriepartner: Bundesamt für Zoll und Grenzsicherheit (BAZG), Schweiz

Diese Bachelorarbeit befasst sich mit der Konzeption, Umsetzung und Evaluation eines AI-gestützten Chat-Assistenten als Proof of Concept (POC) für das Bundesamt für Zoll und Grenzsicherheit (BAZG). Ziel des Projekts ist die Entwicklung eines Systems, das Mitarbeitende des BAZG bei fachspezifischen Anfragen zu Arbeitsanweisungen und Gesetzestexten effizient unterstützt.

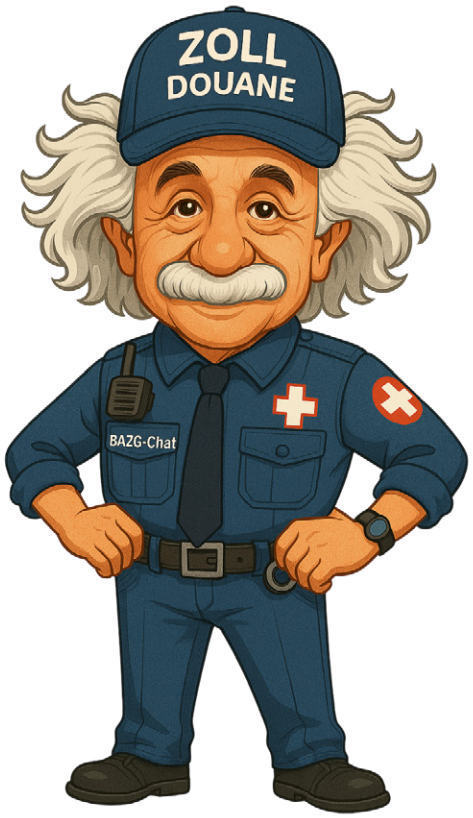
Der Retrieval-Augmented-Generation (RAG)-Ansatz wird verwendet, um ein Large Language Model (LLM) mit internem Fachwissen anzureichern, ohne dabei das Modell selbst zu verändern. Dabei wird die Benutzereingabe (Prompt) um kontextrelevante Informationen aus einer externen Wissensbasis ergänzt, sodass das LLM auch Fragen beantworten kann, deren Grundlagen es während des Trainings nicht gesehen hat.

Zu diesem Zweck wurde eine Applikation entwickelt, die das Einlesen, Strukturieren und Aufteilen umfangreicher Dokumente in kleinere, handhabbare Abschnitte ermöglicht. Besonderes Augenmerk lag auf Datenschutz, der einfachen Aktualisierung der Wissensbasis sowie auf der Evaluation geeigneter Segmentierungsstrategien und Embedding-Modelle. Diese Arbeit reicht bis zur Optimierung der Selektionsmethode für die richtigen Dokumente aus der Wissensdatenbank (Retrieval). Die Wahl des LLM durch das Testen verschiedener Modelle, die Optimierung der Antwortqualität, wenn dem LLM der nötige Kontext zur Verfügung steht, sowie die Verwendung eines privaten LLM zum Schutz sensibler Daten sind für eine Weiterführung dieser Arbeit angedacht.

Zur Qualitätssicherung wurde ein umfassendes Evaluationsframework konzipiert, das sowohl automatisierte Metriken als auch manuelles Feedback durch Fachexperten umfasst. Dabei erwies sich insbesondere die Wahl des Embedding-Modells als zentral für ein effektives Retrieval. Ein Embedding-Modell wandelt den semantischen Gehalt von Texten in numerische Vektoren um, die beim Retrieval dazu dienen, inhaltlich passende Dokumentabschnitte zu identifizieren. In der zweiten und finalen Iteration konnte bei einquelligen Anfragen ein Context Recall von 88 % erreicht werden, das bedeutet, dass dem LLM bei der Beantwortung der Fragen 88 % der dafür benötigten Grundlageninformationen korrekt zur Verfügung gestellt wurden.

Das Projekt liefert einen erweiterbaren POC für den Einsatz eines KI-basierten Assistenzsystems im BAZG. Es bildet eine solide Wissens- und Erfahrungsgrundlage für zukünftige Entwicklungen. Die grösste Herausforderung auf dem Weg zu einem produktiven System dürfte jedoch die strukturierte Aufbereitung der umfangreichen Dokumentensammlungen des BAZG darstellen.

Der entwickelte Prototyp konnte die Machbarkeit des Systems erfolgreich demonstrieren und fand bei den Entscheidungsträgern des BAZG positiven Anklang. Einer Weiterentwicklung und Finanzierung des Projekts steht somit wenig im Wege.



Noël Gueniat
077 416 70 35
noel.gueniat@gmail.com