German Q&A-Gen

Erstellung von Datensätzen für das Fine-Tuning von Information Retrieval Modellen (German Q&A-Gen)

Programm / Ausschreibung	KIRAS, F&E-Dienstleistungen, KIRAS-K-Pass-KMU Innovation AKUT KIA F&E Dienstleistungen (FED KIA_2024)	**Status**	laufend
Projektstart	01.02.2026	**Projektende**	30.09.2026
Zeitraum	2026 - 2026	**Projektlaufzeit**	8 Monate
Projektförderung	€ 99.960		
Keywords	Retrieval-Modelle, Open Source, LLM		

Projektbeschreibung

Im Zuge des Projekts wird eine Open-Source-Software zur automatisierten Generierung spezialisierter Trainingsdaten für Information Retrieval-Modelle erstellt. Diese umfassen 5.000 Frage-Antwort-Paare aus behördlichen BMI-Dienstvorschriften, davon 1.000-2.500 manuell validierte Gold-Standard-Paare. Die Q&A-Paare berücksichtigen erstmals systematisch BMI-Behördensprache mit dualen Sprachregistern (formal/umgangssprachlich) und dienen zur Optimierung des bestehenden RAG-Systems. Innovative Qualitätssicherungsmechanismen wie semantische Duplikatserkennung und Persona-basierte Fragegenerierung gewährleisten hochwertige Datenqualität (Vermeidung von Bias & Duplikaten). Die On-Premise-Implementierung schafft technologische Unabhängigkeit.

Abstract

As part of the project, open-source software for the automated generation of specialized training data for Information Retrieval models will be developed. This includes 5,000 question-answer pairs from BMI administrative service regulations, of which 1,000-2,500 are manually validated gold standard pairs. The Q&A pairs systematically consider BMI administrative language with dual language registers (formal/colloquial) for the first time and serve to optimize the existing RAG system. Innovative quality assurance mechanisms such as semantic duplicate detection and persona-based question generation ensure high data quality (prevention of bias & duplicates). The on-premise implementation creates technological independence.

Projektkoordinator

- JAAS GmbH

Projektpartner

- Bundesministerium für Inneres
- Language Services Semantica e.U.