Optimierung von Supportprozessen mit Retrieval-Augmented Generation und kosteneffizienten LLMs

Studiengang: MAS Data Science

Retrieval-Augmented Generation (RAG) in Kombination mit kosteneffizienten Large Language Models (LLMs) und Multi-Agenten-Systemen (MAS) ermöglicht eine effizientere Bearbeitung von Supportanfragen. Diese Arbeit untersucht, wie durch die lokale Nutzung von RAG und spezialisierten KI-Agenten Supportprozesse optimiert, die Antwortqualität verbessert und Datenschutzanforderungen eingehalten werden können.

Ausgangslage

Unternehmen stehen vor der Herausforderung, eine steigende Anzahl komplexer Supportanfragen effizient zu bearbeiten. Häufig sind Informationen zur Problemlösung in verteilten Wissensquellen wie Wikis oder bereits gelösten Support-Tickets schwer auffindbar, während manuelle Bearbeitungsprozesse zeitintensiv bleiben. Gleichzeitig gewinnen Datenschutz und Datensouveränität an Bedeutung, weshalb cloudbasierte KI-Lösungen nicht immer eine Option sind. Durch den Einsatz von lokal betriebenen LLMs in Kombination mit RAG kann der Zugriff auf bestehendes Unternehmenswissen verbessert und eine automatisierte, datenschutzkonforme Unterstützung im Support ermöglicht werden.

Zielsetzung

Diese Arbeit untersucht, wie RAG und MAS genutzt werden können, um Supportanfragen effizienter zu bearbeiten. Ziel ist es, Anfragen automatisch zu strukturieren, relevante Informationen bereitzustellen und präzise Zusammenfassungen sowie Handlungsempfehlungen zu generieren. Durch den lokalen Betrieb der KI-Modelle soll sichergestellt werden, dass vertrauliche Daten nicht an externe Cloud-Dienste übermittelt werden.

Methodik

Zur Evaluierung wurde ein Proof of Concept (PoC) entwickelt, der interne Support-Tickets sowie Dokumentationen wie Handbücher, Wiki-Einträge und interne Leitfäden nutzt. Die technische Lösung basiert auf ChromaDB als Vektordatenbank, mehreren Open-Source-LLMs (Llama, GLIDER, DeepSeek-R1) sowie einem agentenbasierten RAG-Ansatz (smolagents). Die Multi-Agenten-Struktur orchestriert die Prozesse so, dass DeepSeek-R1 für die Handlungsempfehlungen optimierte Ergebnisse liefern, während Llama parallel dazu die Zusammenfassungen erstellt.
Zusätzlich werden Anhänge mithilfe von Docling und Tesseract OCR verarbeitet, um relevante Informationen aus Dokumenten zu extrahieren und in den Analyseprozess einbinden zu können. Alle Komponenten werden lokal ausgeführt, um volle Kontrolle über die Daten zu gewährleisten. Die Qualität der generierten Inhalte wurde anschliessend anhand von Benutzerfeedback und einem “LLM-as-a-Judge” (GLIDER) evaluiert.

Ergebnisse

Die Evaluation zeigt, dass die Kombination aus RAG, MAS und lokalen LLMs eine Verbesserung der Supportprozesse ermöglicht. Insbesondere führten spezialisierte Agenten für Handlungsempfehlungen zu relevanteren Antworten.

Die automatisch generierten Zusammenfassungen wurden durchweg positiv bewertet, während die Handlungsempfehlungen noch Verbesserungspotenzial aufweisen. Zwar waren sie sprachlich korrekt, erfassen jedoch nicht immer den spezifischen Kontext der Anfrage vollständig. Diesem Problem kann durch ein komplexeres Modell mit erweiterter Kontextverarbeitung entgegengewirkt werden, was jedoch längere Verarbeitungszeiten zur Folge hat.
Darüber hinaus erwies sich der lokale Betrieb als technisch machbar und bot einen entscheidenden Vorteil in Bezug auf Datenschutz und Datensouveränität. Insgesamt bestätigen die Ergebnisse, dass RAG in Kombination mit MAS eine effektive Lösung zur Optimierung von Supportprozessen darstellt.
Zukünftige Weiterentwicklungen sollten die Effizienz der Verarbeitung von Support-Tickets weiter optimieren und die Kontextverständnisfähigkeit der Modelle verbessern. Als nächster Schritt wird angestrebt, die entwickelte Lösung in den aktiven Supportprozess zu integrieren, um sie unter realen Bedingungen zu testen und weiter zu verfeinern.

Tobias Lüthi