Fachhochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Automatisierung des Kundensupports in Softwareunternehmen durch den Einsatz von Large Language Models (LLMs) und Retrieval-Augmented Generation (RAG)

Steve Boris Titinang Sonfack

Masterarbeit • Studiengang Informatik • Fachbereich Informatik und Medien • Unternehmen: HeNaRa • 29.07.2024

Aufgabenstellung

Ziel dieser Arbeit ist es, die Effektivität eines innovativen Chatbots im Bereich des Kundensupports in Unternehmen zu bewerten. Der Chatbot basiert auf einer Kombination aus LLM (Large Language Models) und RAG (Retrieval-Augmented Generation), was eine neue Dimension der Kundeninteraktion ermöglicht. Die Analyse konzentriert sich auf Kriterien wie Relevanz, Genauigkeit und Antwortzeit des Systems in realen Supportszenarien.

Konzept

Entwicklung und Bewertung folgen einem strukturierten Ansatz: Zunächst werden relevante Kundensupport-Dokumente und Informationsquellen gesammelt und mittels OpenAI's text-embedding-ada-002 in Vektoren umgewandelt. Anschließend werden diese Daten in einer Chroma-Vektordatenbank indiziert, um einen effizienten Abruf zu gewährleisten. Durch die Integration von RAG werden die Informationen präzise abgerufen, während LLM für die Generierung natürlicher Sprache sorgt.

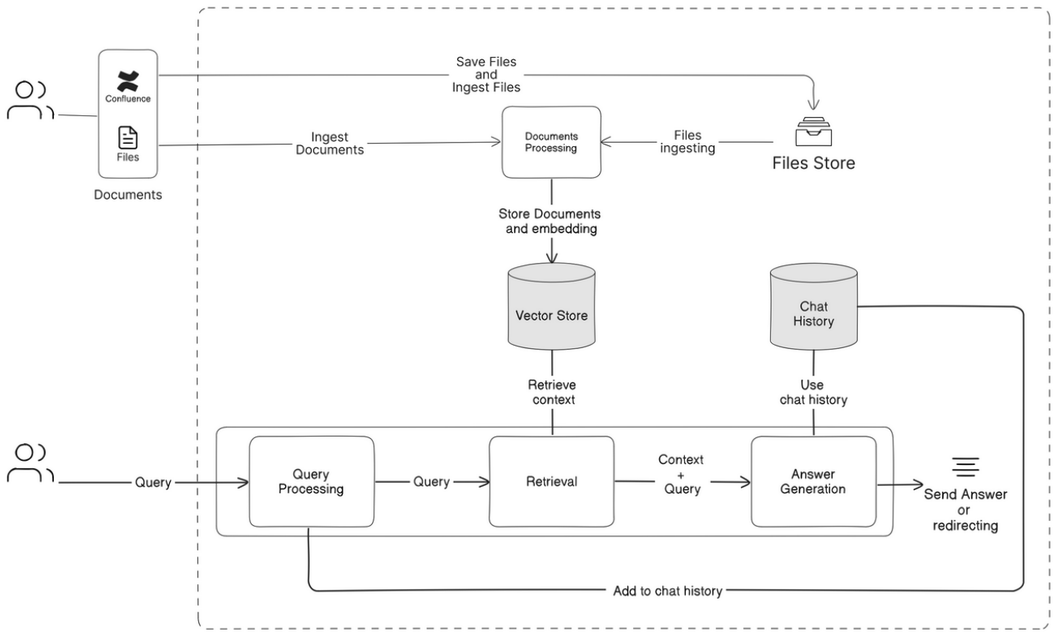

Abb. 1: Chatbot-Architektur

LangChain wird für die Textverwaltung und die Orchestrierung der verschiedenen Komponenten und Prozesse des Chatbots eingesetzt. Für die kontinuierliche Verbesserung dient LangSmith als Evaluierungstool. Durch regelmäßige Evaluierungen werden Schwachstellen identifiziert und behoben.

LLMs

LLMs sind das Kernstück moderner KI-Systeme für Sprachverarbeitung. Diese hochentwickelten Modelle, wie GPT-3, GPT-4, BERT, T5 und LLaMA, verfügen über Millionen oder sogar Milliarden von Parametern und können natürliche Sprache verstehen und generieren [2]. Ihre Stärke liegt in der Fähigkeit zum Few-shot und Zero-shot Learning, wodurch sie neue Aufgaben ohne spezifisches Training bewältigen können. In dieser Arbeit wird speziell GPT-3 eingesetzt, um vom Chatbot abgerufene Informationen in präzise, verständliche Antworten umzuwandeln und so eine menschenähnliche Interaktion zu ermöglichen.

RAG

RAG ist eine innovative Technologie, die die Stärken von Informationsabruf-systemen und generativen Modellen kombiniert [1]. In dieser Arbeit wird dieser Ansatz verwendet, um relevante Informationen aus der Chroma-Wissensdatenbank abzurufen und sie nahtlos in die vom Sprachmodell erzeugten Antworten zu integrieren. Der technische Prozess umfasst folgende Schritte: Anfrage, Abruf relevanter Dokumente, Integration des Kontexts, LLM-Verarbeitung und Generierung der Antwort (siehe Abb. 2). Durch den Einsatz von RAG kann der Chatbot nicht nur Standardanfragen beantworten, sondern auch auf komplexe und kontextbezogene Fragen präzise und differenziert reagieren, da er Zugriff auf aktuelle und relevante Daten hat.

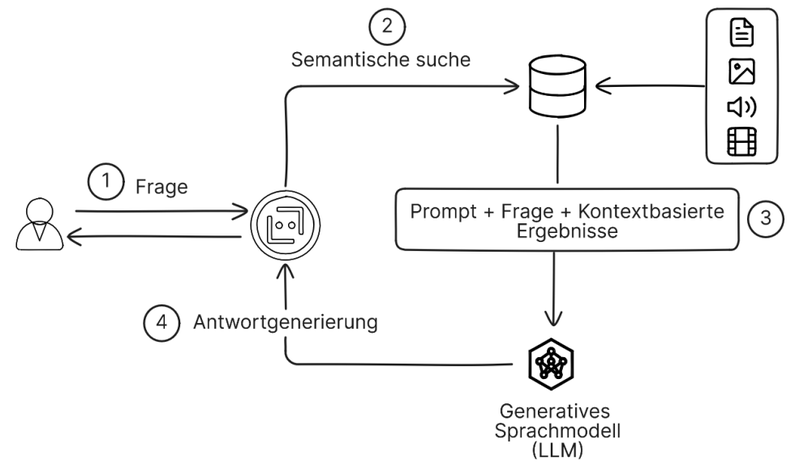

Abb. 2: RAG-LLM Prozessablauf im Chatbot-System

Ergebnisse

Die Ergebnisse zeigen, dass der Chatbot effektiv und effizient auf Kundenanfragen reagieren und eine positive Nutzererfahrung bieten kann. Das Testteam war besonders beeindruckt von der Genauigkeit des Chatbots bei mehreren Fragen.

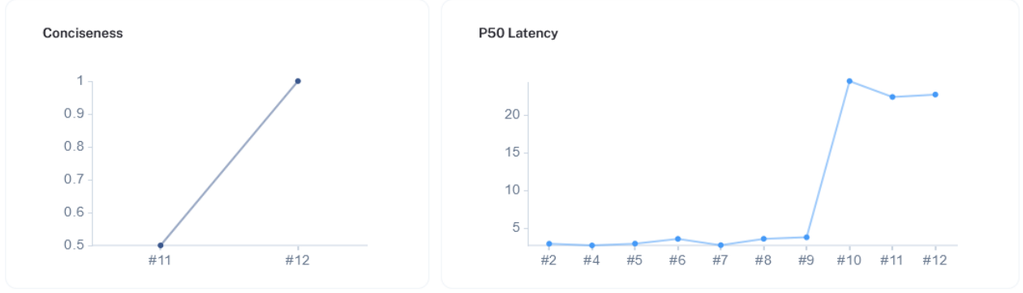

Abb. 3: Diagramm der Leistungsmerkmale

Die Grafiken veranschaulichen zwei Leistungsmetriken des Systems über zwei Testdurchläufe hinweg. Die x-Achse repräsentiert die Testdurchläufe, wobei Test #11 und Test #12 dargestellt sind. Die linke Grafik zeigt die Werte des Conciseness auf der y-Achse. Zwischen Test #11 und #12 ist ein deutlicher Anstieg dieser Metrik zu erkennen. Conciseness bezieht sich in diesem Kontext auf die Klarheit und Verständlichkeit der Ausgaben des Systems. Ein hoher Wert deutet auf eine Verbesserung hin. Die rechte Grafik stellt die p50-Latenz dar. Die p50-Latenz ist ein Maß für die Reaktionszeit des Systems, also wie schnell es auf Eingaben reagiert. Die y-Achse zeigt die Werte der Latenz in Sekunden an. Bei den Tests #10 und #12 wurden eine Reihe von sieben Fragen gestellt, auf die das System im schlimmsten Fall etwa 26 Sekunden benötigt, um sie zu beantworten. Das entspricht durchschnittlich 3,71 Sekunden pro Frage. Im Gegensatz dazu benötigt der manuelle Kundensupport für die gleiche Aufgabe mindestens eine halbe Stunde oder sogar länger.

Fazit

Insgesamt zeigten die Tests, dass der Chatbot deutliche Verbesserungen in Bezug auf die Antwortzeit, die Genauigkeit, die Relevanz der Antworten und die Hilfsbereitschaft erzielte. Diese Ergebnisse unterstreichen das Potenzial moderner KI-Technologien, die Effizienz des Kundensupports zu steigern und eine qualitativ hochwertige Interaktion mit den Nutzern zu gewährleisten. Zukünftige Entwicklungen könnten darauf abzielen, diese positiven Trends zu verstärken und die Anpassungsfähigkeit des Systems an verschiedene Szenarien zu erhöhen.

Quellen

[1] Lewis P., Oguz B. und Rinott R., „Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proceedings of the 34th International Conference on Neural Information Processing Systems (2020). Verfügbar: 10. Dezember 2023 unter https://arxiv.org/abs/2005.114011.

[2] Ding Q. et al., „Unraveling the landscape of large language models: a systematic review and future perspectives," Journal of Electronic Business & Digital Economics, 19. Dezember 2023. Verfügbar: 18. März 2024 unter https://doi.org/10.1108/JEBDE-08-2023-0015.

Gutachter: Dipl.-Inform. Ingo Boersch & Prof. Dr. Georg Merz • Technische Hochschule Brandenburg