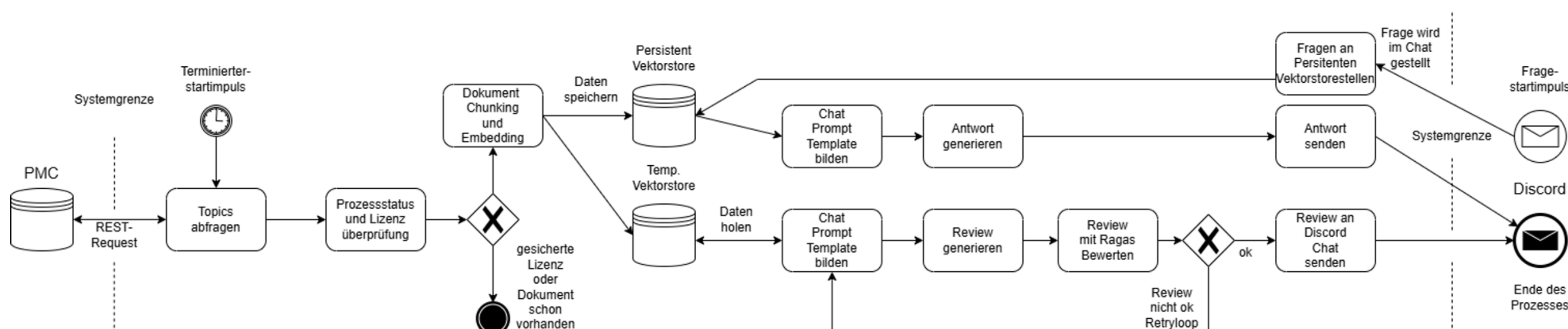


Bachelor-/Master-Thesis Studiengang

Automatisierte Datenextraktion und Berichtserstellung

Ausgangslage

Historisch gesehen wurde die Extraktion relevanter Informationen aus Forschungspublikationen und Fachberichten primär von Domänenexperten manuell durchgeführt. Dieser manuelle Prozess erwies sich jedoch als zeitintensiv, kostspielig und kaum skalierbar. Die ändernde Datenlayouts liessen einfache Systeme zur Extraktion an ihre Grenzen stossen. Der Einsatz von Large Language Models (LLMs) adressiert dieses Problem.



Prozessablaufdiagramm RAG-Pipeline welches den Datenfluss aufzeigt.

Lösungskonzept

Als Lösung wurde eine RAG-Pipeline entwickelt, welche themenspezifische Abfragen an PMC macht. Die neuen Artikel werden in der Pipeline verarbeitet (Transformation, Chunking, Embedding). Ein Sprachmodell generiert aus den Informationen ein Review des Berichtes, welche mit RAGAS auf seine Qualität überprüft wird.

Student: Balzereit Lenard

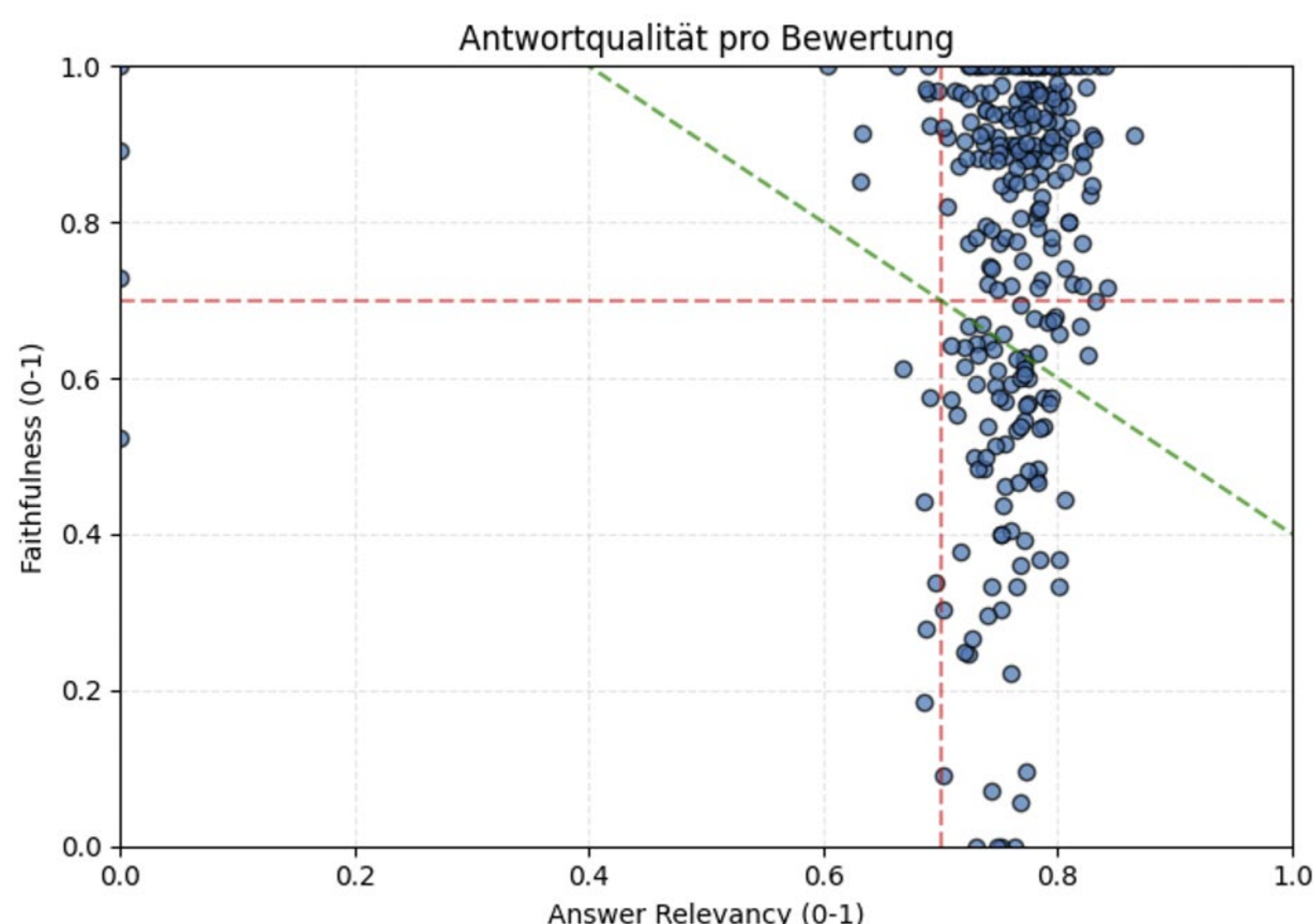
Dozent:
Prof. Dr. Fabian Ille

Experte:
PD Dr. sc. ETH Philipp Stämpfli

Korporationspartner:
Kompetenzzentrum Bioscience and Medical Engineering (CC-BME)

Ergebnisse

Während des Langzeittest wurden über eine Woche 205 Dokumente verarbeitet. Aus 89% aller Dokumente konnte das System Berichte mit zufriedenstellender Qualität erstellen.



Auswertungsplot welcher mittels RAGAS evaluiert wurde. Der Plot Visualisiert die Faithfulness (Wahrheitstreue) und die Relevanz.



Veranschaulichung der verwendeten Frameworks und Bibliotheken