ABSTRACT

Reliable evaluation is a prerequisite for improving retrieval-augmented generation (RAG) systems, yet manual assessment is costly and classical metrics often miss semantic nuances. This thesis investigates LLM-as-a-Judge as a scalable alternative for assessing a German-language RAG chatbot for a seminar webshop.

A human study provided ground truth to evaluate different LLM-as-a-Judge setups and compare them with classical metrics. Across models, *GPT-4.1* and *GPT-4o mini* showed the strongest and most consistent agreement with human ratings, while prompt design had only secondary impact. Building on these insights, a binary pass/fail evaluation method using LLM judges across five criteria, complemented by Recall@k as a classical metric, achieved a macro-F1 of 0.918 against human ratings, with only minor inconsistencies and a general tendency toward overly strict ratings.

The evaluation framework was then applied to 465 test items to guide the iterative optimization of the RAG chatbot. On the retrieval side, performance improved both through stronger embedding models and by extending the baseline approach of BM25 and semantic embeddings with HyQE. On the generation side, structured prompting with *GPT-4.1* eliminated remaining severe failures. End-to-end, the optimized system achieved an 80% reduction in failed cases, fully removing all severe failures.

This study demonstrates that pointwise LLM judges provide robust and cost-effective assessments for task-focused evaluation in German technical domains. They achieved high alignment with human ratings and clearly outperformed classical metrics for semantic judgments. Nonetheless, human oversight remained important, particularly for making fine-grained distinctions between severe and non-severe failures.

ZUSAMMENFASSUNG

Eine verlässliche Evaluation ist Voraussetzung für die Verbesserung von Retrieval-Augmented Generation (RAG)-Systemen. Manuelle Bewertung ist jedoch kostspielig und klassische Metriken erfassen semantische Feinheiten oft nur unzureichend. Diese Arbeit untersucht LLM-as-a-Judge als skalierbare Alternative zur Bewertung eines deutschsprachigen RAG-Chatbots für einen Seminar-Webshop.

Eine Studie mit menschlichen Annotatoren lieferte Referenzbewertungen, um verschiedene LLM-as-a-Judge-Konfigurationen zu evaluieren und mit klassischen Metriken zu vergleichen. Unter den Modellen zeigten *GPT-4.1* und *GPT-4o mini* die stärkste und konsistenteste Übereinstimmung mit menschlichen Bewertungen, während das Prompt-Design nur sekundären Einfluss hatte. Auf Basis dieser Erkenntnisse wurde ein binäres Pass/Fail Bewertungsschema mit LLM-Judges über fünf Kriterien entwickelt, ergänzt durch Recall@k als klassische Metrik. Dieses Verfahren erreichte ein Macro-F1 von 0,918 gegenüber den menschlichen Bewertungen, wobei nur kleinere Inkonsistenzen und eine generelle Tendenz zu übermäßig strikten Bewertungen beobachtet wurden.

Das Bewertungsverfahren diente anschließend als Grundlage für die schrittweise Optimierung des RAG-Chatbots anhand von 465 Testfällen. Auf der Retrieval-Seite verbesserte sich die Leistung sowohl durch stärkere Embedding-Modelle als auch durch die Erweiterung des Baseline-Ansatzes aus BM25 und semantischen Embeddings mit HyQE. Auf der Generierungsseite konnten durch strukturiertes Prompting mit *GPT-4.1* die zuvor aufgetretenen kritischen Fehler eliminiert werden. End-to-End erreichte das optimierte System eine Reduktion der Fehlerfälle um 80 % und beseitigte sämtliche kritische Fehler.

Diese Arbeit zeigt, dass punktweise LLM-Judges (d. h. unabhängige Einzelbewertungen) robuste und kosteneffiziente Ergebnisse für die aufgabenbezogene Evaluation in deutschen technischen Domänen liefern. Sie erzielten eine hohe Übereinstimmung mit menschlichen Bewertungen und übertrafen klassische Metriken für semantische Bewertungen deutlich. Dennoch blieb ergänzende menschliche Begutachtung wichtig, insbesondere für die feine Unterscheidung zwischen kritischen und weniger kritischen Systemfehlern.