Evaluierung und Optimierung von RAG-gestützten Chatbots

Safouan Er-Ryfy

Hochschule Darmstadt
Fachbereiche Mathematik, Naturwissenschaften & Informatik
Referentin: Prof. Dr. Melanie Siegel | Korreferentin: Prof. Dr. Antje Jahn

Einleitung

Large Language Models (LLMs) erzielen beeindruckende Ergebnisse beim Verständnis und der Generierung von Texten. In Unternehmensanwendungen sollten sie jedoch mit Vorsicht eingesetzt werden. Denn sie können fälschlicherweise überzeugende Inhalte erzeugen, die jedoch nicht der Wirklichkeit entsprechen, was zu sogenannten Halluzinationen führt. Zusätzlich ist ihr Wissen statisch und domänenspezifische Fakten sind von Unzuverlässigkeit geprägt. In unternehmensinternen Chatbots kann dies zu Fehlentscheidungen führen.

Retrieval-Augmented Generation (RAG) kombiniert LLMs dagegen mit externen Wissensquellen. Relevante Inhalte werden semantisch abgerufen und in den Prompt integriert. Dadurch werden Halluzinationen reduziert und aktuelle, domänenspezifische Informationen nutzbar gemacht, ohne dass ein aufwendiges Fine-Tuning erforderlich ist.

Vertriebsteams benötigen regelmäßig produkt- und prozessbezogenes Wissen, das in vielen internen Quellen verteilt ist. Der manuelle Zugriff darauf ist zeitaufwändig und ineffizient. Ein RAG-Chatbot soll den Zugriff vereinfachen und Informationen strukturiert bereitstellen. Die zentrale Herausforderung besteht darin, dass es an standardisierten Evaluierungsmethoden mangelt. Unternehmen bewerten häufig manuell, was sich nur schwer skalieren lässt.

Ziel dieser Arbeit ist daher die Entwicklung eines praxisnahen, automatisierten Evaluierungsframeworks, das die Qualität eines bestehenden RAG-Chatbots im Vertriebskontext systematisch messbar macht und Optimierungspotenziale entlang der Pipeline identifiziert.

Forschungsfrage

Die zugehörige Forschungsfrage lautet: „Welche Metriken sind am besten geeignet, um RAG-gestützte Chatbots zu evaluieren?"

Konzept

Das untersuchte System basiert auf einer erweiterten RAG-Pipeline. In der Pre-Retrieval-Phase können Nutzeranfragen durch Query Rewriting und HyDE in robustere Suchanfragen umgewandelt werden. Im Retrieval-Schritt wird ein vektorbasierter Index mit einer hybriden Suche aus Dense Retrieval und BM25 kombiniert. Anschließend werden die abgerufenen Dokumente mittels Re-Ranking nach Relevanz sortiert. In der Generierungsphase werden die ausgewählten Kontexte in den Prompt integriert und von einem LLM zu einer fundierten Antwort zusammengeführt.

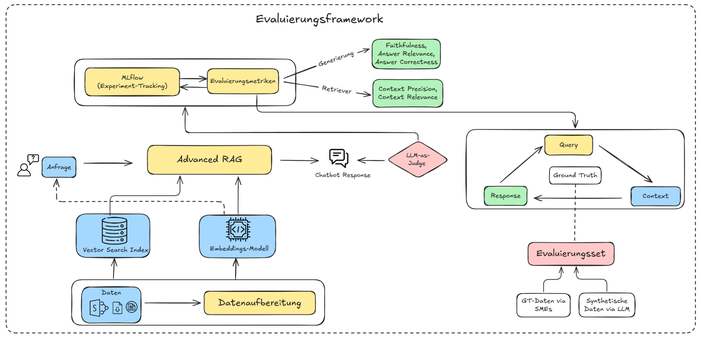

Figure 1. Überblick über das Evaluierungsframework

Methodik

Zielgrößen:

- Retrieval-Qualität
- Antwortqualität
- End-to-End-Systemleistung

Evaluierungsset:

- **Synthetische QA-Paare** für Optimierung und Experimente.
- **SME-validierte Ground Truth** für finale Test-Evaluation.
- **Strikte Trennung** von Entwicklungs- und Testdaten.

Metriken:

- **LLM-as-a-Judge** für automatisierte Bewertung.
- **Separates Judge-Modell** zur Reduktion modellinduzierter Bias (GPT-4o mini).
- **RAGAs**: kontinuierliche Scores (Faithfulness, Answer Relevance, Answer Correctness).
- **MLflow**: binär (Pass/Fail) (Retrieval Relevance, Retrieval Sufficiency, Correctness).

Ergebnisse

Phase 1 - Retrieval: Hybrid Search erzielt den höchsten F2-Score ($\beta = 2$) und bietet die beste Balance aus Relevanz und Vollständigkeit und stellt damit die robusteste Retrieval-Strategie im untersuchten Setting dar. Zusätzliche Komponenten (HyDE, Re-Ranking) erhöhen die Komplexität, liefern jedoch keinen zusätzlichen Qualitätsgewinn.

Phase 2 - Generierung: Unter identischen Retrieval-Bedingungen werden fünf LLMs evaluiert. Die Bewertung erfolgt anhand eines gewichteten Gesamtscores (60% Faithfulness, 40% Answer Relevance). Diese Gewichtung spiegelt den Vertriebskontext wider, in dem faktisch korrekte Antworten wichtiger sind als maximale semantische Ähnlichkeit. Claude Sonnet 4.5 erzielt die höchste Faktentreue bei gleichzeitig stabilen Ergebnissen und wird daher als finales Antwortmodell ausgewählt.

Phase 3 - End-to-End (GT-basiert): Das finale System (Hybrid Search + Claude Sonnet 4.5) erzielt eine höhere Faktentreue und Korrektheit als die Baseline. Der Rückgang der Answer Relevance lässt sich durch semantische Ähnlichkeitsmetriken erklären. Die Context Recall überschätzt die faktische Kontextabdeckung aufgrund der LLM-induzierten Claim-Attribution.

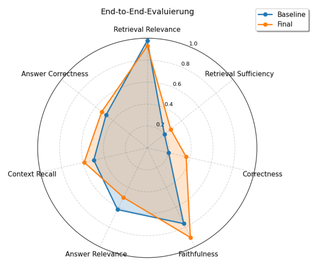

Figure 2. Radar-Diagramm des RAG-System-Vergleichs

Fazit

Wesentliche Erkenntnisse

- Hybrides Evaluierungsframework (MLflow + RAGAs) ermöglicht eine systematische Bewertung von Retrieval-, Generierungs- und End-to-End-Qualität.
- Die Qualität der Wissensbasis und des Retrievals ist ein zentraler Erfolgsfaktor für RAG-Chatbots.
- Answer Relevance und Context Recall zeigen Biases.

Empfohlenes Kern-Metrik-Set.

- **Retrieval**: Retrieval Relevance und Retrieval Sufficiency.
- **Generierung**: Faithfulness sowie Correctness / Answer Correctness.

Limitationen

- Ein systematischer Vergleich der verschiedenen Judge-Modelle mit rein menschlichen Referenzen fand nicht statt.
- In der Optimierungsphase kommen synthetische QA-Paare zum Einsatz.
- Fixed-Size-Chunking ist suboptimal.

Ausblick

- Erweiterung des GT-Datensatzes und systematischer Abgleich zwischen SME- und LLM-Judges.
- Einsatz fortgeschrittener Chunking-Strategien.
- Automatisches Retrieval-Tuning und adaptive Model-Routing-Strategien.
- Ergänzung durch Online-Evaluation im Produktivbetrieb.

Literatur

[1] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16/.

[2] Lei Huang et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

[3] Lianmin Zheng et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[4] Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[5] Xiaohua Wang et al. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.981. URL https://aclanthology.org/2024.emnlp-main.981/.

[6] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey, 2024. URL https://arxiv.org/abs/2312.10997.

[7] MLflow. Evaluating llms/agents with mlflow, 2025. URL https://mlflow.org/docs/latest/genai/eval-monitor.

[8] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4): 333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

[9] Sujoy Roychowdhury, Sumit Soman, H. G. Ranjani, Neeraj Gunda, Vansh Chhabra, and SAI KRISHNA BALA. Evaluation of RAG Metrics for Question Answering in the Telecom Domain. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL https://openreview.net/pdf?id=L74piNoToX.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[11] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. Evaluation of retrieval-augmented generation: A survey. In *Big Data*, pages 102–120, Singapore, 2025. Springer Nature Singapore. ISBN 978-981-96-1024-2. doi: 10.1007/978-981-96-1024-2_8. URL https://doi.org/10.1007/978-981-96-1024-2_8.