Development and Application of an LLM-as-a-Judge Evaluation Method for a RAG-Based Chatbot

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Hanna Dünschede

Supervisors: Prof. Dr. Markus Döhring, Prof. Dr. Jutta Groos
Darmstadt University of Applied Sciences

Introduction

"If you cannot measure it, you cannot improve it" [2]. This principle captures why evaluation is central to progress in NLP. Especially as Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) architectures move from demos to user-facing systems. In practice, however, reliable evaluation is hard: manual assessment is expensive and subjective, while classical automated metrics often fail to capture semantic adequacy. These gaps become more pronounced for RAG chatbots, where performance depends on the interplay between retrieval and generation rather than surface similarity alone.

A promising idea is to use LLMs themselves as evaluators (*LLM-as-a-Judge* [4]). The appeal is scalability and task-adaptability, but open questions remain: prior studies report mixed alignment with human judgment [1] and reveal biases (e.g., verbosity or position bias [3]), leaving uncertainty about reliability. At the same time, most studies focus on English and generic benchmarks, offering limited guidance for domain-specific, non-English RAG applications.

This work investigates LLM-as-a-Judge in a concrete, German-language setting: a RAG-based customer chatbot for the ORDIX® seminar webshop. The aim is to develop a practical, task-focused evaluation approach that can (i) meaningfully reflect end-to-end behavior in a retrieval-grounded chatbot system and (ii) be applied at scale to guide iterative improvement. The study is driven by two questions: (RQ1) How should an LLM-as-a-Judge evaluation framework be configured to align well with human judgments in this context? And (RQ2) which advanced RAG techniques measurably improve a seminar chatbot when assessed under such a framework?

Applied System: Seminar Chatbot

The chatbot answers seminar-related queries by retrieving relevant documents and generating grounded answers.

- **History-aware Query Builder:** Rewrites the latest user turn with recent chat history as context via *GPT-4o mini*, producing a standalone query.
- **Retrieval:** Hybrid approach: semantic search (gbert-large-paraphrase-cosine) + BM25 in parallel, fused with Reciprocal Rank Fusion.
- **Document Relevance Check:** LLM-based (*GPT-4o mini*) filtering of retrieved candidates to retain only relevant documents.
- **Answer Generation:** *GPT-4o* answers using up to 20 turns of history and the filtered context, guided by a task-specific prompt.
- **Human-Handoff Check:** *GPT-4o mini* flags cases that should be escalated to a human.

Evaluation design and limitations. An initial reference-based *LLM-as-a-Judge* setup used *GPT-4o mini* to score *answer correctness* on a 1–4 scale across 72 samples spanning five categories. However, reference creation was difficult, subtle nuances were often missed, the dataset was small, multi-turn category interpretation was inconsistent, and open-ended queries were underrepresented. This motivated the new methodology.

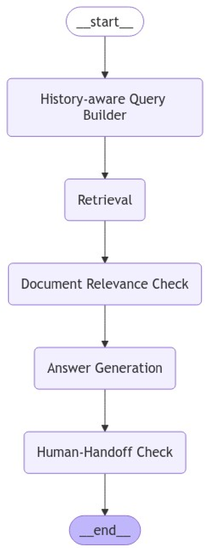

Figure 1. Overview of the RAG chatbot pipeline.

Methodology

- **Dataset:** German, ORDIX®-specific test set spanning six intents (Specific, Seminar Search, Abstract, Human Handoff, Out-of-Scope, Bad Intentions) and single-/multi-turn dialogs.
- **Criteria:** Seven task-aligned criteria: Answer Correctness, Answer Relevance, Context Relevance, Faithfulness, Purpose (scope refusal), Handoff and a pairwise Quality check; with a 1–4 scale for Answer Relevance and Answer Correctness, binary for all others.
- **Human Study:** ORDIX® employees labeled via a lightweight web app; target of three independent ratings per instance; collected 411 complete evaluations.
- **LLM-as-a-Judge:** Staged study: first compare nine models under a baseline prompt, then investigate prompt variants (few-shot, CoT, explanation-first, minimal, LLM-generated, LLM-refined) and a small "LLM jury."
- **Classical metrics:** ROUGE-L, embedding similarity, Recall@k and a regex-based handoff detector to compare against judge outcomes.

Results

Meta-Evaluation of Evaluation Method.

- Best alignment with humans: *GPT-4.1* and *GPT-4o mini*.
- Prompting variants had only secondary effects.
- Comparison with classical metrics: LLM judges outperformed ROUGE-L and embedding similarity; regex-based handoff detection was close to the judge, while Recall@k outperformed it.
- Final setup: binary pass/fail with LLM judges for five criteria (Answer Correctness, Answer Relevance, Faithfulness, Purpose, Handoff) and Recall@k converted into a binary metric for Context Relevance.
- Performance: Macro-F1 = 0.918 vs. human ratings.

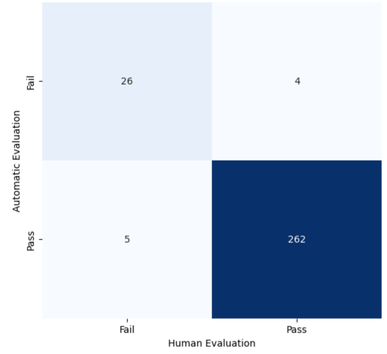

Figure 2. Confusion matrix of binary pass/fail system against human scores.

Chatbot Optimization. Evaluation framework applied to 1,127 test cases guided iterative refinement.

Experiment	Outcome
Query Builder	Refined prompt + **GPT-4.1 mini** → small, consistent recall gains.
Retrieval	**Qwen3-Embedding-0.6B** chosen. HyQE [5] fusion (baseline emb + HyQE emb + BM25) → mean recall **0.907**.
Document Filtering	**GPT-4.1 mini** with an "All-at-Once" prompt (judge all candidates in one pass) → best precision-recall balance.
Answer Generation	**GPT-4.1** + prompting-guide-based prompt → removed severe errors.

Impact. Failures reduced from 88 (21 severe) → 17 (0 severe), i.e. an 80% overall reduction and elimination of critical failures.

Key Takeaways

What worked. The binary *LLM-as-a-Judge* setup (5 criteria + Recall@k) aligned well with human labels and reliably exposed failure modes, it was strong enough to drive iteration (Macro-F1 = 0.918; Accuracy = 97%).

Where it fell short. The judge tended to be slightly strict and lacked granularity for borderline cases. Therefore, a manual post-hoc split into *severe* vs. *non-severe* was introduced to reflect real impact. Judges were usually technically correct but could not anticipate every acceptable edge case.

Conclusion

RQ1 — How to set up LLM-as-a-Judge for high human alignment? Use pointwise judges with concise prompts; *GPT-4.1* and *GPT-4o mini* aligned best. Binary pass/fail proved reliable. Reference-based Recall@k for *context relevance* outperformed LLM judge. Pairwise *quality* was too subjective and excluded. Overall Macro-F1 = 0.918.

RQ2 — Which RAG techniques help under this framework? Better embeddings + hybrid retrieval with HyQE, LLM-based document filtering (all candidates at once), and structured *GPT-4.1* generation (prompting-guide) produced the largest gains; context pruning was ineffective. End-to-end, severe errors dropped to zero.

Implications. Pointwise LLM judges are practical, scalable and reliable for semantic checks in German technical domains, while reference metrics remain valuable for retrieval. Hybrid automation with selective human oversight is still needed for nuanced severity and edge cases. Retrieval quality is central; generation prompt design governs safe behavior.

Limitations. Single domain (ORDIX®), modest corpus size (199 docs), some label imbalance, non-determinism in chatbot answers, and style dimensions (e.g. tone, brevity) not evaluated.

Future Work. Move beyond binary labels (graded scales where human agreement permits), automate severity classification or add judge uncertainty flags to triage to humans, broaden domains/models, explore model routing strategies in chatbot system.

References

[1] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics.

[2] Susan Ratcliffe, editor. *Oxford essential quotations*. Oxford Reference. Oxford University Press, Oxford, 4 ed. edition, 2016.

[3] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*, 2025.

[4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36 of *NIPS '23*, pages 46595–46623. Curran Associates Inc., 2023.

[5] Weichao Zhou, Jiaxin Zhang, Hilaf Hasson, Anu Singh, and Wenchao Li. HyQE: Ranking contexts with hypothetical query embeddings. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13014–13032, Miami, Florida, USA, November 2024. Association for Computational Linguistics.