Identifying Failure Modes in Retrieval-Augmented Generation Under Document Corruption via Unified Evaluation Platforms

Assignee Research

June 11, 2026

Abstract

Evaluating the quality of retrieval-augmented generation (RAG) and document reranking systems remains challenging due to the lack of scalable, user-centric, and multi-perspective evaluation tools. We introduce RankArena, a unified platform for comparing and analysing the performance of retrieval pipelines, rerankers, and RAG systems using structured human and LLM-based feedback as well as for collecting such feedback. RankArena supports multiple evaluation modes: direct reranking visualisation, blind pairwise comparisons with human or LLM voting, supervised manual document annotation, and end-

1 Introduction

This paper examines: RankArena: A Unified Platform for Evaluating Retrieval, Reranking and RAG with Human and LLM Feedback. Research question: Can unified evaluation platforms like RankArena identify specific failure modes in retrieval-augmented generation pipelines under varying levels of document corruption?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.7/10.

3 Results

10 papers retrieved. 15 claims extracted; 13 independently verified. Quality review score: 7.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
RankArena is a unified platform for evaluating retrieval, reranking, and RAG systems using human and LLM feedback.	✓	0.26
RankArena captures fine-grained relevance feedback through pairwise preferences and full-list annotations.	✓	0.27
RankArena records auxiliary metadata including movement metrics, annotation time, and quality ratings.	✓	0.20
RankArena integrates LLM-as-a-judge evaluation to compare model-generated rankings against human ground truth annotation	✓	0.25
Interactions on RankArena are stored as structured evaluation datasets usable for training rerankers, reward models, jud	✓	0.25
The RankArena platform is publicly available at https://rankarena.ngrok.io/.	✓	0.23
A demo video for RankArena is available at https://youtu.be/jIYAP4PaSSI.	✓	0.19
RankArena features five complementary evaluation modes including Reranker Comparison, Manual Annotation, and LLM Judgmen	✓	0.18
The Reranker Comparison mode in RankArena supports direct and blind pairwise battles between ranked document lists.	✓	0.22
The Manual Annotation mode in RankArena supports supervised full-list ranking with quality labels and tracks annotation	×	0.14
RankArena supports a Dataset Collection mode to systematically collect aligned human and LLM preference data for trainin	✓	0.18
In the full-list annotation mode, users can manually reorder or assign relevance grades to documents retrieved from onli	✓	0.19
The agreement rate between models in the provided benchmark table is 53.5%.	×	0.08
The benchmark table includes performance scores for models such as monolith5-large-msmarco-10k, monolith5-base-msmarco-1	✓	0.16
RankArena includes a Comprehensive Reranking Leaderboard, Manual Labeling, End-to-End RAG, Direct Reranker, 1v1 Arena, R	✓	0.26

References

- http://arxiv.org/abs/2508.05512v1
- http://arxiv.org/abs/2502.00306v2
- http://arxiv.org/abs/2402.12317v2