SOVEREIGN: LLM-as-a-Judge: Reassessing the Performance of LLMs in Extractive QA

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Extractive reading comprehension question answering (QA) datasets are typically evaluated using Exact Match (EM) and F1-score, but these metrics often fail to fully capture model performance. With the success of large language models (LLMs), they have been employed in various tasks, including serving as judges (LLM-as-a-judge). In this paper, we reassess the performance of QA models using LLM-as-a-judge across four reading comprehension QA datasets. We examine different families of LLMs and various answer types to evaluate the effectiveness of LLM-as-a-judge in these tasks. Our results show th

1 Introduction

Analysis of: LLM-as-a-Judge: Reassessing the Performance of LLMs in Extractive QA. Research goal: How does the robustness to noisy or irrelevant context in multi-hop HotPotQA questions change when using a large context window (e.g., 128K) versus iterative retrieval with reranking, measured by F1 score and precision under adversarial distractor insertion?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

11 papers retrieved. 5 claims extracted, 5 verified. Tribunal: 7.8/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
LLM-as-a-judge is highly correlated with human judgments and can replace traditional EM/F1 metrics in extractive QA eval	✓	0.39
Using LLM-as-a-judge, the correlation with human judgments improves significantly, from 0.22 (EM) and 0.40 (F1-score) to	✓	0.43
EM and F1 metrics underestimate the true performance of the QA models.	✓	0.35
LLM-as-a-judge is not perfect for more difficult answer types (e.g., job), but it still outperforms EM/F1.	✓	0.37
No bias issues, such as self-preference, are observed when the same model is used for both the QA and judgment tasks.	✓	0.24

References

- http://arxiv.org/abs/2504.11972v2
- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2101.00294v3