SOVEREIGN: How does the F1 score of LLM-as-a-judge evaluation compare to exact match for multi-hop HotPotQA when using it

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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 F1 score of LLM-as-a-judge evaluation compare to exact match for multi-hop HotPotQA when using iterative retrieval with reranking versus 128K-token context windows, across varying numbers of adversarial distractors?.

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

10 papers retrieved. 8 claims extracted, 1 verified. Tribunal: 3.8/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

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
Qwen 2.5 72B exhibits the highest correlation with human judgments among the three LLM judges tested.	×	0.09
Mistral 7B v0.3 has the lowest correlation scores with human judgments among the three LLM judges tested.	×	0.07
The correlation scores between human judgments and EM/F1 scores are smaller than those of all LLM-as-a-judge models.	✓	0.23
A correlation score above 0.80 is generally considered strong.	×	0.03
The study sampled 200 instances (50 per dataset) for human judgments.	×	0.04
After excluding cases where the gold answer was incorrect, 161 valid samples remained, each with 8 predicted answers, re	×	0.03
For the F1-score, a threshold of 0.5 was used to classify values between 0 and 1 for correlation calculation.	×	0.09
NaN values in Table 2 arise when none of the predicted answers by Llama 3.1 70B exactly match the gold answers, resultin	×	0.02

References

- http://arxiv.org/abs/2101.00294v3
- http://arxiv.org/abs/2507.23334v2

- http://arxiv.org/abs/2504.11972v2