SOVEREIGN: To what extent does the choice of LLM-as-a-judge (e.g., GPT-4 vs. Llama-3-70B) affect the relative ranking of

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Most Reading Comprehension methods limit themselves to queries which can be answered using a single sentence, paragraph, or document. Enabling models to combine disjoint pieces of textual evidence would extend the scope of machine comprehension methods, but currently no resources exist to train and test this capability. We propose a novel task to encourage the development of models for text understanding across multiple documents and to investigate the limits of existing methods. In our task, a model learns to seek and combine evidence — effectively performing multihop, alias multi-step, infer

1 Introduction

Analysis of: QAngaroo (MedHop + WikiHop) - Constructing Datasets for Multi-hop Reading Comprehension Across Documents. Research goal: To what extent does the choice of LLM-as-a-judge (e.g., GPT-4 vs. Llama-3-70B) affect the relative ranking of retrieval strategies (iterative reranking vs. long-context) on multi-hop reasoning accuracy in HotPotQA?.

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. 11 claims extracted, 11 verified. Tribunal: 8.5/10 → 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
Most Reading Comprehension methods limit themselves to queries which can be answered using a single sentence, paragraph,	✓	0.33
No resources exist to train and test the capability of combining disjoint pieces of textual evidence.	✓	0.22
The paper proposes a novel task to encourage the development of models for text understanding across multiple documents.	✓	0.22
In the proposed task, a model learns to seek and combine evidence effectively performing multi-hop inference.	✓	0.26
Two datasets from different domains are induced using the proposed methodology.	✓	0.16
Two previously proposed competitive models were evaluated and found to integrate information across documents.	✓	0.19
Both models struggle to select relevant information from provided documents.	✓	0.18
Providing documents guaranteed to be relevant greatly improves model performance.	✓	0.22
Models outperform several strong baselines but their best accuracy reaches 54.5% on an annotated test set.	✓	0.26
Human performance on the task reaches 85.0% accuracy on the annotated test set.	✓	0.15
The models' best accuracy is 54.5% on an annotated test set compared to human performance at 85.0%.	✓	0.23

References

- https://doi.org/10.1162/tacl_a_00021
- https://doi.org/10.48550/arxiv.2308.07107
- https://doi.org/10.18653/v1/2023.emnlp-main.398