SOVEREIGN: How does the performance of instruction-tuned retrievers on multi-hop queries from MuSiQue compare to single-c

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Instruction-tuned language models (LM) are able to respond to imperative commands, providing a more natural user interface compared to their base counterparts. In this work, we present Promptriever, the first retrieval model able to be prompted like an LM. To train Promptriever, we curate and release a new instance-level instruction training set from MS MARCO, spanning nearly 500k instances. Promptriever not only achieves strong performance on standard retrieval tasks, but also follows instructions. We observe: (1) large gains (reaching SoTA) on following detailed relevance instructions (+14.3

1 Introduction

Analysis of: Promptriever: Instruction-Trained Retrievers Can Be Prompted Like Language Model. Research goal: How does the performance of instruction-tuned retrievers on multi-hop queries from MuSiQue compare to single-context adversarial training when evaluated on out-of-domain BEIR subsets (SciFact, TREC-COVID) in terms of MRR@10?.

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

6 papers retrieved. 5 claims extracted, 4 verified. Tribunal: 7.3/10 → 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
Promptriever is the first retrieval model able to be prompted like a language model.	✓	0.30
Promptriever achieves strong performance on standard retrieval tasks.	✓	0.23
Promptriever follows instructions.	×	0.13
Promptriever achieves large gains (reaching SoTA) on following detailed relevance instructions (+14.3).	✓	0.28
The instruction training set curated from MS MARCO spans nearly 500k instances.	✓	0.19

References

- https://doi.org/10.5281/zenodo.20415633
- https://doi.org/10.48550/arxiv.2312.10997
- https://doi.org/10.5281/zenodo.20415634