SOVEREIGN: What is the impact of retriever robustness (measured via BEIR) on LLM reasoning accuracy in multi-hop QA tasks

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Zero-shot evaluation of information retrieval (IR) models is often performed using BEIR; a large and heterogeneous benchmark composed of multiple datasets, covering different retrieval tasks across various domains. Although BEIR has become a standard benchmark for the zero-shot setup, its exclusively English content reduces its utility for underrepresented languages in IR, including Dutch. To address this limitation and encourage the development of Dutch IR models, we introduce BEIR-NL by automatically translating the publicly accessible BEIR datasets into Dutch. Using BEIR-NL, we evaluated a

1 Introduction

Analysis of: BEIR-NL: Zero-shot Information Retrieval Benchmark for the Dutch Language. Research goal: What is the impact of retriever robustness (measured via BEIR) on LLM reasoning accuracy in multi-hop QA tasks when using hybrid retrieval methods compared to purely dense or sparse approaches, and how does this trade-off affect inference throughput?.

2 Methodology

Multi-query arXiv search (2 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

4 papers retrieved. 8 claims extracted, 3 verified. Tribunal: 6.2/10 → REVISE (revision_round=1). 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
BEIR-NL is a Dutch translation of the BEIR benchmark for zero-shot information retrieval evaluation.	✓	0.25
The benchmark BEIR-NL facilitates zero-shot IR evaluation for Dutch.	✓	0.18
BEIR-NL was made available on the Hugging Face hub.	✓	0.19
BEIR-NL uses the same licenses as the original BEIR datasets.	×	0.10
Machine translation can lead to inaccurate benchmark evaluations due to translation quality.	×	0.04
Recent advances in machine translation make large-scale translation feasible.	×	0.03
The dense ranking and reranking models evaluated include e5-multilingual-small, e5-multilingual-base, e5-multilingual-la	×	0.07
The models e5-multilingual-small, e5-multilingual-base, e1-multilingual-large, and e5-multilingual-large-instruct are pa	×	0.05

References

- http://arxiv.org/abs/2412.08329v1
- http://arxiv.org/abs/2104.08663v4
- http://arxiv.org/abs/2408.09437v1