SOVEREIGN: To what extent does fine-tuning on BEIR-NL improve downstream task performance in Dutch legal and news domains

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: To what extent does fine-tuning on BEIR-NL improve downstream task performance in Dutch legal and news domains, and how does this correlate with improvement in R@100 and MRR scores compared to zero-shot baselines?.

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

8 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 6.5/10 → 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.

References

- http://arxiv.org/abs/2412.08329v1
- http://arxiv.org/abs/2104.08663v4
- http://arxiv.org/abs/2504.05181v2