SOVEREIGN: An Efficiency Study for SPLADE Models

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Latency and efficiency issues are often overlooked when evaluating IR models based on Pretrained Language Models (PLMs) in reason of multiple hardware and software testing scenarios. Nevertheless, efficiency is an important part of such systems and should not be overlooked. In this paper, we focus on improving the efficiency of the SPLADE model since it has achieved state-of-the-art zero-shot performance and competitive results on TREC collections. SPLADE efficiency can be controlled via a regularization factor, but solely controlling this regularization has been shown to not be efficient en

1 Introduction

Analysis of: An Efficiency Study for SPLADE Models. Research goal: How does the inference throughput (queries per second) of RAG retrievers degrade under adversarial perturbations in multi-hop vs. single-hop settings when using fine-tuned dense retrieval models (e.g., Contriever) vs. lexical sparse models (e.g., SPLADE) on the MuSiQue dataset?.

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

13 papers retrieved. 5 claims extracted, 5 verified. Tribunal: 7.8/10 $\rightarrow$ 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
SPLADE has achieved state-of-the-art zero-shot performance and competitive results on TREC collections.	✓	0.28
SPLADE efficiency can be controlled via a regularization factor, but solely controlling this regularization has been sho	✓	0.31
The proposed techniques include L1 regularization for queries, a separation of document/query encoders, a FLOPS-regulari	✓	0.32
The proposed models achieve similar latency (less than 4ms difference) as traditional BM25 under the same computing cons	✓	0.26
The proposed models have similar performance (less than 10% MRR@10 reduction) as the state-of-the-art single-stage neura	✓	0.34

References

- http://arxiv.org/abs/1712.07113v2
- http://arxiv.org/abs/2207.03834v1
- https://www.semanticscholar.org/paper/ee3779890d674beb2dd002de3e5bcc6d1e6d0bcb