SOVEREIGN: To what extent does token pruning in SPLADE models degrade retrieval accuracy vs. improve latency on multi-hop

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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: To what extent does token pruning in SPLADE models degrade retrieval accuracy vs. improve latency on multi-hop QA datasets like HotpotQA and MuSiQue?.

2 Methodology

Multi-query arXiv search (1 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

3 papers retrieved. 5 claims extracted, 2 verified. Tribunal: 6.2/10 $\rightarrow$ 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
SPLADE models can be optimized with L1 regularization, separation of query and document encoders, FLOPS-regularized trai	✓	0.17
The proposed adaptations can reduce mono-thread retrieval latency of SPLADE models under PISA and ANSERINI frameworks	×	0.06
The benchmark demonstrates that neural models can achieve similar performance as BM22 and traditional BM25 with minimal	✓	0.16
The proposed adaptations can have similar latency as traditional BM25 systems under the same computing constraints	×	0.14
Sparse SPLADE models with these adaptations can achieve state-of-the-art performance with minimal latency differences co	×	0.13

References

- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2604.09019v2
- http://arxiv.org/abs/2207.03834v1