SOVEREIGN: How does the inference throughput (queries per second) of SPLADE-v3 compare to ColBERT-v2 under controlled spa

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Late interaction neural IR models like ColBERT offer a competitive effectiveness-efficiency trade-off across many benchmarks. However, they require a huge memory space to store the contextual representation for all the document tokens. Some works have proposed using either heuristics or statistical-based techniques to prune tokens from each document. This however doesn't guarantee that the removed tokens have no impact on the retrieval score. Our work uses a principled approach to define how to prune tokens without impacting the score between a document and a query. We introduce three regulari

1 Introduction

Analysis of: Towards Lossless Token Pruning in Late-Interaction Retrieval Models. Research goal: How does the inference throughput (queries per second) of SPLADE-v3 compare to ColBERT-v2 under controlled sparsity regularization on the BEIR benchmark?.

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

9 papers retrieved. 5 claims extracted, 4 verified. Tribunal: 8.0/10 $\rightarrow$ 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
Late interaction neural IR models like ColBERT offer a competitive effectiveness-efficiency trade-off across many benchm	✓	0.41
	×	0.00
Some works have proposed using either heuristics or statistical-based techniques to prune tokens from each document	✓	0.37
Our work uses a principled approach to define how to prune tokens without impacting the retrieval score	✓	0.34
We can preserve ColBERT's performance while using only 30% of the tokens	✓	0.29

References

- https://doi.org/10.1145/3578337.3605142
- https://doi.org/10.1162/tacl_a_00556
- https://doi.org/10.1145/3726302.3730100