Why Advanced Encoders Lag on Sparse Retrieval? The Answer and an Approach to Bridging Vocabulary Gaps

Anonymous Author(s)

Abstract

While advanced foundation models like ModernBERT significantly outperform older architectures in dense retrieval, they surprisingly lag behind the aging BERT-base baseline in learned sparse retrieval (LSR). We identify the root cause as the *Vocabulary Gap*: modern tokenizers utilize raw, case-sensitive vocabularies designed for lossless reconstruction, which map single semantic units to redundant surface forms, wasting model capacity on morphological noise and hindering lexical matching. We formalize this intuition through a theoretical framework, demonstrating that appropriate vocabulary coarse-graining can tighten the generalization bounds by reducing complexity of the hypothesis class, provided that semantic integrity is preserved. To resolve this, we propose **Vocabulary Transfer (VT)**, a model-agnostic framework that migrates advanced encoders to sparse-friendly, normalized vocabularies with minimal computational cost. VT utilizes a novel **Semantic Initialization** via spatial topology to preserve geometric structure and an **Activation Potential Calibration (APC)** mechanism to align pre-trained manifolds with sparsity constraints, preventing the dead neuron and dense collapse observed in standard fine-tuning. Empirically, VT is universally effective: it enables ModernBERT to achieve state-of-the-art performance on the BEIR benchmark (**52.4** nDCG, a **+4.7** improvement), resuscitates failing models like RoBERTa-large, and generalizes seamlessly to inference-free architectures and specialized domains. These results confirm that the performance lag is not an architectural deficiency but a solvable vocabulary mismatch. We've released our code and models.[1]

CCS Concepts

• **Information systems** → **Information retrieval**; **Document representation**.

Keywords

SPLADE, learned sparse representations, passage retrieval

ACM Reference Format:

Anonymous Author(s). 2026. Why Advanced Encoders Lag on Sparse Retrieval? The Answer and an Approach to Bridging Vocabulary Gaps. In *Proceedings of Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

[1]https://anonymous.4open.science/r/vocab-transfer/. All details included.

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

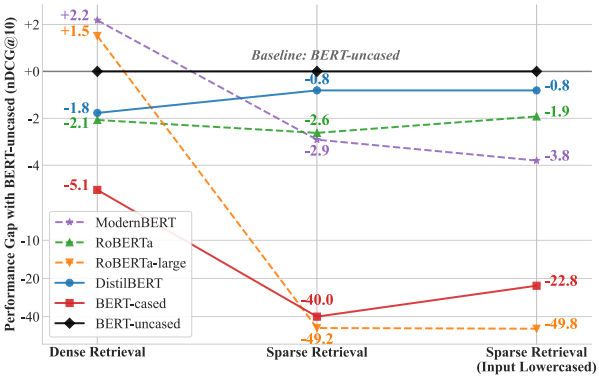

Figure 1: The Vocabulary Gap anomaly. While advanced encoders like ModernBERT significantly outperform BERT in dense retrieval, they lag behind in sparse retrieval under standard fine-tuning.

1 Introduction

The landscape of neural information retrieval has bifurcated into two dominant paradigms: dense retrieval, which encodes queries and documents into continuous low-dimensional embeddings [21, 54], and learned sparse retrieval (LSR), which projects text into high-dimensional, weighted lexical vectors [15, 31]. While dense retrievers excel at capturing semantic nuances, sparse retrievers—exemplified by models like SPLADE [15]—retain the interpretability and efficiency of inverted indices while mitigating the lexical mismatch problem of traditional BM25 [32, 45].

In the dense retrieval paradigm, upgrading the backbone is a proven strategy. Modern foundations like ModernBERT [52] provide not only stronger representations but also architectural advantages like 8k context windows and FlashAttention compatibility.

However, these architectural leaps remain inaccessible to sparse retrieval. We observe a puzzling anomaly: *advanced encoders consistently underperform in sparse settings, often lagging behind the older BERT-base-uncased baseline.* As illustrated in Figure 1, this performance degradation is pervasive. The most intuitive explanation attributes this to the BPE tokenizer differences in modern models. However, we observe that `bert-base-cased`, which uses the same WordPiece tokenizer as the effective `bert-base-uncased` baseline, performs equally poorly. This isolates the degree of vocabulary normalization as the critical variable. This regression persists despite identical training pipelines, suggesting that the architectural advancements of modern backbones are stifled by a fundamental incompatibility with the sparse retrieval objective.

We identify the root cause as the **vocabulary gap**—specifically, the shift in modern tokenization toward raw, case-sensitive vocabularies designed for lossless reconstruction rather than lexical