Throughput Trade-offs of MA-DPR and Quantized Euclidean DPR on BEIR with Edge AI Accelerators

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: What is the throughput trade-off between MA-DPR and quantized Euclidean DPR models when evaluated on the BEIR benchmark using edge AI accelerators. Encoder-only transformer models such as BERT offer a great performance-size tradeoff for retrieval and classification tasks with respect to larger decoder-only models. Despite being the workhorse of numerous production pipelines, there have been limited Pareto improvements to. 6 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Comparative Study of Clinical ModernBERT and BioMedical ModernBERT on the DDXPlus Dataset. Research question: What is the throughput trade-off between MA-DPR and quantized Euclidean DPR models when evaluated on the BEIR benchmark using edge AI accelerators?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.7/10.

3 Results

7 papers retrieved. 6 claims extracted; 6 independently verified. Quality review score: 8.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Encoder-only transformer models such as BERT offer a great performance-size tradeoff for retrieval and classification ta	✓	0.45
There have been limited Pareto improvements to BERT since its release.	✓	0.24
ModernBERT is introduced, bringing modern model optimizations to encoder-only models and representing a major Pareto imp	✓	0.35
ModernBERT models are trained on 2 trillion tokens with a native 8192 sequence length.	✓	0.30
ModernBERT models exhibit state-of-the-art results on a large pool of evaluations encompassing diverse classification ta	✓	0.48
ModernBERT is the most speed and memory efficient encoder and is designed for inference on common GPUs.	✓	0.31

References

- https://doi.org/10.48550/arxiv.2412.13663
- https://doi.org/10.3390/electronics15030541
- https://doi.org/10.1016/j.ins.2026.123181