Dynamic Entity Representation Impact on RAG Inference Latency and Retrieval Effectiveness in MS MARCO

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: To what extent does dynamic entity representation in NER Retriever affect the inference latency of RAG models on the MS MARCO benchmark while maintaining retrieval effectiveness. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MS-Shift: An Analysis of MS MARCO Distribution Shifts on Neural Retrieval. Research question: To what extent does dynamic entity representation in NER Retriever affect the inference latency of RAG models on the MS MARCO benchmark while maintaining retrieval effectiveness?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

11 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
MS MARCO passage dataset contains approximately 8.8M passages and 500k training queries.	×	0.13
Query topics are separated into five semantic clusters (Ci)i$\in$J0,4K.	×	0.03
The t-SNE visualization shows that clusters 3 and 4 are close to each other but correspond to Medical Treatments (red) a	×	0.00
The median query length at the word level for MS MARCO is 6.	×	0.09
Train/test sets for short and long queries contain respectively 10M training triplets and 3500 queries for evaluation.	×	0.06
Each wh-words cluster (Wi)i=0,1,2 contains 6500 queries for evaluation.	×	0.03
The evaluation procedure involves leave-one-out on all the shifts to evaluate the in-domain and zero-shot effectiveness	×	0.07
The performance measure Avg In is the average performance when the distribution of the evaluated cluster is seen at trai	×	0.04
Rel Loss is the relative loss between the average performance measure and the zero-shot performance.	×	0.03

References

- http://arxiv.org/abs/2509.04011v1
- http://arxiv.org/abs/2205.02870v2
- http://arxiv.org/abs/2604.18234v1