

Comparative Accuracy of Dual-Encoder Retrievers and Cross-Encoder Rerankers on Noisy LoCoMo Queries

Assignee Research

June 12, 2026

Abstract

Dense retrieval is becoming one of the standard approaches for document and passage ranking. The dual-encoder architecture is widely adopted for scoring question-passage pairs due to its efficiency and high performance. Typically, dense retrieval models are evaluated on clean and curated datasets. However, when deployed in real-life applications, these models encounter noisy user-generated text. That said, the performance of state-of-the-art dense retrievers can substantially deteriorate when exposed to noisy text. In this work, we study the robustness of dense retrievers against typos in the

1 Introduction

This paper examines: Analysing the Robustness of Dual Encoders for Dense Retrieval Against Misspellings. Research question: What is the comparative accuracy of dual-encoder dense retrievers versus cross-encoder rerankers on LoCoMo when the query text contains intentional misspellings and semantic noise?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

3 Results

14 papers retrieved. 11 claims extracted; 8 independently verified. Quality review score: 7.4/10.

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 |
|--|----------|------------|
| On clean questions, data augmentation, contrastive learning, and their combination do not harm the retrieval performance | ✓ | 0.18 |
| All robustification approaches (Data augm., CL, Data augm. + CL) perform significantly better than the original Dual Enc | × | 0.10 |
| The combined approach of data augmentation and contrastive learning achieves the highest performance among all tested me | × | 0.11 |
| Robustness of dual encoder models deteriorates when typos are restricted to non-stopwords compared to when typos appear | ✓ | 0.15 |
| The most significant performance losses occur when typos appear in discriminative utterances (words overlapping between | ✓ | 0.21 |
| The combined data augmentation and contrastive learning approach remains the best performing method across settings wher | ✓ | 0.23 |
| There is a strong positive correlation between the frequency of typoed words in the training set and retrieval performan | ✓ | 0.28 |
| On the Natural Questions test set with random typos, the original DR model achieves an AR@5 of 49.52. | × | 0.14 |
| On the Natural Questions test set with random typos, the DR + Data augm. + CL (ours) model achieves an AR@5 of 62.13. | ✓ | 0.15 |
| On the Natural Questions test set with typos in discriminative utterances, the original DR model achieves an AR@5 of 10. | ✓ | 0.18 |
| On the Natural Questions test set with typos in discriminative utterances, the DR + Data augm. + CL (ours) model achieve | ✓ | 0.18 |

References

- <http://arxiv.org/abs/2606.01070v1>
- <http://arxiv.org/abs/2303.14991v1>
- <http://arxiv.org/abs/2205.02303v1>