Adversarial Training with Synthetic Misspellings in Zero-Shot Dual-Encoder Retrieval

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does adversarial training on synthetic misspelling datasets affect the zero-shot retrieval accuracy of dual-encoder models on the TriviaQA and Natural Questions benchmarks compared to clean. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Analysing the Robustness of Dual Encoders for Dense Retrieval Against Misspellings. Research question: How does adversarial training on synthetic misspelling datasets affect the zero-shot retrieval accuracy of dual-encoder models on the TriviaQA and Natural Questions benchmarks compared to clean training baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

13 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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	×	0.13
All robustification approaches (Data Augmentation, Contrastive Learning, and Combined) perform significantly better than	×	0.12
The combined approach of data augmentation and contrastive learning achieves the highest performance among all tested me	×	0.08
Robustness of dual encoder models deteriorates when typos are restricted to non-stopwords compared to when typos appear	×	0.11
The most significant performance losses occur when typos appear in discriminative utterances (words overlapping with the	×	0.04
The combined data augmentation and contrastive learning approach remains the best performing method across all typo sett	×	0.07
There is a strong positive correlation between the frequency of typoed words in the training set and retrieval performan	×	0.10
Retrieval performance drops significantly as the frequency of the typoed words in the training set decreases.	×	0.05

References

- http://arxiv.org/abs/2204.00716v2
- http://arxiv.org/abs/2205.02303v1
- http://arxiv.org/abs/2602.12783v2