Dual-Encoder vs. Cross-Encoder Robustness to Misspellings in HELM Benchmarking

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 6 peer-reviewed papers addressing the following research question: How does the robustness of dual-encoder retrieval models to misspellings compare to cross-encoder models, as measured by HELM fairness and robustness metrics across different misspelling severities. 11 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Holistic Evaluation of Language Models. Research question: How does the robustness of dual-encoder retrieval models to misspellings compare to cross-encoder models, as measured by HELM fairness and robustness metrics across different misspelling severities?.

2 Methodology

Systematic literature search across multiple databases yielded 6 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

6 papers retrieved. 11 claims extracted; 7 independently verified. Quality review score: 7.2/10.

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
HELM measures 7 metrics: accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency.	✓	0.23
HELM evaluates 16 core scenarios.	×	0.11
The 7 metrics are measured for each of the 16 core scenarios 87.5% of the time.	✓	0.15
HELM performs 7 targeted evaluations based on 26 targeted scenarios.	✓	0.18
HELM conducts a large-scale evaluation of 30 prominent language models.	✓	0.22
The 30 evaluated models span open, limited-access, and closed models.	✓	0.16
HELM evaluates models on a total of 42 scenarios.	×	0.08
21 of the 42 scenarios used in HELM were not previously used in mainstream LM evaluation.	✓	0.19
Prior to HELM, models on average were evaluated on just 17.9% of the core HELM scenarios.	✓	0.25
Prior to HELM, some prominent models did not share a single evaluation scenario in common.	×	0.14
HELM improves scenario coverage to 96.0% across the 30 evaluated models.	×	0.11

References

- https://doi.org/10.48550/arxiv.2211.09110
- https://doi.org/10.48550/arxiv.2310.14724
- https://doi.org/10.48550/arxiv.2301.12867