Fine-Tuning Multilingual M2Qa Models On Domain-Specific Corpora Performance On Their Adversarial Robustness Scores

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does fine-tuning multilingual M2QA models on domain-specific corpora affect their adversarial robustness scores compared to zero-shot cross-domain transfer. In response to rising concerns surrounding the safety, security, and trustworthiness of Generative AI (GenAI) models, practitioners and regulators alike have pointed to AI red-teaming as a key component of their strategies for identifying and mitigating these risks. However. 8 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Red-Teaming for Generative AI: Silver Bullet or Security Theater?. Research question: How does fine-tuning multilingual M2QA models on domain-specific corpora affect their adversarial robustness scores compared to zero-shot cross-domain transfer?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.3/10.

3 Results

10 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 8.3/10.

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
Practitioners and regulators have pointed to AI red-teaming as a key component of strategies for identifying and mitigat	✓	0.33
Significant questions remain regarding the precise definition of AI red-teaming, its role in regulation, and its relatio	✓	0.21
The authors conducted an extensive survey of relevant research literature and identified recent cases of red-teaming act	✓	0.26
Prior methods and practices of AI red-teaming diverge along axes including the purpose of the activity, the artifact und	✓	0.36
The purpose of AI red-teaming activities is often vague.	✓	0.19
The authors argue that red-teaming is a valuable 'big-tent' idea for characterizing GenAI harm mitigations.	✓	0.29
The authors argue that industry may effectively apply red-teaming and other strategies behind closed doors to safeguard	✓	0.27
The authors argue that gestures towards red-teaming based on public definitions are not a panacea for every possible ris	✓	0.21

References

- https://doi.org/10.48550/arxiv.2304.05613
- https://doi.org/10.48550/arxiv.2307.10169
- https://doi.org/10.1609/aies.v7i1.31647