Mean Shift Clustering Effects on Adversarial Robustness in AdaptToken Models

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the mean shift clustering technique impact the adversarial robustness of AdaptToken-8B vs. AdaptToken-3B when fine-tuned on AdvGLUE tasks, as measured by accuracy under targeted FGSM attacks. 11 claims were extracted from source literature; 10 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Research question: How does the mean shift clustering technique impact the adversarial robustness of AdaptToken-8B vs. AdaptToken-3B when fine-tuned on AdvGLUE tasks, as measured by accuracy under targeted FGSM attacks?.

2 Methodology

Systematic literature search across multiple databases yielded 4 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

4 papers retrieved. 11 claims extracted; 10 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
Deep learning models for medical diagnostics can maintain performance in the presence of adversarial or noisy inputs.	✓	0.25
Model complexity, training data quality, and hyperparameters influence the reliability of deep learning models for medic	✓	0.25
Adversarial attacks aim to deceive deep learning models used in medical diagnostics.	✓	0.21
Privacy attacks seek to extract sensitive information from deep learning models used in medical diagnostics.	✓	0.23
Adversarial training and input preprocessing are defenses to enhance the robustness of deep learning models for medical	✓	0.24
Data augmentation and uncertainty estimation are mechanisms to enhance the robustness of deep learning models for medica	✓	0.23
Tools and packages extend the reliability features of deep learning frameworks such as TensorFlow and PyTorch for medica	✓	0.27
Existing evaluation metrics for robustness are being discussed and evaluated for deep learning models in medical diagnos	✓	0.25
There are limitations in the existing literature on the robustness of deep learning models for medical diagnostics.	✓	0.22
Future research directions aim to enhance the robustness of deep learning models for medical diagnostics.	✓	0.22
The goal is to ensure that AI systems for medical diagnostics are trustworthy, reliable, and stable.	×	0.14

References

- https://doi.org/10.1007/s10462-024-11005-9
- https://doi.org/10.48550/arxiv.2401.05778
- https://doi.org/10.48550/arxiv.2407.21792