Adversarial Robustness and Accuracy Trade-offs in Fine-Tuned S4 and RoBERTa Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the trade-off between adversarial robustness and in-domain accuracy when applying the Growth Bound Matrix to fine-tuned S4 models versus RoBERTa on Adversarial GLUE tasks. 8 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adversarial Robustness of Prompt-based Few-Shot Learning for Natural Language Understanding. Research question: What is the trade-off between adversarial robustness and in-domain accuracy when applying the Growth Bound Matrix to fine-tuned S4 models versus RoBERTa on Adversarial GLUE tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.2/10.

3 Results

11 papers retrieved. 8 claims extracted; 3 independently verified. Quality review score: 6.2/10.

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
Using unlabeled data (iPET) during fine-tuning causes prompting to reduce the drop in adversarial performance with respe	×	0.13
Using multiple prompts to fine-tune multiple models (PET) and ensembling the resultant predictions causes prompting to d	×	0.12
Increasing the number of few-shot examples reduces the relative drop in adversarial performance with respect to in-domai	✓	0.17
Increasing the encoder size reduces the relative drop in adversarial performance with respect to in-domain performance.	×	0.06
RoBERTa encoders are more adversarially robust than ALBERT and BERT encoders of comparable size.	×	0.02
Vanilla FSL methods lead to a notable relative drop in task performance compared to fully fine-tuned models in the face	✓	0.41
The study evaluates four FSL methods: Classic fine-tuning, LM-BFF, PET, and iPET.	×	0.11
FewNLU is a benchmark designed to evaluate the performance of prompt-based few-shot learning capabilities systematically	✓	0.16

References

- http://arxiv.org/abs/2306.11066v2
- http://arxiv.org/abs/1901.08573v3
- http://arxiv.org/abs/2103.15670v3