

# Adversarial Training with Jacobian Regularization for Cross-Domain Language Robustness in AdaptToken-3B

Assignee Research

June 6, 2026

## Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Can adversarial training with Jacobian regularization improve the robustness of AdaptToken-3B on cross-domain language tasks, and how does it compare to mean shift-based analysis in terms of F1 score. 9 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.0/10. 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: Can adversarial training with Jacobian regularization improve the robustness of AdaptToken-3B on cross-domain language tasks, and how does it compare to mean shift-based analysis in terms of F1 score and adversarial perturbation resilience?.

## 2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.0/10.

## 3 Results

12 papers retrieved. 9 claims extracted; 4 independently verified. Quality review score: 6.0/10.

## 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
Using unlabeled data (iPET) during fine-tuning causes prompting to reduce the drop in adversarial performance with respect to	×	0.12
Using multiple prompts to fine-tune multiple models (PET) and ensembling the resultant predictions cause prompting to decrease	×	0.12
Increasing the number of few-shot examples and the encoder size reduces the relative drop in adversarial performance with respect to	✓	0.16
RoBERTa encoders are more adversarially robust than ALBERT and BERT encoders of comparable size.	×	0.02
Few-shot learning aims to train models to perform well on a wide range of natural language understanding tasks with a small number of	×	0.14
Prompt-based learning overcomes the requirement of training task-specific classification heads, matching the fine-tuning	×	0.09
Vanilla FSL methods lead to a notable relative drop in task performance (i.e., are less robust) in the face of adversarial	✓	0.43
Using unlabeled data for prompt-based FSL and multiple prompts flip the trend of reduced robustness in the face of adversarial	✓	0.33
Increasing the number of few-shot examples and model size lead to increased adversarial robustness of vanilla FSL method	✓	0.43

## References

- <http://arxiv.org/abs/2306.11066v2>

- <http://arxiv.org/abs/2504.07887v2>
- <http://arxiv.org/abs/2103.15670v3>