

Correlation between Model Scale and Accuracy Degradation under Adversarial Perturbations in NLU Tasks

Assignee Research

June 14, 2026

Abstract

State-of-the-art few-shot learning (FSL) methods leverage prompt-based fine-tuning to obtain remarkable results for natural language understanding (NLU) tasks. While much of the prior FSL methods focus on improving downstream task performance, there is a limited understanding of the adversarial robustness of such methods. In this work, we conduct an extensive study of several state-of-the-art FSL methods to assess their robustness to adversarial perturbations. To better understand the impact of various factors towards robustness (or the lack of it), we evaluate prompt-based FSL methods against

1 Introduction

This paper examines: Adversarial Robustness of Prompt-based Few-Shot Learning for Natural Language Understanding. Research question: What is the correlation between model scale and accuracy degradation rates under prompt-based few-shot adversarial perturbations in NLU tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.7/10.

3 Results

14 papers retrieved. 11 claims extracted; 10 independently verified. Quality review score: 8.7/10.

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	✓	0.27
Using multiple prompts to fine-tune multiple models (PET) and ensembling the resultant predictions cause prompting to decrease	✓	0.39
Increasing the number of few-shot examples and the encoder size reduces the relative drop in adversarial performance with respect to	✓	0.31
RoBERTa encoders are more adversarially robust than ALBERT and BERT encoders of comparable size.	✓	0.18
Vanilla FSL methods lead to a notable relative drop in task performance compared to fully fine-tuned models in the face of	✓	0.40
Using unlabeled data for prompt-based FSL and multiple prompts flip the trend of reduced robustness in vanilla FSL methods	✓	0.32
Increasing the number of few-shot examples and model size lead to increased adversarial robustness of vanilla FSL method	✓	0.41
The benchmark FewNLU categorizes prompt-based few-shot learning settings into three categories: (i) not using any unlabeled	✓	0.18
The sentiment classification task can be transformed using a template: 'I loved the movie!' \rightarrow 'I loved the movie! It was	✓	0.34
The study evaluates four different FSL methods: Classic fine-tuning, LM-BFF, PET, and iPET. Classic-FT is used as a baseline for fully labeled	✓	0.18
data to give the ceiling performance.	×	0.11

References

- <http://arxiv.org/abs/2403.09832v1>
- <http://arxiv.org/abs/2111.02840v2>
- <http://arxiv.org/abs/2306.11066v2>