Parameter Count and Generalization in Adversarial Few-Shot MBPP Tasks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the correlation between model parameter count and generalization performance on adversarially perturbed MBPP Pro tasks in few-shot learning setups. 10 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.4/10. 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 correlation between model parameter count and generalization performance on adversarially perturbed MBPP Pro tasks in few-shot learning setups?.

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 6.4/10.

3 Results

10 papers retrieved. 10 claims extracted; 5 independently verified. Quality review score: 6.4/10.

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
Using unlabeled data (iPET) during fine-tuning causes prompting to reduce the drop in adversarial performance with respe	×	0.11
Using multiple prompts to fine-tune multiple models (PET) and ensembling the resultant predictions cause prompting to de	×	0.11
Increasing the number of few-shot examples and the encoder size reduces the relative drop in adversarial performance wit	✓	0.16
RoBERTa encoders are more adversarially robust than ALBERT and BERT encoders of comparable size.	×	0.06
Vanilla FSL methods lead to a notable relative drop in task performance (i.e., are less robust) in the face of adversari	✓	0.42
Using unlabeled data for prompt-based FSL and multiple prompts flip the trend of reduced robustness in vanilla FSL metho	✓	0.35
Increasing the number of few-shot examples and model size lead to increased adversarial robustness of vanilla FSL method	✓	0.40
The study evaluates four different FSL methods: Classic fine-tuning, LM-BFF, PET, and iPET.	×	0.07
The study covers three primary settings in state-of-the-art prompt-based FSL methods: no use of unlabeled data for train	✓	0.22
Fine-tuning with fully labeled data is used to give the ceiling performance and contrast the capabilities of the FSL met	×	0.10

References

- http://arxiv.org/abs/2306.11066v2
- http://arxiv.org/abs/2509.21843v1
- http://arxiv.org/abs/2106.12900v3