Diversity of Synthetic Oversampling Techniques and LLM Robustness to Adversarial Prompts in Few-Shot Reasoning

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the diversity of synthetic minority oversampling techniques impact the robustness of large language models against adversarial prompt perturbations in few-shot reasoning benchmarks. 9 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.2/10. 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: How does the diversity of synthetic minority oversampling techniques impact the robustness of large language models against adversarial prompt perturbations in few-shot reasoning benchmarks?.

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. 9 claims extracted; 4 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	×	0.11
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	✓	0.15
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 compared to fully fine-tuned models in the face	✓	0.39
Using unlabeled data for prompt-based FSL and multiple prompts flip the trend of reduced robustness in vanilla FSL metho	✓	0.37
Increasing the number of few-shot examples and model size lead to increased adversarial robustness of vanilla FSL method	✓	0.41
The study evaluates four different FSL methods: Classic fine-tuning, LM-BFF, PET, and iPET.	×	0.08
The study considers fine-tuning with fully labeled data to give the ceiling performance and contrast the capabilities of	×	0.10

References

- http://arxiv.org/abs/2306.11066v2
- http://arxiv.org/abs/2510.22389v2

- http://arxiv.org/abs/2308.10783v2