Rationale-Augmented Preference Data Enhances DPO Model Robustness to Adversarial Prompts

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the inclusion of rationales in preference data influence the robustness of DPO-trained models to adversarial prompts, measured by accuracy on the AdversarialQA benchmark across different. 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. 10 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. 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 inclusion of rationales in preference data influence the robustness of DPO-trained models to adversarial prompts, measured by accuracy on the AdversarialQA benchmark across different perturbation strengths?.

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 5.8/10.

3 Results

10 papers retrieved. 10 claims extracted; 4 independently verified. Quality review score: 5.8/10.

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	×	0.12
Using multiple prompts to fine-tune multiple models (PET) and ensembling the resultant predictions cause prompting to de	×	0.12
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.	×	0.02
Few-shot learning aims to train models to perform well on a wide range of natural language understanding tasks with a sm	×	0.14
Prompt-based learning overcomes the requirement of training task-specific classification heads, matching the fine-tuning	×	0.08
FewNLU is a benchmark designed to evaluate the performance of prompt-based few-shot learning capabilities systematically	×	0.13
Vanilla FSL methods lead to a notable relative drop in task performance (i.e., are less robust) in the face of adversari	✓	0.43
Using unlabeled data for prompt-based FSL and multiple prompts flip the trend of reduced robustness in the face of adver	✓	0.33
Increasing the number of few-shot examples and model size lead to increased adversarial robustness of vanilla FSL method	✓	0.44

References

- http://arxiv.org/abs/2407.14477v4
- http://arxiv.org/abs/2306.11066v2
- http://arxiv.org/abs/2506.10054v4