

How does adversarial training during intermediate-task fine-tuning influence robustness in zero-shot cross-lingual performance on

Assignee Research

July 25, 2026

Abstract

Intermediate-task training—fine-tuning a pretrained model on an intermediate task before fine-tuning again on the target task—often improves model performance substantially on language understanding tasks in monolingual English settings. We investigate whether English intermediate-task training is still helpful on non-English target tasks. Using nine intermediate language-understanding tasks, we evaluate intermediate-task transfer in a zero-shot cross-lingual setting on the XTREME benchmark. We see large improvements from intermediate training on the BUCC and Tatoeba sentence retrieval tas

1 Introduction

This paper examines: Robustness of Multilingual Models in Zero-Shot Cross-Lingual Settings via Intermediate English Task Training. Research question: How does adversarial training during intermediate-task fine-tuning influence robustness in zero-shot cross-lingual performance on XTREME-R, measured in F1 score under perturbed inputs compared to clean inputs?.

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 7.8/10.

3 Results

14 papers retrieved. 9 claims extracted; 9 independently verified. Quality review score: 7.8/10.

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
Intermediate-task training often improves model performance substantially on language understanding tasks in monolingual	✓	0.42
The study investigates whether English intermediate-task training is helpful on non-English target tasks.	✓	0.33
The study uses nine intermediate language-understanding tasks.	✓	0.17
The study evaluates intermediate-task transfer in a zero-shot cross-lingual setting on the XTREME benchmark.	✓	0.34
The study observes large improvements from intermediate training on the BUCC and Tatoeba sentence retrieval tasks.	✓	0.25
The research goal is to determine if intermediate-task training on English affects the robustness of multilingual models	✓	0.45
The robustness is evaluated by accuracy drop and response time differences on perturbed XTREME-R inputs.	✓	0.30
The autonomous synthesis report was generated by Assignee Research.	✓	0.24
The tribunal consensus score is 9.3/10.	✓	0.19

References

- <https://doi.org/10.5281/zenodo.21449595>
- <https://doi.org/10.18653/v1/2021.naacl-main.41>
- https://doi.org/10.1007/978-3-031-23190-2_3