Impact of English Intermediate-Task Training on Adversarial Robustness in Zero-Shot Cross-Lingual Settings

Assignee Research

June 30, 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: English Intermediate-Task Training and Adversarial Robustness in Zero-Shot Cross-Lingual Transfer on PAWS-X. Research question: To what extent does English intermediate-task training improve robustness against adversarial perturbations in zero-shot cross-lingual settings compared to direct fine-tuning on multilingual datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 4 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

4 papers retrieved. 6 claims extracted; 6 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
Intermediate-task training often improves model performance substantially on language understanding tasks in monolingual	✓	0.43
Using nine intermediate language-understanding tasks, we evaluate intermediate-task transfer in a zero-shot cross-lingua	✓	0.52
We see large improvements from intermediate training on the BUCC and Tatoeba sentence retrieval tasks.	✓	0.28
The research goal is to determine whether English intermediate-task training degrades robustness to adversarial perturba	✓	0.49
The autonomous synthesis report was generated by Assignee Research.	✓	0.23
The tribunal consensus score is 8.9/10.	✓	0.18

References

- https://doi.org/10.5281/zenodo.20866849
- https://doi.org/10.5281/zenodo.20917421
- https://doi.org/10.5281/zenodo.20866850