

How does the choice of English intermediate-task (e.g., NLI, QA) impact zero-shot cross-lingual transfer performance on XTREME-R,

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 tasks a

1 Introduction

This paper examines: . Research question: How does the choice of English intermediate-task (e.g., NLI, QA) impact zero-shot cross-lingual transfer performance on XTREME-R, compared to direct fine-tuning, when evaluated by language-specific F1 scores?.

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 6.3/10.

3 Results

11 papers retrieved. 9 claims extracted; 9 independently verified. Quality review score: 6.3/10.

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 improves model performance substantially on language understanding tasks in monolingual English	✓	0.39
Using nine intermediate language-understanding tasks, the study evaluates intermediate-task transfer in a zero-shot cross	✓	0.36
Large improvements from intermediate training are observed on the BUCC and Tatoeba sentence retrieval tasks.	✓	0.26
Moderate improvements from intermediate training are observed on question-answering target tasks.	✓	0.27
MNLI, SQuAD, and HellaSwag achieve the best overall results as intermediate tasks.	✓	0.30
Multi-task intermediate offers small additional improvements.	✓	0.26
Using the best intermediate-task models for each target task, a 5.4 point improvement over XLM-R Large on the XTREME benchmark	✓	0.41
Continuing multilingual MLM during intermediate-task training does not consistently outperform simply performing English	✓	0.30
Using machine-translated intermediate-task data does not consistently outperform simply performing English intermediate-	✓	0.31

References

- <https://doi.org/10.18653/v1/2020.emnlp-main.618>
- <https://doi.org/10.18653/v1/2020.acl-main.653>
- <https://openalex.org/W3116343068>