Scaling Intermediate Tasks for Zero-Shot Cross-Lingual Transfer in LLMs

Assignee Research

July 27, 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: Scaling Intermediate Tasks in Large Language Models for Zero-Shot Cross-Lingual Transfer on XTREME. Research question: What is the impact of scaling the number of intermediate language-understanding tasks on the zero-shot cross-lingual transfer performance of LLMs, as measured by the average accuracy on the XTREME benchmark when using a fixed target task?.

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 7.8/10.

3 Results

4 papers retrieved. 6 claims extracted; 5 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
Using nine intermediate language-understanding tasks, intermediate-task transfer was evaluated in a zero-shot cross-ling	✓	0.42
Large improvements were observed from intermediate training on the BUCC and Tatoeba sentence retrieval tasks.	✓	0.22
The research goal is to determine the impact of scaling the number of intermediate tasks (e.g., 5 vs. 20) on the zero-sh	✓	0.57
The autonomous synthesis report was generated by Assignee Research.	✓	0.21
The tribunal consensus score for the research is 8.5/10.	×	0.13

References

- https://doi.org/10.5281/zenodo.21250468
- https://doi.org/10.5281/zenodo.21250469
- https://doi.org/10.5281/zenodo.21312156