Impact of Intermediate Task Sequencing on Multilingual Fine-Tuning Robustness and Accuracy

Assignee Research

June 23, 2026

Abstract

Pre-trained multilingual language models show significant performance gains for zero-shot cross-lingual model transfer on a wide range of natural language understanding (NLU) tasks. Previously, for zero-shot cross-lingual evaluation, pre-trained models are only fine-tuned on English data and tested on a variety of target languages. In this paper, we do cross-lingual evaluation on various NLU tasks (sentence classification, sequence labeling, question answering) using prompt-tuning and compare it with fine-tuning. The results show that prompt tuning achieves much better cross-lingual transfer t

1 Introduction

This paper examines: Prompt-Tuning Can Be Much Better Than Fine-Tuning on Cross-lingual Understanding With Multilingual Language Models. Research question: Does the sequence of intermediate tasks in multilingual fine-tuning pipelines affect robustness and accuracy across diverse language families in cross-lingual NLU evaluations?.

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 7.5/10.

3 Results

14 papers retrieved. 19 claims extracted; 14 independently verified. Quality review score: 7.5/10.

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
In fine-tuning, all model parameters are tuned on English task data.	✓	0.22
In prompt tuning, only a small ratio of parameters (e.g., prompts or task classifier) are tuned during learning.	✓	0.21
Lester et al. (2021) found that prompt tuning can be better than fine-tuning when the model size is not extremely large	✓	0.29
The experiments use the pre-trained XLM-R checkpoint of LARGE size with about 560M parameters.	✓	0.17
Previous work (Hu et al., 2020) shows XLM-R achieves stronger performance than mBERT.	✓	0.21
In the implementation, continuous prompts are operated as past keys and values in each transformer layer.	✓	0.24
Each transformer layer has separated prompts in the proposed framework.	✓	0.15
During prompt tuning, the multilingual language model parameters are frozen.	✓	0.18
Experiments were performed on four datasets included in XTREME: XNLI, PAWS-X, UD-POS, XQuAD, and TyDiQA-GoldP.	✓	0.20
The prompt length used in experiments is 16, except for the XNLI task where it is 32.	×	0.12
The prompt tuning framework uses only 0.1% to 0.3% additional prompt parameters compared to the original model.	✓	0.15
For both fine-tuning and prompt tuning, models are only fine-tuned on the English training data but evaluated on all tar	✓	0.20
On the XNLI task, the Fine-Tuning method achieved an accuracy of 78.8 (std 0.2) using XLM-R-LARGE.	×	0.09
On the XNLI task, the Prompt Tuning method achieved an accuracy of 79.9 (std 0.1) using XLM-R-LARGE.	×	0.11
On the PAWS-X task, Prompt Tuning achieved an accuracy of 88.4 (std 0.3), while Fine-Tuning achieved 87.9 (std 0.5).	×	0.06
On the UD-POS task, Prompt Tuning achieved an F1 score of 75.4 (std 0.2), while Fine-Tuning achieved 74.4 (std 0.7). 4	×	0.09
On the XQuAD task, Prompt Tuning achieved an F1/EM of 79.0/64.1, while Fine-Tuning achieved 77.3/61.8.	✓	0.20
On the TyDiQA task, Prompt Tuning achieved an F1/EM of 71.5/55.1, while Fine-Tuning achieved 70.1/51.7.	✓	0.16
Table 3 shows the cosine similarity of representa	✓	0.24

References

- http://arxiv.org/abs/2311.07820v1
- http://arxiv.org/abs/2210.12360v2
- http://arxiv.org/abs/2310.00905v2