Cross-lingual Zero-shot Performance with Instruction Tuning Proportion Variations

Assignee Research

July 26, 2026

Abstract

Large language models respond well in high-resource languages like English but struggle in low-resource languages. It may arise from the lack of high-quality instruction following data in these languages. Directly translating English samples into these languages can be a solution but unreliable, leading to responses with translation errors and lacking language-specific or cultural knowledge. To address this issue, we propose a novel method to construct cross-lingual instruction following samples with instruction in English and response in low-resource languages. Specifically, the language mode

1 Introduction

This paper examines: X-Instruction: Aligning Language Model in Low-resource Languages with Self-curated Cross-lingual Instructions. Research question: What is the impact of varying the proportion of high-resource versus low-resource language instructions during instruction tuning on zero-shot cross-lingual performance, evaluated using XTREME-R and accuracy on a held-out low-resource language subset?.

2 Methodology

Systematic literature search across multiple databases yielded 10 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

10 papers retrieved. 10 claims extracted; 7 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
The X-Instruction dataset contains 320k samples for 10 languages.	×	0.15
The seed data for each language includes the first conversation turn of message trees in the Open Assistant dataset.	✓	0.20
The seed data is supplemented by translating 3k English samples into the target language using Google Translate.	✓	0.17
The multilingual corpus for each language is extracted from the CulturaX dataset, which contains web texts in 167 langua	✓	0.20
1M web texts are sampled for each language from the CulturaX dataset due to computation budget constraints.	×	0.15
The diversity of X-Instruction is investigated by parsing English instructions and counting those with a verb-noun struc	✓	0.16
LLaMA-2 models with 7B and 13B parameters are fine-tuned on cross-lingual samples in each language from the X-Instructio	✓	0.29
Alpaca-MT is a baseline where the vanilla Alpaca dataset is translated into 10 languages using Google Translate.	✓	0.18
Bactrian-X is a baseline where LLaMA models are fine-tuned with 3.48M instruction tuning samples in the Bactrian-X datas	✓	0.31
The win rates of different multilingual models on 10 languages are evaluated by GPT-4.	×	0.05

References

- http://arxiv.org/abs/2405.19744v1

- http://arxiv.org/abs/2310.09917v3
- http://arxiv.org/abs/2312.10793v3