WebFAQ Non-English Pre-training for Multilingual Encoder Cross-Lingual Reasoning on XNLI

Assignee Research

June 12, 2026

Abstract

Natural Language Processing systems are heavily dependent on the availability of annotated data to train practical models. Primarily, models are trained on English datasets. In recent times, significant advances have been made in multilingual understanding due to the steeply increasing necessity of working in different languages. One of the points that stands out is that since there are now so many pre-trained multilingual models, we can utilize them for cross-lingual understanding tasks. Using cross-lingual understanding and Natural Language Inference, it is possible to train models whose app

1 Introduction

This paper examines: XNLI 2.0: Improving XNLI dataset and performance on Cross Lingual Understanding (XLU). Research question: Does incorporating WebFAQ's 47 million non-English samples into pre-training improve the cross-lingual reasoning capabilities of multilingual encoders on the XNLI test set relative to English-only pre-training baselines?.

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 7.3/10.

3 Results

14 papers retrieved. 8 claims extracted; 6 independently verified. Quality review score: 7.3/10.

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 XNLI corpus consists of a crowdsourced collection of 5000 test and 2500 development pairs.	✓	0.27
The original XNLI corpus includes data in English, French, Spanish, and German.	×	0.13
The XNLI 2.0 dataset was created by retranslating the MNLI dataset into 14 different languages using Google Translate.	✓	0.25
The recorded accuracy on the original XNLI test dataset is approximately 73% for the model trained in English.	✓	0.31
The model trained in English achieved approximately 76% accuracy on the XNLI 2.0 test set.	✓	0.17
There is a difference in average accuracy in the range of 2.5%-3% between the XNLI and XNLI 2.0 datasets across all 15 l	✓	0.22
The model trained in German has an average accuracy of 78.15% on the new test set.	✓	0.34
Experiments were conducted by training models in all 15 languages available in the dataset.	×	0.11

References

- http://arxiv.org/abs/2603.02208v1
- http://arxiv.org/abs/2504.09645v1
- http://arxiv.org/abs/2301.06527v1