Code-Mixed Pre-Training Enhances Confidence Calibration in Multilingual Entity Recognition

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: Does incorporating code-mixed examples during pre-training improve the calibration of confidence scores for multilingual models on rare entity recognition tasks in the XTREME suite. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: On Verbalized Confidence Scores for LLMs. Research question: Does incorporating code-mixed examples during pre-training improve the calibration of confidence scores for multilingual models on rare entity recognition tasks in the XTREME suite?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

13 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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 evaluation was conducted on 10 datasets, 11 LLMs, and 17 prompts.	×	0.03
The average accuracy over each dataset ranges from 0.5 to 0.9.	×	0.05
The confidence intervals account for the randomness in sampling the 1,000 samples per dataset, but not the randomness in	×	0.04
The accuracy of LLMs increases with increasing model capacity.	×	0.03
Overconfidence is present for LLMs of all sizes.	×	0.03
The ECE values for the benchmark tables are 0.26, 0.21, 0.10, and 0.02.	×	0.01
The evaluation distinguishes between tiny LLMs (gemma1.1-7b, llama3-8b, qwen1.5-7b) and large LLMs (llama3-70b, qwen1.5-	×	0.02
gemma1.1-2b was excluded from the analysis due to issues with confidence scores taken from few-shot examples.	×	0.08
The evaluation aggregates predictions over one or two of the three evaluation dimensions described in Equation (4).	×	0.03
The sampling procedure was repeated across 10 different seeds and the 95% confidence interval for each metric was report	×	0.04
The confidence tends to drop beyond a certain model capacity for some LLM families.	×	0.08
Most of the improvements in calibration come from the increase in accuracy and not the decrease in overconfidence.	×	0.02
The evaluation was conducted on tasks with different difficulties, as assumed in Section 3.2.	×	0.03
The LLMs' confidences remain at a high level leading to a high calibration error, despite decreasing accuracy.	×	0.04
The evaluation used 1,000 samples from each dataset with replacement to mitigate dataset size bias.	×	0.02
The evaluation used 11 models $\times$ 17 methods $\times$ 1,000 samples = 187k predictions per dataset in Figure 3a.	×	0.04

References

- http://arxiv.org/abs/2412.14737v2
- http://arxiv.org/abs/2105.03953v1
- http://arxiv.org/abs/2402.00969v1