LLaMA Model Scale and Adversarial Robustness in Multilingual XTREME Benchmarks

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the correlation between LLaMA model scale (7B to 65B) and robustness against adversarial perturbations in multilingual settings as measured by XTREME performance degradation. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Deep Dive into Adversarial Robustness in Zero-Shot Learning. Research question: What is the correlation between LLaMA model scale (7B to 65B) and robustness against adversarial perturbations in multilingual settings as measured by XTREME performance degradation?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

15 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
The CUB dataset contains 312 attributes, 200 classes, and 11788 images.	×	0.02
The SUN dataset contains 102 attributes, 717 classes, and 14340 images.	×	0.02
The AWA2 dataset contains 85 attributes, 50 classes, and 37322 images.	×	0.02
The standard per-class top-1 accuracy is used for ZSL evaluation.	×	0.06
For GZSL, per-class top-1 accuracy values for seen and unseen classes are used to compute harmonic-scores.	×	0.04
The ResNet-101 feature extractor is merged with the ALE model to make the computational graph end-to-end differentiable.	×	0.02
The feature extractor is frozen and only the ALE model is trained for each dataset.	×	0.02
PyTorch is used for the experiments.	×	0.02
The ALE model is formulated as F(x, y; W) = θ(x)W T φ(y), where θ(x) is the visual and φ(y) is the class embeddings.	×	0.02
The compatibility function F() is parametrized by learnable weights W.	×	0.00
The ALE model is selected because it is one of the earlier studies that showed direct mapping by exploiting data and aux	×	0.03

References

- http://arxiv.org/abs/2104.07412v2
- http://arxiv.org/abs/2008.07651v1

- http://arxiv.org/abs/2007.08428v4