Phonemic-Based NER Robustness in Zero-Shot Multimodal Language Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the robustness of phonemic-based NER generalize to multimodal language models like CLIP or PaLI when tested on zero-shot entity recognition in audio-visual content from low-resource languages. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Multilingual and Multimodal LLMs in the Wild: Building for Low-Resource Languages. Research question: How does the robustness of phonemic-based NER generalize to multimodal language models like CLIP or PaLI when tested on zero-shot entity recognition in audio-visual content from low-resource languages?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.1/10.

3 Results

11 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 4.1/10.

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
FigureQA, CharXiv, ChartQAPro, and DashboardQA are visualization datasets used for reasoning across visual and structure	×	0.03
Multimodal chain-of-thought, ReAct prompting, and structured decoding are reasoning techniques applied to spatial and ta	×	0.03
BLIP-2 Q-Former is an example of an adapter/projector stack architecture for Vision-Language Models (VLMs).	×	0.07
LoRA and QLoRA are Parameter-Efficient Fine-Tuning (PEFT) methods used in practice.	×	0.04
MoME and Uni-MoE are Mixture-of-Experts architectures designed for modality or language specialization.	×	0.03
xGQA, MaRVL, and HaVQA are culture-aware, multilingual benchmarks and diagnostics.	×	0.09
Stress tests for multilingual and multimodal LLMs include evaluations on dialect shifts, noise/occlusion, OCR-heavy inpu	×	0.09
Whisper and Seamless are examples of speech front-end technologies that can be integrated into instruction-tuned LLMs.	×	0.04
Firoj Alam is a Senior Scientist at Qatar Computing Research Institute, HBKU.	×	0.01
Firoj Alam is a senior member of both IEEE and ACM.	×	0.00
Firoj Alam co-organized the BLP-2023 workshop.	×	0.04
Firoj Alam co-organized the GenAI Content Detection shared task at COLING-2025.	×	0.06
Firoj Alam has experience with the CheckThat! Lab at CLEF from 2021 to 2025.	×	0.03

References

- http://arxiv.org/abs/2605.17152v1
- http://arxiv.org/abs/2303.09306v2
- http://arxiv.org/abs/2406.16030v2