Alignment Score Sensitivity of Baichuan 2 and Vicuna-13B in Low-Resource Multimodal Benchmarks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does the alignment score sensitivity of Baichuan 2 and Vicuna-13B vary when evaluated on low-resource language multimodal benchmarks with constrained inference budgets. Multimodal LLMs are evolving from vision-language to tri-modality that see, hear, and read, yet pipelines and benchmarks remain English-centric and compute-heavy. The tutorial offers an overview of this emerging research area for multilingual multimodality across text, speech. 18 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. 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 alignment score sensitivity of Baichuan 2 and Vicuna-13B vary when evaluated on low-resource language multimodal benchmarks with constrained inference budgets?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

8 papers retrieved. 18 claims extracted; 0 independently verified. Quality review score: 4.8/10.

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 tutorial covers reasoning across visual and structured modalities using datasets like FigureQA, CharXiv, ChartQAPro,	×	0.03
Multimodal chain-of-thought, ReAct prompting, and structured decoding are discussed as reasoning techniques for spatial/	×	0.03
Adapter/projector stacks for VLMs (e.g., BLIP-2 Q-Former) and early vs. late fusion are discussed as architectural consi	×	0.03
PEFT techniques like LoRA/QLoRA and quantization for constrained VRAM are covered.	×	0.02
Mixture-of-Experts models like MoME and Uni-MoE are discussed for modality/language specialization.	×	0.03
Speech-centric LLMs are explored, including wiring speech$\rightarrow$text$\rightarrow$LLM for VQA/QA and streaming with VAD/diarization hooks.	×	0.13
Unified speech–text LMMs vs. cascades are compared in terms of deployment trade-offs (latency, robustness, coverage).	×	0.05
Culture-aware, multilingual benchmarks like xGQA, MaRVL, and HaVQA are used for evaluation and benchmarking.	×	0.08
Stress tests include dialect shifts, noise/occlusion, OCR-heavy inputs, and hallucination & grounding checks.	×	0.01
A demo application involves LoRA-tuning a compact multilingual VLM and quick evaluation on a culture-aware slice.	×	0.14
A speech front-end (Whisper/Seamless) is integrated into an instruction-tuned LLM to measure ASR$\rightarrow$task impact.	×	0.02
The tutorial aims to be inclusive and reflect linguistic, cultural, geographic, and disciplinary diversity.	×	0.03
The tutorial centers multilingual, multimodal everyday knowledge and links language, speech, and vision communities.	×	0.14
The tutorial foregrounds low-resource and culturally specific contexts and encourages collaboration across academia, ind 4	×	0.06
The tutorial will be advertised globally with special outreach to under-represented regions and communities.	×	0.02
Firoj Alam is a Senior Scientist at Qatar Computing Research Institute, HBKU, and a senior IEEE and ACM member.	×	0.01
Firoj Alam has co-organized several workshops	×	0.05

References

- http://arxiv.org/abs/2309.10305v4
- http://arxiv.org/abs/2004.07807v2
- http://arxiv.org/abs/2605.17152v1