Zero-Shot Cross-Embodiment Transfer Accuracy of LAP on BridgeData V2

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does LAP's zero-shot cross-embodiment transfer accuracy on BridgeData V2 compare to standard VLA fine-tuning baselines in terms of task success rate. 14 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ToxiFrench: Benchmarking and Enhancing Language Models via CoT Fine-Tuning for French Toxicity Detection. Research question: How does LAP's zero-shot cross-embodiment transfer accuracy on BridgeData V2 compare to standard VLA fine-tuning baselines in terms of task success rate?.

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 3.5/10.

3 Results

8 papers retrieved. 14 claims extracted; 1 independently verified. Quality review score: 3.5/10.

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 ToxiFrench model achieves a 10% increase in balanced accuracy over its baseline.	×	0.12
The ToxiFrench model achieves better performance than GPT-4o and DeepSeek-R1 on the authors' benchmark.	✓	0.20
The full dataset contains less than 5% toxic content.	×	0.08
The ToxiFrench-finetuned model achieved an F1 score of 86% on the example input message shown in Figure 1.	×	0.02
Gemini-2.5-flash achieved an F1 score of 76% on the example input message shown in Figure 1.	×	0.01
Intra-annotator agreement re-annotation of 500 messages yielded a kappa-agreement of 96%.	×	0.02
Inter-annotator agreement re-annotation of 500 messages yielded a kappa-agreement of 81%.	×	0.02
The training set (Strain) contains 52,274 samples with 4% toxicity.	×	0.02
The evaluation and benchmarking set (Sbench) contains 1,388 samples with 50% toxicity.	×	0.04
For Qwen3-4B, accuracy rose from 77% in zero-shot settings to 81% in one-shot settings.	×	0.09
DeepSeek-V3 reached up to 86% accuracy in 4-shot and 10-shot settings.	×	0.06
The best balanced accuracy achieved by the model in Section 4 is 87%.	×	0.05
LLaMA-65B consumes an order of magnitude more energy per generated token than LLaMA-7B.	×	0.00
The introduced dataset contains over 53,000 native French comments.	×	0.06

References

- http://arxiv.org/abs/2508.11281v3
- http://arxiv.org/abs/2602.10556v2
- http://arxiv.org/abs/2404.14700v4