Robustness of Qwen2-72B with TLI Fine-Tuning on Out-of-Distribution Swahili Text in XL-SUM

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the robustness of Qwen2-72B with TLI fine-tuning on out-of-distribution Swahili text (e.g., social media) in XL-SUM compared to Llama3-70B, evaluated via human judgment on fluency and. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Cross-lingual Cross-temporal Summarization: Dataset, Models, Evaluation. Research question: What is the robustness of Qwen2-72B with TLI fine-tuning on out-of-distribution Swahili text (e.g., social media) in XL-SUM compared to Llama3-70B, evaluated via human judgment on fluency and faithfulness?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

12 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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 experiment results are shown in Table 9 for hDe-En (upper) and hEn-De (lower).	×	0.04
Slightly diminishing results from historical text normalization: We observe a decrease in all metrics scores on MemSum p	×	0.07
Norma performs well at token-level normalization, it fails to consider the context of the words, where in our use case s	×	0.02
For hDe-En, intermediate task finetuning with external sources does not improve scores, and sometimes finetuning with mo	×	0.09
For hEn-De, the results show the opposite. We observe higher scores for models trained with more intermediate summarizat	×	0.07
During 5-fold cross-validation, we accumulate the test sets from each fold and in this way, we obtain a final test set c	×	0.02
For example, for direction hDe-En with 328 samples, we accumulate output summaries from each test set consisting of 65-6	×	0.04
In contrast, for zero-shot ChatGPT, we query directly from ChatGPT and collect 328 summaries at once.	×	0.06
We use 5-fold cross-validation to train and test MemSum with our monolingual dataset.	×	0.04
We limit the maximal extracted sentences to 25 per source document.	×	0.03

References

- `http://arxiv.org/abs/1611.04989v2`

- http://arxiv.org/abs/2506.15415v1
- http://arxiv.org/abs/2306.12916v3