ReST-KV Robustness to Attention Redistribution Across Multilingual Benchmarks at Scale

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Does ReST-KV's robustness to attention redistribution generalize to non-English evaluation benchmarks (e.g., BEIR) when scaling context lengths from 128K to 256K. 9 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Qwen2.5 Technical Report. Research question: Does ReST-KV's robustness to attention redistribution generalize to non-English evaluation benchmarks (e.g., BEIR) when scaling context lengths from 128K to 256K?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

4 papers retrieved. 9 claims extracted; 7 independently verified. Quality review score: 7.5/10.

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
Qwen2.5 pre-training datasets were scaled from 7 trillion tokens to 18 trillion tokens.	✓	0.23
Qwen2.5 post-training implements supervised finetuning with over 1 million samples.	✓	0.17
Qwen2.5 post-training utilizes multistage reinforcement learning.	✓	0.15
Qwen2.5 open-weight offerings include base and instruction-tuned models.	✓	0.22
Quantized versions of Qwen2.5 open-weight models are available.	×	0.14
The proprietary Qwen2.5 hosted solutions include two mixture-of-experts (MoE) variants: Qwen2.5-Turbo and Qwen2.5-Plus.	✓	0.22
Qwen2.5-Turbo and Qwen2.5-Plus are available from Alibaba Cloud Model Studio.	✓	0.21
Qwen2.5-72B-Instruct is an open-weight flagship model.	×	0.15
Qwen2.5-72B-Instruct outperforms a number of open and proprietary models on benchmarks.	✓	0.21

References

- https://doi.org/10.48550/arxiv.2307.03170
- https://doi.org/10.48550/arxiv.2412.15115
- https://doi.org/10.48550/arxiv.2311.16867