Reference-Free Factual Consistency Metrics in Long-Context Summarization with Sparse Attention

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How do reference-free factual consistency metrics perform on long-context summaries generated by sparse attention mechanisms like FlowKV compared to full-cache attention across varying document. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Stress Testing Factual Consistency Metrics for Long-Document Summarization. Research question: How do reference-free factual consistency metrics perform on long-context summaries generated by sparse attention mechanisms like FlowKV compared to full-cache attention across varying document lengths?.

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 4.2/10.

3 Results

4 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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 study applies a stress-testing methodology to six widely-used reference-free factual consistency metrics.	✓	0.19
The study evaluates metrics across seven factuality-preserving perturbations.	×	0.14
The seven perturbations include paraphrasing, simplification, synonym replacement, reduced lexical diversity, logically	×	0.14
The experiments were conducted on three benchmark datasets: ScholarQABench, SQuALITY, and LexAbSumm.	×	0.04
The datasets cover science fiction, legal, and scientific domains.	×	0.15
ScholarQABench contains 260 examples used in the study.	×	0.01
SQuALITY contains 351 examples used in the study.	×	0.01
LexAbSumm contains 100 examples used in the study.	×	0.01
The average document length in the LexAbSumm dataset is 14,652 tokens.	×	0.03
All three datasets (ScholarQABench, SQuALITY, LexAbSumm) consist of human-written summaries.	×	0.02
UniEval scores 0.82 on original summaries for the SQuALITY dataset.	×	0.02
UniEval scores drop to 0.39 when source text is added to the summary for the SQuALITY dataset.	×	0.05
MiniCheck shows high robustness to synonym replacement, scoring 0.83 on original and 0.83 on synonym-replaced summaries	×	0.05
BARTScore shows low robustness to synonym replacement, dropping from 0.16 on original to 0.07 on synonym-replaced summar	×	0.05
Most metrics benefit from broader retrieval context windows, though with notable domain-specific variation.	×	0.07
Metric reliability decreases for information-dense claims that overlap semantically with large portions of the source do	×	0.15

References

- http://arxiv.org/abs/2601.15305v1
- http://arxiv.org/abs/2512.07011v1
- http://arxiv.org/abs/2511.07689v2