# FlowKV and Full-Cache Attention in Long-Context Summarization: Performance and Factuality Benchmarks

Assignee Research

June 7, 2026

**Abstract**

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: What is the comparative performance of FlowKV versus full-cache attention on long-context summarization tasks measured by ROUGE-L and factual consistency scores across varying sequence lengths. 15 claims were extracted from source literature; 2 were 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: What is the comparative performance of FlowKV versus full-cache attention on long-context summarization tasks measured by ROUGE-L and factual consistency scores across varying sequence lengths?.

## 2 Methodology

Systematic literature search across multiple databases yielded 9 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

9 papers retrieved. 15 claims extracted; 2 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. | ✓ | 0.17 |
| The study evaluates metrics across seven factuality-preserving perturbations: paraphrasing, simplification, synonym repl | ✓ | 0.20 |
| The experiments were conducted on three benchmark datasets: ScholarQABench, SQuALITY, and LexAbSumm. | × | 0.03 |
| The datasets cover science fiction, legal, and scientific domains. | × | 0.15 |
| ScholarQABench contains 260 examples with an average of 12.5 summary sentences and 6,131 document tokens. | × | 0.03 |
| SQuALITY contains 351 examples with an average of 4.2 summary sentences and 10,840 document tokens. | × | 0.02 |
| LexAbSumm contains 100 examples with an average of 43.2 summary sentences and 14,652 document tokens. | × | 0.02 |
| The six metrics evaluated include BARTScore, MiniCheck, SummaC-Conv, SummaC-ZS, AlignScore, and UniEval. | × | 0.02 |
| For the LexAbSumm dataset, the 'Negated' perturbation resulted in a metric score of 0.560. | × | 0.02 |
| For the ScholarQABench dataset, the 'Negated' perturbation resulted in a metric score of 0.542. | × | 0.02 |
| Under the 'Synonym Replaced' perturbation, AlignScore scored 0.48 on the original metric scale compared to 0.52 for the | × | 0.02 |
| Under the 'Less Diverse' perturbation, UniEval scored 0.39, a decrease from its original score of 0.82. | × | 0.01 |
| The study investigates the impact of retrieval length and evidence dispersion on metric performance. | × | 0.04 |
| 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 | × | 0.14 |

## References


- http://arxiv.org/abs/2602.16843v1
- http://arxiv.org/abs/2511.07689v2
- http://arxiv.org/abs/2210.17378v1