Factuality Metric Robustness in Abstractive Summaries with Optimized Key-Value Cache Eviction

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the robustness of factuality metrics change when applied to abstractive summaries produced by models with optimized key-value cache eviction strategies versus standard full attention. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating the Tradeoff Between Abstractiveness and Factuality in Abstractive Summarization. Research question: How does the robustness of factuality metrics change when applied to abstractive summaries produced by models with optimized key-value cache eviction strategies versus standard full attention?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.4/10.

3 Results

13 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 3.4/10.

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 CNN/DM dataset contains news articles from CNN and DailyMail paired with bullet point summaries.	×	0.06
The XSum dataset contains articles from BBC News, using each article's first sentence as summary.	×	0.04
In Multi-News, each summary is written by a professional editor and paired with a cluster of news articles.	×	0.05
The BART model was pretrained on 160GB of text and gives competitive results on CNN/DM and XSum.	×	0.05
For Multi-News, models were trained with N = 800 and N = 500, called MN-800 and MN-500, respectively.	×	0.03
The MINT score for the test set references for MN-500 is 78.2%, compared to 72.8% for MN-800.	×	0.01
The MINT scores for the CNN/DM and XSum references are 59.6% and 87.8%, respectively.	×	0.05
XSum is the most abstractive dataset among CNN/DM, XSum, and Multi-News.	×	0.08
The CONSTRAINTSFACT dataset contains 17 x 600 human factuality judgements.	×	0.11
Abstractive summaries generated by today's neural models tend to be fluent and well-formed, but lack semantic faithfulne	✓	0.16
Observed rates of factual errors in abstractive summaries have ranged from 30% to over 75%.	×	0.10

References

- http://arxiv.org/abs/2108.02859v2
- http://arxiv.org/abs/2010.08712v2
- http://arxiv.org/abs/2104.13346v2