LongBench Score Consistency Across Sparse and Dense Attention in Llama-3

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does LongBench score consistency vary across different attention mechanisms (e.g., sparse vs. dense) when evaluating Llama-3 under identical few-shot settings and preprocessing. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Dj vu: A Contextualized Temporal Attention Mechanism for Sequential Recommendation. Research question: How does LongBench score consistency vary across different attention mechanisms (e.g., sparse vs. dense) when evaluating Llama-3 under identical few-shot settings and preprocessing?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

15 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
A binary interpretation of time gap as in-session or cross-session is not accurate enough for modeling user behavior.	×	0.03
Time-LSTM assumes that temporal influence takes effect only once during state transition and is fixed regardless of cont	×	0.09
Self-attention mechanisms in sequential recommendation do not explicitly model the ordering of input or segments of hist	×	0.15
The XING dataset contains 64,890 users, 20,662 items, and 1,438,096 actions spanning 80 days.	×	0.02
The UserBehavior dataset contains 68,216 users, 96,438 items, and 4,769,051 actions spanning 9 days.	×	0.02
On the XING dataset, the CTA model achieves a Recall@5 of 0.3217 and an MRR@5 of 0.1849.	×	0.01
On the XING dataset, the CTA model outperforms SASRec, which achieved a Recall@5 of 0.2892 and an MRR@5 of 0.2392.	×	0.02
On the UserBehavior dataset, the CTA model achieves a Recall@5 of 0.1611 and an MRR@5 of 0.0925.	×	0.01
On the UserBehavior dataset, the CTA model outperforms SASRec, which achieved a Recall@5 of 0.1201 and an MRR@5 of 0.079	×	0.02
Increasing the window size (L) from the base configuration to 4, 16, or 32 results in a decrease in Recall@5 performance	×	0.02
Using 4 attention blocks yields a Recall@5 of 0.3217 on the XING dataset, compared to 0.3220 with 1 attention block.	×	0.05
Not sharing embeddings results in a significant performance drop, achieving a Recall@5 of 0.1263 on the XING dataset.	×	0.03
An embedding size of 1000 yields a Recall@5 of 0.3207 on the XING dataset, while a size of 300 yields 0.3147.	×	0.02
The NNL loss function achieves a Recall@5 of 0.3130 on the XING dataset, while the BPR loss function achieves 0.3163.	×	0.04

References

- http://arxiv.org/abs/2601.15305v1
- http://arxiv.org/abs/2002.00741v1
- http://arxiv.org/abs/2510.22389v2