Structured Input Encoding and Sparse Attention in Long-Context Multi-Hop Reasoning

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the impact of structured input encoding in ETC versus sparse attention mechanisms in Reformer on multi-hop reasoning accuracy for long-context question answering. Transformer models have advanced the state of the art in many Natural Language Processing (NLP) tasks. In this paper, we present a new Transformer architecture, Extended Transformer Construction (ETC), that addresses two key challenges of standard Transformer architectures. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ETC: Encoding Long and Structured Inputs in Transformers. Research question: What is the impact of structured input encoding in ETC versus sparse attention mechanisms in Reformer on multi-hop reasoning accuracy for long-context question answering?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

16 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
ETC's performance is comparable to BERT using input length of 512.	×	0.05
BERT's performance is comparable to ETC using input length of 512.	×	0.05
The smaller local radius of ETC (84) puts ETC at a disadvantage with respect to BERT.	×	0.02
ETC's other improvements, such as dynamic whole word masking, seem to compensate for the smaller local radius.	×	0.02
ETC-base with long input length of 4096 tokens, using CPC, hard g2l masking, and separate W Q, W K, and W V matrices for	×	0.11
ETC was tested with the following ablations: shared, no CPC, no hard g2l, and fixed block.	×	0.02
Prior approaches to scale up attention are classified into four categories: sparse attention, recurrence, hierarchical m	×	0.08
Sparse Attention involves limiting each token to attend to a subset of the other tokens.	×	0.09
The Sparse Transformer used predefined attention patterns for both text and image generation.	×	0.05
The Adaptive Attention Span Transformer associates each attention head with a decaying learnable masking function, which	×	0.06

References

- http://arxiv.org/abs/2004.08483v5
- http://arxiv.org/abs/2603.07931v1
- http://arxiv.org/abs/2604.05114v1