Memory Consumption Patterns of ETC, Longformer, and BigBird on HotpotQA Long Sequences

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How do memory consumption patterns differ between ETC, Longformer, and BigBird during training on extended sequence lengths in the HotpotQA benchmark. Transformers-based models, such as BERT, have dramatically improved the performance for various natural language processing tasks. The clinical knowledge enriched model, namely ClinicalBERT, also achieved state-of-the-art results when performed on clinical named entity. 16 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: Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. Research question: How do memory consumption patterns differ between ETC, Longformer, and BigBird during training on extended sequence lengths in the HotpotQA benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 12 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

12 papers retrieved. 16 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
In question answering tasks, Clinical-Longformer and Clinical-BigBird outperformed short-sequence transformer models by	✓	0.20
When evaluated by Exact Match (EM) metric on the emrQA relations subset, Clinical-Longformer and Clinical-BigBird improv	×	0.09
On the emrQA Medication and Heart Disease subsets, Clinical-Longformer and Clinical-BigBird yielded similar Exact Match	×	0.09
In NER tasks, Clinical-Longformer outperformed short-text transformers by more than 2 percent in F1 score across all fou	×	0.06
In NER tasks, Clinical-BigBird performed better than ClinicalBERT and BioBERT in every single experiment across the four	×	0.09
Clinical-Longformer and Clinical-BigBird achieved better results than prior models on the OpenI, MIMIC-AKI, and medNLI d	✓	0.15
BioBERT performed slightly better than Clinical-Longformer and Clinical-BigBird on the OHSUMed dataset.	×	0.09
The maximum sequence lengths for the MedNLI and OpenI datasets are smaller than 512 tokens.	×	0.04
Clinical-Longformer and Clinical-BigBird achieved better results than BERT-like models on MedNLI and OpenI despite these	×	0.11
The i2b2 2014 dataset has the largest averaged sequence length among the four i2b2 NER tasks.	×	0.07
The performance improvement of the proposed models on the i2b2 2014 dataset is almost twice that of the other three i2b2	×	0.03
Clinical-Longformer showed a stronger improvement in F1 score on the emrQA heart disease subset compared to other emrQA	×	0.02
Table 2 reports an F1 score of 0.716 for Clinical-Longformer on the emrQA Medication task.	×	0.03
Table 2 reports an Exact Match (EM) score of 0.911 for Clinical-Longformer on the emrQA Relation task.	×	0.03
Table 3 reports an F1 score of 0.974 for Clinical-Longformer on the i2b2 2006 NER task.	×	0.03
Table 3 reports an F1 score of 0.961 for Clinical-Longformer on the i2b2 2014 NER task.	×	0.03

References

- http://arxiv.org/abs/2211.00974v2
- http://arxiv.org/abs/2004.05150v2
- http://arxiv.org/abs/2201.11838v3