# ETC, Longformer, and BigBird Inference Throughput on HotpotQA Beyond 8K Tokens

Assignee Research

June 1, 2026

**Abstract**

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the inference throughput of ETC compare to Longformer and BigBird on the HotpotQA dataset when sequence lengths exceed 8,000 tokens. 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. 12 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. 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 does the inference throughput of ETC compare to Longformer and BigBird on the HotpotQA dataset when sequence lengths exceed 8,000 tokens?.

## 2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

## 3 Results

10 papers retrieved. 12 claims extracted; 2 independently verified. Quality review score: 4.7/10.

## 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 |
|---|---|---|
| Clinical-Longformer and Clinical-BigBird outperformed the pioneering short-sequence transformer models by 2% across all | ✓ | 0.16 |
| Clinical-Longformer and Clinical-BigBird improved by 5% on the relations subset when evaluated by the stricter EM metric | × | 0.08 |
| Clinical-Longformer consistently led the short-text transformers by more than 2% in all 4 i2b2 datasets for NER tasks. | × | 0.09 |
| Clinical-BigBird performed better than Clinical-BERT and BioBERT in every single NER experiment. | × | 0.07 |
| Clinical-Longformer and Clinical-BigBird achieved better results compared to prior models on OpenI, MIMIC-AKI, and medNL | × | 0.14 |
| BioBERT performed slightly better than Clinical-Longformer and Clinical-BigBird in the OHSUMed dataset. | × | 0.09 |
| Clinical-Longformer and Clinical-BigBird improved the performance of both long and short sequences tasks. | ✓ | 0.17 |
| The maximum sequences of MedNLI and OpenI are smaller than 512 tokens. | × | 0.03 |
| Clinical-Longformer and Clinical-BigBird achieved better results on MedNLI and OpenI despite their sequences being small | × | 0.11 |
| More significant improvement was achieved when applying Clinical-Longformer and Clinical-BigBird to datasets with longer | × | 0.09 |
| The performance improvement on i2b2 2014, which has the largest averaged sequence length, was almost twice that of the o | × | 0.05 |
| Clinical-Longformer more strongly improved the F1 score of the heart disease subset from emrQA. | × | 0.04 |

## References


- http://arxiv.org/abs/2508.06447v2
- http://arxiv.org/abs/2004.05150v2

- http://arxiv.org/abs/2201.11838v3