SOVEREIGN: What are the computational efficiency tradeoffs of sparse attention mechanisms in large-scale language models

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Many real-world applications require the prediction of long sequence time-series, such as electricity consumption planning. Long sequence time-series forecasting (LSTF) demands a high prediction capacity of the model, which is the ability to capture precise long-range dependency coupling between output and input efficiently. Recent studies have shown the potential of Transformer to increase the prediction capacity. However, there are several severe issues with Transformer that prevent it from being directly applicable to LSTF, including quadratic time complexity, high memory usage, and inheren

1 Introduction

Analysis of: Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. Research goal: What are the computational efficiency tradeoffs of sparse attention mechanisms in large-scale language models.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

8 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 8.7/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Long sequence time-series forecasting (LSTF) demands a high prediction capacity of the model, which is the ability to ca	✓	0.47
Recent studies have shown the potential of Transformer to increase the prediction capacity.	✓	0.25
There are several severe issues with Transformer that prevent it from being directly applicable to LSTF, including quadr	✓	0.39
Informer achieves O(L log L) in time complexity and memory usage.	✓	0.23
Informer has comparable performance on sequences' dependency alignment.	✓	0.18
The self-attention distilling in Informer highlights dominating attention by halving cascading layer input, and efficien	✓	0.35
The generative style decoder in Informer predicts the long time-series sequences at one forward operation rather than a	✓	0.40
Extensive experiments on four large-scale datasets demonstrate that Informer significantly outperforms existing methods	✓	0.29

References

- https://doi.org/10.48550/arxiv.2312.00752
- https://doi.org/10.1371/journal.pcbi.1002195
- https://doi.org/10.1609/aaai.v35i12.17325