Novelty Assessment Report

Paper: Critical attention scaling in long-context transformers
PDF URL: https://openreview.net/pdf?id=7SLtElfqCW
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

As large language models scale to longer contexts, attention layers suffer from a fundamental pathology: attention scores collapse toward uniformity as context length N increases, causing tokens to cluster excessively, a phenomenon known as rank-collapse. While $\text{\emph{attention scaling}}$ effectively addresses this deficiency by rescaling attention scores with a polylogarithmic factor β_n, theoretical justification for this approach remains lacking.

We analyze a simplified yet tractable model that magnifies the effect of attention scaling. In this model, attention exhibits a phase transition governed by the scaling factor β_n: insufficient scaling collapses all tokens to a single direction, while excessive scaling reduces attention to identity, thereby eliminating meaningful interactions between tokens. Our main result identifies the critical scaling $\beta_n \asymp \log n$ and provides a rigorous justification for attention scaling in YaRN and Qwen, clarifying why logarithmic scaling maintains sparse, content-adaptive attention at large context lengths.

Core Task Landscape

This paper addresses: **Attention Scaling in Long-Context Transformers**
A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Theoretical Foundations and Analysis**
- **Architecture Design and Attention Mechanisms**
- **Position Encoding Strategies**
- **Systems Optimization and Efficiency**
- **Input Processing and Compression**
- **Domain-Specific Applications**
- **Survey and Comparative Studies**

Complete Taxonomy Tree

- Attention Scaling in Long-Context Transformers Survey Taxonomy
- Theoretical Foundations and Analysis
  - Attention Collapse and Scaling Theory ★ (2 papers)
  - [0] Critical attention scaling in long-context transformers (Anon et al., 2026) View paper
  - [37] Length-induced embedding collapse in transformer-based models (Zhou Yuqi, 2024) View paper
  - Complexity and Approximation Analysis (2 papers)
  - [43] Curse of High Dimensionality Issue in Transformer for Long-context Modeling (Zhang Shu-hai, 2025) View paper
  - [44] HyperAttention: Long-context Attention in Near-Linear Time (Han, 2023) View paper
  - Probabilistic and Bayesian Frameworks (1 papers)
  - [49] Bayesian Attention Mechanism: A Probabilistic Framework for Positional Encoding and Context Length Extrapolation (Barros, 2025) View paper
- Architecture Design and Attention Mechanisms
  - Linear Attention Mechanisms (3 papers)
  - [24] Linrec: Linear attention mechanism for long-term sequential recommender systems (Liu Lang-ming, 2023) View paper
  - [34] Scaling stick-breaking attention: An efficient implementation and in-depth study (Tan, 2024) View paper
  - [45] InAttention: Linear Context Scaling for Transformers (Eisner, 2024) View paper
  - Sparse Attention Mechanisms (5 papers)
  - [2] Longnet: Scaling transformers to 1,000,000,000 tokens (Ding Jia-yu, 2023) View paper
  - [14] The sparse frontier: Sparse attention trade-offs in transformer llms (Nawrot, 2025) View paper
  - [16] Big bird: Transformers for longer sequences (Zaheer, 2020) View paper
  - [25] Longformer: The long-document transformer (Beltagy, 2020) View paper
  - [29] LSG Attention: Extrapolation of pretrained Transformers to long sequences (Condevaux, 2023) View paper
  - Hybrid Attention Architectures (5 papers)
  - [6] LongT5: Efficient text-to-text transformer for long sequences (Xiao-yue, 2022) View paper
  - [11] MoBA: Mixture of Block Attention for Long-Context LLMs (Jiang, 2025) View paper
  - [20] ETC: Encoding long and structured inputs in transformers (Joshua Ainslie, 2020) View paper
  - [21] Every Attention Matters: An Efficient Hybrid Architecture for Long-Context Reasoning (Ling Team, 2025) View paper
  - [41] Hybrid architectures for language models: Systematic analysis and design insights (Bae, 2025) View paper

- Memory-Augmented Attention (4 papers)
  - [1] Leave no context behind: Efficient infinite context transformers with infini-attention (Munkhdalai, 2024) View paper
  - [3] Landmark Attention: Random-Access Infinite Context Length for Transformers (Mohtashami, 2023) View paper
  - [31] Efficient Length-Generalizable Attention via Causal Retrieval for Long-Context Language Modeling (Hu Xiang, 2024) View paper
  - [32] Focused transformer: Contrastive training for context scaling (Tworkowski, 2023) View paper
  - Adaptive and Context-Aware Attention (4 papers)
  - [10] Deep context transformer: bridging efficiency and contextual understanding of transformer models (Shadi Ghaith, 2024) View paper
  - [22] Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference (Luo Yao, 2025) View paper
  - [33] Core Context Aware Transformers for Long Context Language Modeling (Chen, 2024) View paper
  - [46] Learning to Focus: Focal Attention for Selective and Scalable Transformers (Dhananjay Ram, 2025) View paper
  - Block-Based and Hierarchical Attention (4 papers)
  - [8] Ring Attention with Blockwise Transformers for Near-Infinite Context (Liu Hao, 2023) View paper
  - [19] Blockwise parallel transformers for large context models (H Liu, 2023) View paper
  - [27] 2-D Transformer: Extending Large Language Models to Long-Context With Few Memory (Xingyang He, 2025) View paper
  - [35] ScaleFormer: Span Representation Cumulation for Long-Context Transformer (Jiangshu Du, 2025) View paper
- Position Encoding Strategies
  - Positional Encoding-Free Approaches (1 papers)
  - [12] Length generalization of causal transformers without position encoding (Jie Wang, 2024) View paper
  - Rotary and Hybrid Position Encoding (4 papers)
  - [9] Swan-gpt: An efficient and scalable approach for long-context language modeling (Puvvada, 2025) View paper
  - [17] CoCA: Fusing Position Embedding with Collinear Constrained Attention in Transformers for Long Context Window Extending (Jiang Wei, 2024) View paper
  - [23] SWAN: An Efficient and Scalable Approach for Long-Context Language Modeling (Krishna C Puvvada, 2025) View paper
  - [42] Exploring context window of large language models via decomposed positional vectors (Weipeng Chen, 2024) View paper
- Systems Optimization and Efficiency
  - Distributed and Parallel Attention (2 papers)
  - [30] Context parallelism for scalable million-token inference (Yang Amy, 2025) View paper
  - [40] LightSeq: Sequence Level Parallelism for Distributed Training of Long Context Transformers (Li DaCheng, 2023) View paper
  - Hardware-Aware Optimization (3 papers)
  - [7] Efficiently scaling transformer inference (Pope, 2023) View paper
  - [15] Longer Attention Span: Increasing Transformer Context Length with Sparse Graph Processing Techniques (Nathaniel Tomczak, 2025) View paper
  - [18] Lean Attention: Hardware-Aware Scalable Attention Mechanism for the Decode-Phase of Transformers (Bharadwaj, 2024) View paper
  - Attention Quantization and Compression (1 papers)
  - [13] Hamming Attention Distillation: Binarizing Keys and Queries for Efficient Long-Context Transformers (Horton, 2025) View paper
- Input Processing and Compression (2 papers)
  - [28] Leveraging Attention to Effectively Compress Prompts for Long-Context LLMs (Yunlong Zhao, 2025) View paper
  - [47] Scaling In-Context Demonstrations with Structured Attention (Cai, 2023) View paper
- Domain-Specific Applications
  - Biological Sequence Modeling (3 papers)
  - [38] Extending Protein Language Models to a Viral Genomic Scale Using Biologically Induced Sparse Attention (Thibaut Dejean, 2025) View paper
  - [48] Masked Language Modeling for Proteins via Linearly Scalable Long-Context Transformers (Choromanski, 2022) View paper
  - [50] MCWS-Transformers: Towards an Efficient Modeling of Protein Sequences via Multi Context-Window Based Scaled Self-Attention (Ashish Ranjan, 2022) View paper
  - Recommender Systems (1 papers)
  - [39] Longer: Scaling up long sequence modeling in industrial recommenders (Chai Zheng, 2025) View paper
  - Video Generation and Multi-Shot Modeling (1 papers)
  - [26] Long Context Tuning for Video Generation (Guo, 2025) View paper
- Survey and Comparative Studies (3 papers)
  - [4] Efficient attention mechanisms for large language models: A survey (Sun Yutao, 2025) View paper
  - [5] Beyond the limits: A survey of techniques to extend the context length in large language models (Wang Xindi, 2024) View paper
  - [36] X-former Elucidator: Reviving Efficient Attention for Long Context Language Modeling (Andong Li, 2024) View paper

Narrative

Core task: attention scaling in long-context transformers. The field has evolved into a rich landscape organized around several complementary directions. Theoretical Foundations and Analysis examines fundamental phenomena such as attention collapse and scaling behavior, providing the mathematical underpinnings for understanding how transformers behave as context grows. Architecture Design and Attention Mechanisms explores novel attention patterns—ranging from sparse schemes like Big Bird[16] and Longformer[25] to hierarchical approaches such as LongNet[2]—that reduce quadratic complexity. Position Encoding Strategies addresses how models maintain positional awareness over extended sequences, while Systems Optimization and Efficiency tackles practical concerns like memory management and distributed computation, exemplified by Ring Attention[8] and Context Parallelism[30]. Input Processing and Compression investigates methods to condense or selectively retain information, as seen in Landmark Attention[3] and Prompt Compression[28]. Domain-Specific Applications and Survey and Comparative Studies round out the taxonomy by contextualizing these techniques in real-world settings and synthesizing progress across the field, with works like Efficient Attention Survey[4] and Context Extension Survey[5] offering broad perspectives.

Within this landscape, a particularly active line of inquiry centers on understanding and mitigating pathological behaviors that emerge at scale. Critical Attention Scaling[0] sits squarely in the Theoretical Foundations branch alongside Embedding Collapse[37], both investigating how attention distributions degrade or concentrate as context length increases. While Embedding Collapse[37] focuses on representational degradation in embedding spaces, Critical Attention Scaling[0] emphasizes the dynamics of attention weight distributions and their impact on model expressiveness. This theoretical cluster contrasts with more architecture-driven efforts like Infini-

Attention[1] or LongT5[6], which propose new mechanisms to handle long contexts without necessarily dissecting the underlying scaling laws. The interplay between these theoretical insights and architectural innovations remains an open question: understanding when and why attention collapses can inform the design of more robust long-context systems, bridging foundational analysis with practical engineering.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Length-induced embedding collapse in transformer-based models

Authors: Zhou Yuqi, Dai, Sunhao, Yuqi Zhou, Cao Zhanshuo, et al. (11 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Text embeddings from PLM-based models enable a wide range of applications, yet their performance often degrades on longer texts. In this paper, we introduce a phenomenon we call Length Collapse, where embeddings of longer texts tend to cluster together. This clustering results in a distributional inconsistency between the embeddings of short and long texts. We further investigate how these differences contribute to the performance decline observed with longer texts across various downstream task...

Relationship Analysis

Both papers belong to the Attention Collapse and Scaling Theory category, analyzing how attention mechanisms in transformers exhibit collapse phenomena and exploring theoretical justifications for scaling interventions. While the original paper focuses on critical attention scaling factors ($\beta n \asymp \log n$) to prevent rank-collapse as context length increases, examining phase transitions in token clustering behavior, the candidate paper investigates length-induced embedding collapse from a spectral perspective, showing that self-attention acts as a low-pass filter whose strength increases with sequence length. The key difference is that the original paper analyzes attention score scaling to maintain sparse content-adaptive attention, whereas the candidate paper proposes temperature scaling in softmax to preserve high-frequency components and prevent embedding homogenization in long texts.

Contributions Analysis

Overall novelty summary. The paper provides a theoretical analysis of attention scaling in long-context transformers, specifically justifying the logarithmic scaling factor used in models like YaRN and Qwen. It resides in the 'Attention Collapse and Scaling Theory' leaf under 'Theoretical Foundations and Analysis', which contains only two papers including this one. This is a notably sparse research direction within the broader taxonomy of fifty papers, suggesting that rigorous theoretical justification for attention scaling remains an underexplored area despite widespread empirical adoption of these techniques.

The taxonomy reveals that most long-context research concentrates on architecture design (linear attention, sparse patterns, hybrid mechanisms) and systems optimization rather than theoretical foundations. The paper's closest neighbor in its leaf examines embedding collapse from a representational perspective, while nearby branches address complexity analysis and probabilistic frameworks. The work diverges from the dominant empirical trend by providing mathematical grounding for a phenomenon—rank collapse—that practitioners address through heuristic scaling factors, bridging the gap between theoretical understanding and architectural practice.

Among thirty candidates examined across three contributions, none were found to clearly refute the paper's claims. The critical scaling law contribution examined ten candidates with zero refutations, as did the phase transition framework and gradient propagation analysis. This suggests that within the limited search scope, no prior work appears to have rigorously characterized the logarithmic scaling threshold or formalized the phase transition governing attention dynamics. The absence of refutable overlap across all contributions indicates potential novelty, though the search examined a modest candidate pool rather than an exhaustive literature review.

Based on the limited search scope of thirty semantically similar papers, the work appears to occupy a relatively unexplored theoretical niche. The sparse population of its taxonomy leaf and the absence of refuting candidates suggest substantive novelty, though this assessment is constrained by the top-K semantic search methodology and does not constitute comprehensive coverage of all potentially relevant theoretical work in attention mechanisms.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Critical scaling law for attention with logarithmic factor

Description: The authors establish that the critical scaling factor for attention scores is βn proportional to $\log n$, which prevents rank-collapse in long-context transformers. This result provides theoretical justification for empirical methods like YaRN and Qwen that use logarithmic scaling.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Log-Linear Attention

URL: View paper

Brief Assessment

Log-Linear Attention[74] focuses on computational efficiency through logarithmically growing hidden states for sequence modeling, not on theoretical scaling laws for preventing rank-collapse in long-context transformers.

2. Squeezed attention: Accelerating long context length llm inference

URL: View paper

Brief Assessment

Squeezed Attention[71] focuses on accelerating long-context inference through semantic clustering of keys for fixed-context applications, not on theoretical analysis of attention scaling factors or rank-collapse prevention in transformers.

3. The What, Why, and How of Context Length Extension Techniques in Large Language Models--A Detailed Survey

URL: View paper

Brief Assessment

Context Extension Survey[75] focuses on practical techniques for extending context length in LLMs (position interpolation, attention mechanisms, memory augmentation), not on theoretical phase transitions or critical scaling laws for attention mechanisms.

4. SWAN: An Efficient and Scalable Approach for Long-Context Language Modeling

URL: View paper

Brief Assessment

SWAN[23] focuses on architectural design for long-context modeling through hybrid attention mechanisms (sliding-window + global layers), not on theoretical analysis of attention scaling laws or phase transitions in token dynamics.

5. Hierarchical context merging: Better long context understanding for pre-trained llms

URL: View paper

Brief Assessment

Hierarchical Context Merging[76] focuses on a divide-and-conquer approach for handling long contexts through hierarchical merging and token reduction, not on theoretical analysis of attention scaling laws or phase transitions in transformers.

---

6. Gradual Forgetting: Logarithmic Compression for Extending Transformer Context Windows

URL: View paper

Brief Assessment

Gradual Forgetting[78] applies logarithmic compression to input tokens before processing, not to attention score scaling. The original paper analyzes attention mechanism scaling factors (βn ∝ log n), while Gradual Forgetting[78] uses logarithmic temporal filters for input representation.

---

7. â□□ Bench: Extending long context evaluation beyond 100k tokens

URL: View paper

Brief Assessment

Infinity Bench[72] focuses on evaluating LLMs on long-context tasks (100k+ tokens) through benchmark construction, not on theoretical analysis of attention mechanisms or scaling laws. The paper does not address attention score scaling, rank-collapse, or phase transitions in transformers.

---

8. Logarithmic memory networks (lmns): Efficient long-range sequence modeling for resource-constrained environments

URL: View paper

Brief Assessment

Logarithmic Memory Networks[73] focuses on reducing computational complexity through hierarchical memory structures for sequence modeling, not on theoretical analysis of attention score scaling or rank-collapse prevention in transformers.

---

9. Longnet: Scaling transformers to 1,000,000,000 tokens

URL: View paper

Brief Assessment

LongNet[2] focuses on dilated attention for scaling sequence length to billions of tokens, not on theoretical analysis of attention score scaling factors or rank-collapse prevention. The paper does not address critical scaling laws or logarithmic factors in the theoretical sense discussed in the original contribution.

---

10. Towards Efficient Long-Context Natural Language Processing

URL: View paper

Brief Assessment

Efficient Long-Context NLP[77] focuses on efficient text encoding methods (nugget, dodo) for long-context transformers, not on theoretical analysis of attention scaling laws or phase transitions in attention mechanisms.

Contribution 2: Phase transition framework for attention dynamics

Description: The paper introduces a tractable mathematical model demonstrating that attention undergoes a phase transition controlled by βn. Below the critical threshold, tokens collapse to uniformity; above it, attention becomes identity-like, eliminating meaningful token interactions.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Polar Sparsity: High Throughput Batched LLM Inferencing with Scalable Contextual Sparsity

URL: View paper

Brief Assessment

Polar Sparsity[52] focuses on attention head sparsity for batched LLM inference optimization, not on mathematical phase transitions in attention mechanisms controlled by scaling factors. The candidate addresses practical system throughput via selective head activation, while the original analyzes theoretical attention dynamics under βn scaling.

---

2. Weakly supervised change detection using guided anisotropic diffusion

URL: View paper

Brief Assessment

Guided Anisotropic Diffusion[58] focuses on weakly supervised change detection in remote sensing using anisotropic diffusion for label refinement. It does not address phase transitions in attention mechanisms or scaling factors in transformers.

---

3. One Attention, One Scale: Phase-Aligned Rotary Positional Embeddings for Mixed-Resolution Diffusion Transformer

URL: View paper

Brief Assessment

Phase-Aligned Rotary[60] focuses on phase alignment in rotary positional embeddings for mixed-resolution diffusion transformers, not on phase transitions in attention mechanisms controlled by scaling factors βn as studied in the original paper.

---

4. Small-scale proxies for large-scale Transformer training instabilities

URL: View paper

Brief Assessment

Small-scale Proxies[55] focuses on training instabilities in transformers (attention logit growth, output logit divergence) rather than theoretical phase transitions in attention mechanisms controlled by scaling factors like βn.

---

5. A phase transition between positional and semantic learning in a solvable model of dot-product attention

URL: View paper

Brief Assessment

Positional Semantic Transition[53] studies phase transitions between positional and semantic learning mechanisms in attention, not the rank-collapse phenomenon controlled by βn scaling that the original paper addresses.

---

6. Attention to Order: Transformers Discover Phase Transitions via Learnability

URL: View paper

Brief Assessment

Attention to Order[59] studies phase transitions in physical systems (2D Ising model) using transformer learnability as a diagnostic tool, not phase transitions in attention mechanisms themselves. The candidate focuses on detecting thermodynamic phase boundaries through training loss patterns, while the original analyzes how attention scaling parameters control token clustering behavior in language models.

---

7. Lovit: Long video transformer for surgical phase recognition

URL: View paper

Brief Assessment

Lovit[51] focuses on surgical phase recognition using temporal transformers for video analysis, not on mathematical phase transitions in attention mechanisms controlled by scaling factors.

---

8. Dynamical Mean-Field Theory of Self-Attention Neural Networks

URL: View paper

Brief Assessment

Mean-Field Self-Attention[56] studies phase transitions in self-attention neural networks using dynamical mean-field theory from statistical physics, focusing on Hopfield network equivalences and chaotic bifurcations. The original paper examines phase transitions in attention scaling controlled by βn for long-context transformers, addressing rank-collapse through logarithmic scaling factors. These are fundamentally different research directions with distinct mathematical frameworks and objectives.

---

9. Phase Conductor on Multi-layered Attentions for Machine Comprehension

URL: View paper

Brief Assessment

Phase Conductor[57] focuses on multi-layered attention architectures for machine comprehension tasks, not on mathematical phase transitions in attention mechanisms controlled by scaling factors.

---

10. Ultrafast and accurate prediction of polycrystalline hafnium oxide phase-field ferroelectric hysteresis using graph neural networks.

URL: View paper

Brief Assessment

Hafnium Oxide Prediction[54] focuses on predicting ferroelectric hysteresis in polycrystalline materials using graph neural networks for materials science applications, not on attention mechanisms or phase transitions in transformer architectures.

Contribution 3: Gradient propagation analysis across scaling regimes

Description: The authors characterize how the phase transition affects gradient flow during backpropagation. They prove that gradients vanish in the subcritical regime but remain stable in the supercritical regime, connecting forward-pass rank-collapse to backward-pass gradient dynamics.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Geometric dynamics of signal propagation predict trainability of transformers

URL: View paper

Brief Assessment

Geometric Signal Propagation[63] analyzes gradient propagation in transformers at initialization focusing on depth-dependent dynamics and initialization hyperparameters, not on how attention scaling factors affect gradient flow across context-length scaling regimes as studied in the original paper.

---

2. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

---

3. Giadnet: Gradient inspired attention driven denoising network

URL: View paper

Brief Assessment

Giadnet[61] focuses on attention-guided denoising in image processing, not gradient propagation dynamics in transformers across scaling regimes.

---

4. The Mean-Field Dynamics of Transformers

URL: View paper

Brief Assessment

Mean-Field Dynamics[69] focuses on mean-field limits and clustering dynamics of transformers, not on gradient propagation or backpropagation analysis across scaling regimes.

---

5. Mind the gap: a spectral analysis of rank collapse and signal propagation in transformers

URL: View paper

Brief Assessment

Spectral Rank Collapse[65] analyzes gradient propagation in attention-only transformers at initialization using random matrix theory, focusing on spectral gaps and rank collapse. The original paper examines gradient flow during backpropagation across different scaling

regimes (subcritical/supercritical) in the context of attention scaling factors βn, which is a fundamentally different analytical framework and problem setting.

---

6. Attention Retrieves, MLP Memorizes: Disentangling Trainable Components in the Transformer

URL: View paper

Brief Assessment

Attention Retrieves MLP[68] focuses on disentangling trainable components in transformers (attention vs. MLP roles), not on gradient propagation dynamics across scaling regimes in attention mechanisms.

---

7. Mind the Gap: a Spectral Analysis of Rank Collapse and Signal Propagation in Attention Layers

URL: View paper

Brief Assessment

Spectral Analysis Gap[67] focuses on gradient propagation through attention layers via spectral analysis of random matrices at initialization, examining vanishing/exploding gradients. The original paper analyzes gradient flow during backpropagation in relation to forward-pass rank-collapse across different scaling regimes (subcritical/supercritical). These are distinct analytical frameworks addressing different aspects of gradient dynamics.

---

8. Inflection-dependent gradient masking in predictive distribution collapse: A procedural mechanism in large language models

URL: View paper

Brief Assessment

Predictive Distribution Collapse[62] discusses gradient masking and noise scaling in attention mechanisms, but does not address gradient propagation dynamics across different scaling regimes (subcritical vs. supercritical) or connect forward-pass rank-collapse to backward-pass gradient vanishing as characterized in the original work.

---

9. Transformers get stable: An end-to-end signal propagation theory for language models

URL: View paper

Brief Assessment

Stable Transformers[70] analyzes gradient propagation in standard transformers without attention scaling mechanisms, focusing on initialization schemes to prevent vanishing/exploding gradients. The original paper specifically examines how attention scaling factors (βn) affect gradient dynamics across subcritical/supercritical regimes, which is a different technical focus.

---

10. Variable multi-scale attention fusion network and adaptive correcting gradient optimization for multi-task learning

URL: View paper

Brief Assessment

Variable Multi-scale Attention[64] focuses on multi-task learning with attention fusion networks for feature processing, not gradient propagation dynamics in attention mechanisms across scaling regimes.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Critical attention scaling in long-context transformers View paper
- [1] Leave no context behind: Efficient infinite context transformers with infini-attention View paper
- [2] Longnet: Scaling transformers to 1,000,000,000 tokens View paper
- [3] Landmark Attention: Random-Access Infinite Context Length for Transformers View paper
- [4] Efficient attention mechanisms for large language models: A survey View paper
- [5] Beyond the limits: A survey of techniques to extend the context length in large language models View paper
- [6] LongT5: Efficient text-to-text transformer for long sequences View paper
- [7] Efficiently scaling transformer inference View paper
- [8] Ring Attention with Blockwise Transformers for Near-Infinite Context View paper
- [9] Swan-gpt: An efficient and scalable approach for long-context language modeling View paper
- [10] Deep context transformer: bridging efficiency and contextual understanding of transformer models View paper
- [11] MoBA: Mixture of Block Attention for Long-Context LLMs View paper
- [12] Length generalization of causal transformers without position encoding View paper
- [13] Hamming Attention Distillation: Binarizing Keys and Queries for Efficient Long-Context Transformers View paper
- [14] The sparse frontier: Sparse attention trade-offs in transformer llms View paper
- [15] Longer Attention Span: Increasing Transformer Context Length with Sparse Graph Processing Techniques View paper
- [16] Big bird: Transformers for longer sequences View paper
- [17] CoCA: Fusing Position Embedding with Collinear Constrained Attention in Transformers for Long Context Window Extending View paper
- [18] Lean Attention: Hardware-Aware Scalable Attention Mechanism for the Decode-Phase of Transformers View paper
- [19] Blockwise parallel transformers for large context models View paper
- [20] ETC: Encoding long and structured inputs in transformers View paper
- [21] Every Attention Matters: An Efficient Hybrid Architecture for Long-Context Reasoning View paper
- [22] Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference View paper
- [23] SWAN: An Efficient and Scalable Approach for Long-Context Language Modeling View paper
- [24] Linrec: Linear attention mechanism for long-term sequential recommender systems View paper
- [25] Longformer: The long-document transformer View paper
- [26] Long Context Tuning for Video Generation View paper
- [27] 2-D Transformer: Extending Large Language Models to Long-Context With Few Memory View paper
- [28] Leveraging Attention to Effectively Compress Prompts for Long-Context LLMs View paper
- [29] LSG Attention: Extrapolation of pretrained Transformers to long sequences View paper

- [30] Context parallelism for scalable million-token inference View paper
- [31] Efficient Length-Generalizable Attention via Causal Retrieval for Long-Context Language Modeling View paper
- [32] Focused transformer: Contrastive training for context scaling View paper
- [33] Core Context Aware Transformers for Long Context Language Modeling View paper
- [34] Scaling stick-breaking attention: An efficient implementation and in-depth study View paper
- [35] ScaleFormer: Span Representation Cumulation for Long-Context Transformer View paper
- [36] X-former Elucidator: Reviving Efficient Attention for Long Context Language Modeling View paper
- [37] Length-induced embedding collapse in transformer-based models View paper
- [38] Extending Protein Language Models to a Viral Genomic Scale Using Biologically Induced Sparse Attention View paper
- [39] Longer: Scaling up long sequence modeling in industrial recommenders View paper
- [40] LightSeq: Sequence Level Parallelism for Distributed Training of Long Context Transformers View paper
- [41] Hybrid architectures for language models: Systematic analysis and design insights View paper
- [42] Exploring context window of large language models via decomposed positional vectors View paper
- [43] Curse of High Dimensionality Issue in Transformer for Long-context Modeling View paper
- [44] HyperAttention: Long-context Attention in Near-Linear Time View paper
- [45] InAttention: Linear Context Scaling for Transformers View paper
- [46] Learning to Focus: Focal Attention for Selective and Scalable Transformers View paper
- [47] Scaling In-Context Demonstrations with Structured Attention View paper
- [48] Masked Language Modeling for Proteins via Linearly Scalable Long-Context Transformers View paper
- [49] Bayesian Attention Mechanism: A Probabilistic Framework for Positional Encoding and Context Length Extrapolation View paper
- [50] MCWS-Transformers: Towards an Efficient Modeling of Protein Sequences via Multi Context-Window Based Scaled Self-Attention View paper
- [51] Lovit: Long video transformer for surgical phase recognition View paper
- [52] Polar Sparsity: High Throughput Batched LLM Inferencing with Scalable Contextual Sparsity View paper
- [53] A phase transition between positional and semantic learning in a solvable model of dot-product attention View paper
- [54] Ultrafast and accurate prediction of polycrystalline hafnium oxide phase-field ferroelectric hysteresis using graph neural networks. View paper
- [55] Small-scale proxies for large-scale Transformer training instabilities View paper
- [56] Dynamical Mean-Field Theory of Self-Attention Neural Networks View paper
- [57] Phase Conductor on Multi-layered Attentions for Machine Comprehension View paper
- [58] Weakly supervised change detection using guided anisotropic diffusion View paper
- [59] Attention to Order: Transformers Discover Phase Transitions via Learnability View paper
- [60] One Attention, One Scale: Phase-Aligned Rotary Positional Embeddings for Mixed-Resolution Diffusion Transformer View paper
- [61] Giadnet: Gradient inspired attention driven denoising network View paper
- [62] Inflection-dependent gradient masking in predictive distribution collapse: A procedural mechanism in large language models View paper
- [63] Geometric dynamics of signal propagation predict trainability of transformers View paper
- [64] Variable multi-scale attention fusion network and adaptive correcting gradient optimization for multi-task learning View paper
- [65] Mind the gap: a spectral analysis of rank collapse and signal propagation in transformers View paper
- [66] Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation View paper
- [67] Mind the Gap: a Spectral Analysis of Rank Collapse and Signal Propagation in Attention Layers View paper
- [68] Attention Retrieves, MLP Memorizes: Disentangling Trainable Components in the Transformer View paper
- [69] The Mean-Field Dynamics of Transformers View paper
- [70] Transformers get stable: An end-to-end signal propagation theory for language models View paper
- [71] Squeezed attention: Accelerating long context length llm inference View paper
- [72] â□□ Bench: Extending long context evaluation beyond 100k tokens View paper
- [73] Logarithmic memory networks (lmns): Efficient long-range sequence modeling for resource-constrained environments View paper
- [74] Log-Linear Attention View paper
- [75] The What, Why, and How of Context Length Extension Techniques in Large Language Models--A Detailed Survey View paper
- [76] Hierarchical context merging: Better long context understanding for pre-trained llms View paper
- [77] Towards Efficient Long-Context Natural Language Processing View paper
- [78] Gradual Forgetting: Logarithmic Compression for Extending Transformer Context Windows View paper