Novelty Assessment Report

Paper: Quantitative Bounds for Length Generalization in Transformers

PDF URL: https://openreview.net/pdf?id=TLSUIyBIfs

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

We study the problem of length generalization (LG) in transformers: the ability of a model trained on shorter sequences to maintain performance when evaluated on much longer, previously unseen inputs. Prior work by Huang et al. (2024) established that transformers eventually achieve length generalization once the training sequence length exceeds some finite threshold, but left open the question of how large it must be. In this work, we provide the first quantitative bounds on the required training length for length generalization to occur. Motivated by previous empirical and theoretical work, we analyze LG in several distinct problem settings: ℓ_∞ error control vs. average error control over an input distribution, infinite-precision softmax attention vs. finite-precision attention (which reduces to an argmax) in the transformer, as well as for one- or two-layer transformers. In all scenarios, we prove that LG occurs when the internal behavior of the transformer on longer sequences can be ``simulated'' by its behavior on shorter sequences seen during training. Our bounds give qualitative estimates for the required length of training data required for a transformer to generalize, and we verify these insights empirically. These results sharpen our theoretical understanding of the mechanisms underlying extrapolation in transformers, and formalize the intuition that richer training data is required for generalization on more complex tasks.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Length Generalization in Transformers**

A total of **50 papers** were analyzed and organized into a taxonomy with **30 categories**.

Taxonomy Overview

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

- **Theoretical Foundations and Formal Analysis**
- **Positional Encoding Methods**
- **Attention Mechanism Modifications**
- **Training Strategies and Data Augmentation**
- **Algorithmic and Arithmetic Reasoning Tasks**
- **Natural Language Understanding and Generation**
- **Domain-Specific Sequence Modeling**
- **System Optimization and Efficient Implementation**
- **Architectural Surveys and Comparative Studies**
- **Sequence Length as Training Artifact**

Complete Taxonomy Tree

- Length Generalization in Transformers Survey Taxonomy
- Theoretical Foundations and Formal Analysis
  - Identifiability and Learnability Theory ★ (3 papers)
  - [0] Quantitative Bounds for Length Generalization in Transformers (Anon et al., 2026) View paper
  - [8] What algorithms can transformers learn? a study in length generalization (Zhou, 2023) View paper
  - [46] A Formal Framework for Understanding Length Generalization in Transformers (Huang Xinting, 2024) View paper
  - Sparsity and Structural Constraints (1 papers)
  - [9] The Role of Sparsity for Length Generalization in Transformers (Golowich, 2025) View paper
  - Chain-of-Thought and Reasoning Mechanisms (1 papers)
  - [16] Transformers provably learn chain-of-thought reasoning with length generalization (Yu Huang, 2025) View paper
- Positional Encoding Methods
  - Comparative Positional Encoding Analysis (2 papers)
  - [5] Transformers can achieve length generalization but not robustly (Zhou Yongchao, 2024) View paper
  - [20] The Impact of Positional Encoding on Length Generalization in Transformers (Kazemnejad, 2023) View paper
  - Relative and Bias-Based Encodings (2 papers)
  - [7] Length Generalization in Arithmetic Transformers (Jelassi, 2023) View paper
  - [42] Train short, test long: Attention with linear biases enables input length extrapolation (Press, 2021) View paper
  - Conditional and Dynamic Encodings (1 papers)
  - [47] Conditional Positional Encodings for Vision Transformers (Chu, 2021) View paper
  - Position Interpolation and Scaling (1 papers)
  - [24] Length extrapolation of transformers: A survey from the perspective of positional encoding (Liang Zhao, 2024) View paper
- Attention Mechanism Modifications
  - Sparse and Local Attention Patterns (4 papers)
  - [13] The sparse frontier: Sparse attention trade-offs in transformer llms (Nawrot, 2025) View paper

  - [18] Long-Context Generalization with Sparse Attention (Vasylenko, 2025) View paper
  - [33] Star Attention: Efficient LLM Inference over Long Sequences (Acharya, 2024) View paper
  - [44] Longformer: The Long-Document Transformer (Beltagy, 2020) View paper
  - Linear and Low-Rank Attention Approximations (2 papers)
  - [11] Gated Linear Attention Transformers with Hardware-Efficient Training (yang songlin, 2023) View paper
  - [39] Latte: Latent attention for linear time transformers (Maystre, 2024) View paper
  - Dilated and Hierarchical Attention (1 papers)
  - [26] Longnet: Scaling transformers to 1,000,000,000 tokens (Ding Jia-yu, 2023) View paper
  - Hybrid Attention Architectures (2 papers)
  - [40] Swan-gpt: An efficient and scalable approach for long-context language modeling (Puvvada, 2025) View paper
  - [49] Delayed Attention Training Improves Length Generalization in Transformer--RNN Hybrids (Phan, 2025) View paper
- Training Strategies and Data Augmentation
  - Data Format and Representation (1 papers)
  - [2] Universal length generalization with turing programs (Hou, 2024) View paper
  - Training Set Priming and Curriculum Learning (1 papers)
  - [21] Longrecipe: Recipe for efficient long context generalization in large language models (Gu Qing, 2025) View paper
  - Hyperparameter and Configuration Optimization (1 papers)
  - [10] The devil is in the detail: Simple tricks improve systematic generalization of transformers (CsordÃ¡s, 2021) View paper
- Algorithmic and Arithmetic Reasoning Tasks
  - Integer Arithmetic Operations (2 papers)
  - [17] Arithmetic Transformers Can Length-Generalize in Both Operand Length and Count (Cho, 2024) View paper
  - [37] Arbitrary-Length Generalization for Addition in a Tiny Transformer (Patriota, 2024) View paper
  - Logical Reasoning and Theorem Proving (1 papers)
  - [15] Measuring systematic generalization in neural proof generation with transformers (Nicolas Gontier, 2020) View paper
- Natural Language Understanding and Generation
  - Language Modeling and Next-Token Prediction (5 papers)
  - [1] Exploring length generalization in large language models (Anil, 2022) View paper
  - [4] Lm-infinite: Zero-shot extreme length generalization for large language models (Chen Yu, 2024) View paper
  - [6] Lm-infinite: Simple on-the-fly length generalization for large language models (Chi Han, 2023) View paper
  - [22] Long Range Language Modeling via Gated State Spaces (Mehta, 2022) View paper
  - [29] Transformer-XL: Attentive Language Models beyond a Fixed-Length Context (Zihang Dai, 2019) View paper
  - Document Summarization (1 papers)
  - [43] Investigating efficiently extending transformers for long input summarization (Liu Peter, 2023) View paper
  - Task Transfer and Cross-Task Generalization (1 papers)
  - [27] Extrapolation by Association: Length Generalization Transfer in Transformers (Cai Zi-yang, 2025) View paper
- Domain-Specific Sequence Modeling
  - Biological Sequence Analysis (2 papers)
  - [23] GENA-LM: a family of open-source foundational DNA language models for long sequences (Veniamin Fishman, 2024) View paper
  - [45] Evaluation of Coding Schemes for Transformer-based Gene Sequence Modeling (Tian Yuanhe, 2025) View paper
  - Clinical and Medical Text (1 papers)
  - [14] Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences (Li Yikuan, 2022) View paper
  - Time Series Forecasting (3 papers)
  - [30] Skip-Timeformer: Skip-Time Interaction Transformer for Long Sequence Time-Series Forecasting (Wenchang Zhang, 2024) View paper
  - [31] KEDformer: Knowledge extraction seasonal trend decomposition for long-term sequence prediction (Zhenkai Qin, 2024) View paper
  - [34] Timer-XL: Long-Context Transformers for Unified Time Series Forecasting (Liu Yong, 2024) View paper
  - Speech and Audio Processing (1 papers)
  - [3] Exploring Length Generalization For Transformer-based Speech Enhancement (Zhang Qiquan, 2025) View paper
  - Code and Software Artifacts (1 papers)
  - [36] On the Generalizability of Transformer Models to Code Completions of Different Lengths (Nathan Cooper, 2024) View paper
  - Network Traffic and Security (1 papers)
  - [38] Convolutions are Competitive with Transformers for Encrypted Traffic Classification with Pre-training (Lin Chungang, 2025) View paper
  - Educational and Recommender Systems (3 papers)
  - [41] Rethinking and improving student learning and forgetting processes for attention based knowledge tracing models (Youheng Bai, 2025) View paper
  - [48] Longer: Scaling up long sequence modeling in industrial recommenders (Zheng Chai, 2025) View paper
  - [50] Enhancing length generalization for attention based knowledge tracing models with linear biases (Xueyi Li, 2024) View paper
- System Optimization and Efficient Implementation
  - Parallelization and Distributed Training (2 papers)
  - [28] System optimizations for enabling training of extreme long sequence transformer models (Sam Ade Jacobs, 2024) View paper
  - [35] Parallelizing linear transformers with the delta rule over sequence length (Yoon Kim, 2024) View paper
  - Memory Management and Caching (1 papers)
  - [32] Titans: Learning to memorize at test time (Behrouz, 2024) View paper
- Architectural Surveys and Comparative Studies (1 papers)
  - [19] Advancing transformer architecture in long-context large language models: A comprehensive survey (Huang Yunpeng, 2023) View paper
- Sequence Length as Training Artifact (2 papers)
  - [12] A length-extrapolatable transformer (Yutao, 2023) View paper
  - [25] Sequence length is a domain: Length-based overfitting in transformer models (VariÅ¡, 2021) View paper

Narrative

Core task: length generalization in transformers. The field addresses how transformer models can extrapolate to sequence lengths beyond those seen during training, a challenge that spans theoretical analysis, architectural innovation, and practical deployment. The taxonomy organizes research into branches ranging from Theoretical Foundations and Formal Analysis—which examines identifiability, learnability, and the mathematical underpinnings of length extrapolation—to Positional Encoding Methods and Attention Mechanism Modifications that propose concrete architectural changes. Training Strategies and Data Augmentation explore curriculum learning and synthetic data generation, while task-specific branches cover Algorithmic and Arithmetic Reasoning, Natural Language Understanding, and Domain-Specific Sequence Modeling (including genomics and clinical text). System Optimization addresses efficient implementation for long contexts, and surveys provide comparative perspectives across methods. Works like Exploring Length Generalization[1] and Length Extrapolation Survey[24] offer broad overviews, while studies such as Arithmetic Length Generalization[7] and Algorithms Length Study[8] probe specific reasoning domains.

Within the theoretical landscape, a small cluster of works investigates the formal conditions under which transformers can learn to generalize beyond training lengths. Quantitative Length Generalization[0] sits squarely in this identifiability and learnability theory branch, examining the mathematical guarantees and sample complexity bounds that govern length extrapolation. This contrasts with neighboring empirical studies like Algorithms Length Study[8], which focuses on observed performance across algorithmic tasks, and complements formal frameworks such as Formal Length Framework[46], which provides rigorous definitions and proof techniques. While many branches emphasize engineering solutions—positional encodings like ALiBi Linear Biases[42], attention modifications such as Sparse Attention Frontier[13], or training tricks in Simple Tricks Generalization[10]—the theoretical branch to which Quantitative Length Generalization[0] belongs seeks to understand the fundamental limits and necessary conditions for length generalization, offering a principled foundation for the diverse empirical strategies explored elsewhere in the taxonomy.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. What algorithms can transformers learn? a study in length generalization

Authors: Zhou, Hattie, Bradley, Arwen, Littwin, et al. (16 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Large language models exhibit surprising emergent generalization properties, yet also struggle on many simple reasoning tasks such as arithmetic and parity. This raises the question of if and when Transformer models can learn the true algorithm for solving a task. We study the scope of Transformers' abilities in the specific setting of length generalization on algorithmic tasks. Here, we propose a unifying framework to understand when and how Transformers can exhibit strong length generalization...

Relationship Analysis

Both papers belong to the Identifiability and Learnability Theory category, establishing formal conditions for length generalization in transformers. The original paper provides quantitative bounds on the minimum training sequence length required for transformers to generalize to longer sequences, analyzing both finite-precision (hard attention) and infinite-precision (soft attention) settings with explicit complexity measures. The candidate paper introduces the RASP-Generalization Conjecture, proposing that transformers length-generalize when tasks can be solved by short RASP-L programs, focusing on algorithmic task characteristics rather than quantitative training length bounds.

2. A Formal Framework for Understanding Length Generalization in Transformers

Authors: Huang Xinting, Bhattamishra, Satwik, Krebs, Andreas, et al. (10 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

A major challenge for transformers is generalizing to sequences longer than those observed during training. While previous works have empirically shown that transformers can either succeed or fail at length generalization depending on the task, theoretical understanding of this phenomenon remains limited. In this work, we introduce a rigorous theoretical framework to analyze length generalization in causal transformers with learnable absolute positional encodings. In particular, we characterize ...

Relationship Analysis

Both papers belong to the Identifiability and Learnability Theory category, establishing formal conditions for length generalization in transformers. They share the focus on theoretical guarantees for when transformers can learn algorithms that generalize to longer sequences, both introducing 'limit transformer' frameworks and analyzing conditions under which training on shorter sequences enables generalization. The key difference is that the original paper provides quantitative bounds on the minimum training length required for length generalization (scaling with parameters like norms, periodicity, vocabulary size), while the candidate paper focuses on asymptotic identifiability guarantees through an idealized inference procedure and characterizes expressiveness via C-RASP programs without explicit quantitative training length bounds.

Contributions Analysis

Overall novelty summary. The paper provides quantitative bounds on the training sequence length required for transformers to achieve length generalization, analyzing both finite-precision (argmax) and infinite-precision softmax attention across one- and two-layer architectures. It resides in the 'Identifiability and Learnability Theory' leaf, which contains only three papers total, making this a relatively sparse research direction within the broader taxonomy of 50 papers across 30 topics. The sibling papers in this leaf establish formal conditions for learning algorithms that generalize to longer sequences, but the specific focus on quantitative training-length thresholds appears to be a distinguishing characteristic of this work.

The taxonomy reveals that most length generalization research concentrates on architectural modifications (positional encodings, attention mechanisms) and training strategies rather than formal theory. The parent branch 'Theoretical Foundations and Formal Analysis' contains only three leaves with seven papers total, while neighboring branches like 'Positional Encoding Methods' and 'Attention Mechanism Modifications' are substantially more populated. The paper's simulation-based proof technique connects conceptually to work in 'Sparsity and Structural Constraints' and 'Chain-of-Thought and Reasoning Mechanisms', but these neighboring leaves focus on different theoretical aspects—sparse dependencies and multi-step reasoning patterns respectively—rather than training-length bounds.

Among 19 candidates examined across three contributions, only one refutable pair was identified. The first contribution on finite-precision transformers examined five candidates with one potential refutation, suggesting some prior work exists in this specific area. The second contribution on two-layer infinite-precision attention examined four candidates with none refutable, indicating greater novelty in this direction. The third contribution on simulation-based proof techniques examined ten candidates with none refutable, suggesting this methodological approach may be relatively novel within the limited search scope. The analysis explicitly notes this is based on top-K semantic search plus citation expansion, not an exhaustive literature review.

Given the sparse theoretical branch and limited search scope of 19 candidates, the work appears to occupy a relatively underexplored niche within length generalization research. The quantitative bounds for finite-precision attention show some overlap with prior work, while the two-layer infinite-precision analysis and simulation methodology appear more distinctive among the examined candidates. However, the small scale of the literature search means these assessments reflect local rather than comprehensive field coverage.

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

Contribution 1: Quantitative bounds on training length for length generalization in finite-precision transformers

Description: The authors establish explicit upper bounds on the minimum training sequence length N required for single-layer transformers operating at finite precision to generalize to arbitrary-length sequences. These bounds scale with transformer parameter norms, positional embedding periodicity, locality parameter, vocabulary size, and inverse error, covering both worst-case and average-case error control.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Characterizing the Expressivity of Fixed-Precision Transformer Language Models

URL: View paper

Brief Assessment

Fixed-Precision Expressivity[66] focuses on characterizing the expressive power of fixed-precision transformers through formal language theory and logic, not on establishing training length bounds for length generalization.

2. Progress Extrapolating Algorithmic Learning to Arbitrary Sequence Lengths

URL: View paper

Brief Assessment

Algorithmic Sequence Extrapolation[68] focuses on neural network architectures for algorithmic tasks using activation binning and localized memory, not on establishing theoretical bounds for transformer training lengths or finite-precision attention mechanisms.

3. Non-Asymptotic Length Generalization

URL: View paper

Prior Art Analysis

Non-Asymptotic Length[65] demonstrates that similar quantitative bounds on training sequence length for length generalization were established prior to the original paper. Both papers provide explicit upper bounds on minimum training length N required for transformers to generalize to arbitrary-length sequences, with bounds scaling with transformer parameters, positional embeddings, and error tolerance. The candidate paper presents these results in the context of c-rasp functions and provides concrete bounds like $O(t^2)$ for 1-layer and O(t^O(k)) for 2-layer programs, which directly addresses the same problem of establishing 'how long training sequences need to be in order for a transformer to generalize to sequences of arbitrary length.'

Evidence

Evidence 1 - **Rationale**: Both papers explicitly state their goal is to determine minimum training length required for length generalization, establishing that this research question was already being addressed. - **Original**: our goal in this paper is to characterize how long training sequences need to be in order for a transformer to generalize to sequences of arbitrary length. specifically, we adopt the limit transformer formulation from huang et al. (2025), and aim to provide quantitative bounds on the minimum n such ... - **Candidate**: our goal in this paper is to answer the following question: what is the minimum input length required to achieve length generalization as a function of complexity of ground-truth hypothesis, assuming we have infinite computational resources?

Evidence 2 - **Rationale**: Both papers provide explicit quantitative bounds on minimum training length for single-layer transformers/programs to achieve length generalization, with bounds expressed as functions of model parameters and precision. - **Original**: in section 4, we consider limit transformers operating at finite-precision, which matches the setting of huang et al. (2025). this results in a hard attention pattern for sequences of a certain length. our main results are that for one-layer limit transformers, for both worst-case error control (the... - **Candidate**: theorem 5.5 (c-rasp1 length generalization). let f = c-rasp1 and c(f) = max(|a|, |b|, |d|), defined in definition 5.3. then ∀t ∈ n, we have namci (t) ≤ o(t2). that is, the minimum-complexity interpolator, with complexity c and function class f, can length generalize given inputs of length o(t2) when...

Evidence 3 - **Rationale**: Both papers establish quantitative bounds for two-layer transformers/programs, demonstrating that the candidate paper already addressed multi-layer architectures with explicit length complexity bounds. - **Original**: in section 5, we additionally study the setting where the parameters and forward pass are computed at infinite precision. this allows us to establish results independent of the model precision, and is a more suitable model for multi-layer transformers where the inputs to later layers "mix" the first... - **Candidate**: theorem 5.6 (c-rasp2 length generalization, main result) . let f = c-rasp2 and c(f) = t(f)k(f), defined in definition 5.4. then if the ground-truth function f∗ has t(f∗) ≤ t and k(f∗) ≤ k, then the minimum-complexity interpolator, with complexity c and function class f, can length generalize given i...

Evidence 4 - **Rationale**: Both papers formalize the concept of minimum training length (length complexity) with explicit mathematical definitions and bounds, showing the candidate paper established this framework prior to the original. - **Original**: theorem 4.1. there exists an n = o max n 2p/γ, l2∆7|σ|6τ2 ε2 o such that ∥f(x) - g(x)∥ ≤ε for all |x| ≤n implies that ∥f(x) - g(x)∥ = o(ε) for any sequence x. - **Candidate**: definition 4.1 (length complexity of function class) . given a function class f, we define the length complexity of f as the minimum input length that can distinguish any two functions in f: n(f) := min{n ∈ n : ∀f̸= f′ ∈ f, ∃x ∈ {0, 1}≤n s.t. f(x)̸= f′(x)}.

4. Characterizing the expressivity of transformer language models

URL: View paper

Brief Assessment

Expressivity Characterization[64] focuses on characterizing the expressive power of fixed-precision transformers through formal language theory and logic, not on establishing quantitative bounds for training sequence lengths required for length generalization.

5. The Expressivity of Fixed-Precision Transformers without Positional Encoding

URL: View paper

Brief Assessment

Fixed-Precision Without Encoding[67] focuses on expressivity limitations of transformers under fixed precision (recognizing only finite/co-finite languages), not on training length requirements for length generalization.

Contribution 2: Quantitative bounds for two-layer transformers with infinite-precision attention

Description: The authors provide quantitative length generalization bounds for two-layer transformers operating at infinite precision with logarithmically scaled positional embeddings. They introduce a complexity measure and positional margin that govern the minimum training length, establishing results independent of model precision.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Transformer wave function for two dimensional frustrated magnets: Emergence of a spin-liquid phase in the Shastry-Sutherland model

URL: View paper

Brief Assessment

Transformer Wave Function[51] focuses on applying transformer architectures to quantum many-body physics problems (frustrated magnets, Shastry-Sutherland model), not on length generalization bounds or theoretical analysis of transformer attention mechanisms in NLP/ML contexts.

2. Softplus Attention with Re-weighting Boosts Length Extrapolation in Large Language Models

URL: View paper

Brief Assessment

Softplus Attention Reweighting[53] focuses on replacing softmax with softplus activation and introducing a re-weighting mechanism for length extrapolation in LLMs. It does not address quantitative bounds for length generalization in two-layer transformers with infinite precision attention or provide theoretical analysis of training length requirements.

3. Depth Extrapolation of Decoders Trained on Nested Structures

URL: View paper

Brief Assessment

Depth Extrapolation Decoders[54] focuses on depth extrapolation in nested structures (Dyck languages, boolean expressions) rather than length generalization bounds. The candidate studies single-layer models with specific constructions for parentheses completion, not general two-layer transformers with quantitative training length bounds.

4. Representations and computations in transformers that support generalization on structured tasks

URL: View paper

Brief Assessment

Structured Task Representations[52] focuses on analyzing learned representations and attention patterns in transformers solving algorithmic tasks, not on establishing theoretical length generalization bounds or complexity measures for infinite-precision attention.

Contribution 3: Simulation-based proof technique for length generalization

Description: The authors develop a unified proof technique showing that length generalization occurs when the internal behavior of a transformer on longer sequences can be simulated by its behavior on shorter training sequences. This involves constructing shorter strings that approximately preserve sufficient statistics necessary for computing the forward pass.

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. MarketGPT: Developing a Pre-trained transformer (GPT) for Modeling Financial Time Series

URL: View paper

Brief Assessment

MarketGPT[56] focuses on financial time series modeling using GPT for limit order book dynamics, not on theoretical proof techniques for transformer length generalization in formal language tasks.

2. HairFormer: Transformer-Based Dynamic Neural Hair Simulation

URL: View paper

Brief Assessment

HairFormer[55] focuses on transformer-based neural hair simulation for graphics applications, not on theoretical proof techniques for length generalization in transformers or sequence models.

3. Randomized Positional Encodings Boost Length Generalization of Transformers

URL: View paper

Brief Assessment

Randomized Positional Encodings[59] focuses on a practical positional encoding scheme to improve length generalization empirically, not on theoretical proof techniques involving simulation arguments for transformers.

4. Predicting Future Kinetic States of Physicochemical Systems Using Generative Pre-trained Transformer

URL: View paper

Brief Assessment

Kinetic States GPT[61] focuses on predicting future states of physicochemical systems using transformers for molecular dynamics, not on theoretical proof techniques for length generalization in transformers.

5. Music Transformer

URL: View paper

Brief Assessment

Music Transformer[63] focuses on music generation using relative positional attention mechanisms in transformers, not on theoretical proof techniques for length generalization or simulation-based approaches for analyzing transformer behavior on sequences of varying lengths.

6. Dateformer: Time-modeling Transformer for Longer-term Series Forecasting

URL: View paper

Brief Assessment

Dateformer[62] focuses on time series forecasting using patch-wise processing and time representations, not on theoretical proof techniques for transformer length generalization in formal language tasks.

7. Efficacy of Temporal Fusion Transformers for Runoff Simulation

URL: View paper

Brief Assessment

Temporal Fusion Runoff[57] focuses on rainfall-runoff modeling using Temporal Fusion Transformers for hydrological predictions, not on theoretical proof techniques for transformer length generalization.

8. A Three-Dimensional Multiscale Finite Element Framework for Nonlinear Composite Materials Based on Deep Learning

URL: View paper

Brief Assessment

Multiscale Finite Element[60] focuses on computational mechanics for composite materials using neural networks (GRU/Transformer) as surrogate models in finite element simulations, not on theoretical proof techniques for transformer length generalization in NLP tasks.

9. What algorithms can transformers learn? a study in length generalization

URL: View paper

Brief Assessment

Algorithms Length Study[8] focuses on predicting which algorithmic tasks transformers can learn via the RASP-L programming language framework, not on developing simulation-based proof techniques for length generalization bounds.

10. Maritime target intent recognition based on Transformer

URL: View paper

Brief Assessment

Maritime Intent Recognition[58] focuses on maritime target intent recognition using transformers for long-sequence data in maritime security applications, not on theoretical proof techniques for length generalization in transformers.

Appendix: Text Similarity Detection

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

References

- [0] Quantitative Bounds for Length Generalization in Transformers View paper
- [1] Exploring length generalization in large language models View paper
- [2] Universal length generalization with turing programs View paper
- [3] Exploring Length Generalization For Transformer-based Speech Enhancement View paper
- [4] Lm-infinite: Zero-shot extreme length generalization for large language models View paper
- [5] Transformers can achieve length generalization but not robustly View paper
- [6] Lm-infinite: Simple on-the-fly length generalization for large language models View paper
- [7] Length Generalization in Arithmetic Transformers View paper
- [8] What algorithms can transformers learn? a study in length generalization View paper
- [9] The Role of Sparsity for Length Generalization in Transformers View paper
- [10] The devil is in the detail: Simple tricks improve systematic generalization of transformers View paper
- [11] Gated Linear Attention Transformers with Hardware-Efficient Training View paper
- [12] A length-extrapolatable transformer View paper
- [13] The sparse frontier: Sparse attention trade-offs in transformer llms View paper
- [14] Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences View paper
- [15] Measuring systematic generalization in neural proof generation with transformers View paper
- [16] Transformers provably learn chain-of-thought reasoning with length generalization View paper
- [17] Arithmetic Transformers Can Length-Generalize in Both Operand Length and Count View paper
- [18] Long-Context Generalization with Sparse Attention View paper
- [19] Advancing transformer architecture in long-context large language models: A comprehensive survey View paper
- [20] The Impact of Positional Encoding on Length Generalization in Transformers View paper
- [21] Longrecipe: Recipe for efficient long context generalization in large language models View paper
- [22] Long Range Language Modeling via Gated State Spaces View paper
- [23] GENA-LM: a family of open-source foundational DNA language models for long sequences View paper
- [24] Length extrapolation of transformers: A survey from the perspective of positional encoding View paper
- [25] Sequence length is a domain: Length-based overfitting in transformer models View paper
- [26] Longnet: Scaling transformers to 1,000,000,000 tokens View paper
- [27] Extrapolation by Association: Length Generalization Transfer in Transformers View paper
- [28] System optimizations for enabling training of extreme long sequence transformer models View paper
- [29] Transformer-XL: Attentive Language Models beyond a Fixed-Length Context View paper
- [30] Skip-Timeformer: Skip-Time Interaction Transformer for Long Sequence Time-Series Forecasting View paper
- [31] KEDformer: Knowledge extraction seasonal trend decomposition for long-term sequence prediction View paper
- [32] Titans: Learning to memorize at test time View paper
- [33] Star Attention: Efficient LLM Inference over Long Sequences View paper
- [34] Timer-XL: Long-Context Transformers for Unified Time Series Forecasting View paper
- [35] Parallelizing linear transformers with the delta rule over sequence length View paper
- [36] On the Generalizability of Transformer Models to Code Completions of Different Lengths View paper
- [37] Arbitrary-Length Generalization for Addition in a Tiny Transformer View paper
- [38] Convolutions are Competitive with Transformers for Encrypted Traffic Classification with Pre-training View paper
- [39] Latte: Latent attention for linear time transformers View paper
- [40] Swan-gpt: An efficient and scalable approach for long-context language modeling View paper
- [41] Rethinking and improving student learning and forgetting processes for attention based knowledge tracing models View paper
- [42] Train short, test long: Attention with linear biases enables input length extrapolation View paper
- [43] Investigating efficiently extending transformers for long input summarization View paper
- [44] Longformer: The Long-Document Transformer View paper
- [45] Evaluation of Coding Schemes for Transformer-based Gene Sequence Modeling View paper

- [46] A Formal Framework for Understanding Length Generalization in Transformers View paper
- [47] Conditional Positional Encodings for Vision Transformers View paper
- [48] Longer: Scaling up long sequence modeling in industrial recommenders View paper
- [49] Delayed Attention Training Improves Length Generalization in Transformer--RNN Hybrids View paper
- [50] Enhancing length generalization for attention based knowledge tracing models with linear biases View paper
- [51] Transformer wave function for two dimensional frustrated magnets: Emergence of a spin-liquid phase in the Shastry-Sutherland model View paper
- [52] Representations and computations in transformers that support generalization on structured tasks View paper
- [53] Softplus Attention with Re-weighting Boosts Length Extrapolation in Large Language Models View paper
- [54] Depth Extrapolation of Decoders Trained on Nested Structures View paper
- [55] HairFormer: Transformer-Based Dynamic Neural Hair Simulation View paper
- [56] MarketGPT: Developing a Pre-trained transformer (GPT) for Modeling Financial Time Series View paper
- [57] Efficacy of Temporal Fusion Transformers for Runoff Simulation View paper
- [58] Maritime target intent recognition based on Transformer View paper
- [59] Randomized Positional Encodings Boost Length Generalization of Transformers View paper
- [60] A Three-Dimensional Multiscale Finite Element Framework for Nonlinear Composite Materials Based on Deep Learning View paper
- [61] Predicting Future Kinetic States of Physicochemical Systems Using Generative Pre-trained Transformer View paper
- [62] Dateformer: Time-modeling Transformer for Longer-term Series Forecasting View paper
- [63] Music Transformer View paper
- [64] Characterizing the expressivity of transformer language models View paper
- [65] Non-Asymptotic Length Generalization View paper
- [66] Characterizing the Expressivity of Fixed-Precision Transformer Language Models View paper
- [67] The Expressivity of Fixed-Precision Transformers without Positional Encoding View paper
- [68] Progress Extrapolating Algorithmic Learning to Arbitrary Sequence Lengths View paper