Novelty Assessment Report

Paper: Efficient Turing Machine Simulation with Transformers
PDF URL: https://openreview.net/pdf?id=bxVuILo1xx
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Constant bit-size Transformers are known to be Turing complete, but existing constructions require $\Omega(s(n))$ chain-of-thought (CoT) steps per simulated Turing machine (TM) step, leading to impractical reasoning lengths. In this paper, we significantly reduce this efficiency gap by proving that any $(t(n),s(n))$-bounded multi-tape TM can be simulated by a constant bit-size Transformer with an optimal $O(s(n))$-long context window and only $O(s(n)^c)$ CoT steps per TM step, where $c>0$ can be made arbitrarily small by letting the Transformers' head-layer product sufficiently large. In addition, our construction shows that sparse attention with fixed geometric offsets suffices for efficient universal computation. Our proof leverages multi-queue TMs as a bridge. The main technical novelty is a more efficient simulation of multi-tape TMs by synchronous multi-queue TMs, improving both time and space complexity under stricter model assumptions.

Core Task Landscape

This paper addresses: **Turing Machine Simulation with Transformers**

A total of **36 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 Turing Completeness Proofs**
- **Efficient Simulation Constructions**
- **Memory-Augmented Computational Models**
- **Practical Programming Approaches**
- **Learnability and Generalization**
- **Alternative Computational Paradigms**
- **Specialized Applications and Extensions**

Complete Taxonomy Tree

- Turing Machine Simulation with Transformers Survey Taxonomy
- Theoretical Turing Completeness Proofs
  - Hard Attention and Constant Bit-Size Models ★ (3 papers)
  - [0] Efficient Turing Machine Simulation with Transformers (Anon et al., 2026) View paper
  - [2] Constant Bit-size Transformers Are Turing Complete (Li Qian, 2025) View paper
  - [3] Attention is turing-complete (Jorge Perez, 2021) View paper
  - Softmax Attention Completeness (1 papers)
  - [15] Softmax Transformers are Turing-Complete (Hongjian Jiang, 2025) View paper
  - Autoregressive and Decoder-Only Architectures (2 papers)
  - [16] Autoregressive Large Language Models are Computationally Universal (Schuurmans, 2024) View paper
  - [17] How powerful are decoder-only transformer neural models? (Jesse Roberts, 2024) View paper
  - General Computational Power Analysis (2 papers)
  - [18] On the computational power of transformers and its implications in sequence modeling (Bhattamishra, 2020) View paper
  - [22] On the turing completeness of modern neural network architectures (Jorge Eduardo PÃ©rez PÃ©rez, 2019) View paper
- Efficient Simulation Constructions
  - Universal Transformer Variants (2 papers)
  - [7] Sparse universal transformer (Shawn Tan, 2023) View paper
  - [8] Universal transformers (Dehghani, 2018) View paper
  - Solomonoff Induction and Probabilistic Models (3 papers)
  - [1] Transformers as Approximations of Solomonoff Induction (Nathan Young, 2024) View paper
  - [14] Transformers Have the Potential to Achieve AGI (Q Li, 2025) View paper
  - [36] Neural Networks and Solomonoff Induction (J Grau-Moya, n.d.) View paper
- Memory-Augmented Computational Models
  - Neural Turing Machine Extensions (4 papers)
  - [13] Neural Attention Memory (Nam, 2023) View paper
  - [29] Token Turing Machines (Michael S. Ryoo, 2022) View paper
  - [30] Emergent neural turing machine and its visual navigation. (Zejia Zheng, 2019) View paper
  - [33] A provably stable neural network Turing Machine (Stogin, 2020) View paper
  - Attention-Based Memory Architectures (1 papers)

- [21] Memory Augmented Large Language Models are Computationally Universal (Schuurmans, 2023) View paper
- Practical Programming Approaches
  - Prompting-Based Execution (3 papers)
  - [6] Gpt is becoming a turing machine: Here are some ways to program it (Jojic, 2023) View paper
  - [12] Ask, and it shall be given: On the Turing completeness of prompting (Qiu, 2024) View paper
  - [27] Ask, and it shall be given: Turing completeness of prompting (Qiu, 2024) View paper
  - Fine-Tuning for Algorithmic Execution (1 papers)
  - [11] Executing arithmetic: Fine-tuning large language models as turing machines (Lai, 2024) View paper
  - Chain-of-Thought Decomposition Strategies (2 papers)
  - [9] Universal length generalization with turing programs (Hou, 2024) View paper
  - [23] The Imitation Game: Turing Machine Imitator is Length Generalizable Reasoner (Zhang, 2025) View paper
- Learnability and Generalization
  - PAC Learning and Approximation Theory (2 papers)
  - [25] Learning Linear Attention in Polynomial Time (Morris Yau, 2024) View paper
  - [34] Statistically Meaningful Approximation: a Theoretical Analysis for Approximating Turing Machines with Transformers (C Wei, n.d.) View paper
  - Empirical Length and Compositional Generalization (1 papers)
  - [10] Rule extrapolation in language modeling: A study of compositional generalization on OOD prompts (Wieland Brendel, 2024) View paper
  - Shortcut Learning and Recurrent Dynamics (1 papers)
  - [31] Transformers Learn Shortcuts to Automata (Liu Bing-bin, 2022) View paper
- Alternative Computational Paradigms
  - Multi-Transformer and Looped Architectures (2 papers)
  - [28] Looped ReLU MLPs May Be All You Need as Practical Programmable Computers (Liang, 2024) View paper
  - [32] Turing Complete Transformers: Two Transformers Are More Powerful Than One (SK Upadhyay, n.d.) View paper
  - Transfinite and Novel Computational Models (1 papers)
  - [24] Infinite Time Turing Machines and their Applications (Weerawarana, 2025) View paper
- Specialized Applications and Extensions
  - Memory Management and Context Optimization (2 papers)
  - [4] PENCIL: Long Thoughts with Short Memory (Yang Chenxiao, 2025) View paper
  - [19] Momentary Contexts: A Memory and Retrieval Approach for LLM Efficiency (Jaepil, 2024) View paper
  - Domain-Specific Computational Tasks (2 papers)
  - [20] Time series attention based transformer neural Turing machines for diachronic graph embedding in cyber threat intelligence (Binghua Song, 2022) View paper
  - [26] Harnessing Chaos: Predicting the Unpredictable in Cellular Automata (Rodrigo Garrido, 2025) View paper
  - Historical and Conceptual Overviews (2 papers)
  - [5] From turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models (Adrian David Cheok, 2023) View paper
  - [35] Theory for Understanding Transformers: An Overview of Current Research (H Wong, n.d.) View paper

Narrative

Core task: Turing machine simulation with Transformers. The field explores whether and how Transformer architectures can achieve universal computation, organizing itself into several complementary branches. Theoretical Turing Completeness Proofs establish formal conditions under which Transformers can simulate arbitrary Turing machines, often examining variants with hard attention or restricted precision. Efficient Simulation Constructions focus on practical encoding schemes and resource bounds for implementing computation within Transformer layers. Memory-Augmented Computational Models extend basic architectures with external memory mechanisms to enhance computational expressiveness, while Practical Programming Approaches investigate how to leverage existing models for algorithmic tasks through prompting or fine-tuning. Learnability and Generalization studies address whether networks can learn to execute algorithms from data and generalize beyond training distributions, and Alternative Computational Paradigms examine non-standard architectures or computation models. Specialized Applications and Extensions apply these insights to domains like time series or program synthesis.

Within the theoretical landscape, a particularly active line of work examines the minimal architectural requirements for Turing completeness. Attention Turing Complete[3] demonstrated that attention mechanisms alone suffice for universal computation, while Constant Bitsize Completeness[2] refined this by showing completeness even with severely restricted precision. Efficient Turing Simulation[0] sits squarely in this branch, contributing to the understanding of hard attention models with constant bit-size constraints—a setting that balances theoretical elegance with practical relevance. This contrasts with works like Universal Transformers[8] or Sparse Universal Transformer[7], which achieve universality through recurrent depth and architectural modifications rather than minimal precision assumptions. Meanwhile, studies such as Turing to Transformers[5] and Computational Power Transformers[18] provide broader surveys of expressiveness across variants, situating these hard-attention results within the wider spectrum of Transformer computational power.

Related Works in Same Category

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

1. Constant Bit-size Transformers Are Turing Complete

Authors: Li Qian, Wang Yu-yi, Qian Li, Yuyi Wang | **Year/Venue**: 2025 | **URL**: View paper

Abstract

We prove that any Turing machine running on inputs of arbitrary length can be simulated by a constant bit-size transformer, as long as the context window is sufficiently long. This improves previous works, which require scaling up either the model's precision or the number of parameters on longer inputs. Furthermore, we prove that the complexity class SPACE$[s(n)]$ exactly characterizes the expressive power of a constant bit-size transformer with a context window of length $s(n)$. Our approach r...

Relationship Analysis

Both papers belong to the Hard Attention and Constant Bit-Size Models category, proving Turing completeness of transformers with constant precision and bounded context windows. They overlap in establishing that constant bit-size transformers can simulate Turing machines with context windows scaling with space complexity $O(s(n))$ rather than time complexity. The key difference is that the original paper achieves $O(s(n)^c)$ CoT steps per TM step (where c can be made arbitrarily small) using sparse geometric-offset attention and multi-queue TMs as an intermediate bridge, while the candidate paper requires O(s(n)) CoT steps per TM step using Post machines (single-queue automata) as the bridge and fixed-offset relative positional encoding.

---

2. Attention is turing-complete

Authors: Jorge Perez, Pablo BarcelÃ³, Javier MarinkoviÄ | **Year/Venue**: 2021 | **URL**: View paper

Abstract

â¦ the Transformer with hard-attention is Turing complete exclusively based on their capacity to compute â¦ is based on a direct simulation of a Turing machine which we believe to be quite â¦

Relationship Analysis

Both papers belong to the Hard Attention and Constant Bit-Size Models category, establishing Turing completeness through hard attention mechanisms with bounded context. The original paper focuses on efficient simulation with constant bit-size Transformers achieving O(s(n)^c) CoT steps per TM step using sparse geometric-offset attention, while the candidate paper (Pérez et al.) proves Turing completeness of standard Transformers with hard attention and positional encodings, achieving O(t(n)) total CoT length but without the constant bit-size constraint or the efficiency optimizations central to the original work.

Contributions Analysis

Overall novelty summary. The paper contributes an efficient construction for simulating multi-tape Turing machines using constant bit-size Transformers, achieving O(s(n)^c) chain-of-thought steps per TM step where c can be made arbitrarily small. It resides in the 'Hard Attention and Constant Bit-Size Models' leaf, which contains only three papers total including this work. This represents a relatively sparse research direction within the broader taxonomy of 36 papers, suggesting the paper addresses a specialized theoretical question rather than a crowded subfield. The focus on minimal architectural assumptions (constant precision, hard attention) distinguishes this line from more applied or architecturally complex approaches.

The taxonomy reveals several neighboring research directions that contextualize this work. The sibling leaf 'Softmax Attention Completeness' explores similar universality questions but under different attention mechanisms, while 'Efficient Simulation Constructions' branches into Universal Transformer variants and probabilistic models that prioritize practical efficiency over minimal assumptions. The 'Memory-Augmented Computational Models' branch takes an orthogonal approach by adding external memory rather than proving sufficiency of base architectures. The paper's emphasis on geometric-offset sparse attention connects it conceptually to efficiency-focused work, though its theoretical framing keeps it firmly within the completeness-proof tradition rather than practical implementation studies.

Among nine candidates examined, one appears to provide overlapping prior work for the first contribution on efficient CoT overhead, while the second and third contributions (sparse attention sufficiency and improved multi-queue simulation) were not examined against any candidates. This limited search scope means the novelty assessment is necessarily provisional. The first contribution's apparent overlap suggests that efficiency improvements in TM simulation may have been explored previously, though the specific combination of constant bit-size constraints and near-optimal CoT bounds may still offer incremental refinement. The unexamined contributions remain uncertain in their novelty given the absence of comparative analysis.

Based on the top-9 semantic matches examined, the work appears to make incremental theoretical contributions within a specialized niche. The taxonomy structure indicates this is not a heavily populated research area, which could mean either genuine novelty or limited community interest in this particular formulation. The analysis does not cover broader literature on computational complexity or alternative simulation frameworks that might provide additional context. A more comprehensive search would be needed to definitively assess whether the technical innovations (synchronous multi-queue TMs, geometric-offset attention) represent substantial advances over existing theoretical machinery.

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

Contribution 1: Efficient Turing machine simulation with near-optimal CoT overhead

Description: The authors prove that constant bit-size Transformers can simulate multi-tape Turing machines with optimal space (context window O(s(n))) and significantly reduced per-step overhead (O(s(n)^c) CoT steps instead of $\Omega(s(n))$), where the exponent c can be made arbitrarily small by increasing the head-layer product.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Artificial Intelligence and the Computational Power of Neural Networks

URL: View paper

Brief Assessment

Computational Power Networks[39] appears to discuss general computational constraints and reasoning abilities of transformers, but the provided context is too fragmentary to assess whether it addresses the specific efficiency improvements (O(s(n)^c) CoT overhead with arbitrarily small c) claimed in the original paper.

---

2. PENCIL: Long Thoughts with Short Memory

URL: View paper

Prior Art Analysis

PENCIL Long Thoughts[4] demonstrates that prior work exists which achieves efficient Turing machine simulation with optimal complexity bounds. The candidate paper proves that their PENCIL mechanism can 'perform universal efficient computation by simulating any turing machines with optimal time and space complexity', which directly addresses the same problem of efficient TM simulation that the original paper claims as novel. Both papers focus on reducing the overhead of chain-of-thought reasoning while maintaining optimal space complexity, with the candidate achieving this through a reduction mechanism that 'recursively cleans up intermediate thoughts' rather than through multi-queue TMs.

Evidence

Evidence 1 - **Rationale**: Both papers claim to achieve efficient Turing machine simulation with optimal complexity. The candidate explicitly states they prove PENCIL achieves 'optimal time and space complexity' for simulating 'any turing machines', which directly overlaps with the original's claim of achieving 'optimal o(s(n))-long context window' with reduced CoT overhead. - **Original**: constant bit-size transformers are known to be turing complete, but existing constructions require $\omega(s(n))$ chain-of-thought (cot) steps per simulated turing machine (tm) step, leading to impractical reasoning lengths. in this paper, we significantly reduce this efficiency gap by proving that any (t(... - **Candidate**: theoretically, we prove pencil can perform universal efficient computation by simulating any turing machines with optimal time and space complexity, and thus can solve arbitrary computable tasks that are otherwise intractable for vanilla cot.

Evidence 2 - **Rationale**: Both papers identify the same fundamental problem: existing CoT approaches have inefficient memory usage that leads to impractically long reasoning chains. The candidate's observation about 'intermediate computations accumulate indefinitely' directly corresponds to the original's concern about 's(n)-factor slowdown' causing 'total cot length to far exceed' practical needs. -

Original: the constant bit-size construction of li & wang (2025) achieves an optimal context window of length o(s(n)), which reflects the amount of memory required for reasoning, but incurs a cost of ω(s(n)) cot steps to simulate one turing machine (tm) step. here s(n) denotes the space bound of the simulated... - **Candidate**: while state-of-the-art llms have demonstrated great promise of using long chains-of-thought (cot) to boost reasoning, scaling it up to more challenging problems at test-time is fundamentally limited by suboptimal memory usage -- intermediate computations accumulate indefinitely in context even when ...

Evidence 3 - **Rationale**: Both papers claim to solve the efficiency problem through different mechanisms. The original reduces overhead to s(n)^c through multi-queue TMs, while the candidate achieves efficiency through a 'reduction mechanism that recursively cleans up intermediate thoughts'. Both claim to enable solving harder problems with reduced computational overhead. - **Original**: we make significant progress on the open efficiency problem by reducing the per-step slowdown from s(n) to s(n)c, where the exponent c >0 can be made arbitrarily small. - **Candidate**: we introduce pencil, which incorporates a novel reduction mechanism into the autoregressive generation process that recursively cleans up intermediate thoughts based on patterns learned from training. by iteratively generating and erasing thoughts, pencil can think deeper to solve harder problems us...

3. Token Turing Machines

URL: View paper

Brief Assessment

Token Turing Machines[29] focuses on sequential visual understanding tasks (video activity detection, robot learning) using external memory tokens to summarize history, not on theoretical Turing machine simulation with chain-of-thought overhead analysis.

4. Transformers Have the Potential to Achieve AGI

URL: View paper

Brief Assessment

Transformers AGI Potential[14] focuses on proving that transformers can simulate probabilistic Turing machines (PTMs) and discussing their potential for AGI through universal search and induction. The original paper addresses a different problem: reducing the per-step CoT overhead from Ω(s(n)) to O(s(n)^c) for deterministic multi-tape TM simulation with optimal space complexity. These are distinct technical contributions with different objectives and complexity bounds.

5. Towards Understanding Multi-Round Large Language Model Reasoning: Approximability, Learnability and Generalizability

URL: View paper

Brief Assessment

Multi Round Reasoning[38] focuses on approximation, learnability, and generalization of multi-round reasoning processes, not on reducing per-step CoT overhead in Turing machine simulation. The candidate demonstrates that transformers can approximate Turing machines through multiple rounds but does not address the specific efficiency improvement (reducing O(s(n)) to O(s(n)^c) per-step overhead) that is the core novelty of the original contribution.

6. Autoregressive Large Language Models are Computationally Universal

URL: View paper

Brief Assessment

Autoregressive Computationally Universal[16] focuses on proving computational universality through autoregressive decoding and lag systems, not on optimizing chain-of-thought overhead or context window complexity for Turing machine simulation.

7. Eliciting Fine-Tuned Transformer Capabilities via Inference-Time Techniques

URL: View paper

Brief Assessment

Eliciting Transformer Capabilities[37] focuses on proving that fine-tuned capabilities can be approximated via in-context learning without parameter changes, not on Turing machine simulation efficiency or chain-of-thought overhead reduction.

8. Transformers Learn Shortcuts to Automata

URL: View paper

Brief Assessment

Shortcuts to Automata[31] focuses on finite-state automata simulation by transformers using shallow circuits (O(log t) or O(1) depth), not multi-tape Turing machines with space-bounded simulation and CoT overhead analysis. The candidate addresses a fundamentally different computational model (finite vs. unbounded state) and complexity measure (circuit depth vs. CoT steps per TM step).

9. Constant Bit-size Transformers Are Turing Complete

URL: View paper

Brief Assessment

Constant Bitsize Completeness[2] achieves O(t(n)·s(n)) CoT steps with constant bit-size but does not address the per-step overhead reduction to O(s(n)^c) that the original paper claims. The candidate focuses on establishing Turing completeness with constant bit-size rather than optimizing the per-step CoT overhead factor.

Contribution 2: Sparse geometric-offset attention suffices for efficient universal computation

Description: The construction demonstrates that Transformers using sparse attention patterns with geometrically spaced offsets (attending only to tokens at positions 1, 2, 4, 8, ... steps earlier) can achieve efficient Turing-complete computation, avoiding the quadratic overhead of full attention while maintaining universality.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 3: Improved multi-tape to multi-queue TM simulation under synchronous model

Description: The authors develop a more efficient simulation of multi-tape Turing machines using synchronous multi-queue Turing machines, improving upon prior work by reducing space complexity from t(n)^(1+1/k′) to s(n) and time slowdown from t(n)^(1/k′) to s(n)^(1/k′) under a stricter synchronous model where every queue must pop and push exactly one symbol per step.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

Textual similarity detection checked 10 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Constant Bit-size Transformers Are Turing Complete

Detected in: Core Task (sibling), Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Efficient Turing Machine Simulation with Transformers View paper
- [1] Transformers as Approximations of Solomonoff Induction View paper
- [2] Constant Bit-size Transformers Are Turing Complete View paper
- [3] Attention is turing-complete View paper
- [4] PENCIL: Long Thoughts with Short Memory View paper
- [5] From turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models View paper
- [6] Gpt is becoming a turing machine: Here are some ways to program it View paper
- [7] Sparse universal transformer View paper
- [8] Universal transformers View paper
- [9] Universal length generalization with turing programs View paper
- [10] Rule extrapolation in language modeling: A study of compositional generalization on OOD prompts View paper
- [11] Executing arithmetic: Fine-tuning large language models as turing machines View paper
- [12] Ask, and it shall be given: On the Turing completeness of prompting View paper
- [13] Neural Attention Memory View paper
- [14] Transformers Have the Potential to Achieve AGI View paper
- [15] Softmax Transformers are Turing-Complete View paper
- [16] Autoregressive Large Language Models are Computationally Universal View paper
- [17] How powerful are decoder-only transformer neural models? View paper
- [18] On the computational power of transformers and its implications in sequence modeling View paper
- [19] Momentary Contexts: A Memory and Retrieval Approach for LLM Efficiency View paper
- [20] Time series attention based transformer neural Turing machines for diachronic graph embedding in cyber threat intelligence View paper
- [21] Memory Augmented Large Language Models are Computationally Universal View paper
- [22] On the turing completeness of modern neural network architectures View paper
- [23] The Imitation Game: Turing Machine Imitator is Length Generalizable Reasoner View paper
- [24] Infinite Time Turing Machines and their Applications View paper
- [25] Learning Linear Attention in Polynomial Time View paper
- [26] Harnessing Chaos: Predicting the Unpredictable in Cellular Automata View paper
- [27] Ask, and it shall be given: Turing completeness of prompting View paper
- [28] Looped ReLU MLPs May Be All You Need as Practical Programmable Computers View paper
- [29] Token Turing Machines View paper
- [30] Emergent neural turing machine and its visual navigation. View paper
- [31] Transformers Learn Shortcuts to Automata View paper
- [32] Turing Complete Transformers: Two Transformers Are More Powerful Than One View paper
- [33] A provably stable neural network Turing Machine View paper
- [34] Statistically Meaningful Approximation: a Theoretical Analysis for Approximating Turing Machines with Transformers View paper
- [35] Theory for Understanding Transformers: An Overview of Current Research View paper
- [36] Neural Networks and Solomonoff Induction View paper
- [37] Eliciting Fine-Tuned Transformer Capabilities via Inference-Time Techniques View paper
- [38] Towards Understanding Multi-Round Large Language Model Reasoning: Approximability, Learnability and Generalizability View paper
- [39] Artificial Intelligence and the Computational Power of Neural Networks View paper