Barriers to Discrete Reasoning with Transformers: A Survey Across Depth, Exactness, and Bandwidth

Michelle Yuan, Weiyi Sun, Amir H. Rezaeian, Jyotika Singh,
Sandip Ghoshal, Yao-Ting Wang, Miguel Ballesteros, Yassine Benajiba
Oracle AI
Correspondence: michelle.yuan@oracle.com

Abstract

Transformers have become the foundational architecture for a broad spectrum of sequence modeling applications, underpinning state-of-the-art systems in natural language processing, vision, and beyond. However, their theoretical limitations in discrete reasoning tasks, such as arithmetic, logical inference, and algorithmic composition, remain a critical open problem. In this survey, we synthesize recent studies from three theoretical perspectives: circuit complexity, approximation theory, and communication complexity, to clarify the structural and computational barriers that transformers face when performing symbolic computations. By connecting these established theoretical frameworks, we provide an accessible and unified account of why current transformer architectures struggle to implement exact discrete algorithms, even as they excel at pattern matching and interpolation. We review key definitions, seminal results, and illustrative examples, highlighting challenges such as depth constraints, difficulty approximating discontinuities, and bottlenecks in inter-token communication. Finally, we discuss implications for model design and suggest promising directions for overcoming these foundational limitations.

1 Introduction

Transformer architectures (Vaswani et al., 2017) have driven remarkable progress across language understanding (Raffel et al., 2020), generation (Achiam et al., 2023; Guo et al., 2025), and perception tasks (Dosovitskiy et al., 2020; Kirillov et al., 2023), setting new performance benchmarks and enabling a wide range of applications (Tunstall et al., 2022; Singh, 2023). These large models have been coined as "foundation models" because of their consequential impact for society (Bommasani et al., 2022). Despite these advances, a persistent gap remains in their ability to perform reliable discrete reasoning, such as precisely solving arithmetic expressions (Lee et al., 2023), evaluating logical formulas (Dziri et al., 2023), or generalizing systematic rules to longer or unfamiliar examples (Wu et al., 2024; Gao et al., 2024b; Yang et al., 2024; Hong and Park, 2025; Allamanis and Yin, 2025; Singh et al., 2025).

This survey examines the fundamental factors that constrain transformer-based models in discrete reasoning domains. Rather than focusing on empirical evaluation or incremental improvements, we systematically review and synthesize results from three major theoretical disciplines:

1. **Circuit Complexity** frames model capacity in terms of computational depth and parallelism, emphasizing why certain algorithmic processes fundamentally require sequential computation
2. **Approximation Theory** exposes the inherent difficulties in representing discontinuous, rule-based functions as transformers are trained for continuous mappings
3. **Communication Complexity** formalizes the limits of information exchange within self-attention and elucidates how architectural choices restrict multi-hop reasoning and coordination across tokens.

Through this synthesis, we provide readers with a cohesive understanding of why transformers succeed in interpolation tasks (e.g. summarization) but fall short in reliably executing symbolic algorithms. Through outlining key definitions, representative tasks, and landmark theoretical results, we contextualize both the current abilities and intrinsic limitations of transformer architectures. We also identify promising avenues for future research, including hybrid models that combine neural and symbolic components or extend transformer architectures beyond their current design points.

This survey is structured as follows: we first provide necessary background on transformers and

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1966–1978

discrete reasoning tasks, then dedicate individual sections to each of the three theoretical frameworks. Figure 1 shows an outline of the issues and subtopics for each framework. We conclude by drawing connections between these perspectives, surveying related work, and outlining open problems and opportunities for the next generation of reasoning-capable machine learning systems.

2 Background

This section provides essential context on the architectural design of transformers and clarifies why discrete reasoning tasks pose unique challenges for modern neural models. We begin by outlining the key components and mechanisms of transformer architectures, followed by an overview of the types of discrete reasoning problems that motivate our analysis. Finally, we introduce three foundational theoretical frameworks, circuit complexity, approximation theory, and communication complexity, that will anchor the discussion in subsequent sections.

2.1 Transformer Architectures

Transformers (Vaswani et al., 2017) have revolutionized sequence modeling by replacing recurrent layers with a purely attention-based design. The fundamental architecture comprises multiple stacked layers, each containing a multi-head self-attention mechanism and a position-wise feedforward sublayer, typically augmented with residual connections and layer normalization. This parallelized design better captures long-range dependencies compared to previous NLP architectures (Devlin et al., 2019). Later on, researchers have looked towards scaling model size (Hoffmann et al., 2022). This involved increasing both the depth (the number of stacked layers) and the width (hidden size and attention heads), leading directly to the development of modern LLMs (Achiam et al., 2023). However, this scaling strategy quickly exposed two critical limitations (Duman et al., 2023). First, self-attention scales quadratically with the input sequence length. Second, at inference time, the fixed number of stacked layers dictates a fixed sequential depth, which imposes theoretical bounds on the number of computational steps the model can perform in a single forward pass.

To overcome these limitations, LLMs have adopted various architectural and algorithmic innovations. To efficiently scale parameter count while managing inference cost, the Mixture-of-Experts layer (Zhou et al., 2022) allows the creation of models with trillions of parameters while only activating a small fraction for any given input. Memory-efficient techniques like Sparse Attention (Tay et al., 2020) and Flash Attention (Dao et al., 2022) are developed to address the quadratic complexity of long inputs. Furthermore, modifications to positional encoding, such as RoPE, where positional information is encoded by rotating the query and key vectors, enhance the models' ability to maintain and model dependencies effectively over increasingly longer sequences (Su et al., 2024). While these innovations improve efficiency, they do not inherently increase the fixed computation depth (Section 3), guarantee symbolic precision (Section 4), nor fully address memory bandwidth (Section 5).

2.2 Discrete Reasoning Tasks

Discrete reasoning involves operations that require precise, rule-based manipulation of symbolic entities, often with exact correctness required for each computational step. Examples include arithmetic problems, Boolean logic evaluation, and parsing. Unlike tasks involving perceptual or semantic similarity, discrete reasoning is characterized by sharp decision boundaries, combinatorial complexity, and non-smooth target mappings. Transformers often rely on similarity-based heuristics learned from data, rather than internalizing algorithmic procedures or conditional logic (Mirzadeh et al., 2024). This distinction lies at the heart of their observed failure modes on problems that require compositionality, systematic generalization, or step-wise execution of rules (Dziri et al., 2023).

Recent studies continue to report notable shortcomings of foundation models on discrete reasoning tasks. The MathArena Apex leaderboard (Balunović et al., 2025) contains challenging combinatorics problems like:

> **The Zigzagging Chessboard.** Let P be a polygon formed by the edges of an infinite chessboard, which does not intersect itself. Let a_1, a_2, and a_3 denote the number of unit squares that have exactly 1, 2, or 3 edges on the boundary of P, respectively. Find the largest real number k such that the inequality $a_1 + a_2 > ka_3$ holds for each polygon constructed with these conditions.

This problem requires discrete reasoning because

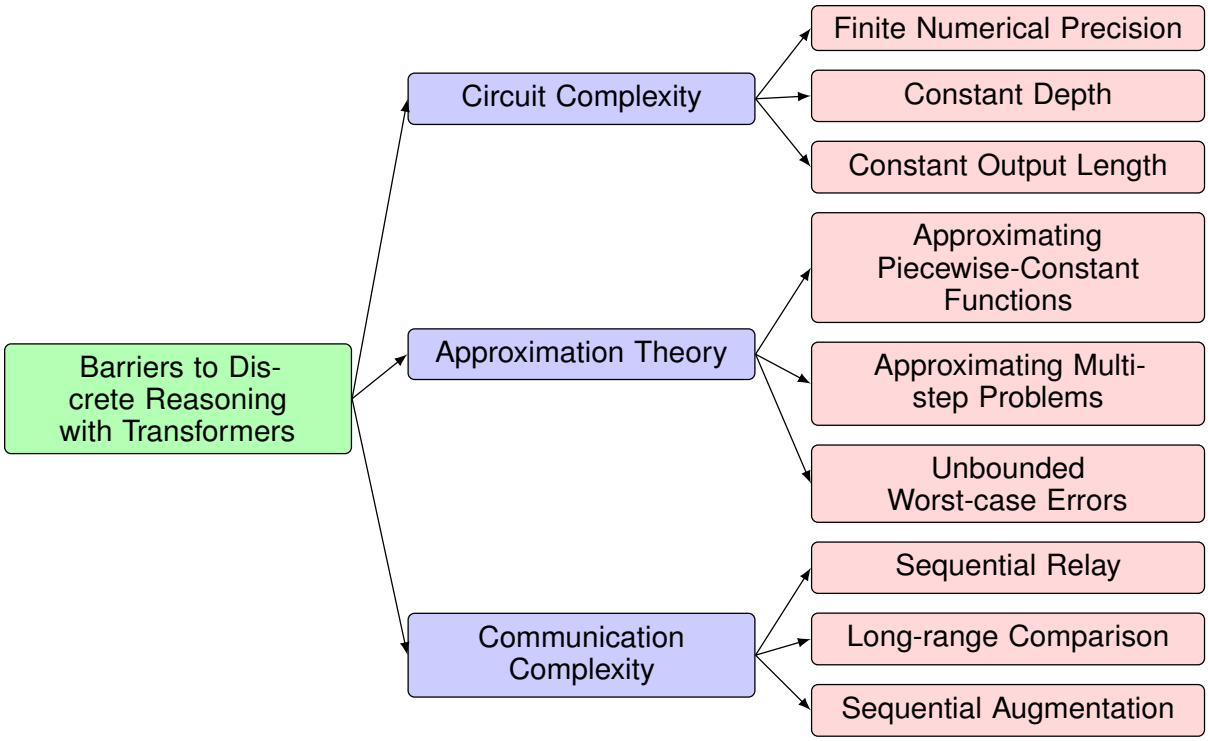

Figure 1: Taxonomy of barriers to discrete reasoning in transformers, organized by theoretical lens: circuit complexity, approximation theory, and communication complexity.

it involves deducing relationships based on the exact combinatorial arrangement and counting of unit squares with specific properties on an infinite chessboard. On the MathArena Apex leaderboard, Gemini 3 Pro currently has the highest accuracy of 23.44%. On other related benchmarks like FrontierMath Tier-4 (Glazer et al., 2024), GPT 5.2 Pro has the highest accuracy of 29.2%. The low performance from these frontier models reflects the intrinsic difficulty posed by combinatorial structures and discrete reasoning in the benchmarks.

2.3 Theoretical Frameworks

Circuit Complexity Circuit complexity is a branch of computational complexity theory concerned with the resources required to compute functions using Boolean circuits. A Boolean circuit can be formally described as a directed acyclic graph in which each internal node represents a logic gate (such as AND, OR, NOT), and the leaves correspond to input variables $x_1, x_2, \ldots, x_n$. The primary measures of a circuit are its *size* (the total number of gates, denoted $S(f)$ for a function f) and its *depth* (the length of the longest path from an input node to the output, denoted $D(f)$). Circuit classes such as AC^0, TC^0, and NC^1 categorize families of functions based on allowable depth, gate types, and fan-in.

The class AC^0 consists of all Boolean circuits with constant depth, unbounded fan-in for AND and NOT gates, and polynomial size. Building on this, the class TC^0 extends AC^0 by including majority gates, which are capable of processing unbounded inputs and outputting 1 if and only if the majority of those inputs are 1. Another important class, NC^1, is similar to AC^0 but restricts circuits to logarithmic depth (specifically O(logn)) and bounded fan-in. The known relationships are $\mathrm{AC}^0 \subset \mathrm{TC}^0 \subseteq \mathrm{NC}^1$ in DLOGTIME-uniform, L-uniform, and non-uniform settings. Key questions in circuit complexity include establishing lower and upper bounds for the resources needed to compute specific Boolean functions (Vollmer, 1999).

Approximation Theory Approximation theory studies how well classes of functions (such as polynomials or neural networks) can approximate target functions, usually with respect to some norm or distance metric. Given a target function $f : X \to Y$ and a family of approximating functions $\mathcal{F}$, the goal is to find an $f^* \in \mathcal{F}$ that minimizes the approximation error $\|f - f^*\|$ according to a specified norm, such as the L^∞ (uniform) or L^2 norm. Classical results, such as the Universal Approximation Theorem (Hornik et al., 1989), assert that for cer-

tain families of functions (e.g., sufficiently wide neural networks with non-polynomial activation functions), it is possible to approximate all continuous functions on compact domains arbitrarily well. However, approximation theory also characterizes the inherent difficulty of approximating discontinuous functions, where the error depends on factors like the smoothness and complexity of the target, and the expressive power or depth of the approximating family (DeVore, 1998).

Communication Complexity Communication complexity analyzes the amount of communication required between two or more parties to collaboratively compute a function whose input is distributed among them. Instead of focusing on computational resources, this framework quantifies the minimal number of bits that must be exchanged for the correct computation of a function $f(x, y)$, where x is known to Alice and y to Bob, and both aim to compute $f(x, y)$ with the least amount of communication. The deterministic communication complexity $D(f)$ of a function f is defined as the minimum number of bits exchanged between the parties in the worst case, using an optimal protocol. Extensions consider randomized protocols, multiple rounds of communication, or multiparty variants. Canonical problems studied in communication complexity include the Equality function, Disjointness, and Greater-Than, each exhibiting different lower and upper bounds for bit and round complexity (Kushilevitz, 1997).

3 Circuit Complexity: Why Depth Becomes a Bottleneck

Recent theoretical work has sought to understand Transformer's expressivity using tools from circuit complexity and formal languages. This line of work provides a principled account of which discrete reasoning tasks transformers can or cannot perform, and clarifies how different transformer variants change the network's representation power. In this section, we focus on how finite precision, constant depth, and constant output constraints each create distinct barriers to discrete reasoning, but we also highlight how relaxing these assumptions can expand transformer capability. We build on unifying frameworks such as Strobl et al. (2024).

3.1 Finite Numerical Precision

In practice, transformer models operate under finite-precision arithmetic, mapping to threshold circuits with limited numerical granularity (Merrill and Sabharwal, 2023b,c; Merrill et al., 2022). These models are fundamentally limited in their ability to represent functions requiring precise bitwise logic, such as parity or majority. For example, encoder-only transformers with hard attention correspond to AC^0 circuits and are unable to recognize parity, majority, or DYCK-k languages (Hao et al., 2022). While softening attention mechanisms (e.g. average-hard or softmax) extends expressivity slightly beyond AC^0 (Merrill et al., 2022; Bhattamishra et al., 2020; Yao et al., 2021; Chiang and Cholak, 2022), and the implementation of embedding schemes such as RoPE do not surpass the upper bound of DLOGTIME-uniform TC^0 (Merrill and Sabharwal, 2023c; Chen et al., 2025). However, if infinite precision were possible, one could in theory go well beyond these bounds, though this is unrealistic in current architectures.

3.2 Constant Depth

Transformers are commonly deployed with a constant number of layers, reflecting a fixed computational depth that sharply limits their ability to perform deeply compositional or sequential tasks. Classic and recent results show that finite-depth, finite-precision transformers without intermediate decoding have small-circuit upper bounds, confining their expressivity to uniform TC^0 or below (Merrill and Sabharwal, 2023c,b; Peng et al., 2024; Merrill et al., 2022; Zubic et al., 2025). They cannot, for instance, compute multi-digit addition, recognize nested formal languages, or resolve long-range dependencies. However, if one allows the number of layers to scale with the input, or augments the model with additional programmatic mechanisms (such as auxiliary memory or scratchpads), then in principle these barriers can be overcome (Pérez et al., 2021; Qiu et al., 2025).

3.3 Constant Output Length (No Sequential Augmentation)

A further constraint arises when transformers are limited to producing outputs of constant length, precluding stepwise, intermediate reasoning. Without chain-of-thought (CoT) or scratchpad mechanisms, all computation must occur in a single forward pass, restricting models to shallow circuits and causing systematic underperformance on multi-step or compositional tasks (Feng et al., 2023). Recent theory shows this is an inherent barrier, but if transformers are permitted to emit longer out-

puts, effectively augmenting depth with CoT tokens, their expressivity hierarchy expands dramatically. Merrill and Sabharwal (2023a) show that logarithmic-length CoT yields modest gains, linear chain-of-thought suffices for all regular languages, and with polynomial CoT or certain architectural variants, transformers Li et al. (2024) demonstrate that with T CoT steps, a constant-depth, constant-precision transformer can simulate any Boolean circuit of size T. Qiu et al. (2025) further prove that, by providing suitable program-prompts, a fixed-size decoder-only transformer can attain Turing-complete expressive capacity by emitting $O(t(n))$ chain-of-thought, making output length a true computational resource.

Summary

Circuit complexity theory clarifies that transformers are provably unable to represent many essential symbolic functions under realistic settings, such as finite precision, constant architectural depth, and no access to chain-of-thought. Nonetheless, relaxing any of these constraints, by increasing precision, depth, or enabling intermediate outputs, yields substantial gains in representational power. This suggests concrete directions for future architecture and prompt design to enhance discrete reasoning.

4 Approximation Theory: Why Exactness is Hard

The Universal Approximation Theorem (UAT) states that sufficiently large neural networks can approximate any continuous function on a compact domain (Hornik et al., 1989). However, this property does not extend to discontinuous functions, which are the very functions that dominate discrete reasoning tasks. Symbolic tasks frequently have outputs that change abruptly at precise input thresholds. Approximating such sharp boundaries with smooth neural networks introduces transition regions, resulting either in systematic errors (if smoothed) or excessive sensitivity (if made steep) (Arora et al., 2022; Chen et al., 2023). Discrete reasoning tasks naturally extend over unbounded spaces, i.e., $f : \mathbb{N}^n \to \mathbb{N}$ for n-digit addition, violating the UAT's compactness assumption. Thus, transformers trained as smooth interpolators cannot guarantee uniform error control as sequence or input length grows (Hahn and Rofin, 2024).

Note that the limitations mentioned in this section are not unique to transformer architectures. In fact, the constraints described here are representative of broader classes of inductively-trained neural networks, whose ability to learn target functions is fundamentally limited by principles established in classical learning theory. For instance, learnability theory formally characterizes which function classes can be efficiently learned and highlights the inherent difficulty of precisely modeling discontinuous or highly complex rules from finite data (Valiant, 1984).

4.1 Difficulty Approximating Piecewise-Constant Functions

Symbolic functions are often *piecewise-constant*. Take n-digit addition as an example. The function mapping $(a, b) \mapsto s$ (where s is the sum sequence) is constant on most of the input space but jumps sharply at digit-wise carry thresholds. Formally, let $c_i = 1$ if $a_i + b_i + c_{i-1} \geq 10$ (carry at position i), where $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$. The set of discontinuity surfaces is

$$\mathcal{D} = \{(a, b) \mid a_i + b_i + c_{i-1} = 10 \text{ for some } i\},$$

which grows linearly with n. For vectors near $\mathcal{D}$, small perturbations in a single digit flip the carry and, recursively, all subsequent digits.

Empirical analyses show that transformer display sharp accuracy drops on discontinuities. For example, Hu et al. (2024) observe regression in carry-over operations on multi-digit addition and multiplication, and conclude that transformers rely on case-based reasoning to match unseen examples to previously seen problems. Visualizing the model failures will represent these discontinuities as "holes". Zhao et al. (2024) specifically design discontinuities as "traps" in math word problems and demonstrate that state-of-the-art LLMs cannot solve these augmented problems.

4.2 Compounding Errors for Multi-step Problems

The complexity intensifies with the number and composition of discontinuities. Let $f^{(k)}$ denote k-fold function composition (e.g., k sequential carries in addition or k nested logical operations). For each composition step, the Lipschitz constant or associated parameter norm must grow to retain pre-

cision at boundaries, i.e.,

$$\left\| f^{(k)}(x) - \hat{f}^{(k)}(x) \right\| \leq L^k \varepsilon$$

for some $L > 1$, so error compounds exponentially in the number of steps unless the per-step error is super-polynomially small. In practice, fixed-depth neural networks cannot achieve this for increasing k (Bubeck and Sellke, 2021).

Lee et al. (2023) run extensive experiments on length generalization for multi-digit arithmetic problems. Despite various training ablations, which includes data sampling, data formatting, and optimization strategies, transformers are unable to generalize to problems with more computational steps. (Dziri et al., 2023) also observe the same regression as they increase the number of steps in multi-step reasoning problems.

4.3 Unbounded Worst-case Error for Discontinuities

Training strategies such as label smoothing (Müller et al., 2019) or calibration (Desai and Durrett, 2020; Chen and Li, 2023) can improve mean squared error locally but do not address the exponential number of discontinuity boundaries for deeper compositions. Let E_{avg} and $E_{\max}$ denote average and worst-case error, respectively:

$$E_{\text{avg}} = \mathbb{E}_{x \sim P_{\text{train}}} \left[|f(x) - \hat{f}(x)| \right]$$

$$E_{\max} = \sup_{x \in \mathcal{X}} |f(x) - \hat{f}(x)|$$

While E_{avg} may be small, $E_{\max}$ remains large, especially for out-of-distribution examples. Wu et al. (2024) introduce a framework to evaluate "counterfactual" reasoning where the logical rules are transferred to a counterfactual world and show that large transformers fail to generalize to alternative domains. There is a whole line of work that shows transformers susceptible to simple, adversarial attacks (Guo et al., 2021; Shayegani et al., 2023; Gao et al., 2024a). In the end, transformers will show a "bimodal" error – many exact outputs, with rare but catastrophic failures on decision boundaries.

Summary

Approximation theory shows that the intrinsic smoothness bias of transformers conflicts with the highly discontinuous, rule-based structure of algorithmic reasoning. Increasing model width or dataset size cannot overcome the compounding effect of boundaries and error growth in deep compositions. New architectures or training methods targeting discontinuity and composition may be necessary for robust symbolic reasoning.

5 Communication Complexity: Why Bandwidth and Rounds Matter

Communication complexity offers a distinct perspective on the foundational limitations of transformers in discrete reasoning. While previous sections explored the role of depth and approximation theory, here we focus on how the limits of message-passing and bandwidth within the self-attention architecture sharply constrain multi-step and symbolic computation.

5.1 Sequential Relay Limits

Transformers implement a highly parallel message-passing scheme: in every self-attention layer ("round"), each token can broadcast information to every other token by updating its representation based on a mixture of all sequence elements. The model's depth directly determines the number of these broadcast rounds it can perform, while the hidden dimension of each token controls the "bandwidth" of each individual message. This structural design is efficient for pattern recognition over moderately-sized contexts but creates bottlenecks for problems requiring several sequential steps of reasoning or information relay along chains of tokens. Practically, increasing hidden size improves the fidelity of exchanged messages but cannot substitute for more architectural rounds. For many sequential tasks there are provable depth-width tradeoffs, namely constant (or small) depth transformers must increase model dimension polynomially in input length to perform k-step sequential composition exactly (Chen et al., 2024; Keles et al., 2023).

Example: Long-carry Propagation as Pointer Chasing. Consider the task of summing two n-bit binary numbers, a and b, where the output most significant bit (MSB) depends on the correct propagation of carries from the least significant bit (LSB)

through all intermediate positions. Each digit position can be mapped to a "party" holding its local inputs (a_i, b_i), with the value of the carry bit at position $i+1$, c_{i+1}, depending on both the carry c_i and the local inputs (Sanford et al., 2023, 2024; Peng et al., 2024). To compute the MSB exactly, information about the final carry must propagate sequentially from the LSB, traversing all $n-1$ positions, a classic "pointer chasing" problem (Ponzio et al., 1999). Within a transformer, each layer corresponds to just one round or "hop" of message passing. Thus, a transformer with M layers can only reliably propagate information M positions; if the required dependency chain (n) exceeds M, exact computation is impossible within a single forward pass, regardless of width. This bottleneck is not mitigated by increasing hidden size. A wider model can increase per-message bandwidth but cannot reduce the minimum number of sequential rounds required for long-range dependency propagation (Sanford et al., 2024; Peng et al., 2024).

5.2 Long-Range Comparison Constraints

Communication complexity theory establishes a range of lower bounds for tasks that require significant information relay across input positions. More generally, classic problems like substring equality, membership in complex formal languages, or enforcing cross-sequence consistency all become exponentially harder as the required number of rounds exceeds model depth. With limited layers, evidence from distant tokens is compressed into lossy summary representations, which leads to brittle, approximate answers and catastrophic errors at decision boundaries (Chen et al., 2024).

Example: Equality Testing Across Distant Substrings. An archetypal communication bottleneck problem is checking whether two distant substrings in a long sequence are equal. Suppose we construct an input $S = u \parallel \text{filler} \parallel v$, where u and v are identical binary strings of length $n/2$, separated by substantial "filler" content. The task is to determine whether $u = v$ (Zhao et al., 2025; Wang et al., 2025). In communication complexity, deterministic protocols for checking equality require transmitting $\Omega(n)$ bits between the parties holding u and v. For transformers, self-attention should allow all-to-all interaction. In practice, computation is bottlenecked by the hidden dimension (bandwidth per token) and the number of available layers (communication rounds). For very long sequences, the information about u must travel through hundreds or thousands of intervening filler tokens to reach v. This is analogous to a channel with severe bandwidth and round constraints. As separation grows, information about u becomes diluted or corrupted. The model needs lossy summarization or hash-like approximations. While randomized protocols (e.g., hashing) can reduce communication, they introduce further error. Thus, transformers risk brittle failures and incorrect equality decisions as sequence length increases. Such effects are observed in long-context tasks, string matching, and symbolic reasoning benchmarks (Bhattamishra et al., 2024).

5.3 Limitations of Sequential Augmentation

CoT prompting and scratchpads extend the effective computation of fixed-depth transformers by spreading reasoning across additional generated tokens. Each token provides an intermediate step, partially compensating for the limited number of architectural rounds. Empirically, this improves performance on discrete reasoning tasks. Formally, such augmentation increases expressivity: scratchpads allow transformers to recognize richer languages than direct input–output mapping, but the benefit is bounded as the required compositional steps increase (Merrill and Sabharwal, 2023a).

Example: Multi-hop Relational Queries in Text. Answering complex questions, like "Is an entity mentioned in section A referenced again under a different alias in section C, and negated in section E?", requires integrating information across multiple distant spans. Each step is a "hop" along a chain of entities and mentions. Each self-attention layer enables one synchronous round of global communication, so resolving a k-hop relational query usually requires at least k rounds. However, what if there are additional resources provided, like scratchpads, increased width, or specialized architectures? Empirical studies still show that transformers often fail at multi-hop reasoning. For example, Peng et al. (2024) demonstrates that increasing hidden size (bandwidth) does not compensate for insufficient depth (rounds) when logical structures are needed. Recent theoretical results (Amiri et al., 2025) show that the length of the scratchpad must scale with task complexity.

Summary

Communication complexity highlights that the limitations are fundamentally tied to the amount of information that can be exchanged across the model's parallel structure. Above examples illustrate that transformer architectures are fundamentally limited not just by width but by lower bounds on sequential information relay. These constraints cannot be overcome by self-attention alone. Tasks requiring information to be propagated, compared, or composed across multiple long-distance dependencies most clearly expose these limitations.

6 Open Questions and Future Directions

The convergence of circuit complexity, approximation theory, and communication complexity illuminates the boundaries of what current transformer architectures can achieve in discrete reasoning. The core findings reveal persistent barriers regarding exact symbolic computation, systematic generalization, and the reliable composition of step-wise logic. Here, we discuss the broader implications for the field, outline open questions, and suggest future directions toward more capable neuro-symbolic and state-aware reasoning systems.

6.1 Implications of Theoretical Barriers

To clarify how these frameworks reinforce each other, we summarize the implications and connections between each section:

1. **Depth–Communication Link**: Circuit complexity and communication complexity both show that transformer depth corresponds directly to the number of sequential computation or information relay steps: limited depth means both fewer computational layers (*circuit*) and fewer rounds for propagating information (*communication*) across input positions. Both views support that fixed-depth transformers have difficulty with long reasoning problems.
2. **Depth–Approximation Link**: The inability to dynamically add more layers to match problem complexity is related to approximation failures. As shown in approximation theory, functions with many discontinuities require either greater depth or exponentially sharper approximators. With fixed transformer depth, both frameworks conclude that errors will compound with each reasoning step.
3. **Approximation–Communication Link**: Approximation theory shows that transformers inherently smooth out their learned functions. Communication complexity demonstrates the difficulty in aggregating the distant pieces of evidence often needed for an exact answer. Thus, failures in exactness can be viewed as both a bottleneck in approximating discontinuities and in collecting all the information required for a reasoning problem.

The above connections illustrate that the three frameworks are not independent. Transformer failures in discrete reasoning are best understood as the intersection and mutual reinforcement of all three frameworks.

6.2 Open Research Questions

While considerable progress has been made in understanding the limitations of transformers for discrete reasoning, a range of important questions remain open. Future research must address both architectural and training-level challenges to extend symbolic and systematic reasoning capabilities. Below, we highlight several central questions guiding exploration in this area:

How can we design architectures that enable symbolic reasoning beyond current transformer limitations? Many symbolic reasoning tasks fundamentally require precise, step-by-step computation and exact logical operations rather than the pattern-matching and smooth interpolations at which transformers excel. An open question is how to develop alternative model architectures, that can match or surpass the interpolation strengths of transformers while robustly supporting symbolic manipulation and exact algorithmic reasoning.

How can models develop an explicit awareness of decision boundaries and discontinuities? Current neural architectures treat most functions as continuous and tend to interpolate over decision thresholds, leading to errors in discrete or rule-based tasks. A major question remains: how can we train models to detect, represent, and reason about sharp boundaries and discontinuous changes, rather than smoothing over them (Feng et al., 2025)?

Which training paradigms most effectively foster compositional generalization and multi-step reasoning? Even with expressive architectures,

generalizing algorithmic behavior to new or longer compositions remains challenging. It is an active area of inquiry to determine what curriculum designs, intermediate supervision, data augmentations, or learning signals best encourage models to internalize the structure of multi-step, compositional rules and to robustly generalize this reasoning to novel or out-of-distribution inputs (Lee et al., 2025; Cai et al., 2025).

6.3 Recommendations for Future Architectures

To address the above research questions, future work may look toward:

- **Neuro-symbolic models:** Combining the pattern recognition strengths of neural networks with explicit, programmable symbolic components (e.g., logic circuits, formal parsers, or differentiable interpreters) may help preserve both algorithmic precision and flexibility (Zhou et al., 2021; Ma et al., 2024; Calanzone et al., 2025).
- **State-aware and memory-augmented architectures:** Mechanisms such as external memory (Hatalis et al., 2023; Hu et al., 2025), structured state-passing (Wu et al., 2023), or state space models (Gu and Dao, 2024) could support unbounded sequential computation and long-range dependency management.
- **Boundary-aware learning and representations:** Developing models with inductive biases or objectives (Ma et al., 2023; Baheri and Alm, 2025) that encourage awareness of sharp decision boundaries, symbolic transitions, or rule changes. The topic is related to manifold learning which has recently gained more attention in the LLM research community (Xie et al., 2025).

Continued integration of machine learning, theoretical computer science, and symbolic AI research will be essential to overcome the intrinsic constraints of the current transformer paradigm.

7 Conclusion

The limitations of transformer-based models on discrete reasoning tasks arise from fundamental architectural properties, not merely from training or data deficiencies. Theoretical insights from circuit complexity, approximation theory, and communication complexity together illuminate why transformers fail to reliably compose logical operations, propagate long-range dependencies, or approximate discontinuous decision boundaries.

Our survey shows that these perspectives are not isolated and instead reinforce one another to reveal a consistent picture of discrete reasoning failures. While techniques such as chain-of-thought prompting or scratchpads offer partial relief, they do not fully overcome the inherent bounds imposed by the transformer's design. Addressing these barriers calls for principled innovations at the intersection of neural, symbolic, and sequential computation, moving beyond scaling alone toward more structurally expressive architectures. By charting these theoretical boundaries and open questions, we hope to provide a clearer foundation for designing models that are not only more capable in discrete reasoning, but also more robust, interpretable, and generalizable across tasks that demand systematic, exact computation.

Limitations

The survey's scope is limited to discussion on transformers in the perspective of three frameworks: circuit complexity, approximation theory, and communication complexity. It is very possible there may be alternative framings of transformer limitations that are not covered in this survey. Moreover, the survey's focus is on discrete reasoning tasks. We understand that there are many interesting tasks beyond discrete reasoning that also require more careful analysis. Finally, the survey may not cover all the variants of transformers in the existing literature. Due to space constraints, we highlight the most important and relevant work for the survey's theme.

Acknowledgments

We immensely thank Dan Roth, Sujith Ravi, and Anna Rumshisky for their insightful comments and suggestions that helped improve the quality of this paper. This work has been funded by Oracle but the views expressed in this paper are those of the authors. They do not reflect the policies or positions of Oracle or their affiliates.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman,

Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Miltiadis Allamanis and Pengcheng Yin. 2025. Disproving program equivalence with llms. *Preprint*, arXiv:2502.18473.

Alireza Amiri, Xinting Huang, Mark Rofin, and Michael Hahn. 2025. Lower bounds for chain-of-thought reasoning in hard-attention transformers. *arXiv preprint arXiv:2502.02393*.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. 2022. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR.

Ali Baheri and Cecilia O Alm. 2025. Hierarchical neuro-symbolic decision transformer. *arXiv preprint arXiv:2503.07148*.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. Matharena: Evaluating llms on uncontaminated math competitions. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmark*.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. 2020. On the ability and limitations of transformers to recognize formal languages. *arXiv preprint arXiv:2009.11264*.

Satwik Bhattamishra, Michael Hahn, Phil Blunsom, and Varun Kanade. 2024. Separations in the representational capabilities of transformers and recurrent architectures. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2022. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

Sébastien Bubeck and Mark Sellke. 2021. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822.

Ziyang Cai, Nayoung Lee, Avi Schwarzschild, Samet Oymak, and Dimitris Papailiopoulos. 2025. Extrapolation by association: Length generalization transfer in transformers. *arXiv preprint arXiv:2506.09251*.

Diego Calanzone, Stefano Teso, and Antonio Vergari. 2025. Logically consistent language models via neuro-symbolic integration. In *The Thirteenth International Conference on Learning Representations*.

Bo Chen, Xiaoyu Li, Yingyu Liang, Jiangxuan Long, Zhenmei Shi, Zhao Song, and Jiahao Zhang. 2025. Circuit complexity bounds for rope-based transformer architecture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11091–11108.

Lijie Chen, Binghui Peng, and Hongxun Wu. 2024. Theoretical limitations of multi-layer transformer. *arXiv preprint arXiv:2412.02975*.

Wenlong Chen and Yingzhen Li. 2023. Calibrating transformers via sparse gaussian processes. In *The Eleventh International Conference on Learning Representations*.

Zixiang Chen, Junkai Zhang, Yiwen Kou, Xiangning Chen, Cho-Jui Hsieh, and Quanquan Gu. 2023. Why does sharpness-aware minimization generalize better than sgd? *Advances in neural information processing systems*, 36:72325–72376.

David Chiang and Peter Cholak. 2022. Overcoming a theoretical limitation of self-attention. *arXiv preprint arXiv:2202.12172*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ronald A DeVore. 1998. Nonlinear approximation. *Acta numerica*, 7:51–150.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Feyza Duman, Pruthuvi Wijewardena, and Chinmay Hegde. 2023. On the computational complexity of self-attention. *Proceedings of Machine Learning Research*.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, and 1 others. 2023. Faith and Fate: Limits of Transformers on Compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:70757–70798.

Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. 2025. Bird: A trustworthy bayesian inference framework for large language models. In *The Thirteenth International Conference on Learning Representations*.

Chenxing Gao, Hang Zhou, Junqing Yu, YuTeng Ye, Jiale Cai, Junle Wang, and Wei Yang. 2024a. Attacking transformers with feature diversity adversarial perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Muhan Gao, TaiMing Lu, Kuai Yu, Adam Byerly, and Daniel Khashabi. 2024b. Insights into llm long-context failures: when transformers know but don't tell. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7611–7625.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, and 1 others. 2024. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Michael Hahn and Mark Rofin. 2024. Why are sensitive functions hard for transformers? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14973–15008.

Yiding Hao, Dana Angluin, and Robert Frank. 2022. Formal language recognition by hard attention transformers: Perspectives from circuit complexity. *Transactions of the Association for Computational Linguistics*, 10:800–810.

Kostas Hatalis, Despina Christou, Joshua Myers, Steven Jones, Keith Lambert, Adam Amos-Binks, Zohreh Dannenhauer, and Dustin Dannenhauer. 2023. Memory matters: The need to improve long-term memory in llm-agents. In *Proceedings of the AAAI Symposium Series*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Kijae Hong and Yeonsu Park. 2025. Large language models for semantic join: A comprehensive survey. *IEEE Access*, 13:184478–184493.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Yi Hu, Xiaojuan Tang, Haotong Yang, and Muhan Zhang. 2024. Case-based or rule-based: How do transformers do the math? In *International Conference on Machine Learning*, pages 19438–19474. PMLR.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, and 1 others. 2025. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*.

Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. 2023. On the computational complexity of self-attention. In *International conference on algorithmic learning theory*, pages 597–619. PMLR.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Eyal Kushilevitz. 1997. Communication complexity. In *Advances in Computers*, volume 44, pages 331–360. Elsevier.

Nayoung Lee, Ziyang Cai, Avi Schwarzschild, Kangwook Lee, and Dimitris Papailiopoulos. 2025. Self-improving transformers overcome easy-to-hard and length generalization challenges. In *Forty-second International Conference on Machine Learning*.

Nayoung Lee, Kartik Sreenivasan, Jason D Lee, Kangwook Lee, and Dimitris Papailiopoulos. 2023. Teaching arithmetic to small transformers. *arXiv preprint arXiv:2307.03381*.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*, 1.

Liheng Ma, Chen Lin, Derek Lim, Adriana Romero-Soriano, Puneet K Dokania, Mark Coates, Philip Torr, and Ser-Nam Lim. 2023. Graph inductive biases in transformers without message passing. In *International Conference on Machine Learning*, pages 23321–23337. PMLR.

Xueqi Ma, Xingjun Ma, Chuang Liu, Sarah Monazam Erfani, and James Bailey. 2024. Exploring high-order message-passing in graph transformers.

William Merrill and Ashish Sabharwal. 2023a. The expressive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*.

William Merrill and Ashish Sabharwal. 2023b. A logic for expressing log-precision transformers. *Advances in neural information processing systems*, 36:52453–52463.

William Merrill and Ashish Sabharwal. 2023c. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545.

William Merrill, Ashish Sabharwal, and Noah A Smith. 2022. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. 2024. On limitations of the transformer architecture. In *First Conference on Language Modeling*.

Jorge Pérez, Pablo Barceló, and Javier Marinkovic. 2021. Attention is turing complete. *Journal of Machine Learning Research*, 22:1–35.

Stephen J. Ponzio, Jaikumar Radhakrishnan, and S. Venkatesh. 1999. The communication complexity of pointer chasing: applications of entropy and sampling. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, STOC '99, page 602–611, New York, NY, USA. Association for Computing Machinery.

Ruizhong Qiu, Zhe Xu, Wenxuan Bao, and Hanghang Tong. 2025. Ask, and it shall be given: On the turing completeness of prompting. In *The Thirteenth International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. 2023. Representational strengths and limitations of transformers. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. 2024. Transformers, parallel computation, and logarithmic depth. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Jyotika Singh. 2023. *Natural Language Processing in the Real World: Text Processing, Analytics, and Classification*. Chapman and Hall/CRC.

Jyotika Singh, Weiyi Sun, Amit Agarwal, Viji Krishnamurthy, Yassine Benajiba, Sujith Ravi, and Dan Roth. 2025. Can LLMs narrate tabular data? an evaluation framework for natural language representations of text-to-SQL system outputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 883–902, Suzhou (China). Association for Computational Linguistics.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. 2024. What formal languages can transformers express? a survey. *Transactions of the Association for Computational Linguistics*, 12:543–561.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. 2020. Sparse sinkhorn attention. In *International conference on machine learning*, pages 9438–9447. PMLR.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media, Incorporated.

Leslie G Valiant. 1984. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Heribert Vollmer. 1999. *Introduction to Circuit Complexity: A Uniform Approach*. Springer-Verlag, Berlin, Heidelberg.

Xilong Wang, Hao Fu, Jindong Wang, and Neil Zhenqiang Gong. 2025. StringLLM: Understanding the string processing capability of large language models. In *The Thirteenth International Conference on Learning Representations*.

Yiran Wu, Tianwei Yue, Shaokun Zhang, Chi Wang, and Qingyun Wu. 2023. Stateflow: Enhancing llm task-solving through state-driven workflows. In *First Conference on Language Modeling*.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Liang Zhao, and 1 others. 2025. mhc: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*.

Kai Yang, Jan Ackermann, Zhenyu He, Guhao Feng, Bohang Zhang, Yunzhen Feng, Qiwei Ye, Di He, and Liwei Wang. 2024. Do efficient transformers really save computation? In *International Conference on Machine Learning*, pages 55928–55947. PMLR.

Shunyu Yao, Binghui Peng, Christos Papadimitriou, and Karthik Narasimhan. 2021. Self-attention networks can process bounded hierarchical languages. *arXiv preprint arXiv:2105.11115*.

Fuheng Zhao, Jiayue Chen, Lawrence Lim, Ishtiyaque Ahmad, Divyakant Agrawal, and Amr El Abbadi. 2025. Llm-sql-solver: Can llms determine sql equivalence? *Preprint*, arXiv:2312.10321.

Jun Zhao, Jingqi Tong, Yurong Mou, Ming Zhang, Qi Zhang, and Xuan-Jing Huang. 2024. Exploring the compositional deficiency of large language models in mathematical reasoning through trap problems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16361–16376.

Honglu Zhou, Asim Kadav, Farley Lai, Alexandru Niculescu-Mizil, Martin Renqiang Min, Mubbasir Kapadia, and Hans Peter Graf. 2021. Hopper: Multi-hop transformer for spatiotemporal reasoning. In *International Conference on Learning Representations*.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, and 1 others. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

Nikola Zubic, Federico Soldà, Aurelio Sulser, and Davide Scaramuzza. 2025. Limits of deep learning: Sequence modeling through the lens of complexity theory. In *The Thirteenth International Conference on Learning Representations*.