SEPARATING QUANTUM CIRCUITS FROM CLASSICAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern large language models – transformers and diffusion language models – are built around two canonical algorithmic tasks: *prediction and generation*. We prove unconditional separations between low-depth quantum computation and the corresponding bounded-resource classical language-model architectures in both regimes. Concretely, we exhibit the following:

1. **Distributional separation.** We give a distribution that is sampleable by QNC^0 circuits (i.e., a family of constant-depth quantum circuits consisting of bounded fan-in gates) that no constant-round diffusion language model (DLM) with shallow scheduling and denoising can sample within constant distance, even when allowed sublinear chain-of-thought and output-token revision/remasking events, the very features modern DLMs rely on.
2. **Functional separation.** We exhibit a function computable in $\wedge \circ \mathsf{QNC}^0[\log\log n]$ (i.e., a family of $O(\log\log n)$-depth QNC^0 circuits, where n is the input length, followed by a single classical AND gate) such that any constant-depth decoder-only transformer computing the function must be *large*: it would have to have width $n^{\Omega(1)}$.

Together, our work initiates the study of quantum advantage in the era of LLMs.

1 INTRODUCTION

A central task in quantum computing is finding problems solvable by quantum computers but not by their classical counterparts. At its strongest, this asks for an *unconditional* complexity-theoretic separation between quantum and classical computing. In the oracular model, exponential separations are well-known, but in the non-oracular model, such separations are far beyond current techniques for general polynomial-time computation. Recently, a fruitful line of work has focused on *restricted* models and on unconditional separations between quantum and classical computation.

One such direction that has received a lot of traction recently is understanding the power of *shallow* circuits, i.e., constant-depth quantum circuits in contrast to their classical counterparts. Shallow quantum circuits are well motivated from a practical point of view, since they are potentially implementable on near-term fault-tolerant quantum computers. Understanding the power of these circuits was spurred by the groundbreaking work of Bravyi et al. (2018), who gave an unconditional relational separation between quantum and classical shallow circuits. Concretely, they considered the class of classical circuits of polynomial size, *constant* depth, unbounded fan-out, and bounded fan-in gates (referred to as NC^0) and the class of quantum circuits of polynomial size and constant depth with single- or two-qubit gates (referred to as *quantum* NC^0, denoted QNC^0). Bravyi et al. (2018) exhibited a relational problem[1] solvable by a QNC^0 circuit but not by any NC^0 circuit. Subsequent works strengthened this result in several directions, showing that the classical lower bound holds in the average case (Coudron et al., 2021; Le Gall, 2019), that the quantum upper bound is robust to noise (Bravyi et al., 2020), that stronger separations can be obtained between QNC^0 vs AC^0 (Bene Watts et al., 2019), QNC^0 vs GC^0 (Grewal & Kumar, 2024; de Oliveira et al., 2025) (we will formally define these classes later), as well as in interactive settings (Grier & Schaeffer, 2020) and sampling separations between QNC^0 and classical classes (Bene Watts & Parham, 2026;

[1]Unlike a *functional* problem which has a unique output for every input, a *relational* problem can have several outputs for an input; the goal is to produce one such valid output for every given input.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Grier et al., 2026a). These results suggest that, in general, shallow quantum circuits can generate and utilize correlations that are inaccessible to several "standard" shallow classical circuit classes.

A natural next question is whether such separations can be shown for more "contemporary" classical computational models. In this direction, given the remarkable success of *large language models* (LLMs) in practice, a rapidly growing body of work has sought to understand the expressive power and computational limitations of various models of LLMs. In this regime, transformer models (Vaswani et al., 2017) have been the dominant architecture for prediction-based language modeling, while diffusion language models (Sohl-Dickstein et al., 2015; Ho et al., 2020) (DLMs) have provided an alternative paradigm for content generation. Although these models can seem superficially different from the traditional circuit complexity classes studied in complexity theory (such as NC^0, AC^0, and TC^0), their theoretical abstractions often retain a key shallow-classical feature: every layer of token updates is generated by bounded-depth classical computation with limited bandwidth. This suggests a new form of quantum-classical comparison: rather than only asking whether one can separate QNC^0 from the "theoretically-motivated" circuit classes NC^0, AC^0, or GC^0, we can now also ask whether shallow quantum circuits can solve computational or sampling tasks that are hard for transformers or DLMs, resulting in a new realm of separation results. Given the recent surge of interest in understanding large language models and the role of quantum computing in machine learning, a particularly timely question is the following:

Is there a task solvable by shallow quantum circuits but not by classical LLMs?

1.1 Main results

We show two results that separate quantum shallow circuits from transformers and DLMs for functional and distributional tasks respectively. Concretely, we exhibit the following:

1. a function computable by a $O(\log \log n)$-depth QNC^0 circuit followed by a single classical AND gate, but hard to compute by a shallow-depth transformer unless it is "large";[2]
2. a distribution sampleable in QNC^0, but hard to approximately sample from a DLM which uses GC^0 denoisers[3], has sublinear chain-of-thought/workspace, and a sublinear number of output-token revision/remasking events (which are the features modern DLMs rely on).

A key distinction from earlier shallow-circuit separations is that our lower bounds are not only against classical circuit classes such as $\mathsf{NC}^0, \mathsf{AC}^0, \mathsf{GC}^0$, but against basic transformer and diffusion language models used in recent theoretical studies of modern language models (Chen et al., 2025; Jiang et al., 2026) to understand the strengths and weaknesses of these models. Although our tasks are deliberately complexity-theoretic, the separating models are among the closest formal proxies currently available for the prediction and generation mechanisms used in practice. To formulate our main results formally, we need a high-level definition of transformers and DLMs, which we provide in the next sections (Sections 2 and 3). For each of these models, we give an overview of some prior *classical* works in the related area, which serve as a starting point for our results; after that, we state our main theorem for that model, and give a detailed proof sketch of the separation in question. In the interest of space, we relegate the technical details and discussions to the corresponding sections (Appendices B and C respectively) in the appendix, along with detailed background and preliminaries (Appendix A).

Classical contribution. Although our main focus is on separating shallow quantum circuits from LLMs, both of our results contain a substantial classical component that may be of independent interest to the machine-learning community. For our DLM separation, we prove new lower bounds on the computational power of diffusion language models equipped with limited chain-of-thought, remasking, and token revision. The proof develops structural properties of the distributions generated by such DLMs, which may be useful beyond the particular quantum-classical separation considered here. For transformers, we revisit a candidate problem introduced in Chen et al. (2025), modify the construction, and show that even this strengthened candidate remains hard for transformer models unless their width is sufficiently large.

[2] We formally define the notion of *small* and *large* transformers in Section 2.

[3] The circuit class $\mathsf{GC}^0[k]$ consists of unbounded fan-in and fanout AND, OR, NOT gates along with the $\mathsf{G}[k]$ gate, which on input $x \in \{0,1\}^n$, can compute an arbitrary function of x if $|x| \leq k$ and evaluates to 0 otherwise. Throughout this work, we will restrict to $k = O(\log n)$.

2 FUNCTIONAL SEPARATION

2.1 TRANSFORMER MODELS (A HIGH LEVEL OVERVIEW)

In this section, we give a high-level overview of transformers, sufficient to understand the proof overview that follows subsequently. In Appendix A.2 we give a detailed overview of the model.

Transformers form the backbone of most modern LLMs. The core transformer architecture is specified by a number of parameters: the input length n (the "context size"), the number L of layers (the "depth"), the number H of attention heads, the embedding dimension d, and a precision parameter p. A transformer can be viewed as a sequence of length-preserving functions on d-dimensional real vectors across a sequence of *layers*, once the input text has been embedded into one such sequence. The transition between every pair of adjacent layers in the transformer architecture is determined by parameterized *query*, *key*, and *value* vectors. At a high level, the so-called *attention* mechanism takes as input a vector $\mathbf{x} \in \mathbb{R}^n$, and then takes a position-wise weighted average of value vectors (where the weights are a function of the key and query vectors), followed by a position-wise piecewise affine transformation. In this work, we exclusively consider *decoder-only* transformers, where each position attends only to positions *preceding* it, reflecting the causal structure of autoregressive generation. We note that almost every prominent modern LLM, such as GPT (Radford et al., 2019), Claude (Elhage et al., 2021), Gemini (Gemini Team et al., 2024) and LlaMa (Meta AI, 2024), use variants of the decoder-only architecture, due to their superiority in next-token prediction.

From the perspective of complexity, the key resource for transformers is not merely the total number of parameters, but the *total amount of information* communicated through the attention mechanism at each layer. Following recent work, this notion of "useful" information transmitted through the layers is captured by the so-called *width* Hdp, which is the product of the number of heads, the embedding dimension, and the precision used to represent each numeric coordinate. Following the convention of Chen et al. (2025); Sanford et al. (2023) we say a transformer is *small* if $Hdp = n^{o(1)}$, and *large* otherwise. In particular, a computational task can be thought of as requiring large width if every constant-depth decoder-only transformer solving the problem requires large internal communication in between the layers. This assumption of constant depth is reasonable in practice.

2.2 PRIOR WORKS AND CONTEXT

We now provide some context and prior works in understanding the complexity of transformers, which will set up the main motivation for our results.

(i) There have been several recent complexity theoretic works characterizing the power of transformer models in terms of circuits (Chiang & Cholak, 2022; Chiang et al., 2023; Yang et al., 2024; Li et al., 2024). Depending on the assumptions on the parameters of the transformers in question (i.e., H, d, p defined earlier), their expressive power can be shown to lie somewhere roughly between AC^0 (e.g., for constant-precision transformers using unique hard attention) and TC^0 (log-precision transformers using softmax attention). Hence, it is conceivable that the results in Bene Watts et al. (2019); Grewal & Kumar (2024) imply that QNC^0 can solve (relational) tasks that the small class of constant-precision transformers cannot solve.

(ii) Chen et al. (2025) were the first to show unconditional limitations of decoder-only transformer models (in the regime of *non-constant* precision, i.e., in the TC^0 regime). Although their main candidate problem certifying hardness is not quite phrased in these terms, it can actually be viewed as a variant of the well-known *index* function. Their lower bound was obtained via a reduction to a novel communication complexity model.

(iii) Finally, a recent result of Grier et al. (2026b) shows how to compute arbitrary TC^0 functions in QAC^0, when the quantum circuit is provided with $O(n^2)$ copies of the input (where n is the input length),[4] and shows that computing this function essentially requires these many copies.

[4]Note that in classical complexity we often assume that the bits and gates have *fanout*, i.e., $x_1, \ldots, x_n$ can be fed to several gates, and so we can get polynomially many copies of the bits $x_1, \ldots, x_n$ for free. However, quantum gates do not have fanout, hence this is not a free operation for quantum circuits. Hence, one needs to account for "copying" the input x and using t copies of the input increases the depth of the circuit by $O(\log t)$; so, creating $\mathsf{poly}(n)$ many copies of the input adds $O(\log n)$ depth to the quantum circuit.

The starting point for our work is the observation that we can avoid the $O(n^2)$ copies for a certain class of problems which are also hard to compute by transformers (which computationally form a subclass of TC^0). In particular, we generalize the candidate function of Chen et al. (2025), and show that this generalized function requires large width for constant-depth transformers computing it, but is computable by the "simplest" non-trivial class $\wedge \circ \mathsf{QNC}^0[\log\log n]$, i.e., an $O(\log\log n)$-depth QNC^0 circuit followed by a single classical AND gate.[5]

2.3 Proof sketch of the functional separation

Why not a relational separation? We first explain why the well-known *relational* separations used by Bravyi et al. (2018); Bene Watts et al. (2019); Grewal & Kumar (2024) do not imply separations between QNC^0 and transformers of the form that we want. To the best of our knowledge, almost all known existing techniques for proving lower bounds on transformer width are via proving communication complexity lower bounds (Sanford et al., 2023; Peng et al., 2024; Chen et al., 2025). In this communication model, the relevant resource is not the total number of bits of communication, but rather the *per-player bandwidth*, which is typically measured *proportional* to the size of that player's input. This creates a basic obstacle for relational problems: the player's output needs to be similar to the input (in order to have a well-formed problem), and if each player is allowed bandwidth proportional to their input length, then enough information can be communicated to solve the corresponding relational communication task in a straightforward way. Thus, the bandwidth regime in which we can hope for a nontrivial lower bound essentially disappears. In contrast, for functional problems, there are meaningful intermediate bandwidth regimes: the output is a single fixed bit, so it makes sense to ask for a bound on the amount of local information that must be transmitted in order to compute the target bit. This is why we focus on a functional separation instead of a relational one.

Our candidate and lower bound. In light of the discussion above, we turn towards functional problems, and in particular to the one considered by Chen et al. (2025), a variant of the *iterated index function*. Before discussing the iterated version, we first define the index function: on input $(i, \mathbf{x}) \in [n] \times \{0,1\}^n$, the goal is to output x_i. It is well-known that this function can be computed in AC^0; our goal is to express this circuit in terms of quantum gates, and in particular through weak QAC^0 circuits, which are QNC^0 together with *many-qubit* gates (potentially acting on n qubits) on top of the usual one- and two-qubit gates. There have been several recent works aimed at understanding the power of multi-qubit gates in computing well-known Boolean functions (Green et al., 2002; Rosenthal, 2021; Vasconcelos & Huang, 2025; Fenner et al., 2025; Gretta et al., 2026a;b; Joshi et al., 2026; Anshu et al., 2025; Bao & Escudero-Gutiérrez, 2026; Nadimpalli et al., 2024); however, a common concern acknowledged in these works is that it is *undesirable* to allow multi-qubit gates in quantum circuits, since they require coherent interaction between arbitrary qubits, which is incompatible with locality and noise constraints in physical architectures of current quantum devices. At this point, the following sequence of results yields our separation:

1. **Index with copies.** We first observe that the index function can be computed in QAC^0 assuming we have *copies* of the input. This is undesirable because: (i) we require copies and (ii) QAC^0 circuits can be trivially simulated by QNC^0 circuits of depth $O(\log n)$. At this point, we observe that one can overcome this trivial simulation as follows. We observe that the index function can be written as a $\mathsf{QNC}^0[\log\log n]$ circuit followed by a (large) AND gate at the end, however we still need copies of the input i in order to compute x_i.
2. **Generalizing to iterated index, still with copies.** Next, consider the *iterated index function* (as defined by Chen et al. (2025)) as follows:

$$\mathsf{IterInx} : [m] \times [N_1]^m \times [N_2]^{\log N_1} \times \cdots \times [N_\ell]^{\log(N_{\ell-1})} \to [N_\ell], \tag{1}$$

$$\mathsf{IterInx}(i, x^1, \ldots, x^\ell) = x^\ell(x^{\ell-1}(x^{\ell-2}(\cdots x^1(i))\cdots)), \tag{2}$$

for a pre-defined choice of parameters $m, N_1, N_2, \ldots, N_{\ell-1}$. In other words, we iteratively compute index on (i, x), look at $x_i \in [N_1]$, then evaluate x_2 on x_1 (where we view x_1 as a $(\log N_1)$-bit string) and repeat this iteratively for ℓ many functions. Using the previous bullet point wherein one can compute index using copies, we show that one can actually even compute $\mathsf{IterInx}$ also in $\wedge \circ \mathsf{QNC}^0[\log\log n]$ given copies of the input.

[5] We call this the *simplest* class here since we show in addition that, if a function is exactly computed by $\wedge \circ \mathsf{QNC}^0[\log\log n]$, then this function is actually computable by AC^0.

3. **Removing the multiple copies.** The next natural step is to remove the assumption of many copies of the input being given. Using a standard trick also used by Grier et al. (2026b), we first define the *multi-index* function as:

$$\mathsf{MultInx} : (i_1, \ldots, i_N, x) = x_i \cdot [i_1 = \cdots = i_N], \tag{3}$$

which is now a total function. Note that by definition, if $i_1 = \cdots = i_N$, then we effectively have access to copies of the input. However, we still need to check the condition $[i_1 = \cdots = i_N]$. We show that this indicator can be checked in $\wedge \circ \mathsf{QNC}^0[\log \log n]$ as well. Furthermore, the same circuit used for checking the index can in fact also compute x_i without increasing the complexity (our construction is subtle in that we ensure that the eventual circuit is not in $\wedge \circ \vee \circ \mathsf{QNC}^0[\log \log n]$, but just a single AND gate at the end of the $\mathsf{QNC}^0[\log \log n]$ circuit).

4. **Generalizing to a hard problem for transformers.** Now that we have a total function $\mathsf{MultInx}$, the goal is to prove a lower bound against transformers. As we observed earlier, this problem is *a priori* easy for decoder-only transformers. To make it hard for the transformer model, we define our final candidate function to be the *iterated multi-index* function, denoted $\mathsf{IterMultInx}$ (putting together the functional computation equation 2 along with the predicate of equation 3). Here, we have copies of truth tables $\{x_i^1, \ldots, x_i^\ell\}_{i \in [t]}$ (the $t = 1$ case is precisely the $\mathsf{IterInx}$ function) and with these copies of the truth table and index position, we show that $\mathsf{IterMultInx}$ can be computed in our desired quantum circuit class. A description of this final candidate is quite dense and technical, and so we refer the reader to Section B.2.2 for the details.

5. **Putting everything together.** The main challenge in putting everything together is keeping track of several parameters which need to be carefully chosen to witness our separation. The current outline is simplified slightly for ease of readability, but the eventual $\mathsf{IterMultInx}$ candidate is actually chosen with a slightly more contrived base index function which is also defined over a larger alphabet. With these technicalities in place, we show that:

   (a) the $\mathsf{IterMultInx}$ can be computed in $\wedge \circ \mathsf{QNC}^0[\log \log n]$;

   (b) the $\mathsf{IterMultInx}$ problem can be reduced to the $\mathsf{IterInx}$ problem, and the hardness for transformers can be therefore imported via the lower bound by Chen et al. (2025).

   We note here that our lower bound is polynomially weaker than the one in Chen et al. (2025), since our problem is more general.

Finally, we remark that our separation above is *essentially tight*. In particular, we can show that for deterministic protocols, any function f that is computed by a $\wedge \circ \mathsf{QNC}^0[\log \log n]$ circuit whose depth is at most $\log \log n$ (i.e., when the implicit constant in the $O(\log \log n)$ depth is at most 1) can also be computed in AC^0, a circuit class that essentially corresponds to transformers with *constant precision* (which is a strong assumption when it comes to transformer computations). Since our goal is to prove lower bounds against non-constant precision transformers, our quantum upper bound of $\wedge \circ \mathsf{QNC}^0[\log \log n]$ (where the implicit constant in the $O(\log \log n)$ depth is strictly greater than 1) allows us to obtain a function that is in QNC^0 but not computable by transformers unless their width is large, i.e., $Hdp = n^{\Omega(1)}$.

3 Distributional Separation

3.1 Diffusion models (a high level overview)

In this section, we give a high-level overview of DLMs, sufficient to understand the proof overview that follows. As with transformers in Section 2, we first give an overview of the model, with more details in Appendix A.3.

Diffusion language models (DLMs) are a central generative paradigm for tasks such as image and text generation. In this work, we will exclusively consider discrete diffusion models. DLMs are organized differently from transformers, and are tailor-made to generate samples from distributions. At a high level, DLMs generate samples by training models via the following process: they introduce random noise to training data, and then train the model to systematically remove the noise, so that the eventual model can generate new noiseless samples from the distribution. More formally,

instead of generating tokens strictly from left to right (as in transformers), a DLM starts with the so-called *masked* sequence $\{\mathsf{mask}\}^\star$ (which one can think of as maximally random) and iteratively *reveals/unmasks* tokens to produce elements from the alphabet Σ (for simplicity in this paper we let $\Sigma = \{0, 1\}$). So, at any point in the process, the DLM maintains a partially revealed sequence whose coordinates are of the form $\Sigma \cup \{\mathsf{mask}\}$, where Σ is the fixed token alphabet and mask is an auxiliary token. The sequence contains input coordinates and output coordinates, together with possible auxiliary *workspace* or *chain-of-thought* (CoT) coordinates. The computation itself begins with several masked coordinates, and proceeds through a number of *denoising rounds* which consists of two phases: In the first phase, a *scheduler* decides on the set of currently masked coordinates to be revealed, a decision that can adaptively depend on all the (non-mask) tokens generated so far, including workspace tokens; and in the second phase, a *denoiser* circuit is applied to each newly revealed coordinate, and a value is sampled.

The defining feature of a standard DLM is *within-round conditional independence*. Once a partially generated sequence is fixed, the coordinates updated in that round are sampled independently (but potentially dependent on the sequence at the beginning of the round), using the corresponding coordinate-wise denoising marginals. Thus, after conditioning on the visible transcript and the current workspace/CoT coordinates, a single denoising step induces a product distribution over the coordinates it updates. Global correlations can arise only through the iterative interaction of scheduling, workspace, and future denoising rounds. DLMs are powerful enough to build correlations adaptively, while imposing a strong structural constraint on what can be generated in a round.

3.2 Prior works and context

The starting point for our result is the work by Jiang et al. (2026), which studies DLMs as parallel samplers and analyzes the strengths and weakness of revision/remasking/chain-of-thought in DLM models. In particular, there is one striking feature in their work. They consider the distribution:

$$U_B^{\oplus} = \mathsf{unif}\{x \in \{0,1\}^B : |x| \equiv 0 \pmod 2\}.$$

They show that standard DLMs, without remasking or revision, cannot sample from this distribution in $O(1)$ steps when the predictor and scheduler are implemented by AC^0 circuits. However, their work cannot handle the scenario where the DLMs are augmented with a (bounded) number of chain-of-thought tokens, and a (bounded) number of token revision/remasking *events*. Note that both of these are features that modern DLMs crucially rely on. Although incomparable, the best known distribution sampling quantum-classical separation we know of is due to Bene Watts & Parham (2026) separating QNC^0 from NC^0; so it is natural to try and prove sampling lower bounds using the candidate problem from Jiang et al. (2026) against DLMs with AC^0 denoisers. Furthermore, several quantum/classical circuit separations (Bravyi et al., 2018; Bene Watts et al., 2019; Bene Watts & Parham, 2026) are in essence based on the observation that one can sample parity-based relations and distributions in QNC^0. With this in mind, our main contributions are as follows:

- (i) we *generalize* the candidate problem of Jiang et al. (2026) and define a *block-parity function*, which still admits an efficient QNC^0 protocol, but is provably hard for DLMs;
- (ii) using the generalized candidate, we are able to prove lower bounds even when we allow *revision/remasking and chain-of-thought* (with the caveat that we only allow at most sublinear number of tokens for these augmented features);
- (iii) our lower bounds allow GC^0 *denoisers*[6], which are significantly stronger than AC^0 denoisers;
- (iv) we show hardness even to *approximately sample* from the new distribution. Along the way, we show several structural properties of DLMs that might be of independent interest.

The main technical contribution of our work is the classical lower bound, which combines several ideas from theoretical computer science. First, we use a direct-product style amplification to turn many local parity constraints into a global obstruction. Second, we view a DLM execution as a query-like process whose visible behavior, after hiding the workspace/CoT coordinates, can be expressed as a convex combination of simple decision-tree-like distributions. Finally, we prove that block-parity functions cannot be approximated by such small convex combinations. We make this candidate problem and proof structure clear in the next section.

[6] We do not define the class GC^0 here and refer the reader to Section A.1; we simply note that it is a classical circuit complexity class that is strictly more powerful than AC^0.

3.3 PROOF SKETCH OF THE DISTRIBUTIONAL SEPARATION

We now give a proof sketch of our DLM lower bound. We wish to sample from the distribution:

$$D_{\text{blk}} = \bigotimes_{b=1}^{M} U_B^{\oplus} = \{(x^1, \ldots, x^M) : x^i \in \{0,1\}^B, |x^i| \equiv 0 \pmod 2 \text{ for all } i\}$$

within total variation distance η for some fixed constant $\eta > 0$. The quantum algorithm to sample from this is easy, and the main ideas have appeared before (Bravyi et al., 2018; Bene Watts et al., 2019; Grewal & Kumar, 2024). We now sketch the classical lower bound. Assume towards a contradiction that there exists a constant-round standard DLM which samples from D_{blk}, with workspace length s, fully adaptive scheduling, bounded revision/remasking, and polynomial-size constant depth $\mathsf{GC}^0[\log n]$ scheduling and denoising circuits.

For each block $b \in [M]$, define its *completion round* to be the first round after which all B output coordinates in that block have been generated. Let r_b be the number of previously ungenerated coordinates of block b that are generated in this round. Each block is completed in one of two ways.

- $r_b \geq 2$, i.e., the DLM completes the block by unmasking many fresh coordinates at once;
- $r_b = 1$, i.e., the DLM completes the block by generating a single final coordinate, with the remaining $B - 1$ coordinates having already been fixed in previous rounds.

The proof is a win-win argument over these two completion modes. Suppose first that many blocks are completed with $r_b \geq 2$. For such a block, conditioned on the coordinates revealed earlier in the block, the r_b fresh coordinates are not freely distributed: under the target distribution, they must satisfy one affine parity constraint. Thus a round that completes many such blocks must simultaneously realize many independent within-block parity constraints. We prove a structural theorem below showing that this is impossible for the conditional laws produced by a bounded-workspace DLM. Hence, it follows that only few blocks can be completed in the multi-bit way and for most blocks, we must have $r_b = 1$. But then the DLM is forced into a "last-bit prediction" regime: when it completes such a block, the final coordinate has to be the parity of the previous $B - 1$ coordinates. This can be used to show that a shallow AC^0 or GC^0 circuit predicts parity with constant bias, contradicting the usual parity lower bound of Håstad (2014) or Kumar (2023).

Formally, the argument proceeds in four main steps.

1. **Mixtures of product distributions induced by DLMs.** Fix a round t, and condition on the revealed output history τ_{out}, i.e., the revealed output coordinates for the final sample together with their values, the scheduling history up to round t, and the previously generated workspace/chain-of-thought tokens. Let w denote the revealed current workspace so far. Now, using the independence properties of DLMs, we have that the output distribution of the t-th round conditioned on τ_{out} and w is given by:

$$P(\text{output of the } t\text{-th round} \mid \tau_{\text{out}}, w) = Q_{\tau_{\text{out}}, w},$$

where $Q_{\tau_{\text{out}}, w}$ is a product distribution over the coordinates revealed in round t. Hence, averaging over the possible current workspaces gives:

$$P(\text{output of the } t\text{-th round} \mid \tau_{\text{out}}) = \sum_w \Pr[w \mid \tau_{\text{out}}] \cdot Q_{\tau_{\text{out}}, w}. \tag{4}$$

Since the token alphabet has constant size, the workspace contains at most s tokens, and so there are at most $2^{O(s)}$ possible values of w. Consequently, conditioned on any visible transcript, a denoising round induces only a small latent mixture of product distributions.

2. **Multi-bit block completions.** Now, fix a transcript branch and consider a block b completed in some round with $r_b \geq 2$. Under the target distribution $U_B^{\oplus}$, the full block satisfies $\bigoplus_{i=1}^{B} x_{b,i} = 0$. Once some coordinates of the block have already been revealed, the remaining unrevealed coordinates are still uniform subject to the residual parity constraint determined by the revealed prefix. Therefore, if the DLM completes the block by generating r_b fresh coordinates simultaneously, the conditional law for these coordinates is:

$$U_{r_b}^{a_b} := \text{unif}\left\{ x \in \{0,1\}^{r_b} : \bigoplus_i x_i = a_b \right\},$$

where a_b depends on the previously revealed coordinates and the parity of their respective coordinates. These distributions are not products over coordinates: the fresh coordinates inside the block must jointly satisfy a parity condition. Since D_{blk} is product across blocks, if many blocks are completed simultaneously, then the target conditional law becomes:

$$V = \bigotimes_b U_{r_b}^{a_b}, \tag{5}$$

namely, a product across blocks of internally parity-constrained factors.

3. **Inapproximability of equation 5.** Our main structural result is that a distribution of the form given in equation 5 cannot be approximated by a small mixture of coordinate-wise product distributions. Quantitatively, if equation 4, written abstractly as $P = \sum_{h=1}^{K} \lambda_h Q_h$, is a mixture of product distributions satisfying $\mathsf{TV}(P, V) \le \eta$, then we necessarily have:

$$\log K \ge \Omega\left(\sum_b (r_b - 1)\right).$$

Intuitively, each block completed with $r_b \ge 2$ contributes an independent intra-block parity constraint: the r_b fresh coordinates must lie in an affine parity coset, i.e., satisfy $\oplus_i x_i = a_b$, rather than in a product distribution. A small latent mixture cannot realize many such constraints simultaneously. Quantitatively, if a round completes a collection of blocks with completion sizes $\{r_b\}$, then approximating the corresponding conditional law requires at least $2^{\Omega(\sum_b (r_b - 1))}$ product components. On the other hand, after conditioning on the visible transcript, the only hidden information is the workspace/CoT coordinates, which gives at most $K = 2^{O(s)}$ mixture components. Hence, we have $\sum_b (r_b - 1) = O(s)$ in any typical denoising round. In particular, since every block with $r_b \ge 2$ contributes at least 1 to this sum, only $O(s)$ blocks can be completed by simultaneous multi-bit generation over all (constantly many) rounds. Therefore, when $s = o(M)$, almost every block must fall in the category where $r_b = 1$. Hence, we have reduced the task of DLM approximating D_{blk} to the task of predicting a parity bit, which will be the basis for our lower bound.

4. **Reduction of final-bit completion to parity prediction.** Suppose now that a block satisfies $r_b = 1$. At its completion round, the DLM has already generated $B - 1$ coordinates and must generate the last one. Under $U_B^{\oplus}$, this last coordinate is forced:

$$X_{b,j} = \bigoplus_{i \ne j} X_{b,i}.$$

Thus, a successful DLM must predict a parity bit. Technically, if the previous $B - 1$ bits generated by the DLM were uniformly random, then we are done since we can use known lower bounds of $\mathsf{GC}^0, \mathsf{AC}^0$ computing parity. However, the first $B - 1$ bits were generated by the DLM itself, while the parity lower bound is for uniform external inputs. This is the most subtle part of the proof, and showing this requires us to account for the entropy of various blocks along the DLM process and show that there exists at least one block among the M blocks in which the pre-completion $B - 1$ bits have large entropy. In this block $b^\star$, we can use the probabilistic method to show that the pre-completion bits are already *close* to uniform, which means that we can replace these bits by a set of independent uniform bits, This changes the execution distribution by only $o(1)$ in total variation, while leaving the final completion round unchanged. After this replacement, the final generated bit predicts the parity of a uniform B-bit input with constant advantage. Since the planted execution can be simulated by a shallow $\mathsf{GC}^0[\log n]$ circuit, we obtain a shallow circuit that computes parity on $B - 1$ bits with constant advantage, contradicting the known parity lower bounds for $\mathsf{GC}^0[\log n]$ circuits (Kumar, 2023; Grewal & Kumar, 2024).

5. **Handling Remasking/Revision.** Finally, we extend the argument to bounded remasking/revision. We observe that this case requires a small modification of the entropy arguments. We cannot unfortunately simply restrict our attention to untouched blocks, since the set of touched blocks may be correlated with the generated values. Instead, we account for the entropy argument by adjusting the probability that it is revised after completion. Since there are at most R revision/remasking events, this adds only an $o(M)$ correction for our

parameters. In fact, the same win-win and planting argument as before goes through now. Each output-token remasking/revision event touches exactly one coordinate, and hence belongs to one block. Therefore, at most R distinct blocks can be touched by all such events. On the remaining blocks, the execution is equivalent to one without output-token remasking/revision, and the previous argument applies. Restricting to the untouched blocks eliminates revision entirely, and the previous argument applies verbatim. Thus the lower bound survives as long as $s + R = o(M)$, which indeed motivates our main choice of parameters.

4 Outlook and Further Directions

We see this work as initiating the task of separating quantum computation and classical LLMs. Naturally, there are several questions arising from this work that are worth further investigating. We hope that this serves as a launching pad for substantial future work from the community.

1. **Strengthening the separations.** Is there a candidate problem for which we can also show DLM-hardness, even in the presence of linear chain-of-thought, and remasking and/or revision? In the case of transformers, we provide one reason why a more typical relational separation (where the input and output length are the same) seems essentially impossible to obtain, but this does not preclude more contrived search-based or relational separations between QNC^0 and decoder-only transformers. Finally, since transformers (without chain-of-thought) are known to be inside TC^0, it is conceivable that one can obtain a width-$\Omega(n)$ lower bound without that implying a complexity-theoretic breakthrough.
2. **Randomized separation.** This work is only concerned with deterministic functions. If we move to randomized problems (where the output needs to be correct with probability $\geq 2/3$, say), can we also show separations, between, e.g., $\wedge \circ \mathsf{QNC}^0$ and transformers?
3. **Quantum-augmented DLMs.** In the model we consider here, new bits are generated using AC^0 or GC^0 denoisers. What if we used quantum circuits to generate new bits? Could such quantum-augmented DLMs be more powerful than using DLMs with classical-circuit based denoisers? Is there a deep connection between quantum dynamics and DLMs?[7]
4. **Copy complexity of transformers.** As mentioned, a recent work of Grier et al. (2026b) shows how to compute TC^0 functions using QAC^0, along with $O(n^2)$ copies of the input. Since transformers can be shown to roughly interpolate between AC^0 and TC^0 (depending on the precision and the type of attention used), it is not inconceivable that computing these functions would require fewer copies for the QAC^0 circuit. Could this be formalized?
5. **Architecture-independent separations.** Our functional and distributional separations rely on different structural limitations of transformers and diffusion language models. Is there a single task solvable by shallow quantum circuits for which every bounded-resource classical language model—whether autoregressive, diffusion-based, or a hybrid of the two—must incur a polynomial cost in at least one of width, chain-of-thought/workspace, sequential generation rounds, or token revision?

Final remarks. Finally, we make a couple of important remarks on our results to contextualize our work fairly. The separations that we exhibit here are *entirely* theoretical. Our motivation is to understand how quantum algorithms could be better than classical LLMs at certain ad hoc tasks that are *specifically* designed to show these separations. Secondly, state-of-the-art AI models rely on significant amounts of chain-of-thought (Wei et al., 2022), which is empirically known to improve the ability to reason by forcing the LLM to perform step-by-step "scratchpad"-style reasoning, which has been shown to yield better accuracy and more expressivity in practice. Our transformer and DLM lower bounds are proven under the assumption of *no* chain-of-thought and *sublinear* chain-of-thought respectively, which are very strong assumptions; the lower bounds do not hold without these restrictions in place, and indeed it is entirely conceivable that our problems are easy for classical LLMs with a reasonable amount of chain-of-thought. The goal of this work, rather, is to initiate the study of the interplay between quantum circuits and classical machine learning models, which we view as being timely and important in the era of large language models.

[7] A recent work of Layden et al. (2025) showed that the dynamics of *continuous*-model DLMs can be efficiently simulated on a quantum computer.

AI USE STATEMENT

We used GPT 5.5 for help in proving Theorem C.4 and the subsequent corollary (a helper result for the DLM lower bound). We also used GPT 5.5 for some very minor corrections in the parameterization of the transformer lower bound, which were since then strengthened by hand. The presentation here has been rewritten and checked by the authors (any further mistakes are our own).

REFERENCES

Anurag Anshu, Yangjing Dong, Fengning Ou, and Penghui Yao. On the computational power of QAC^0 with barely superlinear ancillae. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1476–1487. Association for Computing Machinery, 2025. doi: 10.1145/3717823.3718189. URL https://doi.org/10.1145/3717823.3718189. 4

Jinge Bao and Francisco Escudero-Gutiérrez. Learning junta distributions, quantum junta states, and QAC^0 circuits. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026. URL https://arxiv.org/abs/2410.15822. To appear. 4

Adam Bene Watts and Natalie Parham. Unconditional quantum advantage for sampling with shallow circuits. In Shubhangi Saraf (ed.), *17th Innovations in Theoretical Computer Science Conference (ITCS 2026)*, volume 362 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 17:1–17:12. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026. doi: 10.4230/LIPIcs.ITCS.2026.17. URL https://doi.org/10.4230/LIPIcs.ITCS.2026.17. 1, 6

Adam Bene Watts, Robin Kothari, Luke Schaeffer, and Avishay Tal. Exponential separation between shallow quantum circuits and unbounded fan-in shallow classical circuits. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 515–526. Association for Computing Machinery, 2019. doi: 10.1145/3313276.3316404. URL https://doi.org/10.1145/3313276.3316404. 1, 3, 4, 6, 7, 38

Sergey Bravyi, David Gosset, and Robert König. Quantum advantage with shallow circuits. *Science*, 362(6412):308–311, 2018. doi: 10.1126/science.aar3106. URL https://doi.org/10.1126/science.aar3106. 1, 4, 6, 7

Sergey Bravyi, David Gosset, Robert König, and Marco Tomamichel. Quantum advantage with noisy shallow circuits. *Nature Physics*, 16(10):1040–1045, 2020. doi: 10.1038/s41567-020-0948-z. URL https://doi.org/10.1038/s41567-020-0948-z. 1

Lijie Chen, Binghui Peng, and Hongxun Wu. Theoretical limitations of multi-layer transformer. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 2631–2653. IEEE, 2025. doi: 10.1109/FOCS63196.2025.00136. URL https://doi.org/10.1109/FOCS63196.2025.00136. 2, 3, 4, 5, 16, 17, 18, 25, 26, 33, 34, 35, 50

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7654–7664. Association for Computational Linguistics, May 2022. doi: 10.18653/v1/2022.acl-long.527. URL https://aclanthology.org/2022.acl-long.527/. 3

David Chiang, Peter Cholak, and Anand Pillay. Tighter bounds on the expressivity of transformer encoders. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5544–5562. PMLR, 2023. URL https://proceedings.mlr.press/v202/chiang23a.html. 3

Matthew Coudron, Jalex Stark, and Thomas Vidick. Trading locality for time: Certifiable randomness from low-depth circuits. *Communications in Mathematical Physics*, 382(1):49–86, 2021. doi: 10.1007/s00220-021-03963-w. URL https://doi.org/10.1007/s00220-021-03963-w. 1

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991. ISBN 0-471-06259-6. Chapter 2, "Entropy, Relative Entropy and Mutual Information," pp. 12–49. 14

Michael de Oliveira, Sathyawageeswar Subramanian, Leandro Mendes, and Min-Hsiu Hsieh. Unconditional advantage of noisy qudit quantum circuits over biased threshold circuits in constant depth. *Nature Communications*, 16(1):3559, 2025. 1

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. Transformer Circuits Thread, December 2021. URL https://transformer-circuits.pub/2021/framework/index.html. 3

Stephen A. Fenner, Daniel Grier, Daniel Padé, and Thomas Thierauf. Tight bounds on depth-2 QAC-circuits computing parity, 2025. URL https://arxiv.org/abs/2504.06433. 4

Gemini Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530. 3

Frederic Green, Steven Homer, Cristopher Moore, and Christopher Pollett. Counting, fanout, and the complexity of quantum ACC. *Quantum Information & Computation*, 2(1):35–65, 2002. doi: 10.26421/QIC2.1-3. URL https://doi.org/10.26421/QIC2.1-3. 4

Lucas Gretta, Meghal Gupta, and Malvika Raj Joshi. Parity $\notin$ QAC^0 iff QAC^0 is Fourier-concentrated. In *2026 IEEE 67th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2026a. URL https://arxiv.org/abs/2604.02793. 4

Lucas Gretta, Meghal Gupta, and Malvika Raj Joshi. Super-constant weight Dicke states in constant depth without fanout, 2026b. URL https://arxiv.org/abs/2604.15298. 4

Sabee Grewal and Vinayak M. Kumar. Improved circuit lower bounds and quantum-classical separations, 2024. URL https://eccc.weizmann.ac.il/report/2024/130. ECCC: TR24-130. 1, 3, 4, 7, 8, 15

Daniel Grier and Luke Schaeffer. Interactive shallow Clifford circuits: Quantum advantage against NC^1 and beyond. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 875–888. Association for Computing Machinery, 2020. doi: 10.1145/3357713.3384332. URL https://doi.org/10.1145/3357713.3384332. 1

Daniel Grier, Daniel M. Kane, Jackson Morris, Anthony Ostuni, and Kewen Wu. Quantum advantage from sampling shallow circuits: Beyond hardness of marginals. In Shubhangi Saraf (ed.), *17th Innovations in Theoretical Computer Science Conference (ITCS 2026)*, volume 362 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 73:1–73:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026a. doi: 10.4230/LIPIcs.ITCS.2026.73. URL https://doi.org/10.4230/LIPIcs.ITCS.2026.73. 2

Daniel Grier, Jackson Morris, and Kewen Wu. QAC^0 contains TC^0 (with many copies of the input). In *2026 IEEE 67th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2026b. URL https://arxiv.org/abs/2601.03243. To appear. 3, 5, 9, 24

Johan Håstad. On the correlation of parity and small-depth circuits. *SIAM Journal on Computing*, 43(5):1699–1708, 2014. doi: 10.1137/120897432. URL https://doi.org/10.1137/120897432. 7, 38

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4-Abstract.html. 2

Haozhe Jiang, Nika Haghtalab, and Lijie Chen. Diffusion language models are provably optimal parallel samplers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5bkAbueJwM. 2, 6, 19, 21

Malvika Raj Joshi, Avishay Tal, Francisca Vasconcelos, and John Wright. Improved lower bounds for QAC^0. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing*, pp. 2199–2209. Association for Computing Machinery, 2026. doi: 10.1145/3798129.3800922. URL https://doi.org/10.1145/3798129.3800922. 4

Phokion Kolaitis and Rik Sengupta. On the expressive power of transformers, 2026. URL https://arxiv.org/abs/2608.12671. 16

Vinayak M. Kumar. Tight correlation bounds for circuits between AC^0 and TC^0. In Amnon Ta-Shma (ed.), *38th Computational Complexity Conference (CCC 2023)*, volume 264 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 18:1–18:40. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi: 10.4230/LIPIcs.CCC.2023.18. URL https://doi.org/10.4230/LIPIcs.CCC.2023.18. 7, 8, 15, 38, 39

David Layden, Ryan Sweke, Vojtěch Havlíček, Anirban Chowdhury, and Kirill Neklyudov. Wavefunction flows: Efficient quantum simulation of continuous flow models, 2025. URL https://arxiv.org/abs/2510.08462. 9

François Le Gall. Average-case quantum advantage with shallow circuits. In Amir Shpilka (ed.), *34th Computational Complexity Conference (CCC 2019)*, volume 137 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 21:1–21:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi: 10.4230/LIPIcs.CCC.2019.21. URL https://doi.org/10.4230/LIPIcs.CCC.2019.21. 1

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3EWTEy9MTM. 3, 17, 18

Meta AI. Introducing Meta Llama 3: The most capable openly available LLM to date. Meta AI Blog, April 2024. URL https://ai.meta.com/blog/meta-llama-3/. 3

Shivam Nadimpalli, Natalie Parham, Francisca Vasconcelos, and Henry Yuen. On the Pauli spectrum of QAC_0. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 1498–1506. Association for Computing Machinery, 2024. doi: 10.1145/3618260.3649662. URL https://doi.org/10.1145/3618260.3649662. 4

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 50608–50646. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/48b383b24230e0e6e649d9c98dae4d8c-Abstract-Conference.html. 19, 21

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. On limitations of the transformer architecture. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=KidynPuLNW. 4

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. 3

Gregory Rosenthal. Bounds on the QAC^0 complexity of approximating parity. In James R. Lee (ed.), *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*, volume 185 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 32:1–32:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. doi: 10.4230/LIPIcs.ITCS.2021.32. URL https://doi.org/10.4230/LIPIcs.ITCS.2021.32. 4

Clayton Sanford, Daniel J. Hsu, and Matus J. Telgarsky. Representational strengths and limitations of transformers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36677–36707. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/73bf692447f174984f30499ec9b20e04-Abstract-Conference.html. 3, 4, 16

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html. 2

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, Yuwei Fu, Jing Su, Ge Zhang, Wenhao Huang, Mingxuan Wang, Lin Yan, Xiaoying Jia, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Yonghui Wu, and Hao Zhou. Seed diffusion: A large-scale diffusion language model with high-speed inference, 2025. URL https://arxiv.org/abs/2508.02193. 19, 21

Francisca Vasconcelos and Hsin-Yuan Huang. Learning shallow quantum circuits with many-qubit gates. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5553–5604. PMLR, 2025. URL https://proceedings.mlr.press/v291/vasconcelos25a.html. 4

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. 2

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html. 9

Andy Yang, David Chiang, and Dana Angluin. Masked hard-attention transformers recognize exactly the star-free languages. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10202–10235. Curran Associates, Inc., 2024. doi: 10.52202/079017-0327. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/13d7f172259b11b230cc5da8768abc5f-Abstract-Conference.html. 3