Feedforward Mixing is as Sharp as it is Slow in Reverse

Benedict Aaron Tjandra [*1] **Avi Wigderson** [*2] **João G.M. Araújo** [1] **Alex Vitvitskyi** [1] **Federico Barbero** [1] **Petar Veličković** [1]

Abstract

Deep learning architectures across diverse domains fundamentally rely on feedforward computational graphs. Consequently, optimising these graph topologies is a major research focus. Recent theoretical work evaluates the efficacy of these graphs using two key metrics: forward mixing time $T_{fwd}(G)$ (the speed of information propagation) and normalised minimax fidelity $F(G)$ (the sharpness of information representation). In this work, we prove that, surprisingly, minimax fidelity is tightly upper-bounded by the *backward* mixing time of computational graphs: $F(G) \leq \frac{8}{3} \cdot T_{bwd}(G)$ for all feedforward graphs with a unique source and sink. Practically, this establishes a direct tradeoff, $F(G) \leq \frac{8}{3} \cdot T_{fwd}(G)$, for symmetric graphs where $T_{fwd}(G) = T_{bwd}(G)$. In the uniform setting, because most common architectures—including fully-connected and sliding-window graphs—are symmetric, this exposes an unavoidable architectural limit: optimising for rapid information mixing necessitates a degradation in fidelity, and vice versa. More generally, i.e. if $T_{bwd}(G) = f(T_{fwd}(G))$ and consequently $F(G) \leq f(T_{fwd}(G))$ for some function f, we show that f is well-behaved and at most an $O(n)$ and $O(n(1 + d/c))$ factor of $T_{fwd}(G)$ in uniform and non-uniform graphs respectively, where n is the number of nodes and c and d are constants that bound the weighted adjacency matrix. Our results, ultimately, establish backward mixing time as a key metric for evaluating information propagation in deep learning models.

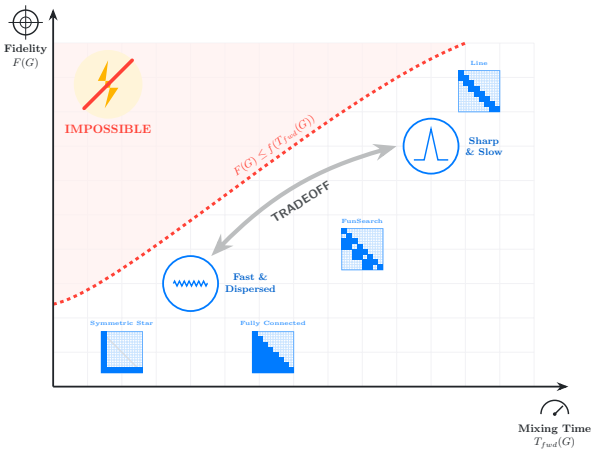


Figure 1. More generally, the normalised minimax fidelity $F(G)$ is upper-bounded by a function of the mixing time $T_{fwd}(G)$. This implies that, across the design space dictated by minimax fidelity and mixing time, feedforward computational graphs lie on or below the red boundary.

1. Introduction

Many problems in deep learning necessitate the processing of sequential inputs (Vitvitskyi et al., 2025). Language modelling (Sutskever et al., 2014), audio and video processing (Wiedemer et al., 2025; van den Oord et al., 2016), time-series forecasting (Kim et al., 2025), and temporal graph modelling (Huang et al., 2023) all employ causal processing of their inputs, and at their core, the forward passes of these sequence models can be abstracted as feedforward computational graphs. Significant research effort has been dedicated to optimising computational graph topologies to facilitate information flow, albeit most of the effort has happened in the *undirected* graph space. Notable examples include rewiring strategies in graph neural networks (GNNs) to avoid over-squashing and under-reaching (Attali et al., 2024; Giovanni et al., 2023; Topping et al., 2022; Deac et al., 2022; Wilson et al., 2024), and the development of dynamic, data-dependent attentional masks (Xia et al., 2022) for vision Transformers. In the feedforward case, several works have designed static attentional masks for decoder-only Transformers (Beltagy et al., 2020; Chen et al., 2025; Child et al., 2019).

Given the ubiquity of computational feedforward graphs, understanding the theoretical limits of their design is paramount. To formally analyse this design space, (Vitvitskyi et al., 2025) introduced two metrics to quantify the efficacy of information propagation: forward *mixing time* (how rapidly information traverses the computational graph; see Espuny Díaz et al. (2024)) and normalised minimax *fidelity* (how distinctly information from any source node is preserved at the sink). Intuitively, an ideal computational graph would simultaneously achieve **low mixing time** (fast global propagation) and **high minimax fidelity** (sharp, distinguishable representations).

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

In this paper, we prove that normalised minimax fidelity is tightly upper-bounded by the **backward** mixing time of computational graphs: $F(G) \leq \frac{8}{3} \cdot T_{bwd}(G)$ for all feedforward graphs with a unique source and a unique sink. Practically, this establishes a direct tradeoff, $F(G) \leq \frac{8}{3} \cdot T_{fwd}(G)$, specifically for graphs that are both **symmetric** (equivalent up to isomorphism when reversed; Section 5.2) and **uniform** (where edge weights are assigned uniformly based on in-degree and out-degree). For symmetric uniform graphs, which include fully-connected and sliding-window graphs, this exposes an unavoidable architectural limit: optimising for rapid information mixing necessitates a degradation in fidelity, and vice versa. In the more general case, i.e. if $T_{bwd}(G) = f(T_{fwd}(G))$ and consequently $F(G) \leq f(T_{fwd}(G))$ for some function f, we show that $T_{bwd}(G)$ and $T_{fwd}(G)$ diverge by at most an $O(n)$ factor for uniform graphs and $O(n(1 + d/c))$ for non-uniform graphs whose weights are bounded between c and d. With these findings, our results establish backward mixing time as a key metric for evaluating information propagation. We summarise our core contributions below:

- **A Duality Between Fidelity and Backward Mixing.** We first establish an equivalence between the forward diffusion of representations and a random walk on the reversed computational graph.
- **Fidelity is Tightly Bounded by the Backward Mixing Time.** Using this duality, we prove our central result: minimax fidelity is tightly upper-bounded by the graph's worst-case backward mixing time for graphs with a unique source and a unique sink. Furthermore, we identify specific topologies (conjoined star graphs) to demonstrate that this bound is tight.
- **Limits on Forward Mixing and Fidelity.** We connect our primary bound back to the forward mixing time $T_{fwd}(G)$. We make the observation that for symmetric graphs—encompassing a wide range of practical architectures—forward and backward mixing times are identical, exposing a tight $F(G) \leq \frac{8}{3} \cdot T_{fwd}(G)$ tradeoff. More generally, $F(G) \leq f(T_{fwd}(G))$ holds, wherethe bounding function f is governed by $f(x) = O(n) \cdot x$ and $f(x) = O(n(1+\frac{d}{c})) \cdot x$ for uniform and non-uniform graphs respectively, where n is the size of the graph G.
- **Empirical Demonstration.** We plot the worst-case forward mixing time and minimax fidelity across a wide range of standard attentional masks in the uniform graph setting, demonstrating that our bounds hold empirically. Moreover, we show that this effect holds in a range of open-source Large Language Models (LLMs) of different sizes, namely the Qwen 3 (Yang et al., 2025), Llama 3 (Grattafiori et al., 2024), Gemma 3 (Gemma et al., 2025), and Gemma 4 (Google DeepMind, 2026) model series.

2. Related Work

The theoretical framing of information flow in deep learning architectures has gained traction as a means to explain empirical bottlenecks. Our work mainly builds upon (Vitvitskyi et al., 2025), who introduced mixing time and minimax fidelity as rigorous metrics to quantify the propagation characteristics of feedforward computational graphs. Building on this framework, (de Ocáriz Borde, 2025) employed these metrics to analyse multi-head attention mechanisms in transformers. Surprisingly, such frameworks are largely unstudied in mathematics, aside from expander graphs (Csóka & Grabowski, 2022) and the spectral analysis of Stanković et al. (2025). (Herasimchyk et al., 2026) and (Chowdhury, 2026) used an *influence* metric to analyse the Lost-In-The-Middle effect in transformers, which is closely related to the normalised minimax fidelity (Barbero et al., 2024).

3. Background

To formalise the study of architecture design in sequence processing, we model the forward pass of a neural network as a directed acyclic graph (DAG) augmented with self-loops, referred to as a feedforward computational graph $G = (\mathcal{V}, \mathcal{E})$. The nodes $\mathcal{V}$ store the feature representations at various layers, and the directed edges $\mathcal{E}$ dictate the flow of computation. Further, we assume that the nodes in $\mathcal{V}$ are topologically sorted such that $\mathcal{V} = \{1, \ldots, |\mathcal{V}|\}$ where node 1 is the unique source and node $|\mathcal{V}|$ is the unique sink n. If a directed edge exists from i to j, then $i \leq j$. The graph is defined by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, where A_{ij} is the directional connection weight from node j to node i.

Following (Vitvitskyi et al., 2025), our analysis relies on certain structural assumptions on G:

- **Self-Loops.** Every node possesses a self-loop, i.e., $(i, i) \in \mathcal{E}$ for all $i \in \mathcal{V}$, representing the property that a node can preserve its own state across computational steps.
- **Unique Sink.** There exists a unique sink node, denoted as $n \in V$, which is the only node with an out-degree of one (its self-loop). This represents the final output representation of the network.
- **Unique Source.** Unlike (Vitvitskyi et al., 2025), we omit the self-similarity assumption and replace it

with a requirement that the graph has a defined starting point. We assume there exists a unique *source*, representing a unified starting point for the flow of information. Consequently, due to the presence of self-loops, the source is the only node with in-degree one.

Below, we recap the (averaged) mixing time and (normalised) minimax fidelity framework introduced by (Vitvitskyi et al., 2025).

3.1. Quantifying Information Propagation Speed: Mixing Time

Let $\mathbf{W} = \text{ColNorm}(\mathbf{A}) \in \mathbb{R}^{n\times n}$ be the column-normalised forward walk matrix, where w_{uv} is the u-th row and v-th column entry of $\mathbf{W}$ and represents the probability that a random walker at node v transitions to node u. Similarly, w^t_{uv} represents that a random walker at node v transitions to node u in t steps. The forward mixing time $T_{fwd}(G)$ is defined as the minimum number of steps t from *any* starting probability distribution $\mathbf{x}$ to converge sufficiently close to the stationary distribution $\boldsymbol{\pi}$, which in our case is one-hot in the unique sink n:

$$T_{fwd}(G) = \min\left\{t \geq 0 : \max_{\mathbf{x}} \left[||\mathbf{W}^t\mathbf{x} - \boldsymbol{\pi}||_1\right] < \frac{1}{2}\right\}$$

We note that while (Vitvitskyi et al., 2025) analyse *averaged* mixing times (Espuny Díaz et al., 2024)—averaging the minimum steps t required to reach the stationary distribution across all starting nodes–our analysis adopts a *worst-case* analysis. This worst-case formulation (represented by the maximisation over $\mathbf{x}$) establishes higher parity with standard Markov chain literature.

3.2. Quantifying Information Sharpness: Normalised Minimax Fidelity

Let $\boldsymbol{\Delta} = \text{RowNorm}(\mathbf{A}) \in \mathbb{R}^{n\times n}$ be the row-normalised diffusion matrix, where Δ_{ij} corresponds to the proportion of influence a prior node j exerts on node i during a single step of information mixing. Over t timesteps, the fidelity for a node i, Δ^t_{ni}, quantifies the total accumulated portion of information from a source node i that successfully reaches the sink n.

The maximal fidelity for a given node i, defined as $\phi_i = \max_t \Delta^t_{ni}$ represents the highest concentration of its features captured at the sink over the entire duration of the diffusion. To capture the robustness bottleneck of the entire graph, the minimax fidelity computes the minimum of these maximum fidelities across all nodes, normalised by the size of the graph n:

$$F(G) = n \cdot \min_{i\in V} \max_t \Delta^t_{ni}$$

Intuitively, this metric identifies the "weakest link" in the graph: the node whose best-case representation at the sink is the most attenuated.

3.3. Edge Weighting Schemes: Uniform vs. Non-Uniform Graphs

Having established the general metrics for evaluating a computational graph, we distinguish between two classes of graphs based on how nodes aggregate incoming information. This categorisation defines how the diffusion matrix $\boldsymbol{\Delta}$ and the random walk matrix $\mathbf{W}$ are populated.

- **Uniform Graphs:** The graph is defined by a binary adjacency matrix where $A_{ij} \in \{0, 1\}$, indicating the absence or presence of an edge. Let $d_{\to i}$ denote the in-degree of node i, and let $d_{j\to}$ denote the out-degree of node j. The diffusion matrix uniformly distributes incoming influence, while the forward walk matrix distributes outgoing transition probabilities equally. For any edge $(j, i) \in \mathcal{E}$:

$$\Delta_{ij} = \begin{cases} 1/d_{\to i} & \text{if } (j,i) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

$$w_{ij} = \begin{cases} 1/d_{j\to} & \text{if } (j,i) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

- **Non-Uniform Graphs:** Nodes assign arbitrary, data-dependent weights to their connections (e.g., standard attention mechanisms). In this setting, the graph is defined by a weighted adjacency matrix A, where $\boldsymbol{\Delta} = \text{RowNorm}(\mathbf{A})$ and $\mathbf{W} = \text{ColNorm}(\mathbf{A})$ as before. $\mathbf{A}$ can therefore be taken as the unnormalised attention weights in transformers. To reflect the natural architectural and numerical constraints of real transformers, we assume non-uniform graphs are (c, d)-constrained, meaning that for every valid edge $(j, i) \in \mathcal{E}$, its weight satisfied $c \leq A_{ij} \leq d$ for positive constants $0 < c \leq d$. This assumption is justified, because raw attention logits are bounded by some interval $[-M, M]$ because of the representation of floating points in machines. Their exponentiation $[e^{-M}, e^M]$ must therefore be bounded away from absolute zero and finite.

4. Equivalence of Fidelity and the 'Backwards' Random Walk

Before presenting our main theoretical bounds, we first establish an equivalence between information diffusion in a feedforward graph and a random walk on its reversed counterpart. Importantly, the results in this section hold for both uniform graphs and non-uniform graphs with arbitrary edge weightings.

Definition 4.1. Let $G = (\mathcal{V}, \mathcal{E})$ be a computational feedforward graph. Define $\tilde{G} = (\mathcal{V}, \tilde{\mathcal{E}})$ to be its reversed counterpart, where nodes are relabeled such that node i in G maps to node $n - i + 1$ in $\tilde{G}$. The edges are correspondingly reversed: $\forall i, j : (i, j) \in \mathcal{E} \iff (n - j + 1, n - i + 1) \in \tilde{\mathcal{E}}$. Let $\tilde{\mathbf{W}} = \text{ColNorm}(\tilde{\mathbf{A}})$ be the column-normalised random walk matrix of its reversed counterpart $\tilde{G}$ and $r(i) = n - i + 1$. Then the edge weights are defined such that, for all $(j, i) \in \mathcal{E}$, we have $\tilde{w}_{r(j),r(i)} = \frac{\tilde{A}_{r(j),r(i)}}{\sum_k \tilde{A}_{k,r(i)}} = \frac{A_{ij}}{\sum_m A_{im}} = \Delta_{ij}$.

By reversing the graph, the unique forward sink of G (node n) becomes the source node of $\tilde{G}$, and the unique source of G (node 1) becomes the sink node of $\tilde{G}$. We now show that computing the fidelity of G follows directly from computing a random walk on $\tilde{G}$.

Theorem 4.2. *For any number of steps t and any nodes i, j, we have:*

$$\Delta_{ij}^t = \tilde{w}_{r(j),r(i)}^t$$

Proof Sketch. The proof proceeds by straightforward induction on t. The base case ($t = 1$) holds by definition. For the inductive step, we delegate the details to Section A. □

Given this, we can arrive at the following straightforward corollary:

Corollary 4.3. *The last row of Δ of any computational feedforward G evolves exactly like the first column of $\tilde{\mathbf{W}}$.*

5. The Backwards Mixing Time - Fidelity Bound

Having established the duality between the forward diffusion of information and the backward random walk, we now present our main theoretical results. We first prove the central result of this paper: the backwards mixing time forms an upper bound over normalised minimax fidelity. We then show that this bound is tight for a certain family of graphs we call *conjoined star* graphs.

5.1. Fidelity is Bounded by Backward Mixing Time

We utilise our equivalence between the fidelity and the backward random walk proven in the earlier section (Theorem 4.2) to show that the backwards mixing time forms an upper bound on the fidelity:

Theorem 5.1 (Fidelity-Backwards-Mixing Bound)**.** *Let G be a feedforward graph with a unique source and a unique sink with $n \geq 2$. Let $F(G)$ be its minimax fidelity and $T_{bwd}(G)$ be the worst-case mixing time of its backward random walk. Then, for both uniform and non-uniform graphs:*

$$F(G) \leq \frac{8}{3} \cdot T_{bwd}(G)$$

Proof. We first recall the definition of the worst-case mixing time $T_{bwd}(G)$ to be the lowest t such that for any starting distribution $\mathbf{x}$ and stationary distribution $\boldsymbol{\pi}$:

$$||\tilde{\mathbf{W}}^t \mathbf{x} - \boldsymbol{\pi}||_1 < \frac{1}{2}$$

where $\tilde{\mathbf{W}}$ is the transition matrix of the reverse random walk. Because there is a unique forward sink n (which becomes the unique source node in the reverse graph), the stationary distribution $\boldsymbol{\pi}$ is exactly 1 at the absorbing node and 0 everywhere else.

Let S be the absorbing set (a singleton due to the unique source assumption) in the reverse walk, and let $P(X_t \notin S)$ be the probability that the walker has not reached the absorbing node after t steps.

By the definition of mixing time, after $T_{bwd}(G)$ steps, the probability of the walker being at any node other than the absorbing node $P(X_{T_{bwd}} \notin S)$ is at most $1/4$. Using the Markov property and the fact that the mixing time holds for any initial distribution, we can see that:

$$\begin{aligned}
&P(X_{(k+1)\cdot T_{bwd}} \notin S) \\
&= \sum_{y \notin S} P(X_{(k+1)\cdot T_{bwd}} \notin S \mid X_{k\cdot T_{bwd}} = y) \cdot P(X_{k\cdot T_{bwd}} = y) \\
&\leq \max_{y \notin S} P_y(X_{T_{bwd}} \notin S) \cdot \sum_{y \notin S} P(X_{k\cdot T_{bwd}} = y) \\
&\leq \frac{1}{4} \cdot P(X_{k\cdot T_{bwd}} \notin S)
\end{aligned}$$

Using this reasoning, we can see that after $k \cdot T_{bwd}(G)$ steps, the probability of the walker not being in the absorbing node is at most $(1/4)^k$. Recall that $\phi_i = \max_t \Delta_{ni}^t$. Because the maximum of any non-negative sequence is bounded by its sum ($\phi_i \leq \sum_{t=0}^{\infty} \Delta_{ni}^t$), we can bound the sum of fidelities for all non-sink nodes in $\tilde{G}$:

$$\begin{aligned}
\sum_{i \notin S} \phi_i &\leq \sum_{i \notin S} \sum_{t=0}^{\infty} \Delta_{ni}^t \\
&= \sum_{t=0}^{\infty} \sum_{i \notin S} \Delta_{ni}^t \\
&= \sum_{t=0}^{\infty} \sum_{i \notin S} \tilde{w}_{r(i),r(n)}^t \\
&= \sum_{t=0}^{\infty} P(X_t \notin S)
\end{aligned}$$

$$= \sum_{k=0}^{\infty} \sum_{r=0}^{T_{bwd}(G)-1} P(X_{k \cdot T_{bwd}(G)+r} \notin S)$$

Because all backward random walks must eventually culminate at the unique sink, $P(X_t \notin S)$ is monotonically decreasing with t. Thus, $P(X_t \notin S) \geq P(X_{t+1} \notin S)$, allowing us to upper-bound the value of each block of size $T_{bwd}(G)$ by its starting value:

$$\begin{aligned}\sum_{i \notin S} \phi_i &\leq \sum_{k=0}^{\infty} T_{bwd}(G) \cdot P(X_{k \cdot T_{bwd}(G)} \notin S) \\ &< \sum_{k=0}^{\infty} T_{bwd}(G) \cdot \left(\frac{1}{4}\right)^k \\ &= T_{bwd}(G) \cdot \frac{1}{1-1/4} = \frac{4}{3} \cdot T_{bwd}(G)\end{aligned}$$

Putting it all together (by bounding the minimum of the maximum node fidelities by its average), we can bound the minimax fidelity $F(G)$:

$$\begin{aligned}F(G) &= n \cdot \min_{i \in V} \max_t \Delta_{ni}^t = n \cdot \min_{i \in V} \phi_i \\ &\leq n \cdot \min_{i \notin S} \phi_i \\ &\leq n \cdot \frac{1}{n-1} \sum_{i \notin S} \phi_i \\ &\leq n \cdot \frac{1}{n-1} \cdot \frac{4}{3} \cdot T_{bwd}(G) \\ &\leq \frac{8}{3} \cdot T_{bwd}(G)\end{aligned}$$

where the final inequality holds because $\frac{n}{n-1} \leq 2$ for $n \geq 2$, and thus $2 \cdot \frac{4}{3} = \frac{8}{3}$ for any graph where $n \geq 2$. □

5.2. Tightness of the Tradeoff

We demonstrate that the above bound is tight. To establish this, it is sufficient to identify a class of uniform graphs where $F(G)$ and $T_{fwd}(G)$ are equivalent up to constants. Because uniform graphs represent a strict subset of non-uniform graphs (specifically, the case where all edge weights are constrained to be equal), proving tightness in the uniform setting proves tightness for the non-uniform setting too – guaranteeing that our tradeoff bounds cannot be improved further.

Theorem 5.2. *The Fidelity-Backward-Mixing Bound is tight up to asymptotic constants.*

Proof. Define the class of *conjoined star* graphs as uniform feedforward graphs where the unique source (node 1) is connected to every other node, and every other node is connected to the unique sink (node n). Intuitively, this is a combination of two star graphs: an *outward* star centred at node 1, and an *inward* star centred at node n (Figure 2). Formally, assuming self-loops, the edge set $\mathcal{E}$ is defined as:

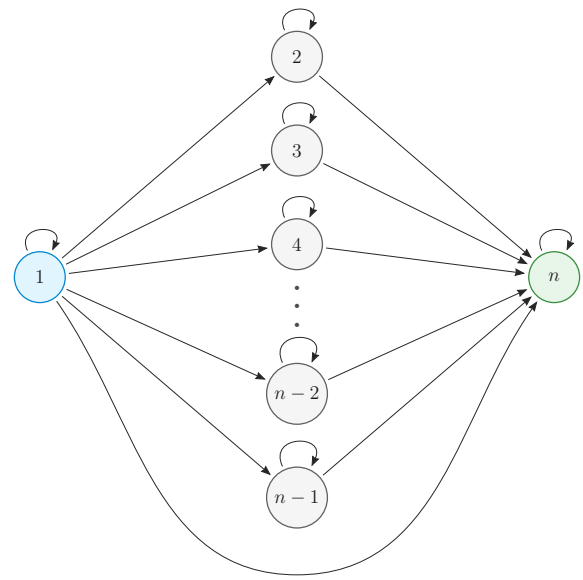


Figure 2. The conjoined star graph.

$$\begin{aligned}\mathcal{E} = &\{(1,i) \mid i \in (1,n]\} \\ &\cup \{(i,n) \mid i \in [1,n)\} \\ &\cup \{(i,i) \mid i \in [1,n]\}\end{aligned}$$

Because this graph is uniform, it is easy to see that $T_{fwd}(G) = T_{bwd}(G)$. For convenience, we will analyse the forward mixing time. To bound $T_{fwd}(G)$, we note that the random walk would be slowest if it started at node 1, the source node. It has a probability of $\frac{n-1}{n}$ of leaving the source node at a single step, and hence will take $\frac{n}{n-1}$ steps on average to leave the source node. Once it leaves the source node, then any node except the source node will take at most 2 steps in expectation to reach the sink, as all intermediary nodes have $d_{out} = 2$. Since $\frac{n}{n-1} \leq 2$ for $n \geq 2$, the expected hitting time $\mathbb{E}[\tau]$ to reach the sink from the source is at most 4. Applying Markov's inequality, the probability of not reaching the sink after t steps is bounded by:

$$\mathbb{P}(\tau \geq t) \leq \frac{\mathbb{E}[\tau]}{t} \leq \frac{4}{t}$$

Setting $t = 16$ ensures that this probability is at most $1/4$. Therefore, the forward mixing time is bounded by a constant ($T_{fwd} \leq 16$): $T_{fwd} = O(1)$. Using Theorem 4.2, we note that the fidelity of a node u evolves exactly like the first column of the reverse random walk matrix. Hence, the maximum fidelity ϕ_u is the maximum probability the random walker starting from n is

at node u during its course of the random walk. The maximum probability for the random walker to be at the node u is at the first step, which is $\frac{1}{n}$. This means that $\phi_u = \frac{1}{n}$ for all intermediary nodes u, $(1 < u < n)$, and normalised minimax fidelity is equal to 1, which implies $F(G) = \Theta(1) = T_{fwd}(G) = T_{bwd}(G)$. □

6. Relating Backwards and Forwards Mixing Time

Despite having a simple proof, we note that the above result is surprising, as it is not that obvious that the backwards mixing time can be connected to fidelity at first glance.

6.1. A Fundamental Tradeoff for Symmetric Graphs

Using this bound, it follows that a fundamental tradeoff bound between the forwards mixing time and the normalised minimax fidelity exists for *symmetric* graphs in the uniform setting, where $T_{fwd}(G) = T_{bwd}(G)$.

In the case where G is isomorphic to $\tilde{G}$, i.e. G is equivalent to the reversed graph $\tilde{G}$ up to node relabeling, we would have an equivalence between the backwards mixing time and forward mixing time ($T_{bwd}(G) = T_{fwd}(G)$) and we deem that the graph is *symmetric*. In the uniform setting, this occurs for a wide range of attentional masks proposed in the literature, i.e. fully-connected attention (Vaswani et al., 2023), sliding window attention (Wang et al., 2019), LogSparse (Li et al., 2020), strided attention (Child et al., 2019), and FiboAttention (Rahimian et al., 2026). We also note that the FunSearch graph proposed by (Vitvitskyi et al., 2025) is symmetric. When the graph is symmetric, we will have $F(G) \leq \frac{8}{3} \cdot T_{fwd}(G)$. Hence, for symmetric graphs, this demonstrates a fundamental tradeoff: a computational graph cannot simultaneously achieve optimal forward mixing speed and maximal representation sharpness.

6.2. General Tradeoffs

More generally, for non-symmetric uniform graphs, e.g. LongFormer (Beltagy et al., 2020), BigBird (Zaheer et al., 2021), and non-uniform graphs, if $T_{bwd}(G)$ is related to $T_{fwd}(G)$ by some f, then we arrive at the following general corollary:

Corollary 6.1 (General Fidelity-Forward-Mixing Bound.)**.** *If the backward mixing time can be bounded by the forward mixing time via some function f, such that $T_{bwd}(G) = f(T_{fwd}(G))$, then the minimax fidelity is bottlenecked by the forward mixing time:*

$$F(G) \leq \frac{8}{3} \cdot f(T_{fwd}(G))$$

This bound assumes that an f that relates T_{fwd} and T_{bwd} must exist. For completeness, we next demonstrate that such a bounding function f must exist for both uniform and non-uniform graphs through a simple argument that upper-bounds both $T_{fwd}(G)$ and $T_{bwd}(G)$.

Remark. From the definition of fidelity and from Theorem 4.2, we can see that $\phi_i \leq 1$ for every node i, as the probability to be at node i at the backward random walk is at most 1 at any timestep t. Hence, we can arrive at a trivial bound for normalised minimax fidelity: $F(G) \leq n$ for all G. We acknowledge that the bounds that are presented below imply a looser bound for $F(G)$ than the trivial $O(n)$ bound. The purpose of including these bounds is to show that a bounding function f relating forward and backward must exist and does not blow up.

6.3. Mixing Time Bounds for Uniform Graphs

For uniform graphs, we find that the forward and backward mixing times are polynomially equivalent, scaling at most linearly with each other with the sequence length n.

Theorem 6.2. *For any uniform feedforward graph with $n \geq 2$ nodes, self-loops, and unique source and sink, the forward and backward mixing times linearly bound one another:*

$$T_{bwd}(G) \leq 8n \cdot T_{fwd}(G)$$
$$T_{fwd}(G) \leq 8n \cdot T_{bwd}(G)$$

Proof Sketch. Because nodes in uniform graphs weight all incoming edges equally, the maximal probability of a self-loop is $1/2$. This allows us to model leaving any state as a geometric random variable. Applying Markov's inequality across all n nodes yields an absolute maximum bound of $T \leq 8n$. The full derivation is provided in Section B.

□

6.4. Mixing Time Bounds for Non-Uniform Graphs

We now extend the above results for non-uniform graphs:

Theorem 6.3. *For any (c, d)-constrained non-uniform feedforward graph with $n \geq 2$, we have*

$$T_{bwd}(G) \leq 4n \left(1 + \frac{d}{c}\right) \cdot T_{fwd}(G)$$
$$T_{fwd}(G) \leq 4n \left(1 + \frac{d}{c}\right) \cdot T_{bwd}(G)$$

Proof Sketch. We analyse the expected hitting times of the random walks. Because the diffusion weights are lower-bounded by c and upper-bounded by d, the sum

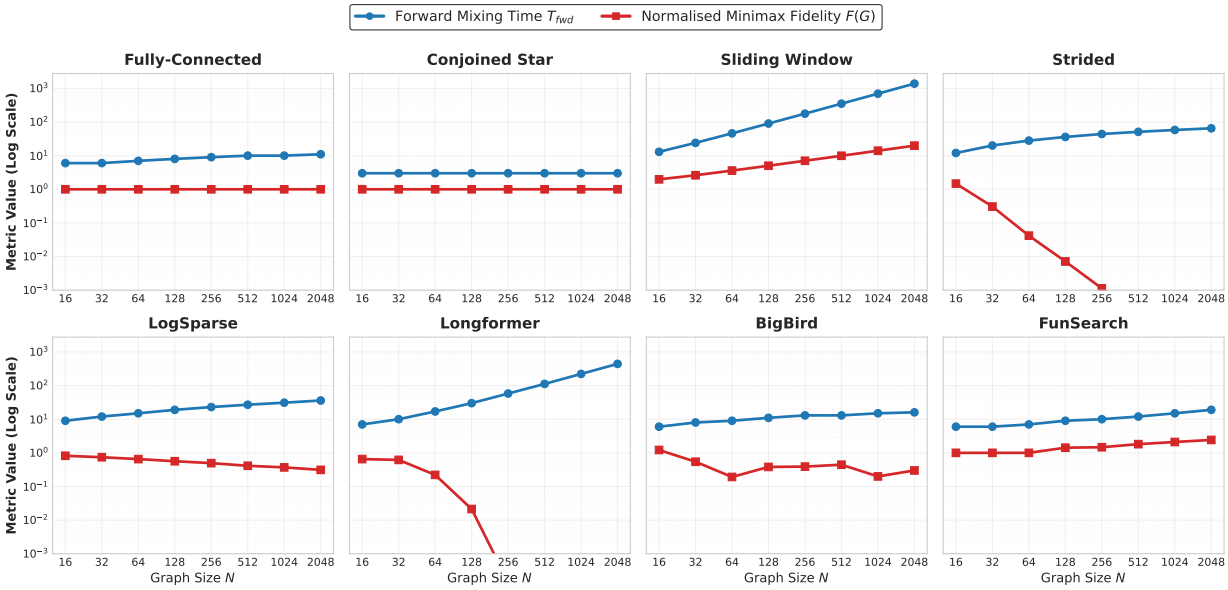


Figure 3. Mixing time and normalised minimax fidelity for a range of attentional masks in the uniform setting.

of fractional coefficients at any node acts as a convex combination bounded by the maximum previous term plus $\left(1+\frac{d}{c}\right)$. Using Markov's inequality on these partial sums yields $T \leq 4n\left(1+\frac{d}{c}\right)$. The complete proof can be found in Appendix C. □

Remark on the Bounding Function f. Using Theorem 6.2 and Theorem 6.3 and plugging it to Theorem 6.1, we can arrive at general upper bounds for fidelity: $F(G) \leq \frac{64}{3} \cdot n \cdot T_{fwd}(G)$ for uniform graphs and $F(G) \leq \frac{32n}{3}\left(1+\frac{d}{c}\right) \cdot T_{fwd}(G)$ for non-uniform graphs. It is crucial to emphasise that these bounds do not serve as the "final say" on the tightness of the tradeoff for all specific architectures, though we observe that this can materialise empirically to some extent in trained LLMs (see Section 7.2) – where the mixing time is larger than a factor of the sequence length w.r.t to the normalised minimax fidelity.

For specific, well-structured classes of computational graphs, the exact bounding function f can be significantly tighter. For example, as discussed earlier, in any symmetric computational graph, $T_{fwd}(G) = T_{bwd}(G)$, making the bounding function the simple identity $f(x) = x$. In such cases, the tradeoff collapses to a drastically tighter, constant-factor bound: $F(G) \leq \frac{8}{3} \cdot T_{fwd}(G)$. Thus, the general bounds above establish a worst-case guarantee, while highlighting that the severity of the tradeoff is dictated by the specific structure of the chosen feedforward graph.

7. Empirical Demonstration

7.1. Uniform Graphs

To demonstrate our bounds hold, in Figure 3, we plot the worst-case forward mixing time $T_{fwd}(G)$ and the minimax fidelity $F(G)$ for a range of attentional masks in the uniform graph setting. Specifically, we plot symmetric masks–fully-connected attention, conjoined star, sliding window attention, strided attention, LogSparse attention, and the FunSearch mask–and non-symmetric masks–LongFormer and BigBird. Each point in the graph corresponds to either minimax fidelity or worst-case mixing time, with varying graph sizes $N \in \{16, 32, 64, \ldots, 2048\}$. We use a default window size of 3 for masks with a sliding window component. For strided attention, we use a stride size of 8, and for LongFormer we use a dilation size of 3. For BigBird, we add 5 random edges to each node.

From Figure 3, we can see for symmetric graphs, it is evident that forward mixing time forms an upper bound on the normalised minimax fidelity, because $T_{fwd}(G) = T_{bwd}(G)$. For non-symmetric graphs, we also observe the same effect due to Theorem 6.2. We also note that in practice, the gap between normalised minimax fidelity and forward mixing time is significantly tighter than the bound in Theorem 6.2.

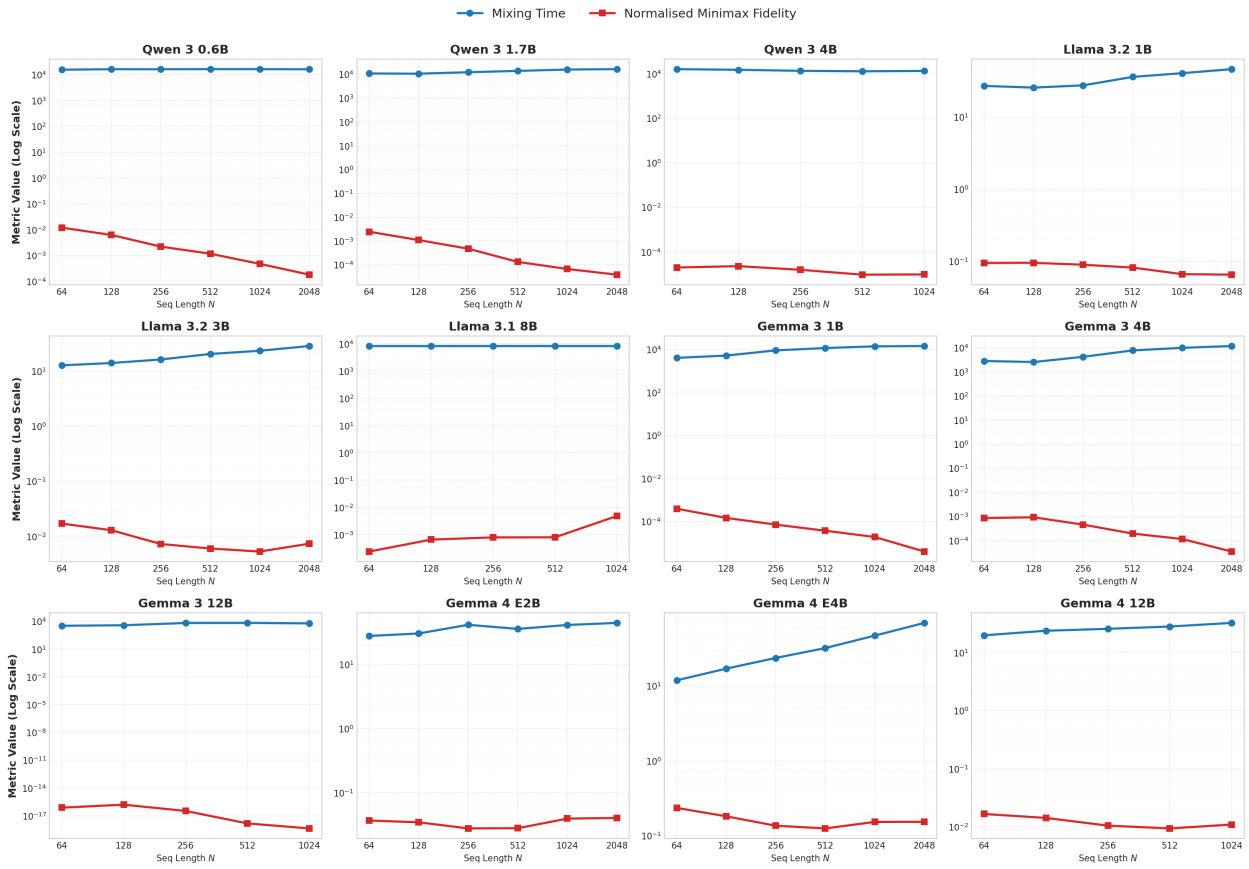


Figure 4. Log-plot of averaged mixing time and normalised minimax fidelity computed by averaging unnormalised attentional masks across a range of LLMs of different sizes.

7.2. Non-Uniform Graphs

To demonstrate our bounds hold in the non-uniform graphs setting, particularly for the attentional masks in LLMs, we require a way to aggregate these masks. Though there exist many ways to do this, we chose the following method: for each sequence, obtain the averaged weighted adjacency matrix $\mathbf{A}$ by averaging the unnormalised attention matrix across layers and across heads. Next, compute the non-uniform diffusion matrix $\mathbf{\Delta}$ and the random walk matrix $\mathbf{W}$ by row-normalising and column-normalising $\mathbf{A}$ respectively. Then, compute the normalised minimax fidelity and mixing time and average this across sequences of different lengths to arrive at Figure 4. In particular, we concatenated the context documents from 256 NQ (Kwiatkowski et al., 2019) samples, truncated them at sequence lengths $N \in \{64, 128, \ldots, 2048\}$, and forwarded these samples across Qwen, Llama 3.2, and Gemma models at various model sizes to compute the metrics. We cap the powers of both matrices to 16,384. If the random walk has not converged by the 16,384-th power, then we stop computation and report 16,384, for example as is the case for all mixing time points in the Qwen 3 0.6B subplot.

In this more realistic setting and as demonstrated in Figure 4, we observe robust empirical confirmation of our theoretical framework: the mean forward mixing time consistently upper-bounds the normalized minimax fidelity across all evaluated models and sequence lengths. We find that these bounds hold layer-wise for each model, but due to space constraints, we leave these plots to the Appendix E.

Crucially, the empirical tightness of the bounding function f varies dramatically across architectures. In certain instances, such as the Qwen series, the forward random walk mixes exceptionally slowly, saturating our computational ceiling of 16,384 steps and exhibiting a wide gap relative to the normalised minimax fidelity. This configuration demonstrates that the pessimistic $O(n(1 + d/c))$ theoretical upper bound can indeed materialize in practice. Conversely, for the Llama and Gemma series, the empirical gap between mixing time and fidelity is much tighter. In the non-uniform setting, this divergence provides evidence that the exact behavior of f is not uniform

across deep learning systems; rather, it is highly sensitive to the model architecture and how it was trained.

8. Conclusion

In this work, we establish an equivalence between the forward diffusion of representations and a random walk on the reversed graph. Using this, we proved that the minimax fidelity of a computational graph is tightly upper-bounded by its backward mixing time, $F(G) \leq \frac{8}{3} \cdot T_{bwd}(G)$–establishing the backwards mixing time as a metric to analyse the sharpness of information propagation.

Next, for symmetric graphs in the uniform setting, we demonstrate a fundamental tradeoff: a computational graph cannot simultaneously achieve optimal forward mixing speed and maximal representation sharpness. In the more general case, we proved that $F(G) \leq f(T_{fwd}(G))$ where $f(T_{fwd}(G))$ is bounded by $O(n) \cdot T_{fwd}(G)$ in the uniform setting and $O(n(1 + d/c)) \cdot T_{fwd}(G)$ in the non-uniform setting.

We note limitations that present clear avenues for future work:

- **Worst-Case vs. Average-Case.** To provide absolute mathematical guarantees on minimax fidelity, our proofs necessitate the use of worst-case mixing times. Because sequence models often operate over distributions where average-case performance is sufficient, a highly relevant extension would be to formally define an *averaged* fidelity metric and investigate whether a parallel tradeoff exists between averaged fidelity and averaged mixing time. Nevertheless, we conjecture the trade-off will still hold in the average case.
- **Analysing Graphs With Multiple Forward Sources.** Our results assume that feedforward graphs have a unique source. Although this captures most masks in practice, some masks have multiple sources, e.g. the blockwise attention in BlockBERT (Qiu et al., 2020). In this case, introducing multiple forward source alters the random walk on the reversed graph by expanding the single absorbing node into a set of multiple absorbing states. A natural next step would be to extend our analysis for such graphs.
- **Deriving Tighter Universal Bounds for Mixing Times.** While we proved that the bounding function relating forward and backward mixing times scales at most by $O(n)$ and $O(n(1 + d/c))$, these serve as universal, worst-case upper bounds across the entire space of valid feedforward topologies. Our empirical findings in Section 7.2 reveal that while these pessimistic ceilings can be approached in practice (e.g., Qwen 3 0.6B), the actual gap between mixing time and fidelity is substantially tighter for other optimized production models like Llama and Gemma. Future work could incorporate finer structural or statistical assumptions—such as bounded graph width, specific attention sparsity patterns, or localized graph expansion properties—to derive tighter, architecture-specific bounds that closely mirror our observed empirical experiments.

Impact Statement

This paper proves the central result that normalised minimax fidelity is upper-bounded by backwards mixing time. In isolation, this work does not present any direct societal risks or negative ethical consequences. However, we predict that our work will inform researchers of this inherent property in computational feedforward graphs and to inspire the design of better models. Hence, any societal risks that our work poses should be attributed to the consequences of accelerating the design of such models.

References

Attali, H., Buscaldi, D., and Pernelle, N. Rewiring techniques to mitigate oversquashing and oversmoothing in gnns: A survey, 2024. URL https://arxiv.org/abs/2411.17429.

Barbero, F., Banino, A., Kapturowski, S., Kumaran, D., Araújo, J. G., Vitvitskyi, A., Pascanu, R., and Veličković, P. Transformers need glasses! information over-squashing in language tasks. *Advances in Neural Information Processing Systems*, 37:98111–98142, 2024.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

Chen, L., Xu, D., An, C., Wang, X., Zhang, Y., Chen, J., Liang, Z., Wei, F., Liang, J., Xiao, Y., and Wang, W. Powerattention: Exponentially scaling of receptive fields for effective sparse attention, 2025. URL https://arxiv.org/abs/2503.03588.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers, 2019. URL https://arxiv.org/abs/1904.10509.

Chowdhury, B. D. Lost in the middle at birth: An exact theory of transformer position bias, 2026. URL https://arxiv.org/abs/2603.10123.

Csóka, E. and Grabowski, Ł. On directed analogues of expander and hyperfinite graph sequences. *Combinatorics, Probability and Computing*, 31(2):184–197, 2022.

de Ocáriz Borde, H. S. Beyond parallelism: Synergistic computational graph effects in multi-head attention, 2025. URL https://arxiv.org/abs/2507.02944.

Deac, A., Lackenby, M., and Veličković, P. Expander graph propagation. In *Learning on Graphs Conference*, pp. 38–1. PMLR, 2022.

Espuny Díaz, A., Morris, P., Perarnau, G., and Serra, O. Speeding up random walk mixing by starting from a uniform vertex. *Electronic journal of probability*, 29: 1–25, 2024.

Gemma, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Giovanni, F. D., Giusti, L., Barbero, F., Luise, G., Lio', P., and Bronstein, M. On over-squashing in message passing neural networks: The impact of width, depth, and topology, 2023. URL https://arxiv.org/abs/2302.02941.

Google DeepMind. Gemma 4 model card, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli,

M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Herasimchyk, H., Labryga, R., Prusina, T., and Laue, S. A residual-aware theory of position bias in transformers, 2026. URL https://arxiv.org/abs/2602.16837.

Huang, S., Poursafaei, F., Danovitch, J., Fey, M., Hu, W., Rossi, E., Leskovec, J., Bronstein, M., Rabusseau, G., and Rabbany, R. Temporal graph benchmark for machine learning on temporal graphs, 2023. URL https://arxiv.org/abs/2307.01026.

Kim, J., Kim, H., Kim, H., Lee, D., and Yoon, S. A

comprehensive survey of deep learning for time series forecasting: Architectural diversity and open challenges, 2025. URL https://arxiv.org/abs/2411.05793.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., and Yan, X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting, 2020. URL https://arxiv.org/abs/1907.00235.

Qiu, J., Ma, H., Levy, O., tau Yih, S. W., Wang, S., and Tang, J. Blockwise self-attention for long document understanding, 2020. URL https://arxiv.org/abs/1911.02972.

Rahimian, A. K., Govind, M. K., Maity, S., Reilly, D., Kümmerle, C., Das, S., and Dutta, A. Fibottention: Inceptive visual representation learning with diverse attention across heads, 2026. URL https://arxiv.org/abs/2406.19391.

Stanković, L., Daković, M., Bardi, A. B., Brajović, M., and Stanković, I. Fourier analysis of signals on directed acyclic graphs (dag) using graph zero-padding. *Digital Signal Processing*, 159: 104995, 2025. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2025.104995. URL https://www.sciencedirect.com/science/article/pii/S105120042500017X.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks, 2014. URL https://arxiv.org/abs/1409.3215.

Topping, J., Giovanni, F. D., Chamberlain, B. P., Dong, X., and Bronstein, M. M. Understanding over-squashing and bottlenecks on graphs via curvature, 2022. URL https://arxiv.org/abs/2111.14522.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio, 2016. URL https://arxiv.org/abs/1609.03499.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Vitvitskyi, A., Araújo, J. G. M., Lackenby, M., and Veličković, P. What makes a good feedforward computational graph?, 2025. URL https://arxiv.org/abs/2502.06751.

Wang, Z., Ng, P., Ma, X., Nallapati, R., and Xiang, B. Multi-passage bert: A globally normalized bert model for open-domain question answering, 2019. URL https://arxiv.org/abs/1908.08167.

Wiedemer, T., Li, Y., Vicol, P., Gu, S. S., Matarese, N., Swersky, K., Kim, B., Jaini, P., and Geirhos, R. Video models are zero-shot learners and reasoners, 2025. URL https://arxiv.org/abs/2509.20328.

Wilson, J., Bechler-Speicher, M., and Veličković, P. Cayley graph propagation. *arXiv preprint arXiv:2410.03424*, 2024.

Xia, Z., Pan, X., Song, S., Li, L. E., and Huang, G. Vision transformer with deformable attention, 2022. URL https://arxiv.org/abs/2201.00520.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. Big bird: Transformers for longer sequences, 2021. URL https://arxiv.org/abs/2007.14062.

A. Proof of Theorem 4.2

Theorem 4.2. *Let $\mathbf{\Delta}$ be the row-stochastic diffusion matrix of an arbitrary feedforward graph G, let $\tilde{\mathbf{W}}$ be the column-normalized random walk matrix of its reversed counterpart $\tilde{G}$, and let $r(i) = |V| - i + 1$. For any number of steps t and any nodes i, j, we have $\Delta_{ij}^t = \tilde{w}_{r(j),r(i)}^t$.*

Proof. We proceed by induction on t.

The base case holds by definition. For the inductive step, assume that the theorem holds for some arbitrary step $t \geq 1$. That is, our inductive hypothesis states that for all $i, j \in \mathcal{V}$:

$$\Delta_{i,j}^t = \tilde{w}_{r(j),r(i)}^t$$

We must show this implies $\Delta_{i,j}^{t+1} = \tilde{w}_{r(j),r(i)}^{t+1}$.

Using the standard definition of matrix multiplication (and the property that matrix powers commute), we can expand $\Delta_{i,j}^{t+1}$ by summing over all possible intermediate nodes k:

$$\Delta_{i,j}^{t+1} = \sum_{k \in V} \Delta_{i,k}^t \cdot \Delta_{k,j} \tag{1}$$

Similarly, we expand $\tilde{w}_{r(j),r(i)}^{t+1}$ by summing over all possible intermediate nodes $m \in \mathcal{V}$:

$$\tilde{w}_{r(j),r(i)}^{t+1} = \sum_{m \in V} \tilde{w}_{r(j),m} \cdot \tilde{w}_{m,r(i)}^t$$

which is equivalent to:

$$\tilde{w}_{r(j),r(i)}^{t+1} = \sum_{k \in V} \tilde{w}_{r(j),r(k)} \cdot \tilde{w}_{r(k),r(i)}^t \tag{2}$$

We can now apply our previously established equivalencies to Equation 2. By the base case, we know $\tilde{w}_{r(j),r(k)} = \Delta_{k,j}$. By the inductive hypothesis, we know $\tilde{w}_{r(k),r(i)}^t = \Delta_{i,k}^t$. Substituting these into the summation gives:

$$\begin{aligned}
\tilde{w}_{r(j),r(i)}^{t+1} &= \sum_{k \in V} \Delta_{k,j} \cdot \Delta_{i,k}^t \\
&= \sum_{k \in V} \Delta_{i,k}^t \cdot \Delta_{k,j} \\
&= \Delta_{i,j}^{t+1}
\end{aligned}$$

B. Proof of Theorem 6.2

To prove Theorem 6.2, we first establish global upper and lower bounds on the mixing times for uniform graphs.

Lemma B.1. *For any uniform feedforward graph with self-loops and unique source and sink, the mixing times satisfy $T_{fwd}(G) \leq 8n$ and $T_{bwd}(G) \leq 8n$.*

Proof. Because the graph is uniform, a node weights all incoming edges equally. The presence of a self-loop means the maximal probability of remaining at the current state is $1/2$, as every non-sink node has an out-degree of at least 2 (a self-loop plus one outgoing edge). Thus, the probability of leaving any state is at least $1/2$. Consequently, leaving a state can be modeled as a geometric random variable, meaning the expected number of steps to leave a node is at most 2. Summing this expectation across a maximum path length of n nodes, the expected number of steps to reach the target node is $\mathbb{E}[T] \leq 2n$. By Markov's inequality, the probability of failing to reach the sink after $8n$ steps is bounded by $\mathbb{P}(T \geq 8n) \leq \frac{2n}{8n} = \frac{1}{4}$. By the definition of mixing time, $T_{fwd} \leq 8n$. This logic holds identically for both the forward and backward random walks. □

Proof of Theorem 6.2. By Lemma B.1, $T_{bwd} \leq 8n$. Since $n \geq 2$, we also have $T_{fwd} \geq 1$ and $T_{bwd} = 1$. Substituting this inequality yields:

$$T_{bwd} \leq 8n \cdot 1 \leq 8n \cdot T_{fwd}$$

The derivation for bounding T_{fwd} by T_{bwd} follows identically. □ □

C. Proof of Theorem 6.3

We define a graph as (c, d)-constrained if for any $(i, j) \in \mathcal{E}$, then $c \leq \Delta_{j,i} \leq d$ for some constant $0 < c \leq d$. We first bound the absolute expected hitting times for these graphs.

Lemma C.1. *For any (c, d)-constrained non-uniform graph, we have*

$$T_{bwd}(G) \leq 4n\left(1 + \frac{d}{c}\right) \quad \text{and} \quad T_{fwd}(G) \leq 4n\left(1 + \frac{d}{c}\right)$$

Proof. Let the total internal row and column sums of A be denoted as $R_i = \sum_k A_{ik}$ and $C_j = \sum_k A_{kj}$ respectively. To analyse the transitions, we split these sums into self-loops and *external weights*:

- Incoming external weight to node i: $I_i = \sum_{j<i} A_{ij}$, meaning $R_i = A_{ii} + I_i$.
- Outgoing external weight from node j: $O_j = \sum_{k>j} A_{kj}$, meaning $C_j = A_{jj} + O_j$.

The backward random walk on the reversed graph $\tilde{G}$ moves with probabilities matching the diffusion matrix, where $\tilde{w}_{r(j),r(i)} = \Delta_{ij} = \frac{A_{ij}}{R_i}$. Let $h_i = \mathbb{E}[T_{\text{bwd}} \mid X_0 = i]$ be the expected number of steps to reach the unique source (node 1), with base case $h_1 = 0$. By the Markov property:

$$h_i = 1 + \Delta_{ii} h_i + \sum_{j<i} \Delta_{ij} h_j$$

$$h_i = 1 + \frac{A_{ii}}{R_i} h_i + \sum_{j<i} \frac{A_{ij}}{R_i} h_j$$

$$
\begin{aligned}
R_i h_i &= R_i + A_{ii} h_i + \sum_{j<i} A_{ij} h_j \\
(R_i - A_{ii}) h_i &= R_i + \sum_{j<i} A_{ij} h_j \\
I_i h_i &= R_i + \sum_{j<i} A_{ij} h_j \\
h_i &= \frac{R_i}{I_i} + \sum_{j<i} \frac{A_{ij}}{I_i} h_j \\
h_i &= \frac{A_{ii} + I_i}{I_i} + \sum_{j<i} \frac{A_{ij}}{I_i} h_j \\
h_i &= \left(1 + \frac{A_{ii}}{I_i}\right) + \sum_{j<i} \frac{A_{ij}}{I_i} h_j
\end{aligned}
$$

First, since $I_i = \sum_{j<i} A_{ij}$ and $\sum_{j<i} \frac{A_{ij}}{I_i} = 1$, the right-hand term is a convex combination of hitting times. Second, we can bound the left term using the fact that:

- The self-loop weight is upper-bounded: $A_{ii} \leq d$.
- Because the graph has a unique source, every non-source node $i > 1$ must have at least one incoming edge, bounding the incoming weight away from zero: $I_i \geq c$.

Therefore:

$$\left(1 + \frac{A_{ii}}{I_i}\right) \leq 1 + \frac{d}{c}$$

Hence, we have:

$$h_i \leq \left(1 + \frac{d}{c}\right) + \sum_{j<i} \frac{A_{ij}}{I_i} h_j$$

Unrolling this recurrence would give us $h_n \leq n \cdot \left(1 + \frac{d}{c}\right)$. By Markov's inequality, enforcing $\mathbb{P}(T > t) \leq \frac{\mathbb{E}[T]}{t} \leq 1/4$ requires $t \geq 4n\left(1 + \frac{d}{c}\right)$, and hence $T_{bwd} \leq 4n\left(1 + \frac{d}{c}\right)$.

For the forward walk, let $g_j = \mathbb{E}[T_{\text{fwd}} \mid X_0 = j]$ be the expected number of steps to reach the unique sink (node n), where $g_n = 0$. The forward walk matrix is governed by $w_{kj} = \frac{A_{kj}}{C_j}$. Setting up the identical Markov property statement:

$$
\begin{aligned}
g_j &= 1 + w_{jj} g_j + \sum_{k>j} w_{kj} g_k \\
C_j g_j &= C_j + A_{jj} g_j + \sum_{k>j} A_{kj} \\
O_j g_j &= C_j + \sum_{k>j} A_{kj} g_k \\
g_j &= \frac{A_{jj} + O_j}{O_j} + \sum_{k>j} \frac{A_{kj}}{O_j} g_k
\end{aligned}
$$

$$g_j = \left(1 + \frac{A_{jj}}{O_j}\right) + \sum_{k>j} \frac{A_{kj}}{O_j} g_k$$

Because the graph contains a unique sink, every non-sink node $j < n$ must possess at least one forward outgoing edge, meaning $O_j \geq c$. Given $A_{jj} \leq d$, the forward step cost obeys the same ceiling constraint:

$$\left(1 + \frac{A_{jj}}{O_j}\right) \leq 1 + \frac{d}{c}$$

Unrolling this recurrence backward via induction from the absorbing base case $g_n = 0$ yields the absolute worst-case forward expected hitting time:

$$g_1 \leq n\left(1 + \frac{d}{c}\right)$$

Via the same argument using Markov's inequality, we have $T_{fwd} \leq 4n\left(1 + \frac{d}{c}\right)$. □

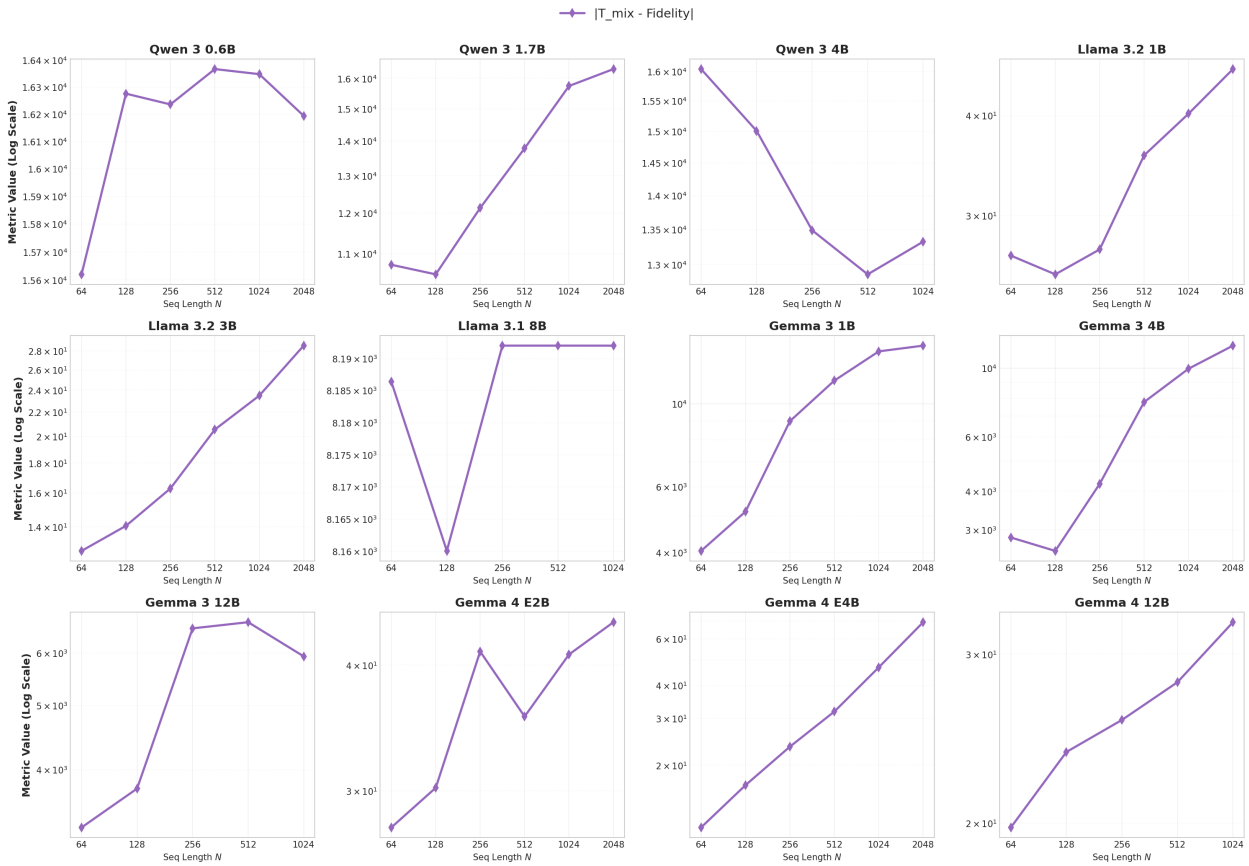
Proof of Theorem 6.3. Hence, observe that for any $n \geq 2$, instantaneous mixing is impossible. Therefore, both $T_{bwd} \geq 1$ and $T_{fwd} \geq 1$. Utilising Lemma C.1, we can bound the backward time linearly by the forward time:

$$T_{bwd} \leq 4n\left(1 + \frac{d}{c}\right) \cdot 1 \leq 4n\left(1 + \frac{d}{c}\right) \cdot T_{fwd}$$

The derivation for bounding T_{fwd} by T_{bwd} follows identically. □

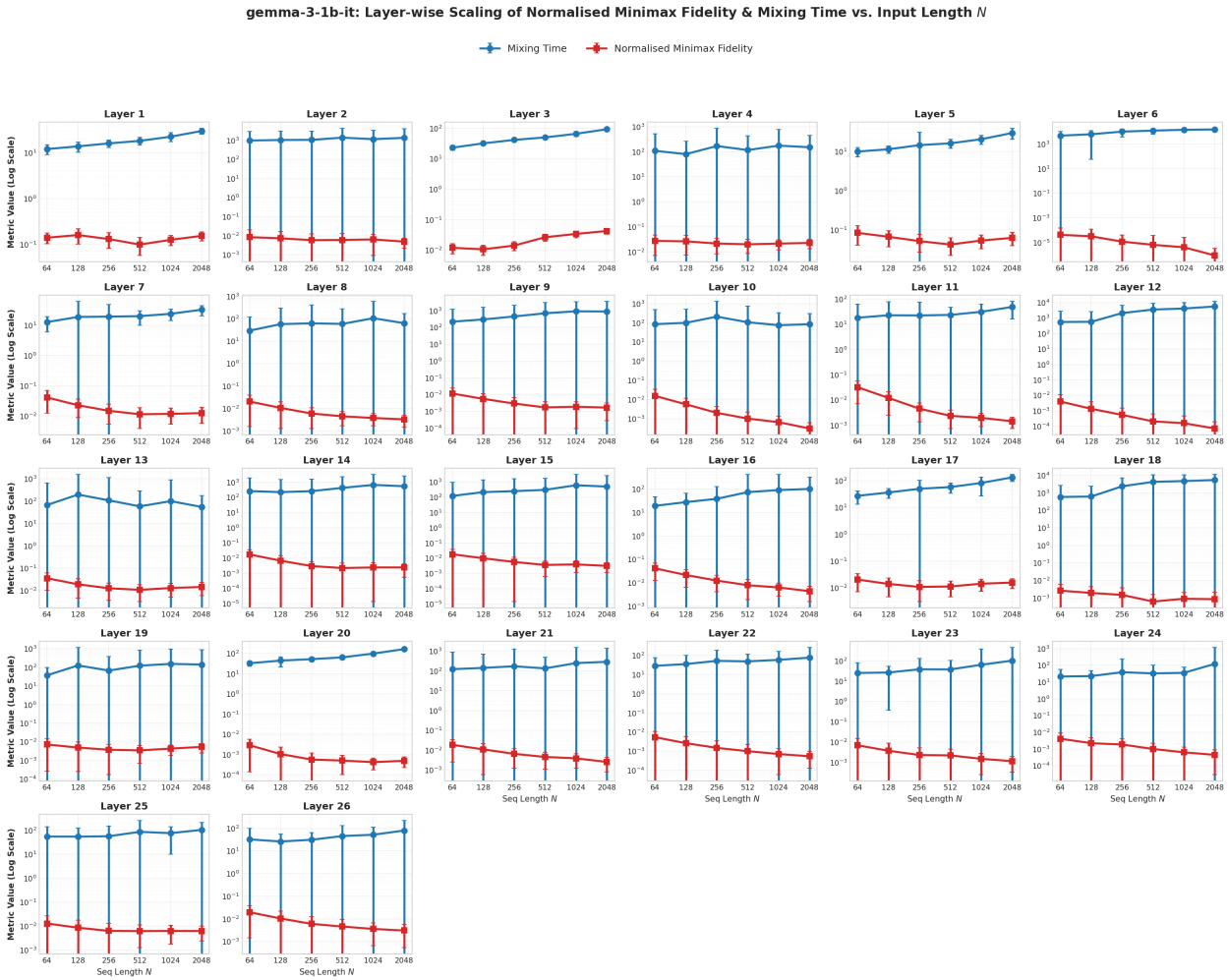
D. Difference between Mixing Time and Normalised Minimax Fidelity Plot

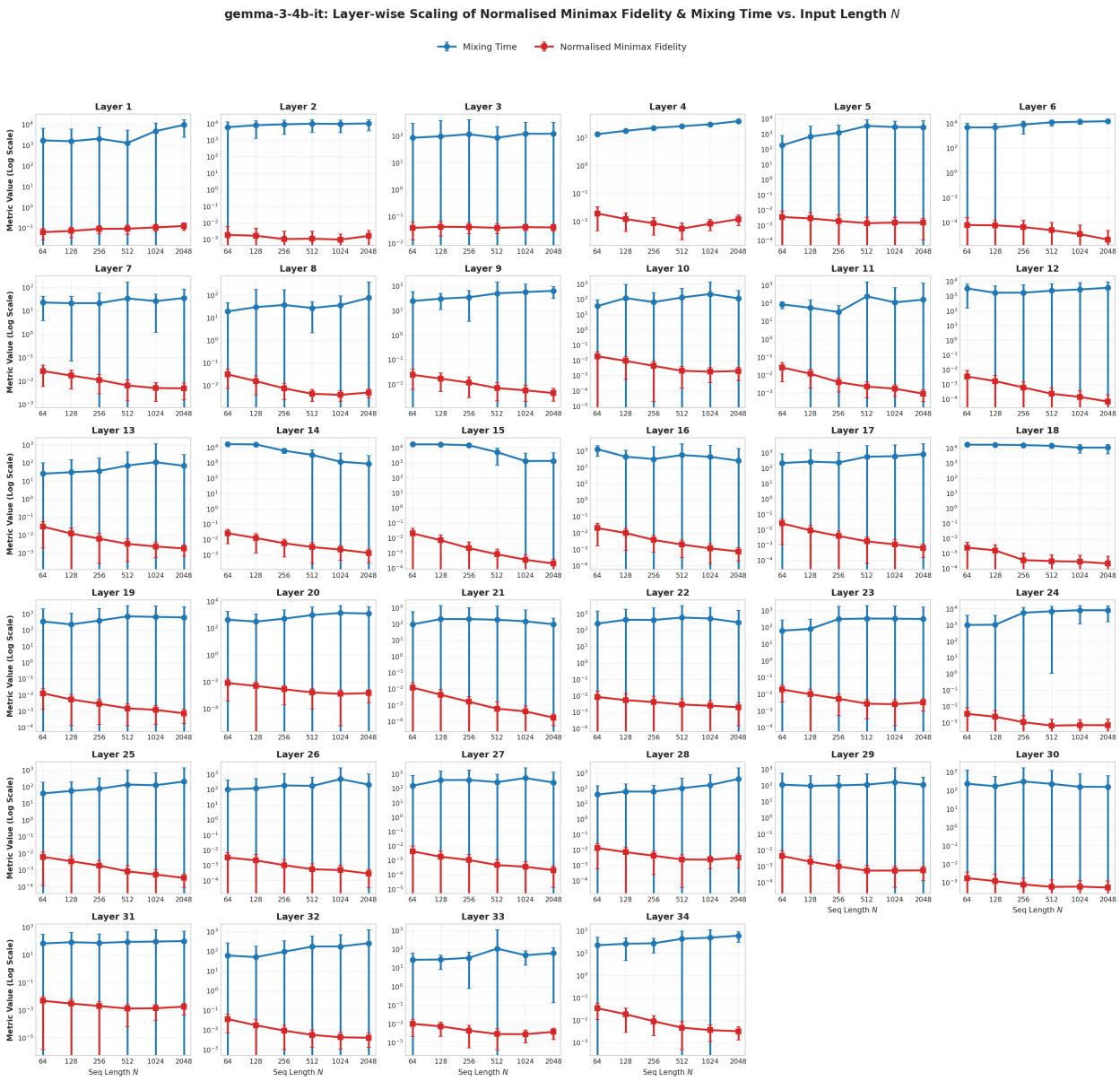
We plot the difference between the averaged mixing time and the normalised minimax fidelity, offering a different view to Figure 4.

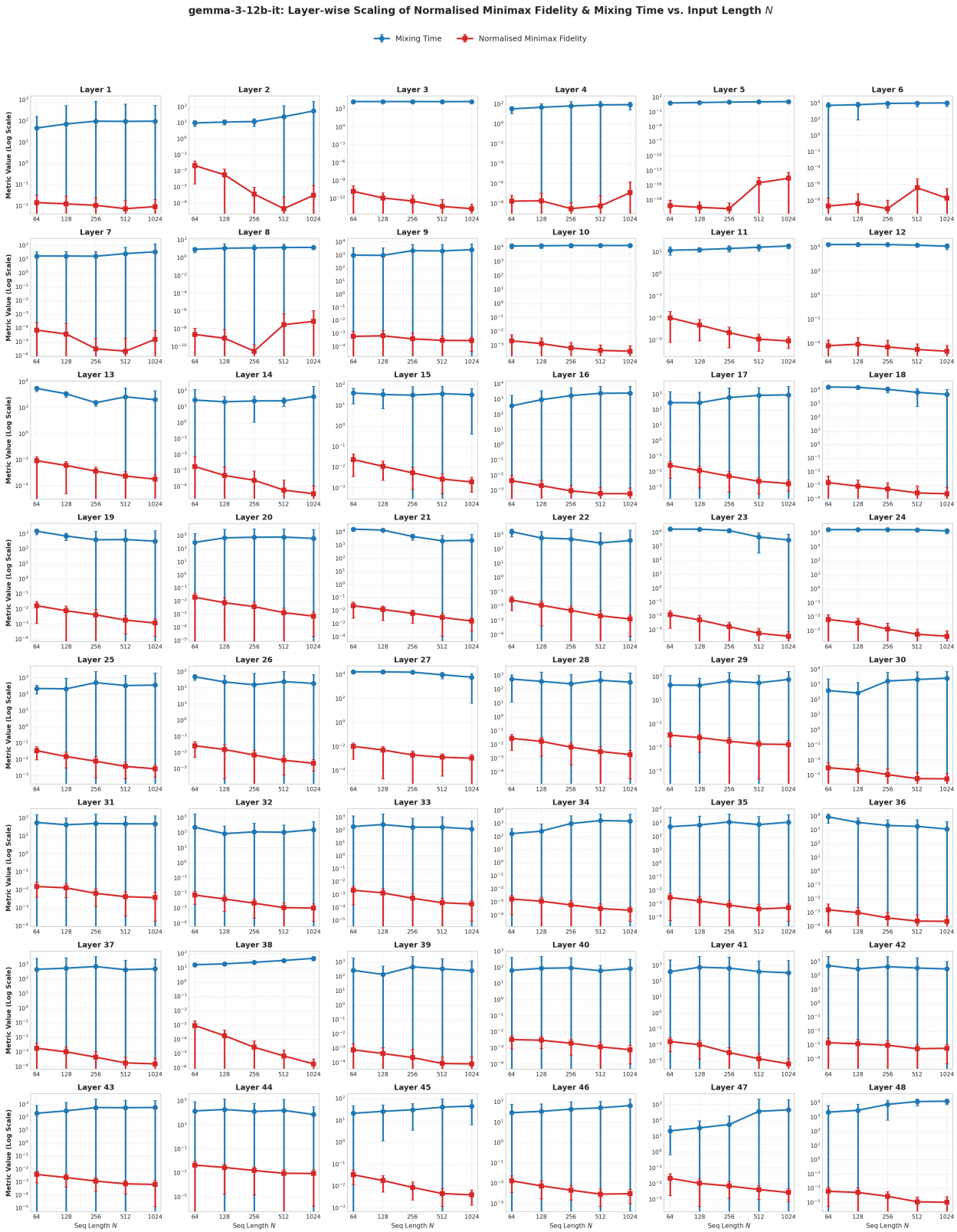


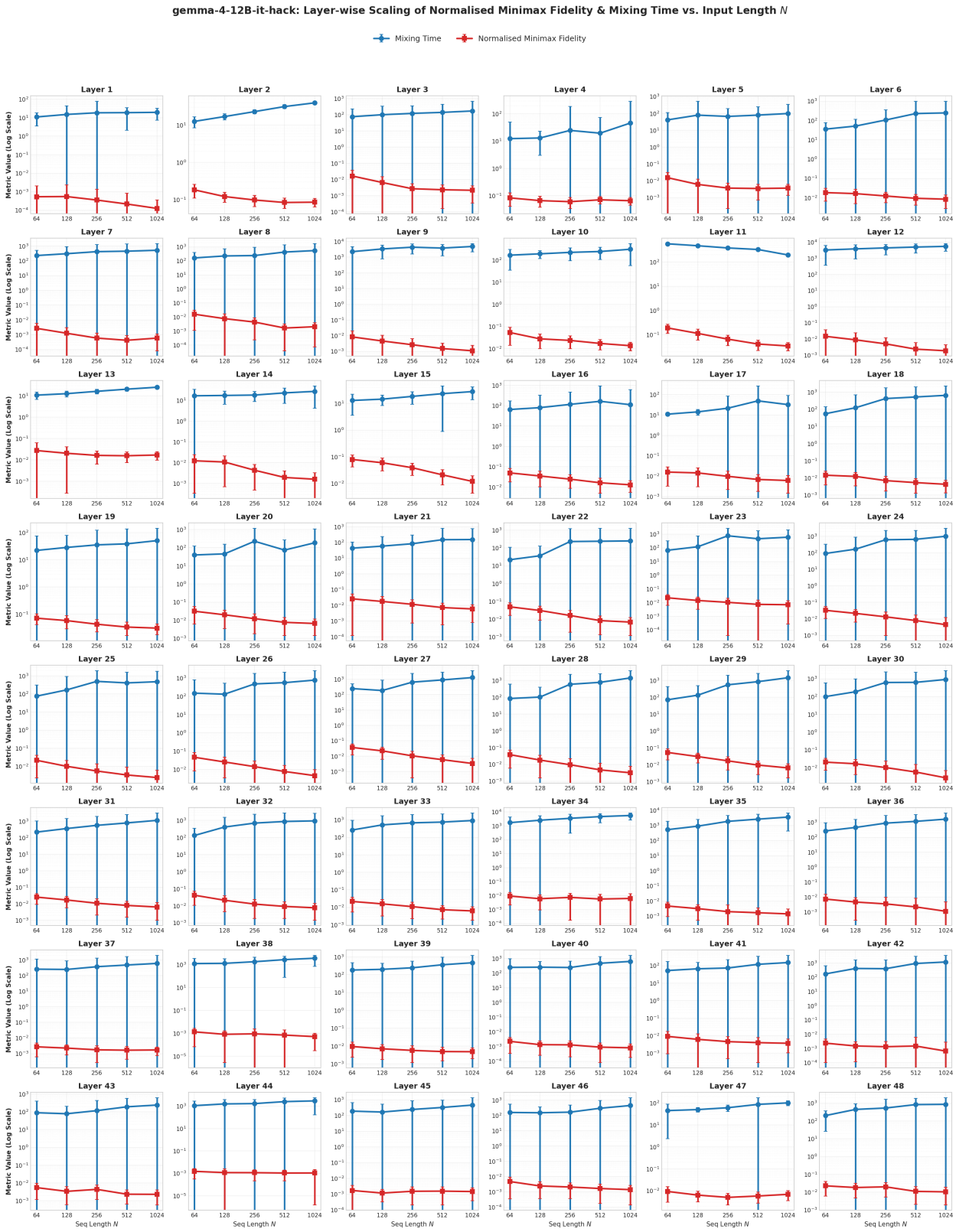
E. Mixing time and Minimax Fidelity: Per Layer Plots

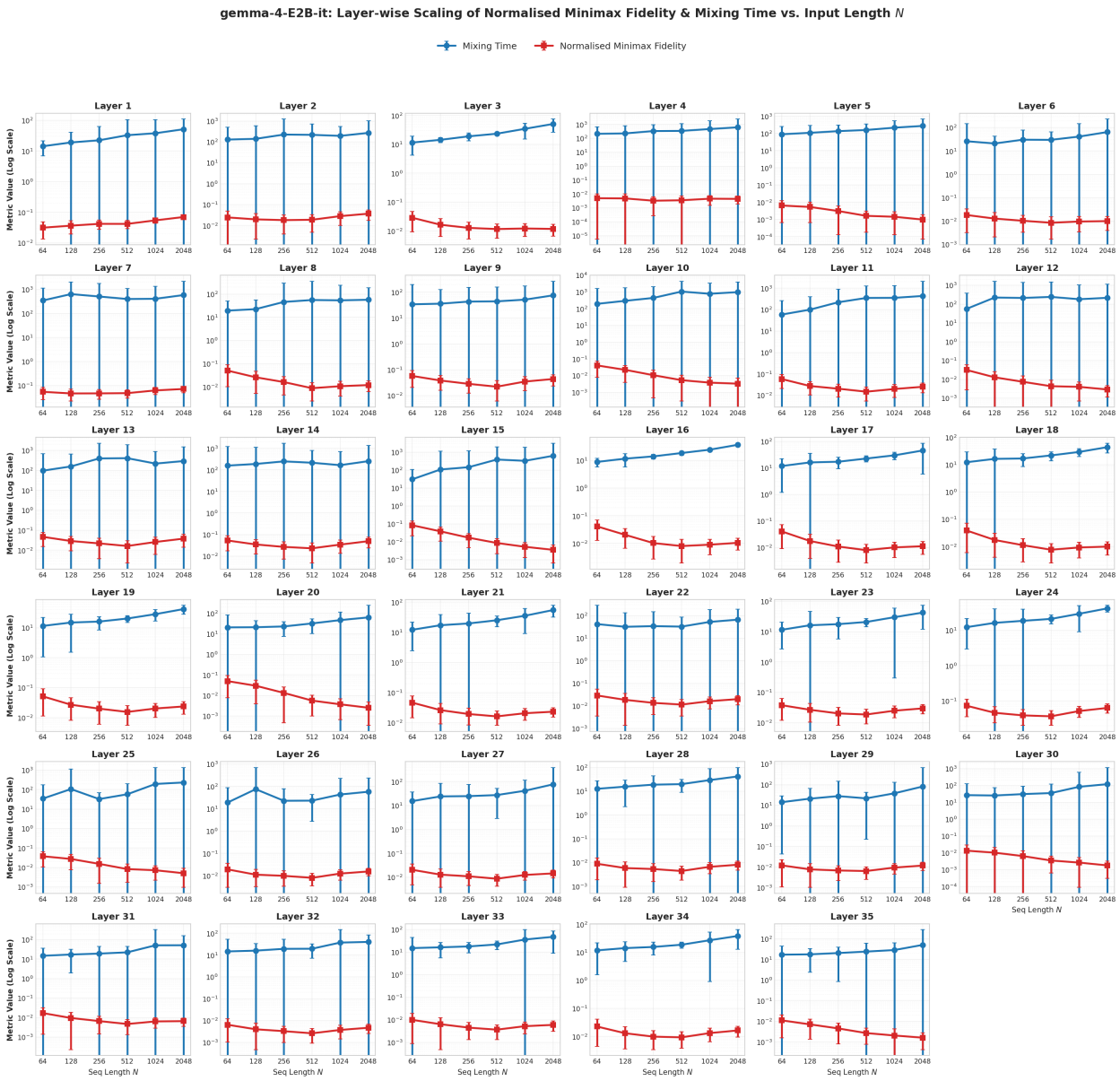
Instead of averaging the unnormalised attention masks, we can also compute per-layer plots for each model. In these plots, we show both the mean and standard deviation that result from aggregating multiple sequences. Interestingly, we also see the same pattern here, where the mean mixing time dominates normalised minimax fidelity. Note that the standard deviation error bars are asymmetrical because we're doing a log-plot.

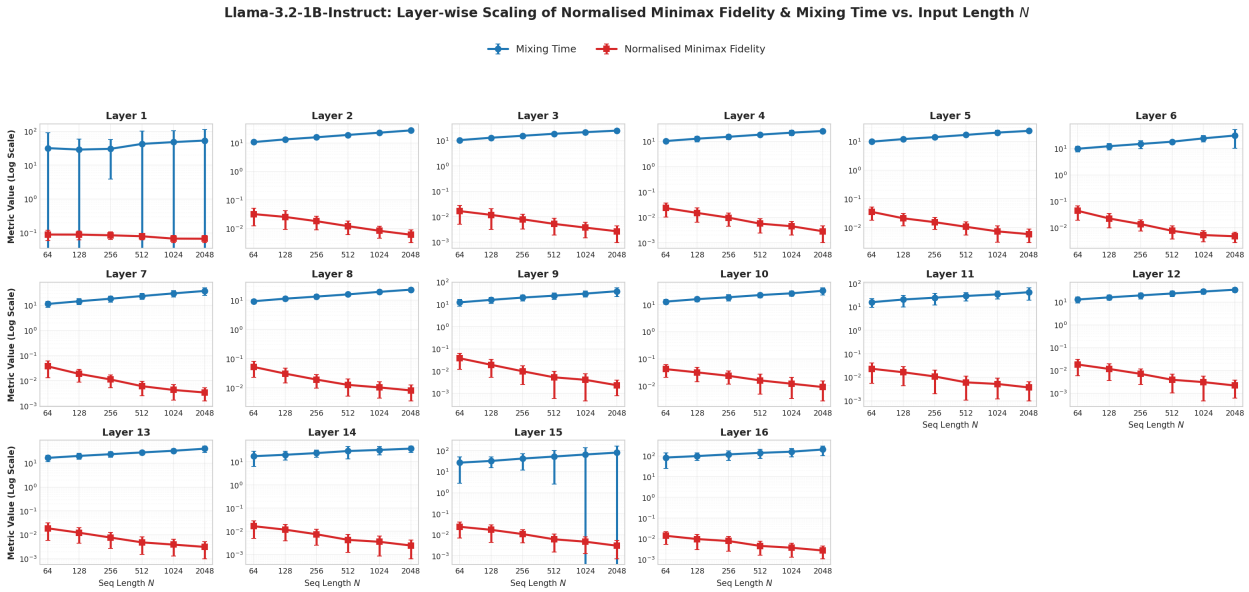

gemma-3-1b-it: Layer-wise Scaling of Normalised Minimax Fidelity & Mixing Time vs. Input Length N
Mixing Time
Normalised Minimax Fidelity
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Layer 13
Layer 14
Layer 15
Layer 16
Layer 17
Layer 18
Layer 19
Layer 20
Layer 21
Layer 22
Layer 23
Layer 24
Layer 25
Layer 26
Metric Value (Log Scale)
Seq Length N
64
128
256
512
1024
2048

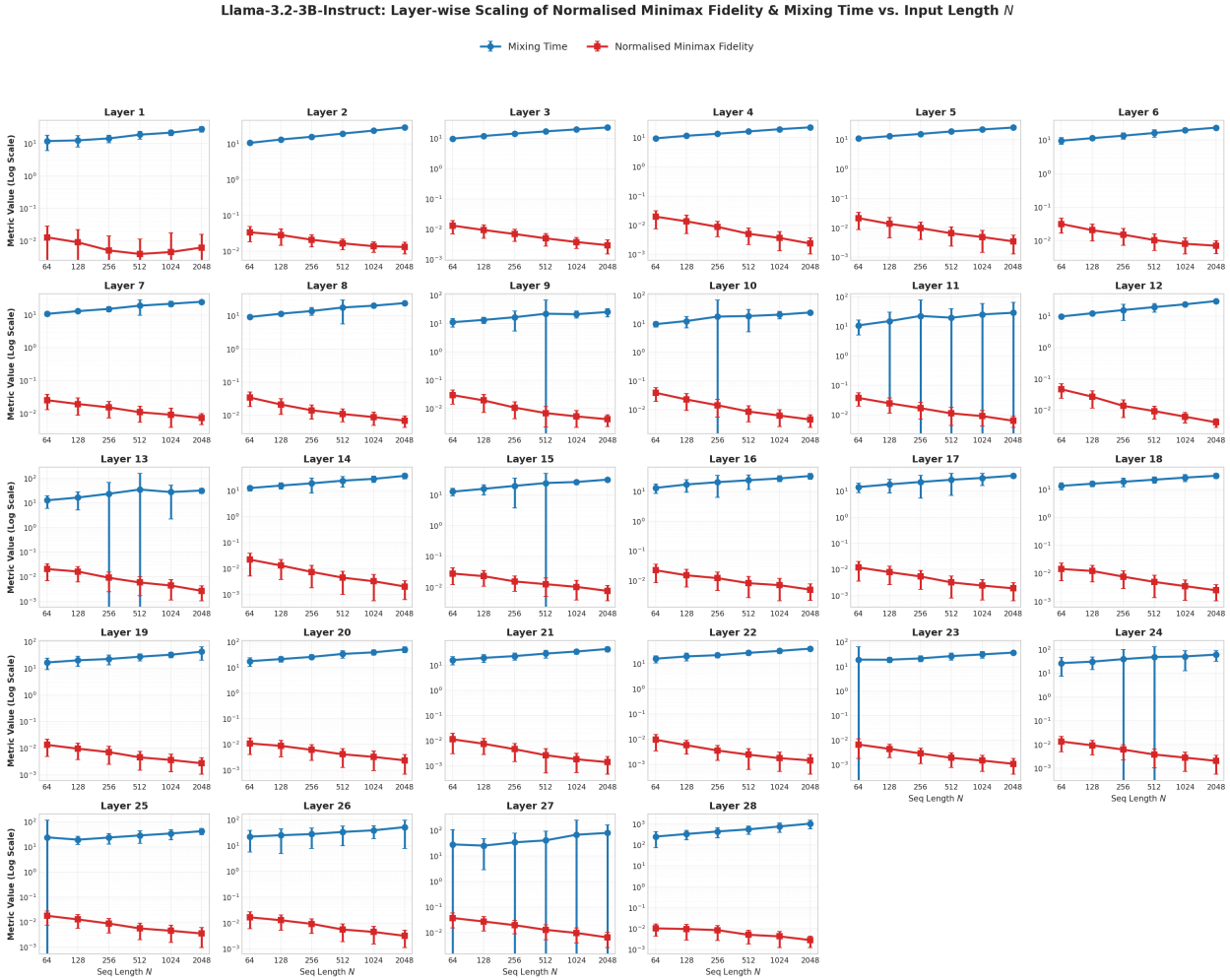

gemma-3-4b-it: Layer-wise Scaling of Normalised Minimax Fidelity & Mixing Time vs. Input Length N
Mixing Time
Normalised Minimax Fidelity
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Layer 13
Layer 14
Layer 15
Layer 16
Layer 17
Layer 18
Layer 19
Layer 20
Layer 21
Layer 22
Layer 23
Layer 24
Layer 25
Layer 26
Layer 27
Layer 28
Layer 29
Layer 30
Layer 31
Layer 32
Layer 33
Layer 34
Metric Value (Log Scale)
Seq Length N
64
128
256
512
1024
2048

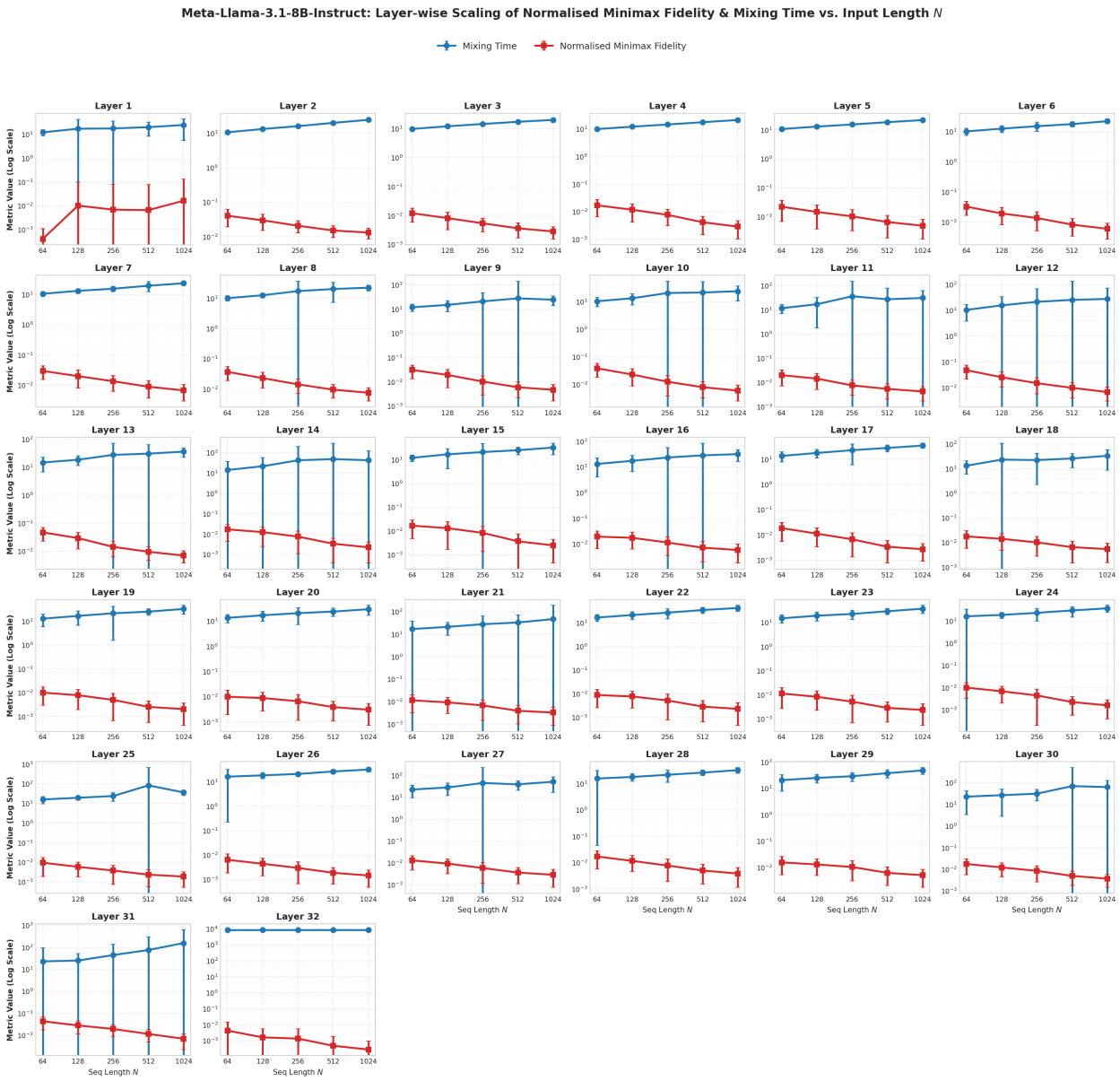

gemma-3-12b-it: Layer-wise Scaling of Normalised Minimax Fidelity & Mixing Time vs. Input Length N
Mixing Time
Normalised Minimax Fidelity
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Layer 13
Layer 14
Layer 15
Layer 16
Layer 17
Layer 18
Layer 19
Layer 20
Layer 21
Layer 22
Layer 23
Layer 24
Layer 25
Layer 26
Layer 27
Layer 28
Layer 29
Layer 30
Layer 31
Layer 32
Layer 33
Layer 34
Layer 35
Layer 36
Layer 37
Layer 38
Layer 39
Layer 40
Layer 41
Layer 42
Layer 43
Layer 44
Layer 45
Layer 46
Layer 47
Layer 48
Metric Value (Log Scale)
Seq Length N
64
128
256
512
1024

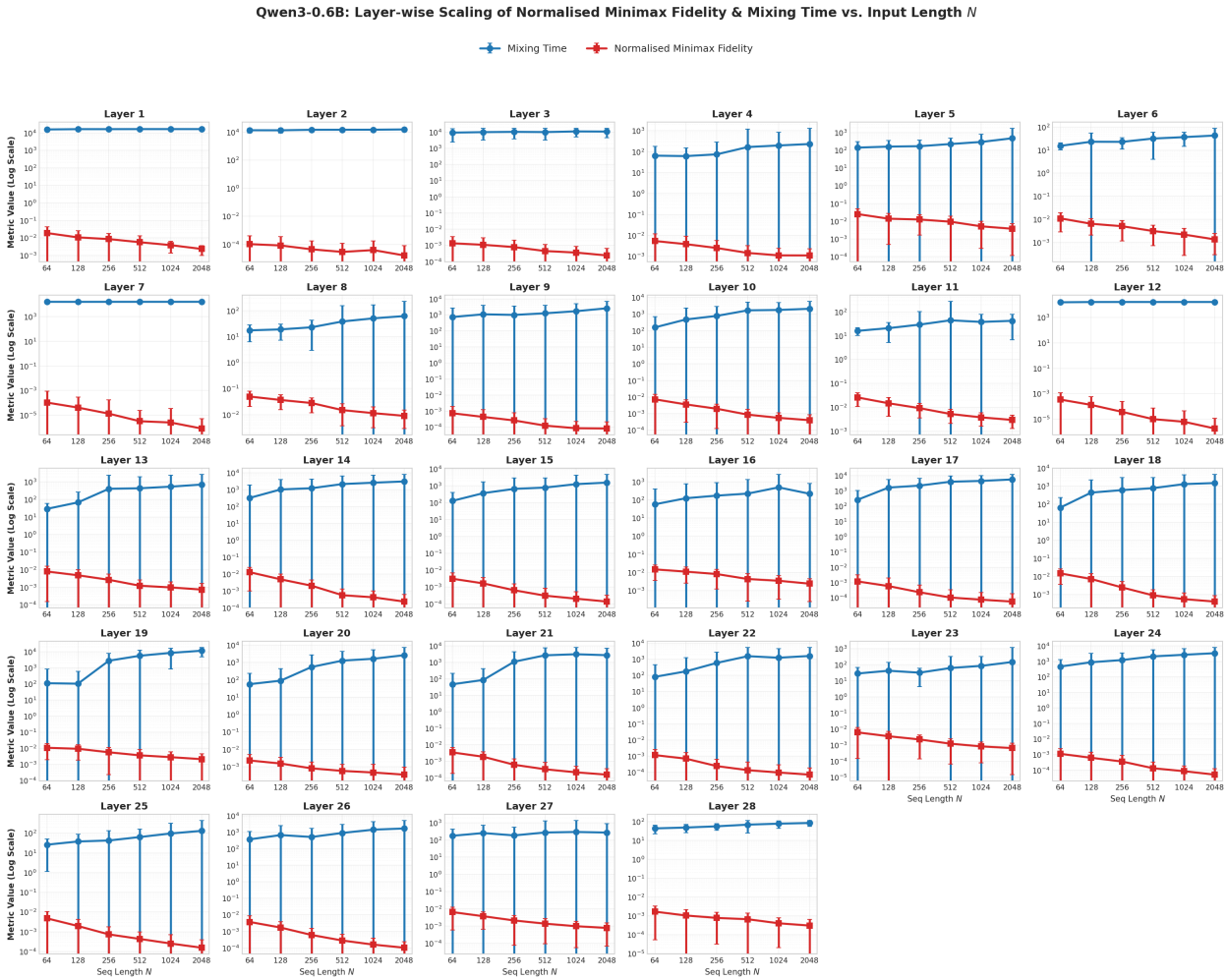

gemma-4-12B-it-hack: Layer-wise Scaling of Normalised Minimax Fidelity & Mixing Time vs. Input Length N
Mixing Time
Normalised Minimax Fidelity
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Layer 13
Layer 14
Layer 15
Layer 16
Layer 17
Layer 18
Layer 19
Layer 20
Layer 21
Layer 22
Layer 23
Layer 24
Layer 25
Layer 26
Layer 27
Layer 28
Layer 29
Layer 30
Layer 31
Layer 32
Layer 33
Layer 34
Layer 35
Layer 36
Layer 37
Layer 38
Layer 39
Layer 40
Layer 41
Layer 42
Layer 43
Layer 44
Layer 45
Layer 46
Layer 47
Layer 48
Metric Value (Log Scale)
Seq Length N
64
128
256
512
1024

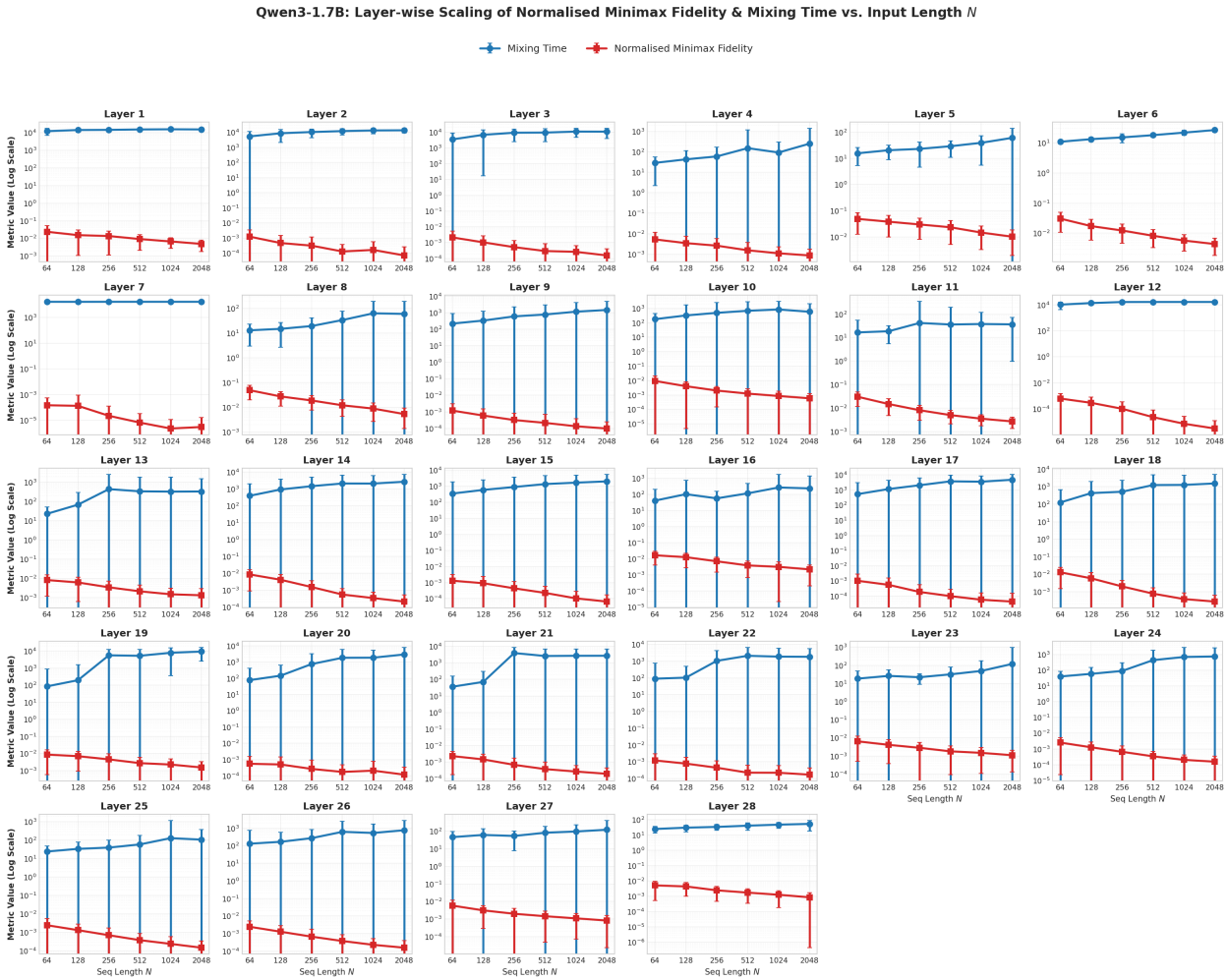

gemma-4-E2B-it: Layer-wise Scaling of Normalised Minimax Fidelity & Mixing Time vs. Input Length N
Mixing Time
Normalised Minimax Fidelity
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Layer 13
Layer 14
Layer 15
Layer 16
Layer 17
Layer 18
Layer 19
Layer 20
Layer 21
Layer 22
Layer 23
Layer 24
Layer 25
Layer 26
Layer 27
Layer 28
Layer 29
Layer 30
Layer 31
Layer 32
Layer 33
Layer 34
Layer 35
Metric Value (Log Scale)
Seq Length N
64
128
256
512
1024
2048

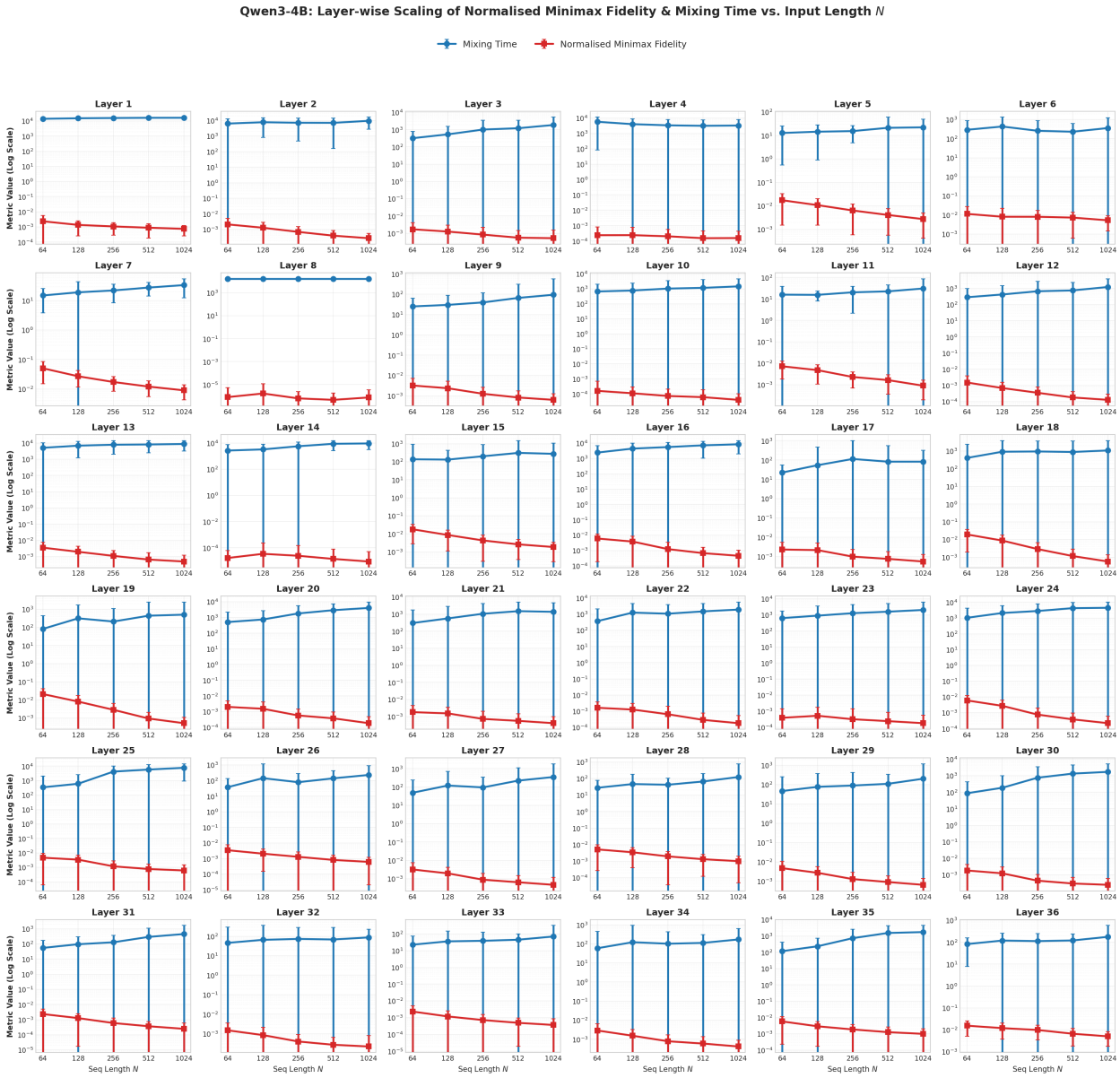

Llama-3.2-1B-Instruct: Layer-wise Scaling of Normalised Minimax Fidelity & Mixing Time vs. Input Length N
Mixing Time
Normalised Minimax Fidelity
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Layer 13
Layer 14
Layer 15
Layer 16
Metric Value (Log Scale)
Seq Length N


Llama-3.2-3B-Instruct: Layer-wise Scaling of Normalised Minimax Fidelity & Mixing Time vs. Input Length N
Mixing Time
Normalised Minimax Fidelity
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Layer 13
Layer 14
Layer 15
Layer 16
Layer 17
Layer 18
Layer 19
Layer 20
Layer 21
Layer 22
Layer 23
Layer 24
Layer 25
Layer 26
Layer 27
Layer 28
Metric Value (Log Scale)
Seq Length N


Meta-Llama-3.1-8B-Instruct: Layer-wise Scaling of Normalised Minimax Fidelity & Mixing Time vs. Input Length N
Mixing Time
Normalised Minimax Fidelity
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Layer 13
Layer 14
Layer 15
Layer 16
Layer 17
Layer 18
Layer 19
Layer 20
Layer 21
Layer 22
Layer 23
Layer 24
Layer 25
Layer 26
Layer 27
Layer 28
Layer 29
Layer 30
Layer 31
Layer 32
Metric Value (Log Scale)
Seq Length N
64
128
256
512
1024


Qwen3-0.6B: Layer-wise Scaling of Normalised Minimax Fidelity & Mixing Time vs. Input Length N
Mixing Time
Normalised Minimax Fidelity
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Layer 13
Layer 14
Layer 15
Layer 16
Layer 17
Layer 18
Layer 19
Layer 20
Layer 21
Layer 22
Layer 23
Layer 24
Layer 25
Layer 26
Layer 27
Layer 28
Metric Value (Log Scale)
Seq Length N
64
128
256
512
1024
2048


Qwen3-1.7B: Layer-wise Scaling of Normalised Minimax Fidelity & Mixing Time vs. Input Length N
Mixing Time
Normalised Minimax Fidelity
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Layer 13
Layer 14
Layer 15
Layer 16
Layer 17
Layer 18
Layer 19
Layer 20
Layer 21
Layer 22
Layer 23
Layer 24
Layer 25
Layer 26
Layer 27
Layer 28
Metric Value (Log Scale)
Seq Length N
64
128
256
512
1024
2048


Qwen3-4B: Layer-wise Scaling of Normalised Minimax Fidelity & Mixing Time vs. Input Length N
Mixing Time
Normalised Minimax Fidelity
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Layer 13
Layer 14
Layer 15
Layer 16
Layer 17
Layer 18
Layer 19
Layer 20
Layer 21
Layer 22
Layer 23
Layer 24
Layer 25
Layer 26
Layer 27
Layer 28
Layer 29
Layer 30
Layer 31
Layer 32
Layer 33
Layer 34
Layer 35
Layer 36
Metric Value (Log Scale)
Seq Length N
64
128
256
512
1024