# A STRUCTURAL THEORY OF POSITION BIAS IN TRANSFORMERS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Transformer models systematically favor certain token positions, yet the architectural origins of this position bias remain poorly understood. This bias is closely connected to the Lost-in-the-Middle phenomenon, where models underutilize information placed in the middle of the context. We show that causal Transformer architectures induce a systematic position-bias prior that can help explain the influence profiles associated with Lost-in-the-Middle. To characterize this prior, we develop a structural theory based on cumulative attention rollout with layerwise residual mixing. At finite depth, we isolate the competing positional drifts induced by causal masking, residual mixing, and positional encoding. Their interaction produces broad, often U-shaped profiles. At infinite depth, prior work predicts inevitable collapse under attention-only theory and identifies the discrepancy with observed Transformer behavior as an open problem. Under the stated boundedness assumptions, our residual-aware analysis replaces inevitable collapse with a dichotomy governed by cumulative attention mixing. Empirically, across 30 model-dataset settings, our adaptive residual-aware rollout closely matches measured input-token influence and reduces the average profile discrepancy by approximately $85\%$ relative to both the attention-only and fixed-residual rollout baselines.

## 1 INTRODUCTION

Despite their widespread success, the mechanisms by which Transformers propagate information across depth remain poorly understood. In particular, modern Transformer models exhibit strong and systematic position biases that persist across tasks, inputs, and even lexical scrambling (Rahimi et al., 2026). These include primacy bias (preference for early tokens), recency bias (preference for recent tokens), and the Lost-in-the-Middle (LiM) phenomenon, where information in the center of long contexts is underutilized (Liu et al., 2024). A growing body of empirical evidence suggests that these effects are not primarily semantic, but instead reflect architectural structure.

Recent theoretical work has sought to explain position bias through attention rollout, interpreting stacked attention layers as a stochastic information propagation process across tokens. In a recent fundamental analysis, Wu et al. (2025) show that, in attention-only causal Transformers, cumulative attention must collapse onto the first token as depth increases. This asymptotic prediction is at odds with the non-collapsed profiles observed in deep pretrained Transformers, a theory-observation gap that Wu et al. (2025) explicitly identify as an open problem.

Existing rollout models either omit the residual path or mix it with attention at a fixed one-half weight. We instead model their relative strength with an adaptive layerwise mixing schedule estimated from forward branch geometry. We analyze the resulting cumulative rollout and isolate three architectural forces at finite depth: causal masking induces primacy, stronger residual preservation induces recency, and any positional mechanism that shifts its transport kernels toward recent keys strengthens recency. In the models studied, these forces combine into broad, often U-shaped influence profiles, forming a structural prior consistent with Lost-in-the-Middle. At infinite depth, the cumulative mixing schedule determines whether rollout retains mass at later tokens or collapses onto the first token.

Our analysis compares attention-only and fixed-residual rollouts with our adaptive residual-aware rollout, which yields broad, typically U-shaped profiles. Comparison to empirical input-output token influence in pretrained language models shows close agreement, providing an architecture-faithful framework for understanding position bias.



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.

Our contributions are as follows:

- **Adaptive residual-aware attention rollout.** We model identity and attention paths through an adaptive layerwise mixing schedule, estimated from forward branch geometry without using the influence target.
- **Architectural origin of position bias.** We show that the architectural transport mechanisms studied here produce broad, often U-shaped influence profiles across ALiBi- and RoPE-based models, establishing a structural prior consistent with empirically observed position-dependent effects.
- **Resolution of the analytical open problem (Wu et al., 2025).** Within the residual-aware transport model, we generalize the prior inevitability result to a dichotomy: under the stated boundedness assumptions, cumulative attention mixing determines whether collapse occurs.
- **Empirical validation against measured token influence.** Across 30 model-dataset settings, our rollout reduces the mean Wasserstein distance to measured token influence from $0.62$ for Wu et al. to $0.09$, an $85\%$ reduction, while also outperforming fixed residual mixing.

## 2 Related work

**Attention rollout and theoretical position bias.** Attention rollout and information-flow perspectives have long been used to analyze Transformer behavior (Abnar & Zuidema, 2020; Chefer et al., 2021; Barbero et al., 2024). Most closely related to our work, Wu et al. (2025) provide a rigorous analysis of attention-only causal Transformers and prove that cumulative attention must collapse onto the first token as depth increases. They also examine residual effects empirically, but their collapse theorem uses an attention-only propagation model. Empirically, such collapse is not observed in modern large-scale Transformers, motivating a theory that tracks effective residual mixing across layers. Concurrent work derives closed forms for a uniform-Cesàro, constant-mixing specialization of residual rollout, which also exhibits fixed-length infinite-depth primacy collapse (Chowdhury, 2026); Appendix B gives the precise comparison. In a simplified high-dimensional model, Kim et al. (2026) show that LayerNorm induces recency in second-layer attention scores, including in the presence of residual connections; their target is local attention-score ordering rather than cumulative input-output influence across depth. ALTI shows that raw attention weights can miss value- and output-projection effects by constructing vector-aware token contributions (Ferrando et al., 2022). Our RoPE kernel uses an ALTI-style scalarization for corpus-averaged transport, while our target is systematic position bias rather than input-specific attribution.

**Empirical position bias.** A substantial body of empirical work documents systematic position bias in Transformer attention, including primacy bias or "attention sinks" at early tokens (Zheng et al., 2023; Xiao et al., 2024; Yin et al., 2024; Guo et al., 2024; Kaul et al., 2025; Cobbina & Zhou, 2025) and recency bias favoring late tokens (Zhao et al., 2021; Sun et al., 2021; He et al., 2023; Liu et al., 2024). These effects are robust across inputs and tasks, persist under lexical scrambling (Rahimi et al., 2026), vary across positional encodings such as RoPE (Su et al., 2024) and ALiBi (Press et al., 2022; Kazemnejad et al., 2023; Yang et al., 2025), and arise even without explicit encodings (Guo et al., 2024), suggesting architectural rather than semantic origins. Building on the multilayer account of causal masking and positional encodings in Wu et al. (2025), we characterize how adaptive, layer-dependent residual mixing changes these dynamics at finite and infinite depth. The resulting corpus-level rollouts closely match gradient and blank-out token influence in pretrained Transformers and substantially improve upon both the attention-only and fixed-residual rollout baselines.

**Lost-in-the-Middle phenomenon.** The coexistence of primacy and recency biases gives rise to the Lost-in-the-Middle effect, where model performance degrades for information located in the center of long contexts (Liu et al., 2024; Guo & Vosoughi, 2025; Menschikov et al., 2025). Prior work characterizes this effect empirically and offers functional interpretations (Veseli et al., 2025; Salvatore et al., 2025). Several approaches aim to control such biases through architectural modifications or alternative attention mechanisms by reformulating softmax (Kaul et al., 2025) or introducing multi-scale positional encodings (Zhang et al., 2024). While effective in practice, neither line of work provides a unified architectural account of why U-shaped profiles arise or why such interventions are necessary. Our theory offers this principled perspective and complements existing mitigations.

## 3 NOTATION AND PROBLEM SETUP

We introduce notation and define a theory for analyzing how architectural components of Transformer models shape cumulative cross-token information propagation across layers, without modeling training objectives or optimization dynamics. Our framework is analytically tractable yet close to practical architectures. It explicitly models cross-token transport, residual connections, attention masking, and relative positional encodings, while abstracting away token-wise components such as MLP blocks and layer normalization.

**Tokens and layers.** We consider a sequence of $n$ tokens indexed by $i, j \in \{1, \dots, n\}$ and a Transformer of depth $T$, with layers $t \in \{1, \dots, T\}$.

**Masks.** Attention is constrained by a binary mask $\mathcal{M} \subseteq \{1, \dots, n\}^2$, where $(i, j) \in \mathcal{M}$ indicates that token $i$ may attend to token $j$. We consider causal masking: $(i, j) \in \mathcal{M}$ iff $j \leq i$.

At layer $t$, let $A^{(t)} \in \mathbb{R}^{n \times n}$ be a nonnegative, row-stochastic matrix supported on $\mathcal{M}$, representing cross-token propagation. We instantiate this kernel for ALiBi and RoPE as follows.

**ALiBi kernel.** For ALiBi, the kernel is the attention matrix, computed via a masked row-wise softmax

$$A^{(t)}_{i\cdot} = \mathrm{softmax}(\ell^{(t)}_{i\cdot}), \quad \ell^{(t)}_{ij} = \begin{cases} s^{(t)}_{ij} + b^{(t)}_{ij} & \text{if } (i,j) \in \mathcal{M}, \\ -\infty & \text{otherwise,} \end{cases}$$

where $b^{(t)}_{ij}$ captures position-dependent biases and $s^{(t)}_{ij}$ captures content-dependent contributions. ALiBi uses $b^{(t)}_{ij} = -m(i-j)$ with positional slope $m \geq 0$. This expression uses single-head notation.

**RoPE kernel.** For RoPE, we likewise write $A^{(t)}_{i\cdot} = \mathrm{softmax}(\ell^{(t)}_{i\cdot})$. Here $\ell^{(t)}_{ij}$ denotes an effective transport logit incorporating the raw attention, which already includes RoPE's positional information, and the value and output projections. For $(i, j) \notin \mathcal{M}$, set $\ell^{(t)}_{ij} = -\infty$. The effective logits are specified in Section F.

**Residual connections and mixing.** Each attention layer is embedded in a residual block

$$X^{(t)} \;=\; X^{(t-1)} + \mathrm{Attn}(X^{(t-1)}),$$

combining identity propagation with an attention-driven update. At the level of cross-token propagation, we represent the attention-driven update by the row-stochastic transport kernel $A^{(t)}$ defined above. For ALiBi this is the attention matrix; for RoPE the effective kernel also accounts for the value and output projections.

**Residual-aware transition with adaptive mixing.** Existing formulations account for residuals by augmenting attention with identity and renormalizing; e.g., Abnar & Zuidema (2020) replace $A^{(t)}$ with a fixed convex combination $\frac{1}{2}I + \frac{1}{2}A^{(t)}$, where $I \in \mathbb{R}^{n \times n}$ is the identity matrix. Residual networks also admit an interpretation as forward-Euler discretizations of continuous-time dynamics (Haber & Ruthotto, 2017; Chen et al., 2018). Since numerical step sizes may vary across steps (Hairer et al., 1993), this viewpoint motivates layer- and input-dependent attention strength rather than a fixed one-half weight. Since our analysis concerns only the relative contribution of the identity and attention paths, let $\lambda_t \in [0, 1]$ denote the attention-branch weight. The resulting convex combination induces the row-stochastic residual-aware transition matrix

$$R^{(t)} \;:=\; (1 - \lambda_t)I + \lambda_t A^{(t)}. \tag{1}$$

The theory allows any specified schedule $\{\lambda_t\}_{t=1}^{T}$; its empirical construction is given in Section 6.1.

**Cumulative rollout.** Cumulative information propagation to depth $T$ is captured by the rollout matrix

$$P^{(T)} := R^{(T)} R^{(T-1)} \cdots R^{(1)}.$$

The entry $P^{(T)}_{ij}$ quantifies the rollout influence of input token $j$ on token $i$ within the transport model. Each $R^{(t)}$ is row-stochastic; hence $P^{(T)}$ is row-stochastic as well. Under causal masking, the last-row distribution $p^{(T)}(j) := P^{(T)}_{nj}$ describes this influence on the final position used to predict the next token. This distribution is the central object: we analyze how attention masking, positional encodings, and the residual-mixing schedule shape $p^{(T)}$ at finite depth and whether it collapses in the infinite-depth limit.

## 4 FINITE-DEPTH ATTENTION DYNAMICS: THREE FORCES

To characterize systematic architectural effects at finite depth that shape the rollout distribution $p^{(T)}$, our analysis relies on the following weak positional monotonicity assumption on the transport kernels, reflecting the ordering induced by causal masking while remaining agnostic to content and parameter values.

**Assumption 4.1** (Stochastically monotone transport kernels)**.** For each layer $t \in \{1, \dots, T\}$, let $A^{(t)} \in \mathbb{R}^{n \times n}$ denote the transport kernel. We assume that $A^{(t)}$ is stochastically monotone (Definition A.1) with respect to the natural order on positions: for all $1 \le i < i' \le n$ and all prefix cutoffs $k \in \{1, \dots, n\}$,

$$\sum_{j=1}^{k} A_{ij}^{(t)} \;\ge\; \sum_{j=1}^{k} A_{i'j}^{(t)}.$$

Equivalently, for $i < i'$, the row distribution $A_{i,\cdot}^{(t)}$ is first-order stochastically smaller than $A_{i',\cdot}^{(t)}$.

This condition is naturally compatible with a broad class of transport kernels induced by causal masking and smoothly decaying or symmetric kernels. It is especially natural when $A^{(t)}$ is interpreted as an effective (head-aggregated and corpus-averaged) kernel that abstracts away content-dependent fluctuations and captures net architectural tendencies.

For the corpus-averaged ALiBi kernels, strict prefix-monotonicity is not always satisfied exactly, but deviations are numerically negligible. Across models, layers, and datasets, violations are extremely rare, occurring for at most $10^{-6}$ of all triples $(i < i', k)$. The mean prefix-mass shortfall conditional on a violation is approximately $2 \times 10^{-6}$, yielding an unconditional expected shortfall on the order of $10^{-12}$. For the projection-aware RoPE kernels, violations occur for approximately $0.42\%$ of all triples. The mean prefix-mass shortfall conditional on a violation is approximately $8.3 \times 10^{-5}$, yielding an unconditional expected shortfall of approximately $3.5 \times 10^{-7}$. These results indicate that $A^{(t)}$ is extremely close to prefix-monotone in practice, supporting Assumption 4.1 as a stable regularity condition for the finite-depth force decomposition.

Under this regularity condition, the following propositions isolate three architectural forces represented in our transport model.

**Causal masking induces primacy drift.** When composed across layers, the directional asymmetry of causal masking accumulates into a drift toward earlier positions in the rollout.

**Proposition 4.2** (Primacy drift under causal masking)**.** *Assume causal masking and Assumption 4.1. Let $p^{(t)}$ for $t \in \{1, \dots, T-1\}$ denote the last-row rollout distribution $p^{(t)}(j) = P_{nj}^{(t)}$. Then for every prefix length $k < n$, the prefix mass is monotonically increasing:*

$$\sum_{j=1}^{k} p^{(t+1)}(j) \;\ge\; \sum_{j=1}^{k} p^{(t)}(j).$$

*Consequently, at finite depth the last-row rollout distribution exhibits a systematic drift of mass toward earlier positions.*

**Residual mixing controls recency drift.** Decreasing $\lambda_t$ strengthens identity propagation and reduces cross-token mixing. Information is then retained more strongly at its current position rather than propagated backward, biasing finite-depth rollout toward recent tokens.

**Proposition 4.3** (Residual strength induces recency drift)**.** *Assume causal masking and Assumption 4.1. Fix transport kernels $\{A^{(t)}\}_{t=1}^{T}$ and consider two residual-mixing schedules $\{\lambda_t\}$ and $\{\lambda'_t\}$ with $\lambda_t \le \lambda'_t$ for all $t$. Let $p^{(T)}$ and $p'^{(T)}$ denote the corresponding last-row rollout distributions. Then for every prefix length $k < n$,*

$$\sum_{j=1}^{k} p^{(T)}(j) \;\le\; \sum_{j=1}^{k} p'^{(T)}(j).$$

*Equivalently, strengthening residual connections (i.e., decreasing $\lambda_t$) induces a recency drift in the last-row rollout distribution.*

**Positional encodings induce recency drift.** When positional encodings systematically prefer recent keys, they induce a recency bias that compounds across layers through rollout.

**Proposition 4.4** (Positional encodings induce recency drift)**.** *Assume causal masking. For each layer $t$, let*

$$A_0^{(t)} := \mathrm{softmax}(s^{(t)}), \qquad A_{\mathrm{PE}}^{(t)} := \mathrm{softmax}(s^{(t)} + b^{(t)}),$$

*where positional logits $b_{ij}^{(t)}$ are recency-favoring in the sense that, for all admissible keys $j < k$,*

$$b_{ij}^{(t)} \le b_{ik}^{(t)}.$$

*Assume that $\{A_0^{(t)}\}_{t=1}^T$ satisfies Assumption 4.1.*

*Let*

$$R_0^{(t)} := (1-\lambda_t)I + \lambda_t A_0^{(t)}, \qquad R_{\mathrm{PE}}^{(t)} := (1-\lambda_t)I + \lambda_t A_{\mathrm{PE}}^{(t)},$$

*and denote the corresponding rollout distributions by $p_0^{(T)}$ and $p_{\mathrm{PE}}^{(T)}$.*

*Then for every prefix length $k < n$,*

$$\sum_{j=1}^{k} p_{\mathrm{PE}}^{(T)}(j) \;\le\; \sum_{j=1}^{k} p_0^{(T)}(j).$$

*Equivalently, positional encodings induce a recency drift in the last-row rollout distribution.*

**Example (ALiBi).** Under causal masking, ALiBi assigns larger biases to more recent keys for each query position and therefore satisfies the recency-favoring condition of Proposition 4.4. When the baseline kernels satisfy Assumption 4.1, ALiBi consequently induces a recency drift in the final rollout distribution.

**Structural forces and Lost-in-the-Middle.** The propositions identify competing architectural drifts at finite depth. In the models studied below, their combination yields broad, non-collapsed profiles in which early and recent tokens receive more influence than intermediate positions. This gives a structural account of the U-shaped influence prior associated with the Lost-in-the-Middle phenomenon, without claiming that every admissible kernel and mixing schedule must be U-shaped. Appendix C gives explicit sufficient conditions and a computable finite-mixing certificate under the common-kernel approximation.

## 5 Infinite-depth attention dynamics

We now turn from finite- to infinite-depth attention dynamics. For the attention mechanism and its boundedness conditions, we follow the setup of Wu et al. (2025). We additionally allow bounded additive positional-bias logits. They show that attention-only causal rollout collapses to the first token in the infinite-depth limit. We study the corresponding residual-aware dynamics, with their attention-only setting recovered by taking $\lambda_t \equiv 1$.

Let $W_Q^{(t)}$, $W_K^{(t)}$, and $W_V^{(t)}$ denote the query, key, and value projections. For admissible keys, the content logits are $s_{ij}^{(t)} = (X_{i,:}^{(t-1)} W_Q^{(t)})(X_{j,:}^{(t-1)} W_K^{(t)})^\top$, and $A^{(t)}$ is the causal row-wise softmax of $s^{(t)} + b^{(t)}$. The normalized dynamics are

$$X^{(t)} = (1-\lambda_t) X^{(t-1)} + \lambda_t A^{(t)} X^{(t-1)} W_V^{(t)}.$$

Without additive positional-bias logits, the bounded-bias assumption below is trivially satisfied and the collapse branch of our theorem reduces to their result.

**Assumption 5.1** (Bounded query and key operators)**.** As in Wu et al. (2025), we assume there exists a constant $C > 0$ such that $\|W_Q^{(t)}\|_2 \le C$ and $\|W_K^{(t)}\|_2 \le C$ for all $t \ge 1$, where $\|\cdot\|_2$ is the operator norm.

**Assumption 5.2** (Bounded value accumulation)**.** We assume the sequence of value transformations is well-behaved, such that the layer-wise operator norms $C_t := \|W_V^{(t)}\|_2$ satisfy $\sum_{t=1}^{\infty} \lambda_t \max(0, C_t - 1) < \infty$. This is the residual-aware analogue of the bounded value-accumulation condition used by Wu et al. (2025). Because residual connections introduce skip paths that bypass contiguous prefix products, stability requires constraining the excess growth of the layer-wise transformations directly.

**Assumption 5.3** (Bounded positional-bias range)**.** When additive positional-bias logits $b_{ij}^{(t)}$ are present, we assume that their admissible row-wise range is uniformly bounded: there exists $B < \infty$ such that $|b_{ij}^{(t)} - b_{ik}^{(t)}| \le B$ for all layers $t$, query positions $i$, and admissible keys $j, k$.

**Lemma 5.4** (Uniform stability and attention lower bound)**.** *Suppose Assumptions 5.1-5.3 hold and the initial input is bounded such that $\|X^{(0)}\|_2 \le C$ under causal masking. Then there exist constants $C_X > 0$ and $\varepsilon > 0$, both independent of the depth $t$, such that for all $t \ge 1$:*

(i) *The hidden states are uniformly bounded:* $\max_i \|X_{i,:}^{(t)}\|_2 \le C_X$.

(ii) *The attention weights are uniformly bounded away from zero:* $A_{ij}^{(t)} \ge \varepsilon$ *for all* $(i, j) \in \mathcal{M}$.

**Theorem 5.5** (Residual-aware infinite-depth collapse dichotomy)**.** *Under causal masking and Assumptions 5.1-5.3, the asymptotic behavior of the residual-aware attention rollout $P^{(T)} := R^{(T)} \cdots R^{(1)}$ is fully determined by the summability of the mixing coefficients $\{\lambda_t\}_{t \ge 1}$.*

*(i)* ***Finite total mixing*** *⇒* ***no collapse.*** *If $\sum_{t=1}^{\infty} \lambda_t < \infty$, then for every token $i \in \{1, \dots, n\}$,*

$$\liminf_{T \to \infty} P_{ii}^{(T)} \;\ge\; \prod_{t=1}^{\infty} \big(1 - (1 - \varepsilon)\lambda_t\big) \;>\; 0.$$

*Thus, for every $i > 1$, the rollout does not collapse to the first token.*

*(ii)* ***Infinite total mixing*** *⇒* ***collapse.*** *If $\sum_{t=1}^{\infty} \lambda_t = \infty$, then for every token $i \in \{1, \dots, n\}$, $\lim_{T \to \infty} P_{i1}^{(T)} = 1$. Moreover, for all $1 < j \le i$ and all $T \ge 1$, there exists a constant $\varepsilon > 0$ such that*

$$P_{ij}^{(T)} \;\le\; \exp\Big(-(j-1)\varepsilon \sum_{s=1}^{T} \lambda_s\Big),$$

*and consequently $P_{ij}^{(T)} \to 0$ for all $j > 1$ as $T \to \infty$.*

**Discussion.** Thus, in the asymptotic analytical setting of Wu et al. (2025), cumulative residual mixing, rather than depth alone, governs collapse: fixed positive mixing still collapses. The depth-uniform lower bound $\varepsilon > 0$ supports both the quantitative bounds and the collapse conclusion. The theorem neither establishes that measured finite-model schedules are summable nor extends its positive-attention bound to empirical projection-aware RoPE kernels. Its finite-$T$ bounds quantify how cumulative mixing controls the approach to the asymptotic limit.

**Multi-head extension.** We model the effective attention operator at layer $t$ as the uniform head average over the $H$ causal row-stochastic head kernels $A^{(t,h)}$, $A^{(t)} := \frac{1}{H} \sum_{h=1}^{H} A^{(t,h)}$. Since prefix masses and residual-aware transitions depend linearly on the underlying attention kernel, all finite- and infinite-depth results established for a single head apply verbatim to $A^{(t)}$.

## 6 Empirical validation

After describing the experimental setup, we construct corpus-level rollouts and measure residual mixing schedules in pretrained LLMs. We then compare three rollout rules using gradient-based input-token influence: (a) the attention-only rollout of Wu et al. (2025), with $\lambda_t = 1$; (b) the fixed residual augmentation of Abnar & Zuidema (2020), with $\lambda_t = \frac{1}{2}$; and (c) our adaptive residual-aware rollout, with a measured layerwise $\lambda_t$.

**Experimental setup.** We evaluate a suite of pretrained ALiBi and RoPE-based language models summarized in Table 1. The ALiBi models comprise `BLOOM` (Scao et al., 2022), `MPT` (Mosaic, 2023a), and `Falcon` (Penedo et al., 2023); the RoPE-based models comprise `Qwen2.5` (Qwen Team, 2024), `Qwen3` (Qwen Team, 2025), and the hybrid RoPE/NoPE `SmolLM3` (Hugging Face Smol Models Team, 2025). At sequence length $n = 256$, both the ALiBi and RoPE results average over 1,000 FineWeb-Edu prompts (Penedo et al., 2024). Appendix G provides the available results on DCLM-Baseline and Wikipedia (Li et al., 2024; Wikimedia Foundation, 2026); Appendix I provides longer-context results at $n = 2048$. The complete code is available here[1].

[1] `https://anonymous.4open.science/r/position_bias_submit-7406`

### 6.1 CORPUS-LEVEL ROLLOUT CONSTRUCTION

Our empirical target is the systematic input-token influence profile of a pretrained Transformer across a corpus. Let $\mathcal{D}$ denote the evaluation-corpus distribution and allow the kernel and mixing coefficient to depend on an input $x$, with $R^{(t)}(x) = (1 - \lambda_t(x))I + \lambda_t(x)A^{(t)}(x)$. The exact corpus-level prediction of the transport model is

$$\bar{P}^{(T)} := \mathbb{E}_{x\sim\mathcal{D}}\left[R^{(T)}(x)\cdots R^{(1)}(x)\right], \qquad \bar{p}^{(T)} = e_n^\top \bar{P}^{(T)}. \tag{2}$$

The influence profile is not directly observable; we use gradient influence as the main empirical proxy and validate it with token blank-out in Appendix I.3.

We approximate the exact corpus-level transport prediction in Equation (2) in two steps. First, we interchange the expectation of the ordered product with the ordered product of expectations:

$$\mathbb{E}_{x\sim\mathcal{D}}\left[R^{(T)}(x)\cdots R^{(1)}(x)\right] \approx \mathbb{E}_{x\sim\mathcal{D}}[R^{(T)}(x)]\cdots\mathbb{E}_{x\sim\mathcal{D}}[R^{(1)}(x)]. \tag{3}$$

This neglects cross-layer dependence between the input-dependent transitions while retaining the average transition at each layer.

As a second approximation, we apply a mean-field factorization within each layer. Writing $\bar{\lambda}_t = \mathbb{E}_{x\sim\mathcal{D}}[\lambda_t(x)]$ and $\bar{A}^{(t)} = \mathbb{E}_{x\sim\mathcal{D}}[A^{(t)}(x)]$, we replace $\mathbb{E}_{x\sim\mathcal{D}}[\lambda_t(x)A^{(t)}(x)]$ with $\bar{\lambda}_t\bar{A}^{(t)}$. This gives

$$\mathbb{E}_{x\sim\mathcal{D}}[R^{(T)}(x)]\cdots\mathbb{E}_{x\sim\mathcal{D}}[R^{(1)}(x)] \approx P_{\mathrm{mf}}^{(T)} := \left[(1-\bar{\lambda}_T)I + \bar{\lambda}_T\bar{A}^{(T)}\right]\cdots\left[(1-\bar{\lambda}_1)I + \bar{\lambda}_1\bar{A}^{(1)}\right]. \tag{4}$$

Thus each expected transition is represented using the separately averaged mixing coefficient and transport kernel. Appendix B distinguishes the two approximations and develops further rollout simplifications. Appendix I.1 evaluates them separately. Across all tested models and corpora, neither introduces more than $6.5 \times 10^{-3}$ normalized final-row 1-Wasserstein error, well below the distances between rollout and measurement reported below. The resulting corpus-level model prediction is $\bar{p}_{\mathrm{mf}}^{(T)} = e_n^\top P_{\mathrm{mf}}^{(T)}$.

This construction is designed to isolate architectural transport rather than reproduce corpus-level influence through learned content associations. We keep the model's depth, causal mask, positional mechanism, and residual mixing, while removing input-dependent content as far as the positional mechanism permits. The residual mixing schedule is still measured from the pretrained model, as defined next, but it is estimated independently of the influence profiles used for evaluation. The two approximations above are applied after choosing the transport kernels. Removing the ALiBi content logits below is a controlled architectural intervention, not another mean-field approximation.

For each input and layer, we measure the attention-branch weight

$$\lambda_t(x) = \frac{\|X^{(t)}\| + \|\mathrm{Attn}(X^{(t-1)})\| - \|X^{(t-1)}\|}{2\|X^{(t)}\|}, \tag{5}$$

using Frobenius norms over tokens and features. This two-branch Shapley allocation (Shapley, 1953), applied to the coalition value $v(S) = \|\sum_{b\in S} b\|$, accounts for reinforcement and cancellation between the residual stream and attention update. No influence measurement enters this estimate. For nonzero $\|X^{(t)}\|$, the triangle and reverse-triangle inequalities give $0 \le \lambda_t(x) \le 1$, as required by the transport model. In practice, we estimate all corpus expectations by empirical averages.

Appendix H reports the measured norm-Shapley schedules for both model families; Appendix I.5 evaluates sensitivity to the alternative relative-norm calibration.

It remains to construct the kernels $\bar{A}^{(t)}$. All kernels match the depth, masking pattern, and positional encoding of the corresponding pretrained model. For ALiBi, we can perform this isolation directly: we remove content by setting $s_{ij}^{(t)} = 0$ in the attention logits. The resulting kernels retain only the causal mask and the model's fixed positional slopes, and therefore contain no input-dependent content.

The same intervention is unavailable for RoPE: setting its content logits to zero would also remove the rotary positional signal. We therefore construct projection-aware forward-contribution kernels

Table 1: Causal Transformer models included in the experimental study.

| Model | Encoding | Params. | Depth |
|---|---|---|---|
| `falcon-rw-7b` | ALiBi | 7B | 36 |
| `mpt-7b` | ALiBi | 7B | 32 |
| `mpt-30b` | ALiBi | 30B | 48 |
| `bloom-7b1` | ALiBi | 7.1B | 30 |
| `bloom-176b` | ALiBi | 176B | 70 |
| `qwen2.5-7b` | RoPE | 7B | 28 |
| `qwen2.5-32b` | RoPE | 32B | 64 |
| `qwen2.5-72b` | RoPE | 72B | 80 |
| `qwen3-8b-base` | RoPE | 8B | 36 |
| `smollm3-3b-base` | RoPE | 3B | 36 |

Table 2: Normalized 1-Wasserstein distances between rollout profiles and gradient influence.

| Model | (a) Wu $\lambda = 1$ | (b) Abnar $\lambda = \frac{1}{2}$ | (c) Ours adaptive |
|---|---|---|---|
| `falcon-rw-7b` | 0.66 | 0.63 | **0.03** |
| `mpt-7b` | 0.62 | 0.58 | **0.15** |
| `mpt-30b` | 0.61 | 0.61 | **0.02** |
| `bloom-7b1` | 0.63 | 0.58 | **0.13** |
| `bloom-176b` | 0.65 | 0.65 | **0.11** |
| `qwen2.5-7b` | 0.63 | 0.59 | **0.12** |
| `qwen2.5-32b` | 0.59 | 0.59 | **0.11** |
| `qwen2.5-72b` | 0.60 | 0.60 | **0.12** |
| `qwen3-8b-base` | 0.58 | 0.57 | **0.09** |
| `smollm3-3b-base` | 0.60 | 0.60 | **0.11** |

and average them over the corpus. These kernels retain attention values and head-specific output projections but use neither gradients nor blank-out measurements; Appendix F gives the construction. This yields a strongly compressed, corpus-averaged transport model that preserves RoPE's positional signal.

We compute $\bar{p}_{\mathrm{mf}}^{(T)}$ using the same kernels under three mixing rules. Variant (a), following Wu et al. (2025), sets $\bar{\lambda}_t = 1$ and removes the residual path. Variant (b), following the residual-augmented rollout of Abnar & Zuidema (2020), sets $\bar{\lambda}_t = \frac{1}{2}$. Variant (c), ours, uses the adaptive measured schedule $\{\bar{\lambda}_t\}_{t=1}^{T}$. The comparison in Figure 1 shows variants (a)–(c) alongside measured influence.

### 6.2 Gradient input-token influence

For an input $x$ with embeddings $\mathbf{e}$ and predicted next token $y(x) = \arg\max_v z_v(x)$, the local Jacobian-norm score is

$$g_j(x) = \left\| \nabla_{\mathbf{e}_j} P(y(x) \mid x) \right\|_2$$

(Simonyan et al., 2014; Liu et al., 2026). We average raw scores across the corpus and then normalize across positions:

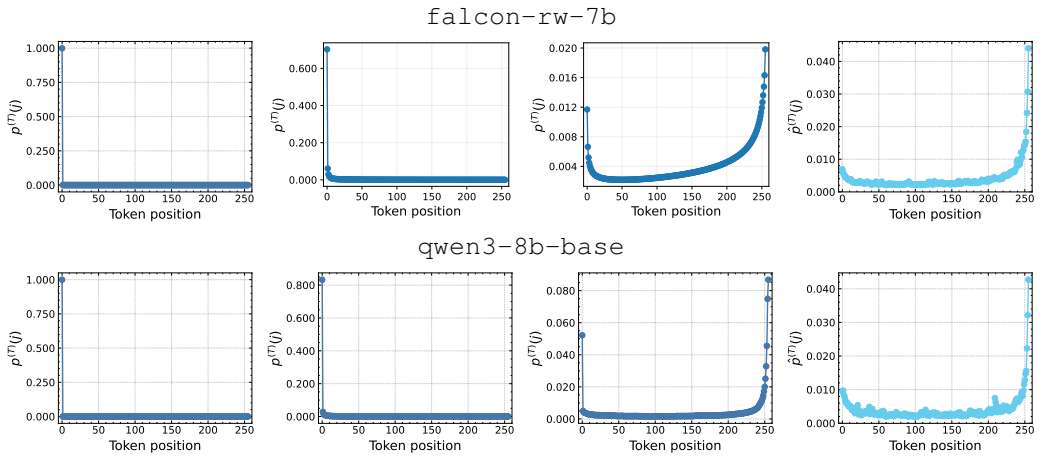
$$q_{\mathrm{grad}}(j) = \frac{\mathbb{E}_{x\sim\mathcal{D}}[g_j(x)]}{\sum_k \mathbb{E}_{x\sim\mathcal{D}}[g_k(x)]}. \tag{6}$$

We write $\hat{q}_{\mathrm{grad}}$ for the corresponding empirical profile. Appendix I.3 validates the comparison using token blank-out as an alternative, gradient-free influence measure. Figure 1 shows $\hat{q}_{\mathrm{grad}}$ for `falcon-rw-7b` and `qwen3-8b-base`. In both cases, the empirical distributions exhibit a U-shaped profile.

**Qualitative validation.** Consistent with the controlled analysis above, our adaptive residual-aware rollout reproduces the dominant U-shaped structure observed in the measured distribution. For the ALiBi model shown here, this agreement is obtained without input-dependent content logits. By contrast, the attention-only rollout (a) collapses to the first token and fails to reproduce the empirical profile.

**Quantitative validation.** Table 2 reports normalized 1-Wasserstein distances between the rollout variants and the measured influence distribution at sequence length $n = 256$, using ground metric $d(i, j) = |i - j|/(n - 1)$. Appendix G reports additional model-dataset results. Appendix I contains the corpus-level approximation, context-length, and mixing-coefficient ablations, and Appendix I.3 replaces gradient influence by token blank-out. Across all 30 model-dataset comparisons, our adaptive rollout (c) attains a lower Wasserstein distance than both the attention-only rollout (a) and the fixed residual rollout (b). As an independent intervention test, the rollout correctly predicts the point-estimate direction of primacy and recency changes under attention-branch rescaling in all 24 comparisons across six models (Section I.4).

Together, these results show that adaptive layerwise residual mixing recovers the broad, often U-shaped positional profiles missed by both attention-only and fixed-residual rollout.




(a) Attention-only rollout ($\lambda_t = 1$) (b) Fixed residual rollout ($\lambda_t = \frac{1}{2}$) (c) Adaptive residual rollout (d) Measured input token influence

Figure 1: Final-token profiles at sequence length $n = 256$ for the 36-layer `falcon-rw-7b` (top) and the 36-layer `qwen3-8b-base` (bottom), representing ALiBi and RoPE, respectively. Panels (a)–(c) show the Wu et al. attention-only rollout ($\lambda_t = 1$), the Abnar–Zuidema fixed residual rollout ($\lambda_t = \frac{1}{2}$), and our adaptive rollout with measured schedules $\{\bar{\lambda}_t\}$, respectively. Panel (d) shows the gradient influence proxy $\hat{q}_{\text{grad}}$.

## 7 Discussion and Limitations

**Architectural scope.** Our analysis isolates architectural position bias induced by attention, masking, positional encodings, and residual mixing. The rollout uses only forward-pass quantities, independently of the influence profiles used for evaluation, and characterizes token-mixing priors rather than full representation dynamics. We do not model normalization as a separate force; nevertheless, the content-free ALiBi rollout closely matches measured cumulative influence, indicating that an explicit normalization term is unnecessary for obtaining this predictive approximation.

**Positional encodings.** The finite-depth comparison theorem is stated for ordered transport kernels, with an analytic additive-logit corollary for ALiBi. The projection-aware RoPE rollout retains corpus-averaged forward contributions rather than removing content entirely, but does not use the measured influence target.

**Lost-in-the-Middle.** Our analysis characterizes a U-shaped influence prior through theoretical rollout and gradient-based input-token attribution. Lost-in-the-Middle, as introduced by Liu et al. (2024), is a downstream task-performance phenomenon; directly connecting this prior to retrieval-style degradation is left for future work.

## 8 Conclusion

We developed a residual-aware theory of cumulative information propagation in which an adaptive layerwise schedule controls the balance between identity and attention paths. At finite depth, the theory isolates the competing positional drifts induced by causal masking, residual mixing, and positional encoding. In pretrained ALiBi- and RoPE-based models, their composition predicts broad, often U-shaped influence profiles. Across 30 model-dataset settings, our rollout reduces the mean Wasserstein distance to measured token influence from $0.62$ for attention-only rollout and $0.60$ for fixed residual mixing to $0.09$, an approximately $85\%$ reduction relative to either baseline.

At infinite depth, cumulative mixing, rather than the mere presence of a residual path, determines whether later-token mass is retained or collapses to the first token under the stated boundedness conditions. Together, these results establish adaptive residual mixing as a key component connecting cumulative attention transport to the systematic position bias observed in causal Transformers.

## AI USE STATEMENT

Generative AI tools were used to assist with developing and refining proofs, implementing code for experiments, retrieving and summarizing related work, and drafting and editing the manuscript. The authors independently checked the mathematical claims, proofs, references, implementation, and reported results, and take responsibility for the final content of the paper.

## REPRODUCIBILITY STATEMENT

The appendix provides complete proofs, kernel constructions, experimental details, and all ablations. The complete code is available at `https://anonymous.4open.science/r/position_bias_submit-7406`. It includes a locked software environment, experiment manifests, and commands for reproducing the measurements, tables, and figures. Additional protocol, uncertainty, and compute details are given in Appendix J.

## REFERENCES


Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, 2020.

Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João Madeira Araújo, Oleksandr Vitvitskyi, Razvan Pascanu, and Petar Veličković. Transformers need glasses! information over-squashing in language tasks. In *NeurIPS*, 2024.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

Borun D. Chowdhury. Lost in the middle at birth: An exact theory of transformer position bias. *arXiv preprint arXiv:2603.10123*, 2026.

Kwesi Adu Cobbina and Tianyi Zhou. Where to show demos in your prompt: A positional bias of in-context learning. In *EMNLP*, 2025.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In *EMNLP*, 2022.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs. *arXiv preprint arXiv:2410.13835*, 2024.

Xiaobo Guo and Soroush Vosoughi. Serial position effects of large language models. In *ACL*, 2025.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34 (1):014004, 2017.

Ernst Hairer, Syvert Paul Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, 2 edition, 1993.

Bobby He, James Martens, Guodong Zhang, Aleksandar Botev, Andrew Brock, Samuel L. Smith, and Yee Whye Teh. Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation. In *ICLR*, 2023.

Cheng Huan and Hongwei Yuan. An adjoint-sensitivity framework for lost-in-the-middle phenomena in causal residual transformers. *arXiv preprint arXiv:2607.17696*, 2026.

Hugging Face Smol Models Team. Smollm3: Smol, multilingual, long-context reasoning llm, 2025. URL `https://huggingface.co/blog/smollm3`.

Prannay Kaul, Chengcheng Ma, Ismail Elezi, and Jiankang Deng. From attention to activation: Unravelling the enigmas of large language models. In *ICLR*, 2025.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *NeurIPS*, 2023.

Junu Kim, Xiao Liu, Zhenghao Lin, Lei Ji, Yeyun Gong, and Edward Choi. Layernorm induces recency bias in transformer decoders. In *Findings of ACL*, pp. 28638–28652. ACL, 2026.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models. In *NeurIPS*, 2024.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *TACL*, 12:157–173, 2024.

Toni J. B. Liu, Baran Zadeoğlu, Nicolas Boullé, Raphaël Sarfati, and Christopher J. Earls. Jacobian scopes: token-level causal attributions in LLMs. *arXiv preprint arXiv:2601.16407*, 2026.

Mikhail Menschikov, Alexander Kharitonov, Maiia Kotyga, Vadim Porvatov, Anna Zhukovskaya, David Kagramanyan, Egor Shvetsov, and Evgeny Burnaev. Beyond early-token bias: Model-specific and language-specific position effects in multilingual LLMs. *arXiv preprint arXiv:2505.16134*, 2025.

Mosaic. Introducing MPT-7b: A new standard for open-source, commercially usable LLMs, May 2023a. URL `https://www.databricks.com/blog/mpt-7b`. Databricks Blog.

Mosaic. MPT-30b: Raising the bar for open-source foundation models, June 2023b. URL `https://www.databricks.com/blog/mpt-30b`. Databricks Blog.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. In *NeurIPS*, 2024.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Maryam Rahimi, Mahdi Nouri, and Yadollah Yaghoobzadeh. Layer-wise positional bias in short-context language modeling. *arXiv preprint arXiv:2601.04098*, 2026.

Nikolaus Salvatore, Hao Wang, and Qiong Zhang. Lost in the middle: An emergent property from information retrieval demands in LLMs. *arXiv preprint arXiv:2510.10276*, 2025.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Lloyd S. Shapley. A value for $n$-person games. In *Contributions to the Theory of Games, Volume II*, pp. 307–318. Princeton University Press, 1953.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2014.

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

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. Do long-range language models actually use long-range context? In *EMNLP*, 2021.

Blerta Veseli, Julian Chibane, Mariya Toneva, and Alexander Koller. Positional biases shift as inputs approach context window limits. In *COLM*, 2025.

Wikimedia Foundation. Wikimedia downloads, January 2026. URL `https://dumps.wikimedia.org`.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In *ICML*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *ICLR*, 2024.

Bowen Yang, Bharat Venkitesh, Dwarak Talupuru, Hangyu Lin, David Cairuz, Phil Blunsom, and Acyr Locatelli. Rope to nope and back again: A new hybrid attention strategy. In *NeurIPS*, 2025.

Qingyu Yin, Xuzheng He, Xiang Zhuang, Yu Zhao, Jianhua Yao, Xiaoyu Shen, and Qiang Zhang. Stablemask: Refining causal masking in decoder-only transformer. In *ICML*, 2024.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, Zhangyang Wang, et al. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. In *NeurIPS*, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *ICML*, 2021.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS*, 2023.

## APPENDIX OVERVIEW

The appendix is organized as follows.
Appendix A contains the proofs of the finite- and infinite-depth results.
Appendix B gives finite-depth rollout representations and simplifications.
Appendix C gives conditional finite-mixing U-shape results.
Appendix D gives further shape and approximation results.
Appendix E gives the fixed-kernel exponential sanity check.
Appendix F specifies the ALiBi and RoPE kernel constructions.
Appendix G provides additional qualitative and quantitative rollout results.
Appendix H reports the measured residual-mixing schedules.
Appendix I reports the corpus-level approximation, context-length, influence-measurement, intervention, and mixing-coefficient studies.
Appendix J provides protocol, uncertainty, and compute details.
Appendix K summarizes license names, links, and citations for the models and datasets used in this paper.

## A PROOFS OF THEORETICAL RESULTS

This appendix contains the proofs of the finite-depth propositions from Section 4 and the infinite-depth theorem stated in Section 5.

### A.1 PROOFS OF SECTION 4 (FINITE-DEPTH CASE)

First, we need to state a few definitions and lemmas that we will use in the proofs.

**Stochastic order on positions.** Let $\mu, \nu$ be probability distributions on $\{1, \dots, n\}$. We write

$$\mu \succeq_{\mathrm{FOSD}} \nu \qquad \text{(first-order stochastic dominance)} \quad \Longleftrightarrow \quad \sum_{j=1}^{m} \mu(j) \;\le\; \sum_{j=1}^{m} \nu(j) \qquad \forall m \in \{1, \dots, n\}.$$

Thus $\mu \succeq_{\mathrm{FOSD}} \nu$ means that $\mu$ is shifted toward *larger indices* (more recent tokens). Equivalently, for every monotonically increasing test function $\varphi : \{1, \dots, n\} \to \mathbb{R}$,

$$\mu \succeq_{\mathrm{FOSD}} \nu \quad \Longleftrightarrow \quad \mathbb{E}_{X\sim\mu}[\varphi(X)] \;\ge\; \mathbb{E}_{X\sim\nu}[\varphi(X)]. \tag{7}$$

(Here $\mathbb{E}_{X\sim\mu}[\varphi(X)] = \sum_{j=1}^{n} \mu(j)\varphi(j)$, and similarly for $\nu$.)

**Kernels and rollout.** A row-stochastic matrix $K \in \mathbb{R}^{n\times n}$ is viewed as a Markov kernel on $\{1, \dots, n\}$. For a distribution $\mu$ (row vector), $\mu K$ denotes the pushforward distribution. In our setting, the layer operator is

$$R^{(t)} \;=\; (1 - \lambda_t) I + \lambda_t A^{(t)},$$

where $A^{(t)}$ is row-stochastic (attention) and $\lambda_t \in [0, 1]$. The rollout distribution of the final token after $T$ layers is

$$\mu_T^\top \;=\; e_n^\top P^{(T)} \;=\; e_n^\top R^{(T)} R^{(T-1)} \cdots R^{(1)}.$$

Throughout Appendix A.1, the kernels $A^{(t)}$ and mixing coefficients $\lambda_t$ are any specified deterministic quantities satisfying the assumptions of the corresponding result. Corpus-averaged kernels and mixing schedules are one empirical instantiation, introduced only in Section 6.1.

**Definition A.1** (Stochastically monotone (isotone) kernel)**.** A row-stochastic kernel $K$ is *stochastically monotone* if its row laws are monotone in FOSD:

$$i < i' \quad \Longrightarrow \quad K(i, \cdot) \;\preceq_{\mathrm{FOSD}}\; K(i', \cdot).$$

Equivalently, for every monotonically increasing $\varphi$, the function $g_\varphi(i) := \sum_{j=1}^{n} K(i, j)\varphi(j)$ is monotonically increasing in $i$.

**Lemma A.2** (Order preservation under a stochastically monotone kernel)**.** *Let $K$ be stochastically monotone. If $\mu \succeq_{\mathrm{FOSD}} \nu$, then*

$$\mu K \succeq_{\mathrm{FOSD}} \nu K.$$

*Proof.* Fix any monotonically increasing $\varphi$. Let $g(i) := \sum_j K(i,j)\varphi(j)$. By Definition A.1, $g$ is monotonically increasing. Using Equation (7),

$$\mathbb{E}_{X\sim\mu K}[\varphi(X)] = \sum_i \mu(i)g(i) \;\geq\; \sum_i \nu(i)g(i) = \mathbb{E}_{X\sim\nu K}[\varphi(X)].$$

Since this holds for all monotonically increasing $\varphi$, $\mu K \succeq_{\mathrm{FOSD}} \nu K$ by Equation (7). □

**Lemma A.3** (Residual kernels inherit stochastic monotonicity)**.** *If $A$ is stochastically monotone, then for any $\lambda \in [0,1]$ the residual kernel $R_\lambda = (1-\lambda)I + \lambda A$ is also stochastically monotone.*

*Proof.* Fix a monotonically increasing $\varphi$ and define $g_A(i) := \sum_j A(i,j)\varphi(j)$. Since $A$ is stochastically monotone, $g_A$ is monotonically increasing. For $R_\lambda$,

$$g_{R_\lambda}(i) := \sum_j R_\lambda(i,j)\varphi(j) = (1-\lambda)\varphi(i) + \lambda g_A(i),$$

which is monotonically increasing as a convex combination of monotonically increasing functions. Hence $R_\lambda$ is stochastically monotone. □

**Lemma A.4** (Rowwise dominance implies dominance after one step)**.** *Let $K, K'$ be row-stochastic kernels. Assume row-wise dominance, i.e., for all $i$,*

$$K(i,\cdot) \;\succeq_{\mathrm{FOSD}}\; K'(i,\cdot).$$

*Then for any distribution $\mu$,*

$$\mu K \succeq_{\mathrm{FOSD}} \mu K'.$$

*Proof.* Fix a monotonically increasing $\varphi$ and define $g(i) := \sum_j K(i,j)\varphi(j)$ and $g'(i) := \sum_j K'(i,j)\varphi(j)$. By the row-wise dominance assumption and Equation (7), $g(i) \geq g'(i)$ for all $i$. Therefore,

$$\mathbb{E}_{X\sim\mu K}[\varphi(X)] = \sum_i \mu(i)g(i) \;\geq\; \sum_i \mu(i)g'(i) = \mathbb{E}_{X\sim\mu K'}[\varphi(X)].$$

□

**Lemma A.5** (Monotone-vector propagation)**.** *If $K$ is stochastically monotone and $v \in \mathbb{R}^n$ is monotonically decreasing in its index, then $Kv$ is monotonically decreasing.*

*Proof.* For $i < i'$, $K(i,\cdot) \preceq_{\mathrm{FOSD}} K(i',\cdot)$. Taking $\varphi = -v$ (which is monotonically increasing since $v$ is monotonically decreasing) and using Equation (7) yields $\sum_j K(i,j)(-v_j) \leq \sum_j K(i',j)(-v_j)$, i.e. $(Kv)_i \geq (Kv)_{i'}$. □

### A.1.1 Primacy drift under causal masking

*Proof of Proposition 4.2.* Fix $k < n$ and define $\mathbf{1}_k \in \mathbb{R}^n$ by $(\mathbf{1}_k)_j = \mathbb{1}_{\{j\leq k\}}$. Let $m^{(t)} := P^{(t)}\mathbf{1}_k$, so that the final-token prefix mass equals

$$\sum_{j\leq k} \mu_t(j) \;=\; e_n^\top P^{(t)}\mathbf{1}_k \;=\; m_n^{(t)}.$$

Since each $A^{(t)}$ is stochastically monotone by Assumption 4.1, Lemma A.3 implies each $R^{(t)}$ is also stochastically monotone.

Since $\mathbf{1}_k$ is monotonically decreasing and $m^{(t+1)} = R^{(t+1)}m^{(t)}$, Lemma A.5 implies $m^{(t)}$ is monotonically decreasing for every $t$. Hence, $m_i^{(t)} \geq m_n^{(t)}$ for all $i$. Using row-stochasticity,

$$m_n^{(t+1)} = \sum_{i=1}^n R_{ni}^{(t+1)}\, m_i^{(t)} \;\geq\; \sum_{i=1}^n R_{ni}^{(t+1)}\, m_n^{(t)} = m_n^{(t)},$$

which proves that the final-token prefix mass $\sum_{j\leq k}\mu_t(j)$ is monotonically increasing in depth. □

### A.1.2 Stronger residual (smaller $\lambda$) induces recency drift

*Proof of Proposition 4.3.* Fix a layer and write $R_\lambda = (1-\lambda)I + \lambda A$, where $A$ is row-stochastic and causal, so that $\mathrm{supp}(A(i,\cdot)) \subseteq \{1,\dots,i\}$ for each $i$. Under our convention that larger indices are more recent, for any distribution $\nu$ supported on $\{1,\dots,i\}$ we have $\delta_i \succeq_{\mathrm{FOSD}} \nu$, where $\delta_i$ is the Kronecker delta. Therefore, for any $0 \le \lambda \le \lambda' \le 1$ and each $i$,

$$R_\lambda(i,\cdot) = (1-\lambda)\delta_i + \lambda A(i,\cdot) \;\succeq_{\mathrm{FOSD}}\; (1-\lambda')\delta_i + \lambda' A(i,\cdot) = R_{\lambda'}(i,\cdot). \tag{8}$$

Now compare two schedules $\{\lambda_t\}$ and $\{\lambda'_t\}$ with $\lambda_t \le \lambda'_t$ for all $t$. To respect the product order $P^{(T)} = R^{(T)} \cdots R^{(1)}$, set $\mu^{(T+1)} = \mu'^{(T+1)} = e_n^\top$ and, for $t = T, T-1, \dots, 1$, define

$$\mu^{(t)} = \mu^{(t+1)} R^{(t)}_{\lambda_t}, \qquad \mu'^{(t)} = \mu'^{(t+1)} R^{(t)}_{\lambda'_t}.$$

Suppose $\mu^{(t+1)} \succeq_{\mathrm{FOSD}} \mu'^{(t+1)}$. Rowwise dominance from Equation (8), followed by order preservation under the stochastically monotone kernel $R^{(t)}_{\lambda'_t}$, gives

$$\mu^{(t+1)} R^{(t)}_{\lambda_t} \succeq_{\mathrm{FOSD}} \mu^{(t+1)} R^{(t)}_{\lambda'_t} \succeq_{\mathrm{FOSD}} \mu'^{(t+1)} R^{(t)}_{\lambda'_t}.$$

Backward induction therefore yields $e_n^\top P^{(T)} = \mu^{(1)} \succeq_{\mathrm{FOSD}} \mu'^{(1)} = e_n^\top P'^{(T)}$. □

### A.1.3 Recency-favoring positional encodings induce recency drift

*Proof of Proposition 4.4.* Fix a layer and row, and let

$$\pi_\theta(j) \propto \exp(s_{ij} + \theta b_{ij}), \qquad 0 \le \theta \le 1,$$

on the row's causal support. For a prefix indicator $h_m(j) = \mathbb{1}_{\{j \le m\}}$, the softmax derivative identity yields

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \mathbb{E}_{\pi_\theta}[h_m] = \mathrm{Cov}_{\pi_\theta}(h_m, b_{i\cdot}) \le 0.$$

The inequality holds because $h_m$ is nonincreasing in $j$, whereas $b_{ij}$ is nondecreasing. Hence every prefix mass decreases with $\theta$, which proves $A^{(t)}_{\mathrm{PE}}(i,\cdot) \succeq_{\mathrm{FOSD}} A^{(t)}_0(i,\cdot)$. Mixing both kernels with the same identity row gives $R^{(t)}_{\mathrm{PE}}(i,\cdot) \succeq_{\mathrm{FOSD}} R^{(t)}_0(i,\cdot)$. By Assumption 4.1 and Lemma A.3, every $R^{(t)}_0$ is stochastically monotone. Set $\mu^{(T+1)}_{\mathrm{PE}} = \mu^{(T+1)}_0 = e_n^\top$ and define $\mu^{(t)}_a = \mu^{(t+1)}_a R^{(t)}_a$ for $t = T, T-1, \dots, 1$. Rowwise dominance and order preservation then give inductively

$$\mu^{(t+1)}_{\mathrm{PE}} R^{(t)}_{\mathrm{PE}} \succeq_{\mathrm{FOSD}} \mu^{(t+1)}_{\mathrm{PE}} R^{(t)}_0 \succeq_{\mathrm{FOSD}} \mu^{(t+1)}_0 R^{(t)}_0.$$

Hence $p^{(T)}_{\mathrm{PE}} \succeq_{\mathrm{FOSD}} p^{(T)}_0$. For ALiBi, $b_{ij} = -m(i-j)$ with $m \ge 0$, which is nondecreasing in $j$. □

## A.2 Proof of Theorem 5.5 (infinite-depth case)

We recall the residual-aware transition $R^{(t)} = (1-\lambda_t)I + \lambda_t A^{(t)}$ and the rollout $P^{(T)} = R^{(T)} R^{(T-1)} \cdots R^{(1)}$.

*Proof of Lemma 5.4.* To prove (i), let $X^{(t)}$ denote the states of the normalized analytical dynamics corresponding to the transition $R^{(t)}$, $X^{(t)} = (1-\lambda_t)X^{(t-1)} + \lambda_t A^{(t)} X^{(t-1)} W_V^{(t)}$, which tracks the relative identity and attention contributions. Using the triangle inequality and submultiplicativity of the operator norm, and noting that the row-stochastic attention matrix $A^{(t)}$ does not increase the maximum row norm (i.e., its induced $\infty$-norm on the rows is 1):

$$\max_i \|X^{(t)}_{i,:}\|_2 \le (1-\lambda_t) \max_i \|X^{(t-1)}_{i,:}\|_2 + \lambda_t \max_i \|X^{(t-1)}_{i,:}\|_2 \|W_V^{(t)}\|_2.$$

Letting $M^{(t)} = \max_i \|X^{(t)}_{i,:}\|_2$ and $C_t = \|W_V^{(t)}\|_2$, we have:

$$M^{(t)} \le M^{(t-1)}\big((1-\lambda_t) + \lambda_t C_t\big) = M^{(t-1)}\big(1 + \lambda_t(C_t - 1)\big) \le M^{(t-1)}\big(1 + \lambda_t \max(0, C_t - 1)\big).$$

Unrolling this recursion through $T$ gives:

$$M^{(T)} \le M^{(0)} \prod_{s=1}^{T} \big(1 + \lambda_s \max(0, C_s - 1)\big) \le M^{(0)} \exp\left( \sum_{s=1}^{T} \lambda_s \max(0, C_s - 1) \right).$$

By Assumption 5.2, the sum in the exponent converges to some finite value. Let $K = \exp\big(\sum_{s=1}^{\infty} \lambda_s \max(0, C_s - 1)\big) < \infty$. Since the initial input is bounded by $C$, we have $M^{(T)} \le C \cdot K$. Setting $C_X = C \cdot K$ establishes the uniform bound on the trajectories.

For (ii), we utilize the result of part (i). The attention logits are $\ell_{ij}^{(t)} = s_{ij}^{(t)} + b_{ij}^{(t)}$, where $s_{ij}^{(t)} = (X_{i,:}^{(t-1)} W_Q^{(t)})(X_{j,:}^{(t-1)} W_K^{(t)})^\top$. By Cauchy-Schwarz and Assumption 5.1, $|s_{ij}^{(t)}| \le (C_X C)^2$. Thus, for any admissible query $i$ and keys $j, k$, the content logit range is uniformly bounded: $|s_{ij}^{(t)} - s_{ik}^{(t)}| \le 2(C_X C)^2$. Combined with Assumption 5.3, the full admissible logit range is bounded:

$$|\ell_{ij}^{(t)} - \ell_{ik}^{(t)}| \le |s_{ij}^{(t)} - s_{ik}^{(t)}| + |b_{ij}^{(t)} - b_{ik}^{(t)}| \le 2(C_X C)^2 + B =: M.$$

Since softmax is invariant to row-wise additive constants, we can lower bound any admissible entry $A_{ij}^{(t)}$ by writing it as $(\sum_k \exp(\ell_{ik}^{(t)} - \ell_{ij}^{(t)}))^{-1}$. Because $\ell_{ik}^{(t)} - \ell_{ij}^{(t)} \le M$, we obtain $A_{ij}^{(t)} \ge \frac{1}{n e^M} =: \varepsilon > 0$, which is strictly positive for any fixed sequence length $n$. □

*Proof of Theorem 5.5(i).* Assume $\sum_{t=1}^{\infty} \lambda_t < \infty$. We first lower bound the diagonal entries of $R^{(t)}$ using Lemma 5.4:

$$R_{ii}^{(t)} = (1 - \lambda_t) + \lambda_t A_{ii}^{(t)} \ge 1 - \lambda_t + \lambda_t \varepsilon = 1 - (1 - \varepsilon)\lambda_t.$$

Since each $A^{(t)}$ (and thus $R^{(t)}$) is lower-triangular due to causal masking, the diagonal entries of their product are the products of their diagonal entries:

$$P_{ii}^{(T)} = (R^{(T)} \cdots R^{(1)})_{ii} = \prod_{t=1}^{T} R_{ii}^{(t)} \ge \prod_{t=1}^{T} \big(1 - (1 - \varepsilon)\lambda_t\big).$$

Since $\sum \lambda_t < \infty$, the infinite product $\prod_{t=1}^{\infty}(1 - (1 - \varepsilon)\lambda_t)$ converges to a strictly positive constant. Thus, $\liminf_{T \to \infty} P_{ii}^{(T)} > 0$, and the attention mass cannot collapse entirely to the first token. □

*Proof of Theorem 5.5(ii).* Assume $\sum_{t=1}^{\infty} \lambda_t = \infty$. For any token $i$, we track the total weight $W_{i,>1}^{(T)}$ assigned to all positions excluding the first token. Since each attention matrix $A^{(t)}$ must assign at least $\varepsilon$ mass to the first token, it reduces the remaining weight by a factor of at least $(1 - \varepsilon)$.

The residual $R^{(t)} = (1 - \lambda_t)I + \lambda_t A^{(t)}$ acts as a weighted average between maintaining the current weight distribution and applying this reduction. Consequently, the weight assigned to positions $j > 1$ shrinks at each layer by a factor of $(1 - \varepsilon\lambda_t)$. Concretely, using $A_{jj}^{(t)} \le 1 - (j-1)\varepsilon$ from Lemma 5.4, the diagonal entries of the single-layer transition satisfy:

$$R_{jj}^{(t)} = (1 - \lambda_t) + \lambda_t A_{jj}^{(t)} \le 1 - \lambda_t + \lambda_t(1 - (j-1)\varepsilon) = 1 - (j-1)\varepsilon\lambda_t.$$

Compounding these contractions through the rollout $P^{(T)} = R^{(T)} \cdots R^{(1)}$, and noting that these are lower-triangular matrices so the diagonal entries of the product are the products of the diagonal entries, we obtain:

$$P_{jj}^{(T)} = \prod_{s=1}^{T} R_{jj}^{(s)} \le \prod_{s=1}^{T} \big(1 - (j-1)\varepsilon\lambda_s\big) \le \exp\left( -(j-1)\varepsilon \sum_{s=1}^{T} \lambda_s \right).$$

For $j > 1$, the divergence of $\sum \lambda_s$ implies $P_{jj}^{(T)} \to 0$ as $T \to \infty$.

To establish the bound for all $j \le i \le n$, we observe that for any $j \in \{2, \ldots, n\}$, the lower-right submatrix spanning indices $\{j, \ldots, n\}$ is closed under multiplication. Let $R_{\ge j}^{(s)}$ denote this submatrix. Since $R^{(s)}$ is row-stochastic, for any row $i \ge j$, the sum of its entries in the submatrix is

$$\sum_{k=j}^{i} R_{ik}^{(s)} = 1 - \sum_{k=1}^{j-1} R_{ik}^{(s)} = 1 - \lambda_s \sum_{k=1}^{j-1} A_{ik}^{(s)}.$$

Using Lemma 5.4(ii), we have $A_{ik}^{(s)} \geq \varepsilon$, so $\sum_{k=1}^{j-1} A_{ik}^{(s)} \geq (j-1)\varepsilon$. Thus,

$$\|R_{\geq j}^{(s)}\|_\infty = \max_{i \geq j} \sum_{k=j}^{i} R_{ik}^{(s)} \leq 1 - (j-1)\varepsilon\lambda_s.$$

By submultiplicativity of the induced $\infty$-norm, the rollout submatrix $P_{\geq j}^{(T)} = R_{\geq j}^{(T)} \cdots R_{\geq j}^{(1)}$ satisfies

$$\|P_{\geq j}^{(T)}\|_\infty \leq \prod_{s=1}^{T} \|R_{\geq j}^{(s)}\|_\infty \leq \prod_{s=1}^{T} \big(1 - (j-1)\varepsilon\lambda_s\big) \leq \exp\left(-(j-1)\varepsilon \sum_{s=1}^{T} \lambda_s\right).$$

Since $P_{ij}^{(T)}$ for $i \geq j$ is an entry in this submatrix, it is bounded by the same value, which proves the exponential decay for all $j > 1$.

As $T \to \infty$, the divergence of $\sum \lambda_t$ forces this upper bound to zero. This implies that the attention mass for all tokens $j > 1$ vanishes, leaving the first token to absorb the entire unit of row mass: $\lim_{T\to\infty} P_{i1}^{(T)} = 1$. □

## B Finite-Depth Representations and Simplifications

We now make input dependence explicit: $A^{(t)}(x)$ and $\lambda_t(x)$ are the kernel and mixing coefficient for input $x$, and $R^{(t)}(x) = (1 - \lambda_t(x))I + \lambda_t(x)A^{(t)}(x)$. No particular calibration formula is needed.

**Corpus averages.** Write $\bar{\lambda}_t = \mathbb{E}_{x\sim\mathcal{D}}\lambda_t(x)$ and $\bar{A}^{(t)} = \mathbb{E}_{x\sim\mathcal{D}}A^{(t)}(x)$. Separate averaging defines $R_{\mathrm{mf}}^{(t)} = (1 - \bar{\lambda}_t)I + \bar{\lambda}_t\bar{A}^{(t)}$. The joint mean instead satisfies

$$\mathbb{E}_{x\sim\mathcal{D}}R^{(t)}(x) = (1 - \bar{\lambda}_t)I + \bar{\lambda}_t A_{\mathrm{eff}}^{(t)}, \qquad A_{\mathrm{eff}}^{(t)} = \frac{\mathbb{E}_{x\sim\mathcal{D}}[\lambda_t(x)A^{(t)}(x)]}{\bar{\lambda}_t}, \tag{9}$$

when $\bar{\lambda}_t > 0$. If $\bar{\lambda}_t = 0$, the mean operator is $I$ and $A_{\mathrm{eff}}^{(t)}$ may be chosen arbitrarily. Both constructions are stochastic kernels. They coincide only under an appropriate factorization of the within-layer expectation; neither automatically commutes with expectation over the full product.

**Two mean-field steps.** For input-dependent operators, the approximations

$$\mathbb{E}_{x\sim\mathcal{D}}[R^{(T)}(x)\cdots R^{(1)}(x)] \approx (\mathbb{E}_{x\sim\mathcal{D}}R^{(T)}(x))\cdots(\mathbb{E}_{x\sim\mathcal{D}}R^{(1)}(x)), \tag{10}$$

$$\mathbb{E}_{x\sim\mathcal{D}}[\lambda_t(x)A^{(t)}(x)] \approx (\mathbb{E}_{x\sim\mathcal{D}}\lambda_t(x))(\mathbb{E}_{x\sim\mathcal{D}}A^{(t)}(x)) \tag{11}$$

discard different dependencies. Equation (10) concerns cross-layer composition; Equation (11) concerns the two factors within one layer. Neither is an identity in general. In the rest of this section, $A^{(t)}, \lambda_t$ denote a specified deterministic construction, using either separate means or the exact joint-mean kernel in Equation (9).

**Proposition B.1** (Ordered attention-step representation)**.** *Let $Z_t$ be independent Bernoulli variables with parameters $\lambda_t$. With $\widehat{R}^{(t)} = I$ when $Z_t = 0$ and $\widehat{R}^{(t)} = A^{(t)}$ when $Z_t = 1$,*

$$P^{(T)} = \mathbb{E}_Z[\widehat{R}^{(T)}\cdots\widehat{R}^{(1)}].$$

*If $A^{(t)} = A$ for all layers, this becomes*

$$P^{(T)} = \mathbb{E}[A^S] = \sum_{k=0}^{T} w_k A^k, \quad S = \sum_t Z_t, \quad \sum_{k=0}^{T} w_k z^k = \prod_{t=1}^{T}(1 - \lambda_t + \lambda_t z). \tag{12}$$

*Proof.* Independence permits entrywise expectation of the ordered product to factor as $(\mathbb{E}\widehat{R}^{(T)})\cdots(\mathbb{E}\widehat{R}^{(1)}) = R^{(T)}\cdots R^{(1)}$. When every attention kernel equals $A$, each realization reduces to $A^S$. The generating function of the independent Bernoulli sum gives the weights in Equation (12). □

Identity steps disappear from each sampled path. For different kernels, the number of attention steps alone does not determine their ordered product.

**A common kernel.** For $\tau = \sum_t \lambda_t > 0$, define

$$A_* = \frac{\sum_t \lambda_t A^{(t)}}{\tau}, \qquad P_{\text{common}} = [(1-\lambda_T)I + \lambda_T A_*] \cdots [(1-\lambda_1)I + \lambda_1 A_*]. \tag{13}$$

Then $A_*$ is stochastic and $\sum_t \lambda_t (A^{(t)} - I) = \tau(A_* - I)$. Thus the integrated generator is preserved, but generally $P_{\text{common}} \neq P^{(T)}$. For $\tau = 0$ both rollouts equal $I$ and the common kernel is immaterial. Concurrently, Chowdhury (2026) derive closed forms for the uniform causal (Cesàro) kernel $M_{ij} = i^{-1}\mathbf{1}\{j \le i\}$ with constant residual mixing $\alpha$. This kernel assigns equal logits to all admissible positions and therefore contains neither content dependence nor an ALiBi or RoPE contribution. Their rollout $[(1-\alpha)I + \alpha M]^T$ is exactly the constant-kernel, constant-schedule special case of Equation (13); at $\alpha = 1/2$, it is also the Abnar–Zuidema fixed-residual rule specialized to the Cesàro kernel. Their sequence-length continuum yields a decreasing primacy density and an atom at the final position, rather than the broad finite U-shaped profiles studied here. It is also distinct from the exponential construction below, which recovers this rollout only under $\alpha = \tau/T$ as $T \to \infty$. At fixed sequence length and any constant $\alpha > 0$, the residual atom has mass $(1-\alpha)^T \to 0$ and the infinite-depth rollout collapses to the first token. Related continuous-depth work by Huan & Yuan (2026) studies positional adjoint sensitivity in a causal residual flow. Its controlled simulations use a fixed Cesàro kernel and the propagator $[I + (T\alpha/M)(C \otimes S)]^M$. In the scalar-feature specialization, normalization removes a common growth factor and the positional transport converges to $\exp[T\alpha(C - I)]$, the fixed-kernel exponential model in Equation (14). Chowdhury expresses this through a discrete fixed-residual rollout. Huan and Yuan express it through a finite-time continuous-depth flow and a right-biased terminal adjoint. After scalarization and normalization, both belong to our common-kernel exponential picture. Increasing $M$ refines a finite-total-mixing flow rather than introducing a measured or adaptive layerwise mixing schedule. Their experiments are small controlled simulations with 48 positions and four-dimensional features, and do not compare predicted profiles with measured influence in pretrained Transformers.

**Proposition B.2** (Exponential simplification)**.** *For a stochastic kernel $A$ and $\tau \ge 0$,*

$$P_{\exp}(\tau) = \exp[\tau(A - I)] = \mathbb{E}[A^{N_\tau}], \qquad N_\tau \sim \text{Poisson}(\tau). \tag{14}$$

*Moreover, $[I + \tau(A - I)/M]^M \to P_{\exp}(\tau)$ as $M \to \infty$. For integers $M \ge \tau$ each factor is stochastic.*

*Proof.* Since $A$ commutes with $I$,

$$e^{\tau(A-I)} = e^{-\tau} \sum_{k=0}^{\infty} \frac{\tau^k}{k!} A^k,$$

which is the stated Poisson mixture. The Euler limit is the standard matrix exponential limit. For $M \ge \tau$, each factor can be written as $(1 - \tau/M)I + (\tau/M)A$ and is therefore stochastic. □

Applied to $A_*$, this replaces the finite Bernoulli count by a Poisson count at the same total mixing. It is a finite-$\tau$ approximation, not the infinite-total-mixing limit. The associated row-vector ODE is

$$\frac{d}{d\tau} p(\tau) = p(\tau)(A - I), \qquad p(0) = e_n^\top.$$

The exponential form solves this ODE exactly. It does not assert that the original layer-varying rollout is an exact solution of a constant-generator ODE.

**Computing only the final profile.** Set $v^{(0)} = e_n^\top$ and $v^{(k+1)} = v^{(k)} A$. Then Equation (12) needs only weighted row vectors, not full matrix powers. For Equation (14), truncation after $K$ powers leaves a nonnegative remainder of total mass $\mathbb{P}(N_\tau > K)$, giving an explicit $\ell_1$ truncation error. This requires $K$ matrix-vector products and storage of the kernel and a constant number of row vectors.

## C Conditional Finite-Mixing U-Shapes

Let $n \ge 3$, let $A$ be any causal stochastic kernel, and put $p(\tau) = e_n^\top \exp[\tau(A - I)]$. Two notions of shape will be used. *Endpoint elevation* means that $p_1$ and $p_n$ both exceed the mean over positions

$2, \dots, n-1$. For a partition into a left prefix $L$, a nonempty middle interval $M$, and a right suffix $D$, *block U-shape* means

$$\min\{p(L)/|L|, p(D)/|D|\} > p(M)/|M|.$$

Neither definition asserts a unique minimum or excludes local reversals.

**Theorem C.1** (Endpoint elevation from direct primacy and survival)**.** *Let $b = 1 - A_{nn} > 0$ and $\eta = A_{n1}/b$. If*

$$\frac{1}{n-1} < \eta < 1, \qquad 0 < \tau < \frac{1}{b} \log \frac{n-1-\eta}{1-\eta},$$

*then $p(\tau)$ has endpoint elevation. If $\eta = 1$, the conclusion holds for every finite $\tau > 0$.*

*Proof.* The generator $A - I$ defines a continuous-time Markov chain. Its exit rate from state $n$ is $b$, and causality prevents return after exit, so $p_n(\tau) = q := e^{-b\tau}$. Conditional on the first exit, the chain moves directly to state 1 with probability $\eta$. Since state 1 is absorbing, $p_1(\tau) \geq \eta(1-q)$, while the total mass on positions $2, \dots, n-1$ is at most $(1-\eta)(1-q)$. The left endpoint exceeds the interior average when $\eta > 1/(n-1)$. The right endpoint exceeds it when $q > (1-\eta)/(n-1-\eta)$, which is equivalent to the stated upper bound on $\tau$. If $\eta = 1$, the interior mass is zero and both endpoints have positive mass for every finite $\tau > 0$. □

Only two entries of $A$ enter this sufficient condition. It does not require stochastic monotonicity. It can fail for recency-weighted rows even when indirect paths produce primacy, which motivates the next certificate.

**Theorem C.2** (Finite path certificate for block U-shape)**.** *Let $L = \{1, \dots, \ell\}$, $D = \{n-r+1, \dots, n\}$, and $m = n - \ell - r > 0$, with $\ell, r \geq 1$. For $k \geq 0$, set*

$$v^{(k)} = e_n^\top A^k, \quad u_k = v^{(k)}(L), \quad s_k = v^{(k)}(D), \quad w_k = e^{-\tau} \frac{\tau^k}{k!}.$$

*For any integer $K \geq 0$, define $\epsilon_K = \mathbb{P}(N_\tau > K)$ and*

$$L_K = \sum_{k=0}^{K} w_k u_k + \epsilon_K u_K,$$

$$D_K = \max\left\{ e^{-(1-A_{nn})\tau}, \sum_{k=0}^{K} w_k s_k \right\}.$$

*If*

$$\min\left\{ \frac{L_K}{\ell}, \frac{D_K}{r} \right\} > \frac{1 - L_K - D_K}{m}, \tag{15}$$

*then $p(\tau)$ has block U-shape for this partition. Conversely, whenever the profile has strict block U-shape for this fixed partition, some finite $K$ satisfies Equation* (15)*.*

*Proof.* The discrete chain with kernel $A$ cannot leave the causal prefix $L$, so $u_k$ is nondecreasing. In the Poisson mixture, the tail contribution to $L$ is at least $\epsilon_K u_K$, hence $p(L) \geq L_K$. Both the partial Poisson sum and the probability of never leaving state $n$ lower-bound $p(D)$, so $p(D) \geq D_K$. It follows that $p(M) \leq 1 - L_K - D_K$, and Equation (15) proves sufficiency. As $K \to \infty$, $L_K$ and $D_K$ converge to $p(L)$ and $p(D)$, respectively. Therefore any strict block U-shape for the fixed partition satisfies the certificate at some finite $K$. □

This is a computable certificate for the exponential model, not a claim that every causal kernel has a U-shaped rollout. It depends only on the kernel and total mixing, not an output-influence measurement. The same path calculation applies to the exact common-kernel finite product by using its finite weights $w_k$ from Equation (12); taking all $T$ powers gives its exact block masses.

**Limits of shape transfer.** For normalized distributions, define

$$W_1(p,q) = \frac{1}{n-1}\sum_{k=1}^{n-1} |\mathsf{F}_p(k) - \mathsf{F}_q(k)|.$$

Small $W_1$ alone does not preserve pointwise shape inequalities. Bounds on coordinate or block-mass errors, together with positive shape margins, are needed to transfer a certificate to the original finite rollout. An operator perturbation bound is given in Theorem D.2.

# D Additional Shape and Approximation Results

## D.1 A structural single-valley special case

**Proposition D.1** (Decreasing rows and a terminal recency spike)**.** *Suppose every row of the causal stochastic kernel $A$ is nonincreasing in key position on its support. Let $p(\tau) = e_n^\top e^{\tau(A-I)}$, $a = A_{n,n-1}$, $d = A_{n-1,n-1} - A_{nn}$, and*

$$\phi_d(\tau) = \begin{cases} (e^{d\tau}-1)/d, & d \neq 0, \\ \tau, & d = 0. \end{cases}$$

*Then $p_1 \geq \cdots \geq p_{n-1}$ and $p_{n-1}/p_n = a\phi_d(\tau)$. Hence $a\phi_d(\tau) < 1$ gives a decreasing prefix followed by a strict terminal rise. If all causal entries are positive and $\tau > 0$, the prefix inequalities are strict.*

*Proof.* Pad every row by zeros to length $n$. Each padded row is decreasing, and so is every probability-weighted mixture of rows. Thus $v^{(k)} = e_n^\top A^k$ is decreasing for $k \geq 1$. The $k = 0$ term in the Poisson mixture affects only the last coordinate. If all causal entries are positive, for $j < n$ the contribution of row $j$ to $v_j^{(2)} - v_{j+1}^{(2)}$ is $A_{nj}A_{jj} > 0$, and all other contributions are nonnegative. The positive Poisson weight for $k = 2$ makes the prefix inequalities strict. Finally the last two ODEs are

$$p_n' = -(1 - A_{nn})p_n, \quad p_{n-1}' = ap_n - (1 - A_{n-1,n-1})p_{n-1}.$$

With initial values 1 and 0, solving gives the stated ratio. □

For uniform causal attention, $A_{ij} = 1/i$ when $j \leq i$, this gives

$$0 < \tau < n(n-1)\log\frac{n}{n-1} \quad \Longrightarrow \quad p_1 > \cdots > p_{n-1} < p_n.$$

This is a special-case shape with a one-token right arm. The decreasing-row hypothesis differs from Theorem 4.1 and is not asserted for general ALiBi or RoPE kernels.

## D.2 Why drift assumptions alone do not imply U-shape

The causal, positive-on-support, stochastically monotone kernel

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0.01 & 0.99 & 0 \\ 0.01 & 0.98 & 0.01 \end{pmatrix}$$

gives at $\tau = 2$

$$p(2) = (1 - e^{-0.02},\ e^{-0.02} - e^{-1.98},\ e^{-1.98}).$$

The middle coordinate exceeds both endpoints. Thus stochastic monotonicity, causal support, and positivity alone do not imply a finite-mixing U-shape.

### D.3 Perturbation and shape margins

**Proposition D.2** (Common-kernel perturbation bound)**.** *For stochastic kernels $A, B$ and $\tau \geq 0$,*

$$\left\| e_n^\top e^{\tau(A-I)} - e_n^\top e^{\tau(B-I)} \right\|_1 \leq \tau \max_i \sum_j |A_{ij} - B_{ij}|.$$

*Proof.* Duhamel's identity expresses the exponential difference as

$$\int_0^\tau e^{s(A-I)}(A-B)e^{(\tau-s)(B-I)}\, ds.$$

The leftmost row distribution is stochastic. Its product with $A - B$ has $\ell_1$ norm at most the maximum absolute row sum of $A - B$. Right multiplication by a stochastic matrix contracts the $\ell_1$ norm of any signed row vector. Integration proves the bound. □

If $\|p - q\|_1 \leq \delta$, every adjacent difference changes by at most $\delta$. Thus a strict single-valley sign pattern with every gap larger than $\delta$ transfers. For the block partition in Theorem C.2, the left-minus-middle average changes by at most $\delta \max\{1/\ell, 1/m\}$, and the right-minus-middle average by at most $\delta \max\{1/r, 1/m\}$. Positive margins exceeding these bounds transfer block U-shape. These bounds are sufficient, not necessary.

### D.4 Necessity of effective-mixing and diagonal conditions

For $n = 2$, let $\lambda_t = 1$ and

$$A^{(t)} = \begin{pmatrix} 1 & 0 \\ 2^{-t-1} & 1 - 2^{-t-1} \end{pmatrix}, \qquad t \geq 1.$$

Every state can reach 1 at every layer, but $P_{22}^{(T)} = \prod_{t=1}^{T}(1 - 2^{-t-1})$ has a positive limit. Thus $\sum_t \lambda_t = \infty$ and layerwise reachability do not suffice for collapse. Conversely, one layer with $\lambda_1 = 1$ and $A_{22}^{(1)} = 0$, followed by $\lambda_t = 0$, removes the second diagonal path despite finite total mixing. This explains the nonzero-factor condition in Theorem 5.5(i).

## E Fixed-kernel exponential sanity check

**Corollary E.1** (Fixed-kernel exponential sanity check)**.** *For a fixed causal stochastic $A$, every finite $\tau \geq 0$ satisfies*

$$[e^{\tau(A-I)}]_{ii} = e^{-\tau(1-A_{ii})} > 0.$$

*If every state can reach 1 along positive entries of $A$, then $e_i^\top e^{\tau(A-I)} \to e_1^\top$ as $\tau \to \infty$.*

This recovers the finite-mixing retention and infinite-mixing collapse behavior of the simplified model. It does not identify the finite-$\tau$ exponential approximation with the infinite-depth limit of a trained network.

*Proof.* The diagonal of the exponential of a triangular matrix is the exponential of its diagonal. This gives the finite-$\tau$ equality. For the limit, reachability gives a finite path from every state to 1. Since $A$ is fixed and the state space finite, there are an integer $L \geq 1$ and $\rho > 0$ such that $(A^L)_{i1} \geq \rho$ for all $i > 1$; shorter paths can be padded at the absorbing state. The nonabsorbed mass after $k$ discrete steps is at most $(1 - \rho)^{\lfloor k/L \rfloor}$. In the Poisson mixture, $N_\tau \to \infty$ in probability. Averaging this bound, which is bounded by one and tends to zero with $k$, proves the limit. □

## F Kernel construction and aggregation

**Content-free ALiBi kernel.** For head slopes $m_h \geq 0$ and causal mask $\mathcal{M}$, setting content logits to zero gives

$$B_{ij}^h = \frac{\mathbf{1}_{\{(i,j)\in\mathcal{M}\}} \exp[-m_h(i-j)]}{\sum_{k:(i,k)\in\mathcal{M}} \exp[-m_h(i-k)]}, \qquad A_{ij} = \frac{1}{H}\sum_{h=1}^{H} B_{ij}^h. \tag{16}$$

When the slopes, head aggregation, and mask are the same at every layer, this kernel is layer-independent. Otherwise the layer index is retained.

**Forward contribution kernel for RoPE.** RoPE rotates content-dependent queries and keys. Zeroing those content vectors removes the rotation-dependent signal as well; a masked-uniform kernel remains, not an ALiBi-like positional prior. To retain forward contributions, let $B^h_{t,ij}(x)$ be raw attention for input $x$ and use column-vector feature notation to define

$$c_{t,ij}(x) = \sum_h B^h_{t,ij}(x) W^h_{O,t} V_{t,hj}(x), \qquad a_{t,i}(x) = \sum_j c_{t,ij}(x). \tag{17}$$

For bias-free output projections, these are the source vectors of the attention update. For rows with positive total score, we use an ALTI-style contribution score (Ferrando et al., 2022) to scalarize the attention branch:

$$S_{t,ij}(x) = \left[\|a_{t,i}(x)\|_1 - \|a_{t,i}(x) - c_{t,ij}(x)\|_1\right]_+, \qquad A^{(t)}_{ij}(x) = \frac{S_{t,ij}(x)}{\sum_k S_{t,ik}(x)}. \tag{18}$$

If $\sum_k S_{t,ik}(x) = 0$, we set $A^{(t)}_{i\cdot}(x) = e_i^\top$. Equivalently, the effective logits in the setup are $\ell^{(t)}_{ij}(x) = \log S_{t,ij}(x)$, with $\log 0 = -\infty$, for rows with positive total score. Their softmax gives exactly the normalized scores above, not the softmax of the scores themselves. Here $[z]_+ = \max\{z, 0\}$. Inadmissible entries are zero. Head vectors are summed before scalarization, so directions and cancellation are retained at this step. The residual is not included in $c_{t,ij}$; it enters only through the residual transition in Section 3. The resulting kernel is causal and stochastic, as required by the rollout theory.

This definition uses no gradient, output-attribution target, or input blank-out score. It remains dependent on the model, corpus, and chosen feature norm. No explicit MLP contribution is assigned to the transport kernel; position-wise processing can nevertheless affect subsequent attention and measured influence.

# G Additional rollout and influence analysis

This appendix provides additional qualitative and quantitative results for the controlled rollout experiments discussed in Section 6. We report final-token rollout distributions $p^{(T)}$ together with the corresponding gradient influence proxies $\hat{q}_{\mathrm{grad}}$ across ALiBi- and RoPE-based Transformer models.

## G.1 Quantitative agreement between theory and measurement

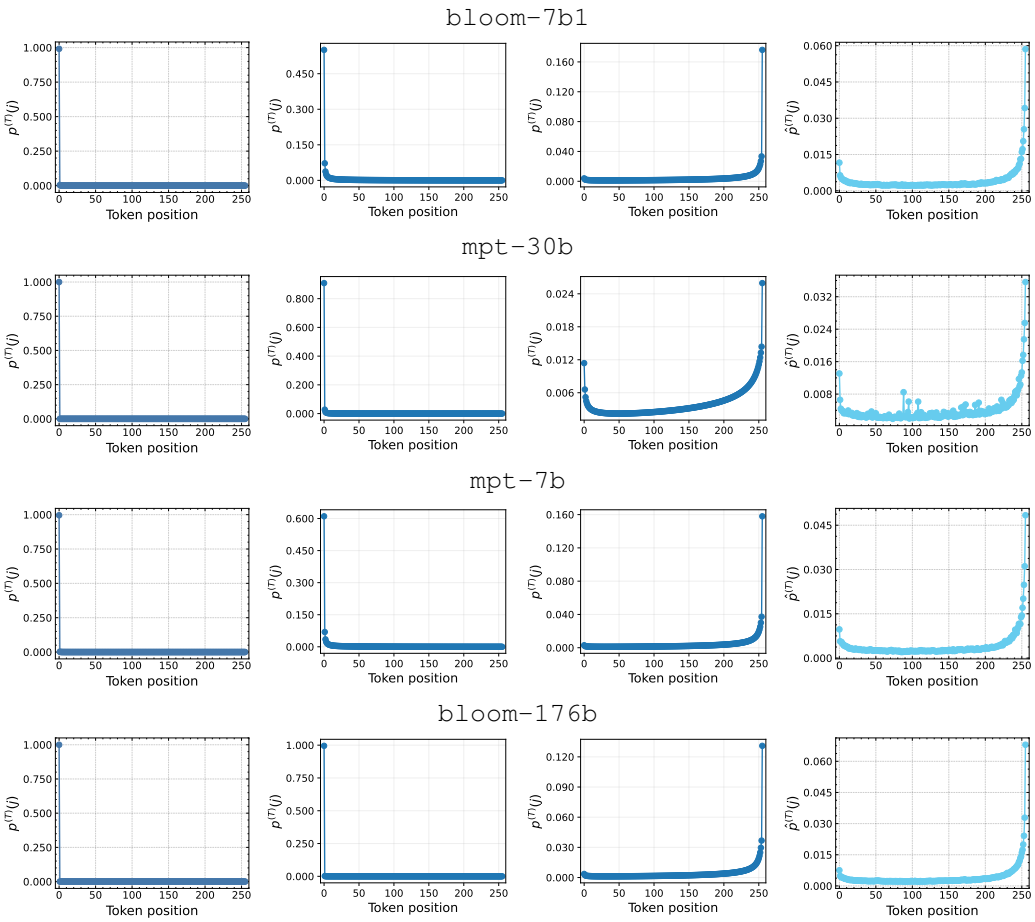
To quantify the agreement between the predicted influence distributions $p^{(T)}$ and the gradient influence proxy $\hat{q}_{\mathrm{grad}}$, we report normalized 1-Wasserstein distance on token positions, using the ground metric $d(i, j) = |i - j|/(n - 1)$ (Table 3). The $n = 256$ results use the norm-Shapley coefficient from the main experiment for both ALiBi and RoPE. The adaptive rollout is substantially closer to measured influence than the attention-only rollout in all 30 model-dataset settings. We ablate the mixing coefficient and context length separately below.

## G.2 Qualitative comparison of rollout and measured influence

Figures 2 and 3 extend the central qualitative comparison. Across the examined models, adaptive residual rollout recovers the broad, often U-shaped measured profiles, whereas attention-only rollout concentrates most mass near the first token.

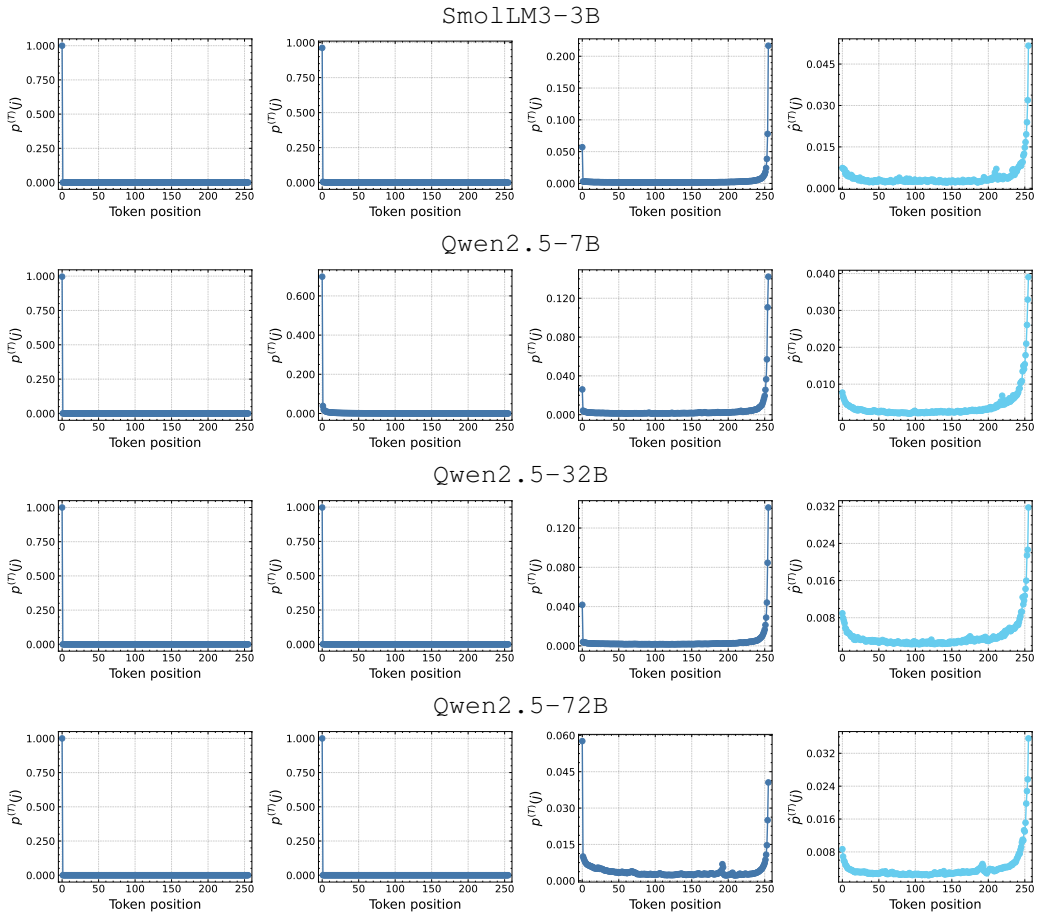
Table 3: Normalized 1-Wasserstein distances at $n = 256$ (↓ lower is better) between controlled rollout variants and gradient influence. Columns (a), (b), and (c) use $\lambda_t = 1$ following Wu et al. (2025), $\lambda_t = \frac{1}{2}$ following Abnar & Zuidema (2020), and our adaptive measured norm-Shapley coefficient $\lambda_t$, respectively.

| Model | FineWeb-Edu | | | DCLM-Baseline | | | Wikipedia | | |
|---|---|---|---|---|---|---|---|---|---|
| | (a) Wu $\lambda = 1$ | (b) Abnar $\lambda = \frac{1}{2}$ | (c) Ours adaptive | (a) Wu $\lambda = 1$ | (b) Abnar $\lambda = \frac{1}{2}$ | (c) Ours adaptive | (a) Wu $\lambda = 1$ | (b) Abnar $\lambda = \frac{1}{2}$ | (c) Ours adaptive |
| `falcon-rw-7b` | 0.66 | 0.63 | **0.03** | 0.66 | 0.64 | **0.02** | 0.68 | 0.65 | **0.05** |
| `mpt-7b` | 0.62 | 0.58 | **0.15** | 0.59 | 0.55 | **0.19** | 0.66 | 0.62 | **0.11** |
| `mpt-30b` | 0.61 | 0.61 | **0.02** | 0.58 | 0.58 | **0.04** | 0.65 | 0.65 | **0.02** |
| `bloom-7b1` | 0.63 | 0.58 | **0.13** | 0.61 | 0.56 | **0.15** | 0.67 | 0.61 | **0.10** |
| `bloom-176b` | 0.65 | 0.65 | **0.11** | 0.63 | 0.63 | **0.06** | 0.69 | 0.69 | **0.11** |
| `qwen2.5-7b` | 0.63 | 0.59 | **0.12** | 0.59 | 0.55 | **0.14** | 0.65 | 0.61 | **0.10** |
| `qwen2.5-32b` | 0.59 | 0.59 | **0.11** | 0.56 | 0.56 | **0.12** | 0.64 | 0.64 | **0.07** |
| `qwen2.5-72b` | 0.60 | 0.60 | **0.12** | 0.56 | 0.56 | **0.10** | 0.62 | 0.62 | **0.11** |
| `qwen3-8b-base` | 0.58 | 0.57 | **0.09** | 0.57 | 0.56 | **0.10** | 0.64 | 0.62 | **0.05** |
| `smollm3-3b-base` | 0.60 | 0.60 | **0.11** | 0.60 | 0.59 | **0.11** | 0.65 | 0.64 | **0.07** |




(a) Attention-only rollout ($\lambda_t = 1$) (b) Fixed residual rollout ($\lambda_t = \frac{1}{2}$) (c) Adaptive residual rollout (d) Measured input token influence

Figure 2: Final-token profiles at sequence length $n = 256$ on FineWeb-Edu for the four ALiBi models not shown in Figure 1. The columns show attention-only rollout, fixed residual rollout, adaptive residual rollout, and measured gradient influence.

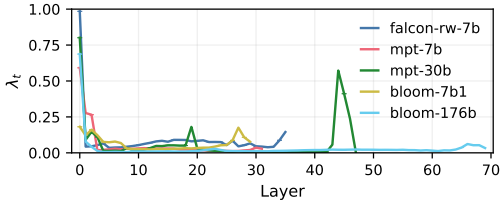



(a) Attention-only rollout ($\lambda_t = 1$) (b) Fixed residual rollout ($\lambda_t = \frac{1}{2}$) (c) Adaptive residual rollout (d) Measured input token influence

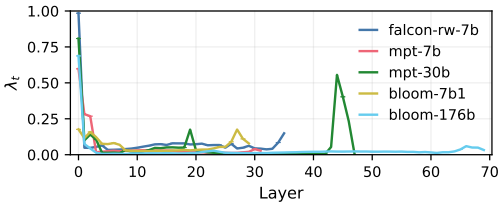
Figure 3: Final-token profiles at sequence length $n = 256$ on FineWeb-Edu for the four RoPE models not shown in Figure 1. The columns show attention-only rollout, fixed residual rollout, adaptive residual rollout, and measured gradient influence.

## H MEASURED RESIDUAL-MIXING SCHEDULES

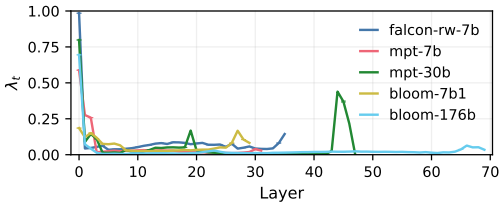
Figures 4 and 5 report the depth-wise norm-Shapley coefficient from Equation (5) for the ALiBi and RoPE models, respectively. These are the measured schedules used by our adaptive rollouts; neither gradients nor blank-out influence enter their construction.




(a) FineWeb-Edu

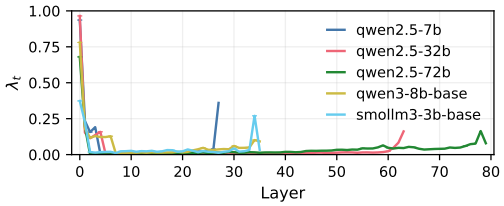



(b) DCLM-Baseline

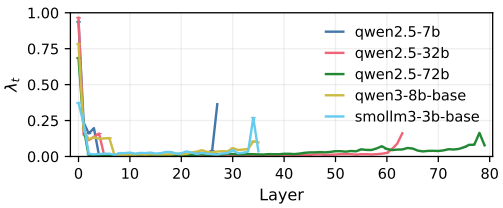



(c) Wikipedia

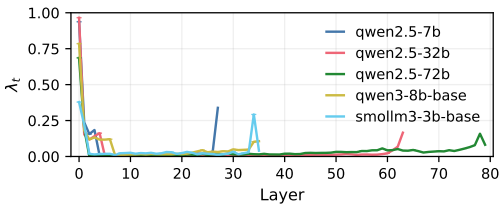
Figure 4: Depth-wise norm-Shapley mixing coefficients for the ALiBi models across three datasets. Mean $\pm$ 95% confidence interval at sequence length $n = 256$; the intervals are narrow relative to the plotted range.




(a) FineWeb-Edu




(b) DCLM-Baseline




(c) Wikipedia

Figure 5: Depth-wise norm-Shapley mixing coefficients for the RoPE models across three datasets. Mean $\pm$ 95% confidence interval at sequence length $n = 256$; the intervals are narrow relative to the plotted range.

# I Ablation studies

## I.1 Corpus-level rollout approximations

We evaluate the two approximations in Equations (3) and (4) separately under the same norm-Shapley protocol as the main results. Let $P_{\text{input}}$ denote the average of input-specific rollouts, $P_{\text{joint}}$ the product of the expected layer transitions, and $P_{\text{mf}}$ the rollout that additionally averages $\lambda_t$ and $A^{(t)}$ separately within each layer. Cross-layer compares $P_{\text{input}}$ with $P_{\text{joint}}$ and tests the product of expectations against the expectation of the product. Within-layer compares $P_{\text{joint}}$ with $P_{\text{mf}}$ and tests the mean-field factorization. Combined compares $P_{\text{input}}$ directly with $P_{\text{mf}}$. Table 4 reports normalized 1-Wasserstein distances between their final rows.

Table 4: Corpus-level rollout approximation ablation at $n = 256$ under the norm-Shapley protocol. Entries are normalized final-row 1-Wasserstein distances in **units of $10^{-3}$** (lower is better).

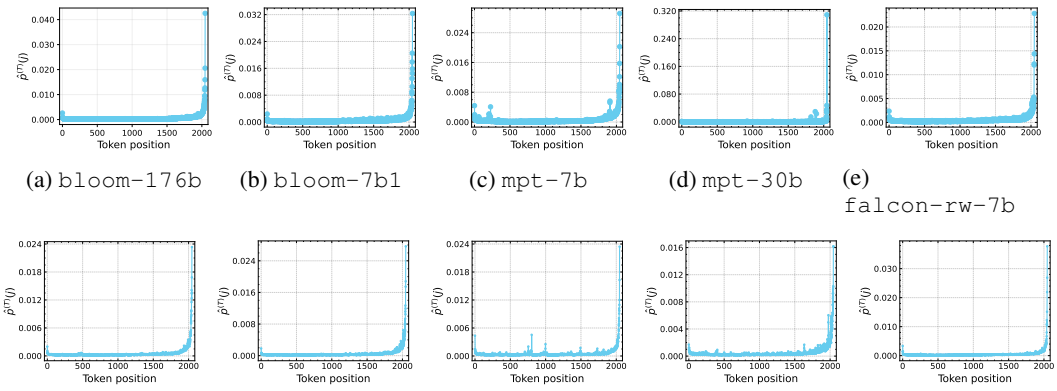
| Model | FineWeb-Edu | | | DCLM-Baseline | | | Wikipedia | | |
|---|---|---|---|---|---|---|---|---|---|
| | Cross-layer | Within-layer | Combined | Cross-layer | Within-layer | Combined | Cross-layer | Within-layer | Combined |
| `falcon-rw-7b` | 0.17 | 0.00 | 0.17 | 0.31 | 0.00 | 0.31 | 0.20 | 0.00 | 0.20 |
| `mpt-7b` | 0.01 | 0.00 | 0.01 | 0.02 | 0.00 | 0.02 | 0.01 | 0.00 | 0.01 |
| `mpt-30b` | 0.10 | 0.00 | 0.10 | 0.17 | 0.00 | 0.17 | 0.39 | 0.00 | 0.39 |
| `bloom-7b1` | 0.03 | 0.00 | 0.03 | 0.03 | 0.00 | 0.03 | 0.03 | 0.00 | 0.03 |
| `bloom-176b` | 0.02 | 0.00 | 0.02 | 0.02 | 0.00 | 0.02 | 0.02 | 0.00 | 0.02 |
| `qwen2.5-7b` | 0.72 | 0.60 | 0.13 | 3.58 | 0.24 | 3.71 | 3.27 | 0.23 | 3.10 |
| `qwen2.5-32b` | 0.07 | 0.04 | 0.04 | 5.27 | 0.41 | 5.48 | 4.36 | 0.35 | 4.27 |
| `qwen2.5-72b` | 0.01 | 0.01 | 0.01 | 5.25 | 1.05 | 6.13 | 5.65 | 1.03 | 6.42 |
| `qwen3-8b-base` | 0.52 | 0.47 | 0.08 | 5.42 | 0.12 | 5.50 | 5.62 | 0.37 | 5.25 |
| `smollm3-3b-base` | 0.14 | 0.09 | 0.04 | 4.95 | 0.25 | 4.76 | 4.01 | 0.50 | 3.62 |

Both approximation errors remain small across models, positional mechanisms, and corpora. The cross-layer error is at most $5.65 \times 10^{-3}$ and the within-layer mean-field error is at most $1.05 \times 10^{-3}$; their combined effect is at most $6.42 \times 10^{-3}$. These discrepancies are well below the distances between rollout and measurement in Table 3. For the content-free ALiBi construction, the within-layer error is zero because $A^{(t)}$ is input-independent, so $\mathbb{E}[\lambda_t A^{(t)}] = \mathbb{E}[\lambda_t]A^{(t)}$ exactly, and the cross-layer error is at most $3.91 \times 10^{-4}$. For RoPE, neither factorization is exact, but all reported discrepancies remain well below the distances between rollout and measurement. This supports using the tractable mean-field rollout $P_{\text{mf}}^{(T)}$ as the corpus-level prediction.

## I.2 Context length

We repeat the analysis at $n = 2048$. Figure 6 shows the gradient influence profiles on Wikipedia, and Table 5 reports the rollout comparison on Wikipedia. This tests whether the conclusions persist at longer context.

The adaptive rollout remains the best method in every comparison at $n = 2048$. Its RoPE distances range from $0.04$ to $0.12$, compared with $0.67$–$0.76$ for attention-only rollout and $0.55$–$0.74$ for fixed residual rollout. The ALiBi results retain the same qualitative ordering as at $n = 256$, with adaptive distances between $0.07$ and $0.16$. Their agreement is somewhat less tight because the controlled ALiBi construction sets all content logits to zero, increasing the approximation burden over a much longer sequence. The advantage over fixed mixing is particularly pronounced for the deeper models. This is consistent with the infinite-depth theory: fixed positive mixing makes cumulative attention mixing grow with depth and eventually produces primacy collapse under the theorem's conditions, whereas adaptive mixing makes collapse depend on cumulative mixing rather than depth alone and, in these finite models, preserves residual paths to recent tokens.




(a) `bloom-176b` (b) `bloom-7b1` (c) `mpt-7b` (d) `mpt-30b` (e) `falcon-rw-7b`

(f) `qwen2.5-7b` (g) `qwen2.5-32b` (h) `qwen2.5-72b` (i) `qwen3-8b-base` (j) `smollm3-3b-base`

Figure 6: Gradient input-token influence $\hat{q}_{\text{grad}}$ at sequence length $n = 2048$ on Wikipedia for ALiBi (top) and RoPE (bottom). The profiles retain both primacy and recency peaks, although their prominence varies across models; in particular, they do not collapse onto the first token at the longer context length.

Table 5: Context-length ablation at $n = 2048$ on Wikipedia: normalized 1-Wasserstein distances (↓ lower is better) to gradient influence.

| Model | (a) Wu $\lambda = 1$ | (b) Abnar $\lambda = \frac{1}{2}$ | (c) Ours adaptive |
|---|---|---|---|
| `falcon-rw-7b` | 0.46 | 0.20 | **0.14** |
| `mpt-7b` | 0.46 | 0.18 | **0.16** |
| `mpt-30b` | 0.82 | 0.45 | **0.07** |
| `bloom-7b1` | 0.45 | 0.15 | **0.14** |
| `bloom-176b` | 0.74 | 0.49 | **0.16** |
| `qwen2.5-7b` | 0.72 | 0.55 | **0.05** |
| `qwen2.5-32b` | 0.76 | 0.74 | **0.06** |
| `qwen2.5-72b` | 0.67 | 0.66 | **0.12** |
| `qwen3-8b-base` | 0.70 | 0.60 | **0.04** |
| `smollm3-3b-base` | 0.73 | 0.71 | **0.09** |

### I.3 Validation beyond gradient attribution

We repeat the comparison using token blank-out (Abnar & Zuidema, 2020). For an input $x$, let $x^{(-j)}$ denote the same input with embedding $j$ set to zero while positions and the causal mask remain unchanged. Holding the original predicted token $y(x)$ fixed, we define

$$b_j(x) = \left| z_{y(x)}(x) - z_{y(x)}(x^{(-j)}) \right|, \qquad q_{\text{blank}}(j) = \frac{\mathbb{E}_{x\sim\mathcal{D}}[b_j(x)]}{\sum_k \mathbb{E}_{x\sim\mathcal{D}}[b_k(x)]}.$$

As for gradient influence, raw scores are averaged over the corpus before normalization across positions.

All entries use FineWeb-Edu. The adaptive rollout is closer to blank-out than either baseline in every case. The blank-out and gradient profiles have a mean distance of 0.06, and replacing gradient influence by blank-out preserves the ordering of all three rollout variants. Across the ten models, the mean adaptive distance to blank-out is 0.10, compared with 0.64 for attention-only rollout and 0.63 for fixed residual rollout; the adaptive distances are also on the same scale as the corresponding gradient-based results in Table 3. Thus the result is not specific to gradient attribution or gradient saturation. Moreover, the adaptive error of 0.10 is of the same order as the 0.06 disagreement between the two influence measures, placing the rollout within the empirical measurement-sensitivity scale of the target profiles. This proxy disagreement provides a calibration for the remaining discrepancy between rollout and measurement, rather than a formal lower bound on it.

Table 6: Influence-measurement ablation at $n = 256$: normalized 1-Wasserstein distances (↓ lower is better) to blank-out influence. Columns (a), (b), and (c) use $\lambda_t = 1$ following Wu et al. (2025), $\lambda_t = \frac{1}{2}$ following Abnar & Zuidema (2020), and our adaptive measured norm-Shapley coefficient $\lambda_t$, respectively. The final column directly compares blank-out and gradient influence.

| Model | (a) Wu $\lambda = 1$ | (b) Abnar $\lambda = \frac{1}{2}$ | (c) Ours adaptive | Blank-out vs. gradient |
|---|---|---|---|---|
| `falcon-rw-7b` | 0.72 | 0.69 | **0.09** | 0.10 |
| `mpt-7b` | 0.66 | 0.62 | **0.11** | 0.08 |
| `mpt-30b` | 0.60 | 0.59 | **0.04** | 0.03 |
| `bloom-7b1` | 0.68 | 0.63 | **0.08** | 0.06 |
| `bloom-176b` | 0.59 | 0.59 | **0.17** | 0.06 |
| `qwen2.5-7b` | 0.65 | 0.62 | **0.09** | 0.06 |
| `qwen2.5-32b` | 0.65 | 0.65 | **0.06** | 0.06 |
| `qwen2.5-72b` | 0.67 | 0.67 | **0.20** | 0.08 |
| `qwen3-8b-base` | 0.61 | 0.59 | **0.08** | 0.03 |
| `smollm3-3b-base` | 0.61 | 0.61 | **0.11** | 0.03 |

### I.4 Predicting an intervention on residual mixing

The rollout makes an intervention prediction that can be tested without fitting to the resulting influence profile. For each model, we first record the unmodified attention and residual branches, the corpus-level transport kernels, and the norm-Shapley mixing schedule. We then multiply every attention-branch output by the fixed factor $\beta$, leaving the residual branch unchanged. The original prompts and next-token targets are held fixed across conditions. Neither the transport kernels nor the mixing schedule is remeasured after intervention.

Let $r_t$ and $a_t$ denote the residual and attention branches recorded in the unmodified model. The norm-Shapley definition predicts the intervened coefficient

$$\lambda_t(\beta) = \frac{\|r_t + \beta a_t\| + \beta\|a_t\| - \|r_t\|}{2\|r_t + \beta a_t\|}. \tag{19}$$

We substitute the corpus average of Equation (19) into the rollout while keeping the unmodified corpus-level kernels fixed. Thus the prediction uses only forward quantities measured before intervention; no gradient influence from an intervened model enters its construction. We use the moderate interventions $\beta = 0.8$ and $\beta = 1.2$ on three ALiBi and three RoPE models, using $1{,}024$ matched prompts per model. Prefix and recency mass refer to the first and last 26 positions, respectively.

The prefix prediction has the correct sign in all twelve model-intervention settings. The complementary recency prediction also has the correct point estimate in all twelve settings, although its paired interval excludes zero in ten. The prefix interval likewise excludes zero in ten of twelve settings. The agreement is not a restatement of the static profile comparison: the coefficient in Equation (19) and the fixed rollout predict the direction of change before observing the intervened influence curve. Measured changes are generally smaller than predicted. The rollout isolates the direct transport effect while holding the original kernels fixed, whereas the intervention also changes downstream representations and subsequent attention computations. These unmodeled feedbacks can attenuate the response without reversing its direction. Accordingly, all 24 point estimates have the predicted sign, while four confidence intervals include zero because the corresponding measured shifts are small.

The interventions preserve the original top prediction on $83\%$–$94\%$ of prompts, and the mean next-token KL remains at most $0.22$. These diagnostics support interpreting $\beta = 0.8$ and $\beta = 1.2$ as moderate perturbations around the trained models. Larger factors are not used in this analysis because applying the same scale at every layer compounds with depth and can move the network outside this local regime.

Table 7: Predicted and measured changes under attention-branch scaling, relative to the unmodified model ($\beta = 1$). Changes are percentage points of normalized prefix or recency mass. Brackets give paired 95% bootstrap confidence intervals for measured changes.

| | | Prefix mass | | Recency mass | |
|---|---|---|---|---|---|
| Model | $\beta$ | Predicted | Measured [95% CI] | Predicted | Measured [95% CI] |
| bloom-7b1 | 0.8 | $-1.21$ | $-0.74\,[-0.89, -0.58]$ | $+8.44$ | $+2.31\,[+1.99, +2.62]$ |
| | 1.2 | $+1.41$ | $+0.78\,[+0.69, +0.88]$ | $-7.47$ | $-2.12\,[-2.37, -1.84]$ |
| mpt-7b | 0.8 | $-1.04$ | $-0.51\,[-1.39, +0.53]$ | $+8.92$ | $+1.61\,[+0.46, +2.81]$ |
| | 1.2 | $+1.08$ | $+1.41\,[+0.59, +2.39]$ | $-7.14$ | $-4.62\,[-5.89, -3.46]$ |
| falcon-rw-7b | 0.8 | $-2.12$ | $-0.79\,[-1.10, -0.44]$ | $+5.99$ | $+0.17\,[-0.48, +0.83]$ |
| | 1.2 | $+2.34$ | $+0.96\,[+0.70, +1.23]$ | $-4.95$ | $-0.71\,[-1.37, -0.04]$ |
| qwen2.5-7b | 0.8 | $-1.57$ | $-1.20\,[-1.99, -0.63]$ | $+5.30$ | $+2.21\,[+1.07, +3.41]$ |
| | 1.2 | $+1.54$ | $+0.90\,[+0.20, +1.39]$ | $-4.67$ | $-1.66\,[-2.98, -0.35]$ |
| qwen3-8b-base | 0.8 | $-2.65$ | $-1.09\,[-1.87, -0.47]$ | $+7.02$ | $+2.28\,[+1.32, +3.30]$ |
| | 1.2 | $+2.68$ | $+0.37\,[-0.38, +0.94]$ | $-6.00$ | $-1.85\,[-3.09, -0.65]$ |
| smollm3-3b-base | 0.8 | $-2.84$ | $-0.71\,[-1.02, -0.43]$ | $+9.55$ | $+0.91\,[+0.03, +1.98]$ |
| | 1.2 | $+2.94$ | $+0.44\,[+0.15, +0.71]$ | $-8.37$ | $-0.41\,[-1.31, +0.76]$ |

Table 8: Intervention stability. KL is the mean next-token $D_{\mathrm{KL}}(p_\beta \| p_1)$; Top-1 is the percentage of prompts retaining the unmodified next-token prediction.

| | $\beta = 0.8$ | | $\beta = 1.2$ | |
|---|---|---|---|---|
| Model | KL | Top-1 | KL | Top-1 |
| bloom-7b1 | 0.03 | 91% | 0.02 | 93% |
| mpt-7b | 0.22 | 83% | 0.08 | 85% |
| falcon-rw-7b | 0.09 | 88% | 0.05 | 89% |
| qwen2.5-7b | 0.08 | 88% | 0.03 | 92% |
| qwen3-8b-base | 0.05 | 90% | 0.03 | 94% |
| smollm3-3b-base | 0.08 | 88% | 0.06 | 91% |

### I.5 Mixing coefficient

We replace the norm-Shapley coefficient by the alternative relative-norm coefficient, computed using the Frobenius norm,

$$\lambda_t^{\mathrm{rel}}(x) := \frac{\|\mathrm{Attn}(X^{(t-1)})\|_F}{\|X^{(t-1)}\|_F + \|\mathrm{Attn}(X^{(t-1)})\|_F}.$$

Table 9: Mixing-coefficient ablation at $n = 256$: normalized 1-Wasserstein distances (↓ lower is better) to gradient influence on FineWeb-Edu. Columns (a) and (b) use $\lambda_t = 1$ following Wu et al. (2025) and $\lambda_t = \frac{1}{2}$ following Abnar & Zuidema (2020). Columns (c) and (d) use our adaptive rollout with the relative-norm and norm-Shapley coefficients, respectively. Bold indicates the lowest distance in each row.

| Model | (a) Wu $\lambda = 1$ | (b) Abnar $\lambda = \frac{1}{2}$ | (c) Ours relative norm | (d) Ours norm-Shapley |
|---|---|---|---|---|
| falcon-rw-7b | 0.66 | 0.63 | 0.19 | **0.03** |
| mpt-7b | 0.62 | 0.58 | **0.09** | 0.15 |
| mpt-30b | 0.61 | 0.61 | 0.16 | **0.02** |
| bloom-7b1 | 0.63 | 0.58 | **0.04** | 0.13 |
| bloom-176b | 0.65 | 0.65 | **0.08** | 0.11 |
| qwen2.5-7b | 0.63 | 0.59 | 0.29 | **0.12** |
| qwen2.5-32b | 0.59 | 0.59 | 0.38 | **0.11** |
| qwen2.5-72b | 0.60 | 0.60 | 0.48 | **0.12** |
| qwen3-8b-base | 0.58 | 0.57 | 0.24 | **0.09** |
| smollm3-3b-base | 0.60 | 0.60 | 0.42 | **0.11** |
| **Mean** | 0.62 | 0.60 | 0.24 | **0.10** |

Norm-Shapley reduces the mean distance from $0.24$ to $0.10$ relative to the relative-norm calibration and is better for every RoPE model. The comparison is mixed within the ALiBi family, but norm-Shapley provides the stronger common calibration across both positional mechanisms while accounting for reinforcement and cancellation between the two residual branches.

## J Reproducibility details

**Protocol.** All experiments evaluate public pretrained checkpoints; no model is trained or fine-tuned, and no rollout parameter is fitted to the reported influence profiles. We use seed 42 and batch size one. Samples are taken in deterministic streaming order from the dataset configurations and splits listed in the supplementary manifest. Each text is tokenized with the corresponding model tokenizer, truncated to the required length, and retained only when it contains exactly $n$ tokens. Models run in evaluation mode in bfloat16, while accumulated profiles, kernels, and reported distances use double precision. Gradient influence is the input-embedding gradient norm of the model's predicted next-token probability. Normalized 1-Wasserstein distance uses ground cost $|i-j|/(n-1)$.

**Uncertainty.** The main Wasserstein tables report point estimates between corpus-mean profiles; there is no variation from training or parameter initialization. The bands in Appendix H are pointwise 95% intervals computed as the sample mean plus or minus $1.96$ standard errors over prompts. The intervention study uses the same prompts and targets in every condition and reports paired percentile-bootstrap intervals from 1,000 prompt resamples with seed 42. Its directional result holds for all 24 reported point estimates. The blank-out study provides an additional measurement-proxy check independent of gradients.

**Compute and artifacts.** The experiments require inference and attribution only. They were run with bfloat16 model weights on NVIDIA RTX PRO 6000 Blackwell GPUs with 96 GB of device memory, using one GPU for the smaller models and model parallelism for the largest checkpoints, up to two GPUs for Qwen2.5-72B and four for BLOOM-176B. Aggregation, Wasserstein evaluation, bootstrap resampling, and plotting run on CPU. The complete code is available at `https://anonymous.4open.science/r/position_bias_submit-7406`. It contains the configuration files, locked software environment, and commands required to reproduce the experiments, tables, and figures. It also specifies the exact model and dataset identifiers, sample counts, sequence lengths, and random seed. Model weights and datasets are downloaded from their public repositories.

## K LICENSES AND USAGE NOTES

Table 10 summarizes license names, links, and citations for the models and datasets used in this paper.

Table 10: License summary for models and datasets used in this work.

| Resource | License | Link | Notes / Citation |
|---|---|---|---|
| Falcon | Apache 2.0 | Model repository | (Penedo et al., 2023) |
| MPT | Apache 2.0 | Model repository | (Mosaic, 2023a;b) |
| BLOOM | BigScience RAIL License v1.0 | Model repository | (Scao et al., 2022) |
| Qwen2.5 (7B, 32B) | Apache 2.0 | Model repository | (Qwen Team, 2024) |
| Qwen2.5 (72B) | Qwen License Agreement | Model repository | (Qwen Team, 2024) |
| Qwen3-8B-Base | Apache 2.0 | Model repository | (Qwen Team, 2025) |
| SmolLM3-3B-Base | Apache 2.0 | Model repository | (Hugging Face Smol Models Team, 2025) |
| FineWeb-Edu | ODC-By-1.0 | Dataset repository | (Penedo et al., 2024) |
| DCLM-Baseline | CC-by-4.0 | Dataset repository | (Li et al., 2024) |
| Wikipedia | CC-by-SA-3.0 | Dataset repository | (Wikimedia Foundation, 2026) |