

SpiralFormer: Looped Transformers Can Learn Hierarchical Dependencies via Multi-Resolution Recursion

Chengting Yu^{†1,2}, Xiaobo Shu^{†1}, Yadao Wang¹, Yizhen Zhang¹,
Haoyi Wu^{1,3}, You Wu^{1,3}, Rujiao Long¹, Ziheng Chen¹,
Yuchi Xu¹, Wenbo Su^{✉1}, Bo Zheng¹

¹Future Living Lab of Alibaba ²Zhejiang University ³ShanghaiTech University

[†]Equal contribution [✉]Corresponding author

Abstract

Recursive (looped) Transformers decouple computational depth from parameter depth by repeatedly applying shared layers, providing an explicit architectural primitive for iterative refinement and latent reasoning. However, early looped Transformers often underperform non-recursive baselines of equal compute. While recent literature has introduced more effective recursion mechanisms to mitigate this gap, existing architectures still operate at a fixed, full-token resolution, neglecting the potential efficiency of computing over compressed latent representations. In this paper, we propose **SpiralFormer**, a looped Transformer that executes recurrence under a **multi-resolution recursion** schedule. We provide probing evidence that multi-resolution recursion enables the model to learn hierarchical dependencies by inducing iteration-wise functional specialization across different scales. Empirically, SpiralFormer achieves better parameter and compute efficiency than both looped and non-looped baselines across model scales from 160M to 1.4B, establishing sequence resolution as a potential axis for scaling recursive architectures.

1 Introduction

Scaling model parameters and data has been the dominant driver behind the rapid progress of large language models (LLMs) (Brown et al., 2020; Chowdhery et al., 2023). Yet this axis faces growing headwinds: high-quality text is finite, training and deployment costs are substantial, and memory/communication overheads scale unfavorably with model size (Villalobos et al., 2022; Patterson et al., 2021). These constraints have motivated efficient architectures that improve *capability per parameter* by allocating more computation within a fixed parameter budget.

A promising alternative is **recursive (looped) Transformers**, which decouple computational

depth from parameter depth by repeatedly applying a shared set of layers (Dehghani et al., 2018; Geiping et al., 2025; Zhu et al., 2025b). Beyond parameter efficiency, looping provides an explicit substrate for *iterative refinement*: the model can update its hidden representation multiple times before emitting the next token, resembling “thinking time” in latent space (Geiping et al., 2025; Zeng et al., 2025b,a). This perspective aligns looped Transformers with recent progress on *latent reasoning*, where models perform additional computation in continuous space (e.g., latent thoughts) instead of externalizing every intermediate step as explicit text (Hao et al., 2024; Zhu et al., 2025a; Li et al., 2025b).

However, despite this conceptual appeal, early looped Transformers often underperform equally expensive non-recursive baselines. A growing literature mitigates these issues by enriching the recursion mechanism (Yu et al., 2025; Bae et al., 2025; Wu et al., 2025; Heo et al., 2025). While effective, these approaches keep every iteration at the *full-length* sequence, operating at a single, fixed resolution and leaving the resolution axis of recursion largely unexplored suggesting from latent reasoning that many reasoning steps can be carried out on compressed representations.

Complementarily, latent reasoning work (Hao et al., 2024; Cheng and Van Durme, 2024) suggests that intermediate computation need not operate at full token granularity. Many approaches can be viewed as introducing a limited number of *high-capacity latent slots* (e.g., chunk-level summaries) that compress multi-token information and support richer computation per step, improving token efficiency (Li et al., 2025b; Chen et al., 2025a; Shi et al., 2025). This points to a missing architectural axis for recursive Transformers: if loop iterations may involve early global processing followed by later local refinement, then re-running full-resolution attention at every loop may be com-

putationally wasteful and misaligned with a hierarchical dependency structure.

Inspired by the view that latent reasoning benefits from operating over a small set of high-capacity latent slots, we ask whether the same compression form can be made an explicit architectural primitive *within* recursion. We propose **SpiralFormer**, a looped Transformer that executes a single shared core across a **multi-resolution recursion** schedule. Each iteration constructs a compressed latent sequence by downsampling current hidden states into chunk-level slots, applies the shared Transformer core to capture interactions at that specific resolution, and then causally upscales the result back to token resolution as an update. Empirically, we show that multi-resolution recursion enables looped Transformers to learn hierarchical dependencies. Using attention-based probes, we observe coherent shifts in attention statistics across loop iterations and head-level specialization under changes in resolution. We also show that SpiralFormer achieves **superior parameter and compute efficiency** over both looped and non-looped baselines. Our main contributions are:

- We introduce **multi-resolution recursion**, a novel architectural primitive for looped Transformers that operationalizes multi-token latent compression directly within the recursion.
- We provide probing evidence that multi-resolution recursion enables looped Transformers to learn **hierarchical, scale-dependent dependencies** by exhibiting iteration-wise functional specialization across different resolutions.
- We show that SpiralFormer achieves superior **parameter and compute efficiency**, consistently outperforming looped and non-looped baselines across model scales (160M–1.4B) with less compute or fewer parameters.

2 Method

2.1 Preliminaries: Recursive (Looped) Transformers

Recursive (looped) Transformers repeatedly apply a loop-shared block for multiple iterations, enabling deeper computation depth with fewer parameters.

Architectural backbone (Middle-cycle). We adopt the widely used Middle-cycle architecture (Geiping et al., 2025; Bae et al., 2025; Yu et al., 2025). It consists of (i) a pre-loop block f_{pre} with

N_{pre} Transformer layers, (ii) a core block f_{loop} with N_{loop} layers shared across loop iterations, and (iii) a post-loop block f_{post} with N_{post} layers.

Let $\mathbf{x} \in \mathbb{R}^{L \times d}$ denote token embeddings. The model computes

$$\mathbf{h}^{(0)} = f_{\text{pre}}(\mathbf{x}), \quad (1)$$

$$\mathbf{h}^{(t+1)} = \Phi_{\text{loop}}^{(t)}(\mathbf{h}^{(t)}), \quad t = 0, \dots, T-1, \quad (2)$$

$$\mathbf{h}^{\text{out}} = f_{\text{post}}(\mathbf{h}^{(T)}). \quad (3)$$

where we use $\Phi_{\text{loop}}^{(t)}$ to denote the *entire* loop-step transition that maps $\mathbf{h}^{(t)}$ to $\mathbf{h}^{(t+1)}$, combining the computation driven by the shared block f_{loop} with a state-update topology (defined below).

Topological variants. Recursive architectures can differ in how each iteration fuses newly computed information into the running state. We abstract this by a topology update operator $\mathcal{U}$ that takes (i) the current loop state $\mathbf{h}^{(t)}$, (ii) a step-produced token-level update $\tilde{\mathbf{u}}^{(t)}$ (defined in §2.2), and (iii) an optional topology carrier $\mathcal{H}^{(t)}$ that stores cross-iteration information, as:

$$(\mathbf{h}^{(t+1)}, \mathcal{H}^{(t+1)}) = \mathcal{U}(\tilde{\mathbf{u}}^{(t)}, \mathbf{h}^{(t)}, \mathcal{H}^{(t)}; t). \quad (4)$$

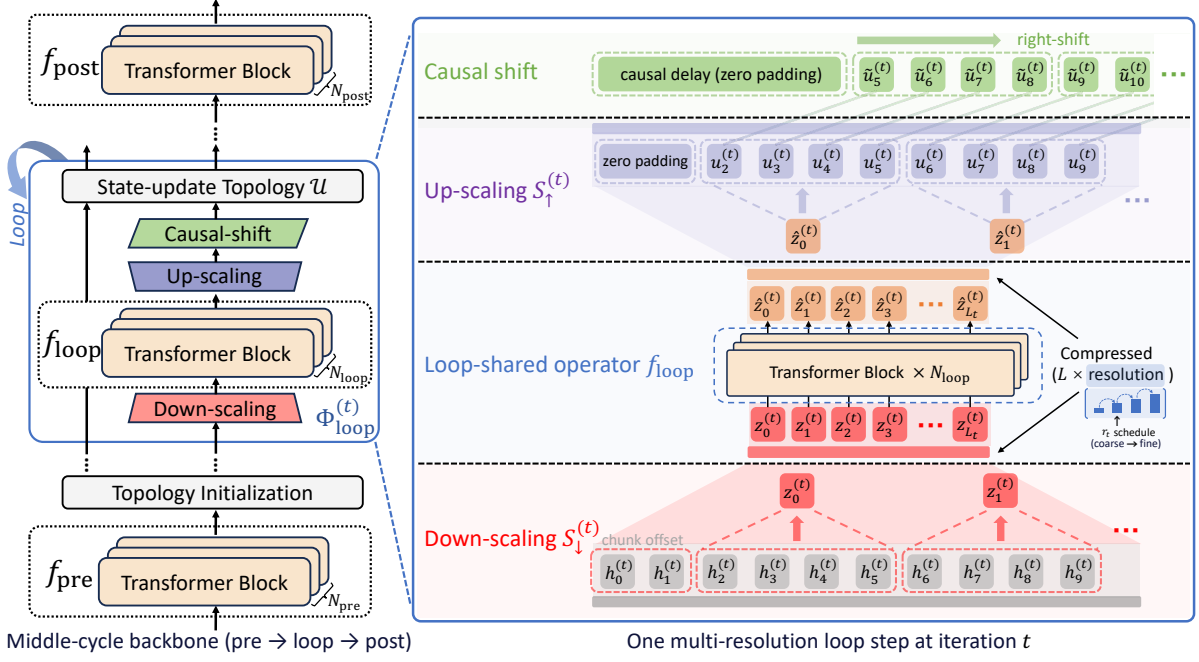
In our experiments we instantiate $\mathcal{U}$ as either Anchor or MeSH, defined below.

Anchor topology. Each iteration is attached to a fixed anchor state $\mathbf{h}^{(0)} = f_{\text{pre}}(\mathbf{x})$. Under the unified interface, the topology carrier is a constant container $\mathcal{H}^{(t)} \equiv \mathbf{h}^{(0)}$, and the topology update is

$$\begin{aligned} \mathcal{U}_{\text{anchor}}(\tilde{\mathbf{u}}^{(t)}, \mathbf{h}^{(t)}, \mathcal{H}^{(t)}; t) \\ = (\tilde{\mathbf{u}}^{(t)} + \mathbf{h}^{(0)}, \mathbf{h}^{(0)}). \end{aligned} \quad (5)$$

MeSH topology. MeSH replaces direct state passing with an explicit multi-slot memory buffer and lightweight step-wise read/write routers (Yu et al., 2025). Under MeSH, the topology carrier is $\mathcal{H}^{(t)} \equiv \mathbf{M}^{(t)} = \{\mathbf{m}_b^{(t)}\}_{b=0}^{B-1}$, where each memory slot $\mathbf{m}_b^{(t)} \in \mathbb{R}^{L \times d}$ has the same shape as the loop state. At each iteration t , MeSH (i) *writes* the causality-corrected update $\tilde{\mathbf{u}}^{(t)}$ into the buffer using a learned write router, and (ii) *reads* from the updated buffer using a learned read router to synthesize the next loop state $\mathbf{h}^{(t+1)}$:

$$(\mathbf{h}^{(t+1)}, \mathbf{M}^{(t+1)}) = \mathcal{U}_{\text{mesh}}(\tilde{\mathbf{u}}^{(t)}, \mathbf{h}^{(t)}, \mathbf{M}^{(t)}; t). \quad (6)$$



We follow Yu et al. (2025) and use a transitional write–read initialization to reorganize the pre-loop output into the buffer before entering the main loop. Full update equations are provided in Appendix B.

2.2 Multi-Resolution Recursion: Executing One Shared Core Across Scales

Our main idea is to execute different loop iterations at different sequence *resolutions*. Early iterations compute on a short (coarse) sequence to cheaply build global interactions; later iterations progressively increase resolution to refine token-level representations. This produces iteration-wise functional specialization without adding per-iteration core parameters.

Resolution schedule. We define a resolution schedule $\{r_t\}_{t=0}^{T-1}$ with $r_t \in (0, 1]$, where iteration t uses an effective length

$$L_t = \lfloor r_t L \rfloor. \quad (7)$$

One iteration as downscale → core → upscale.

At loop iteration t , we compute an update by:

$$\mathbf{z}^{(t)} = \mathcal{S}_{\downarrow}^{(t)}(\mathbf{h}^{(t)}; r_t) \in \mathbb{R}^{L_t \times d}, \quad (8)$$

$$\hat{\mathbf{z}}^{(t)} = f_{\text{loop}}(\mathbf{z}^{(t)}) \in \mathbb{R}^{L_t \times d}, \quad (9)$$

$$\mathbf{u}^{(t)} = \mathcal{S}_{\uparrow}^{(t)}(\hat{\mathbf{z}}^{(t)}; \mathbf{h}^{(t)}, r_t) \in \mathbb{R}^{L \times d}. \quad (10)$$

We detail the down- and up-scaling operators in §2.3.2 and §2.3.3, respectively. We then apply a causality-correcting right-shift operator (defined in §2.3.4) to obtain $\tilde{\mathbf{u}}^{(t)}$ and update the loop state:

$$(\mathbf{h}^{(t+1)}, \mathcal{H}^{(t+1)}) = \mathcal{U}(\tilde{\mathbf{u}}^{(t)}, \mathbf{h}^{(t)}, \mathcal{H}^{(t)}; t). \quad (11)$$

2.3 Causal Down- and Up-scaling Operators

The design of the $\mathcal{S}_{\downarrow}^{(t)}$ and $\mathcal{S}_{\uparrow}^{(t)}$ operators is critical for maintaining the *strict causality* required by autoregressive models: the update written to position i must not depend on tokens $> i$. The main challenge is that chunk-level aggregation summarizes an entire chunk, which includes future tokens relative to earlier positions in the same chunk. We enforce strict causality by a right-shift that also induces an interpretable overlap structure.

2.3.1 Chunking and position maps

At iteration t , we downscale the length- L sequence using blockwise chunking with chunk size $g_t = \lfloor \frac{1}{r_t} \rfloor \in \mathbb{N}$. We introduce an integer *chunk offset* $\omega_t \in \{0, 1, \dots, g_t - 1\}$, which shifts the chunk boundaries. We define the (offset) *chunk index*

map and in-chunk map as

$$\pi_t(i) = \left\lfloor \frac{i + \omega_t}{g_t} \right\rfloor, \quad \rho_t(i) = (i + \omega_t) - g_t \pi_t(i), \quad (12)$$

where $i \in \{0, \dots, L-1\}$ and $\rho_t(i) \in \{0, \dots, g_t-1\}$.

Dropping the last incomplete chunk. We realize the target resolution r_t using an integer chunk size $g_t = \lfloor \frac{1}{r_t} \rfloor$, which yields an effective number of chunks

$$L_t = \left\lfloor \frac{L}{g_t} \right\rfloor \approx \lfloor r_t L \rfloor. \quad (13)$$

The shifted partition may produce an incomplete last chunk; we simply drop it. Equivalently, we only keep positions

$$\mathcal{V}_t = \{i \in \{0, \dots, L-1\} : \pi_t(i) \leq L_t - 1\}, \quad (14)$$

and define the j -th chunk as

$$\mathcal{I}_{t,j} = \{i \in \mathcal{V}_t : \pi_t(i) = j\}, \quad j = 0, \dots, L_t - 1. \quad (15)$$

Default choice. Unless otherwise specified, we use the half-chunk offset $\omega_t = \lfloor g_t/2 \rfloor$. We provide further motivation and analysis in Appendix D. The full training-time procedure is summarized in Algorithm 1 in Appendix I.

2.3.2 Down-scaling operators $\mathcal{S}_\downarrow^{(t)}$

To create the latent vector $\mathbf{z}_j^{(t)}$, we apply a learnable weighted combination of states within the chunk. A lightweight, iteration-specific scorer $\mathcal{A}^{(t)} : \mathbb{R}^d \rightarrow \mathbb{R}$ (one linear layer in our implementation) computes importance weights $\alpha^{(t)}$ via Softmax normalization:

$$\alpha_{j,i}^{(t)} = \text{Softmax}_{i \in \mathcal{I}_{t,j}}(\mathcal{A}^{(t)}(\mathbf{h}_i^{(t)})) \quad (16)$$

The latent vector is then the aggregated sum:

$$\mathbf{z}_j^{(t)} = \sum_{i \in \mathcal{I}_{t,j}} \alpha_{j,i}^{(t)} \mathbf{h}_i^{(t)} \quad (17)$$

2.3.3 Up-scaling operators $\mathcal{S}_\uparrow^{(t)}$

Given $\hat{\mathbf{z}}^{(t)} \in \mathbb{R}^{L_t \times d}$, up-scaling writes updates back to token-level positions. We use an allocation vector $\beta_j^{(t)} \in \mathbb{R}^{g_t}$ together with a resolution-dependent gain λ_t for each valid position $i \in \mathcal{V}_t$:

$$\mathbf{u}_i^{(t)} = \lambda_t \beta_{\pi_t(i), \rho_t(i)}^{(t)} \cdot \hat{\mathbf{z}}_{\pi_t(i)}^{(t)}. \quad (18)$$

We apply an output-dependent router $\mathcal{B}^{(t)} : \mathbb{R}^d \rightarrow \mathbb{R}^{g_t}$ to predict allocation weights $\beta^{(t)}$ (one linear layer):

$$\beta_j^{(t)} = \text{Softmax}(\mathcal{B}^{(t)}(\hat{\mathbf{z}}_j^{(t)})) \in \mathbb{R}^{g_t}, \quad \sum_{\rho=0}^{g_t-1} \beta_{j,\rho}^{(t)} = 1. \quad (19)$$

Gain scaling. We set the up-scaling gain as $\lambda_t = \sqrt{g_t}$ to calibrate the token-level update magnitude across different chunk sizes.

2.3.4 Right-shift for strict causality and overlap regimes

A naive use of $\mathbf{u}^{(t)}$ would violate causality, as the representation $\hat{\mathbf{z}}_j^{(t)}$ depends on the entire chunk $\pi_t(j)$. We therefore apply a **causal right-shift** of size s_t to the update tensor, resulting in the final update $\tilde{\mathbf{u}}^{(t)}$:

$$\tilde{\mathbf{u}}^{(t)}[i] = \begin{cases} 0, & i < s_t, \\ \mathbf{u}^{(t)}[i - s_t], & i \geq s_t, \end{cases} \quad s_t \in \mathbb{N}. \quad (20)$$

Default configuration. SpiralFormer uses a coarse-to-fine schedule ($r_t = 2r_{t-1}$), the overlap-preserving causal shift $s_t = g_t - 1$, and the half-chunk offset $\omega_t = \lfloor g_t/2 \rfloor$. The shift $s_t = g_t - 1$ is the smallest value that guarantees strict causality under within-chunk aggregation and induces a single-token overlap between the source and target chunks (Proposition 1, Appendix C). The offset changes chunk boundaries and the decoding-time triggering pattern; details are given in Appendix D and Algorithm 2. We also discuss an alternative no-overlap setting ($s_t \geq g_t$) for inference-time pipelining in Appendix E.

3 Experiments

3.1 Setup

We evaluate SpiralFormer by pretraining decoder-only Transformers on the Pythia family (160M–1.4B) (Biderman et al., 2023). All models are trained from scratch for one epoch on the same deduplicated 250B-token subset of the Pile (Gao et al., 2020) and are evaluated by (i) language modeling perplexity and (ii) few-shot downstream accuracy. Additional implementation details are provided in Appendix F.

Models and configurations. We evaluate three model families at each Pythia scale:

Table 1: Main results on Pythia-suite (160M–1.4B) comparing SpiralFormer with (i) the non-recursive BASELINE (PYTHIA) and (ii) full-resolution LOOPEDFORMER. We report total / non-embedding parameters, prefill FLOPs for 4096-token, validation perplexity and downstream accuracy (0-shot / 5-shot). Shaded cells indicate reduced parameters or FLOPs relative to the BASELINE at the same scale. *denotes Anchor topology and †denotes MeSH topology. Best results in each scale are in **bold**, and second-best are underlined.

MODEL	CONFIG	PARAMS (M) (TOTAL/NON-EMB)	FLOPS (1E12) (4096 PREFILL)	PERPLEXITY ↓				TASK ACC ↑	
				PILE	WIKI	LD-O	LD-S	0-SHOT	5-SHOT
Pythia-160M									
BASILINE (PYTHIA)	12 LAYERS	163.5 / 85.1	1.65	11.31	30.32	42.86	175.62	39.88	40.54
LOOPEDFORMER*	2+4×{1,1}+2	135.2 / 56.7	1.65	11.63	31.69	50.38	195.11	38.81	40.15
LOOPEDFORMER†	2+4×{1,1}+2	135.2 / 56.7	1.65	11.37	30.43	46.60	178.77	39.41	40.60
SPIRALFORMER-B†	2+4×{1/8, 1/4, 1/2, 1}+2	135.2 / 56.8	1.48	11.29	30.27	43.27	155.78	39.73	41.02
SPIRALFORMER-L†	4+4×{1/16, 1/8, 1/4, 1/2}+4	163.6 / 85.1	1.49	10.94	28.85	41.24	147.52	39.30	41.37
Pythia-410M									
BASILINE (PYTHIA)	24 LAYERS	407.4 / 302.3	4.59	9.07	21.79	19.48	65.86	43.87	45.31
LOOPEDFORMER*	4+8×{1,1}+4	306.7 / 201.5	4.59	9.19	22.12	20.37	52.55	43.70	45.68
LOOPEDFORMER†	4+8×{1,1}+4	306.7 / 201.6	4.59	9.09	21.84	19.63	42.51	44.12	45.56
SPIRALFORMER-B*	4+8×{1/8, 1/4, 1/2, 1}+4	306.7 / 201.6	4.10	9.13	22.04	21.96	47.33	43.87	46.30
SPIRALFORMER-B†	4+8×{1/8, 1/4, 1/2, 1}+4	306.8 / 201.6	4.11	9.00	21.48	19.11	39.78	44.31	46.75
SPIRALFORMER-L†	8+8×{1/16, 1/8, 1/4, 1/2}+8	407.5 / 302.4	4.16	8.73	20.55	20.38	47.89	44.97	47.06
Pythia-1B									
BASILINE (PYTHIA)	16 LAYERS	1020.2 / 805.7	9.67	7.96	17.66	13.53	33.65	46.95	49.07
LOOPEDFORMER*	3+5×{1,1}+3	768.4 / 553.9	9.67	8.10	18.15	13.32	32.34	46.73	48.83
LOOPEDFORMER†	3+5×{1,1}+3	768.4 / 554.0	9.67	7.90	17.54	12.19	26.71	47.53	49.51
SPIRALFORMER-B†	3+5×{1/8, 1/4, 1/2, 1}+3	768.6 / 554.1	8.95	7.80	17.21	11.96	25.55	48.14	50.25
SPIRALFORMER-L†	5+6×{1/16, 1/8, 1/4, 1/2}+5	1020.4 / 805.9	8.96	7.64	16.73	11.94	23.90	48.97	51.83
Pythia-1.4B									
BASILINE (PYTHIA)	24 LAYERS	1423.0 / 1208.6	14.08	7.44	15.97	10.51	22.81	49.50	51.93
BASILINE†	24 LAYERS	1423.1 / 1208.7	14.08	7.26	15.25	9.46	16.31	50.67	53.23
LOOPEDFORMER*	4+8×{1,1}+4	1020.2 / 805.7	14.08	7.51	16.25	10.71	19.37	49.39	51.27
LOOPEDFORMER†	4+8×{1,1}+4	1020.2 / 805.8	14.08	7.39	15.84	9.72	19.39	50.56	52.79
SPIRALFORMER-B†	4+8×{1/8, 1/4, 1/2, 1}+4	1020.4 / 805.9	12.92	7.30	15.61	9.06	15.30	51.48	53.22
SPIRALFORMER-L†	8+8×{1/16, 1/8, 1/4, 1/2}+8	1423.2 / 1208.8	13.13	7.14	15.03	9.73	14.42	51.75	54.37

(i) the *standard* Pythia BASELINE; (ii) the *full-resolution recursive* LOOPEDFORMER that matches the unrolled compute depth by looping a shared core; and (iii) SPIRALFORMER, which replaces each loop step by *multi-resolution recursion*. We denote recursive layer allocations as $N_{\text{pre}} + N_{\text{loop}} \times \{r_0, \dots, r_{T-1}\} + N_{\text{post}}$, where $\{r_t\}$ is the resolution schedule of the T loop iterations (§2.2). In particular, LOOPEDFORMER corresponds to the full-resolution schedule $\{1, 1\}$. For example, $4+8 \times \{1, 1\}+4$ is a recursive model with a 4-layer pre-loop block, an 8-layer loop-shared core executed twice at full resolution, and a 4-layer post-loop block, while $4+8 \times \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1\}+4$ executes the same shared core over a coarse-to-fine schedule.

3.2 Main results.

Table 1 compares SpiralFormer against (i) full-resolution recursive baselines (LOOPEDFORMER) and (ii) standard Pythia (BASELINE). To disentangle compute and parameter effects, we report two variants: SPIRALFORMER-B (*base*), which keeps

the same recursive layer allocation as LOOPEDFORMER but uses a coarse-to-fine multi-resolution schedule, and SPIRALFORMER-L (*large*), which matches the parameter count of the non-recursive baseline while replacing the middle full-resolution computation with coarse-to-fine recursions.

Multi-resolution recursion improves recursive baselines. Under the same layer allocation, SPIRALFORMER-B consistently improves perplexity and few-shot accuracy over full-resolution LOOPEDFORMER, while also reducing FLOPs. Across sizes, SPIRALFORMER-B reduces FLOPs by $\sim 7\text{--}11\%$ relative to LOOPEDFORMER (e.g., $4.59 \rightarrow 4.11$ at 410M, $9.67 \rightarrow 8.95$ at 1B), while maintaining or improving downstream accuracy, which demonstrates that multi-resolution recursion strengthens looped Transformers without increasing parameters.

Surpassing non-recursive baselines with less compute. SPIRALFORMER-L matches the parameter count of Pythia but reduces FLOPs by executing most loop computation on compressed se-

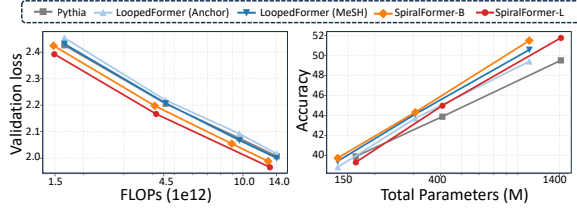


Figure 2: **Scaling behavior of SpiralFormer.** **Left:** validation loss versus computing FLOPs. **Right:** downstream 0-shot accuracy versus total parameters.

quences, while improving quality. Across sizes, SPIRALFORMER-L reduces FLOPs by $\sim 3\text{--}10\%$ at matched parameters (e.g., $1.65 \rightarrow 1.49$ at 160M, $4.59 \rightarrow 4.16$ at 410M, $14.08 \rightarrow 13.13$ at 1.4B), and improves perplexity and few-shot performance. At 1.4B, it reduces FLOPs ($14.08 \rightarrow 13.13$) while improving 5-shot accuracy ($51.93 \rightarrow 54.37$). Overall, coarse-to-fine recursion improves capability per compute and serves as a drop-in efficiency upgrade to standard Transformer scaling. Appendix G reports a completed 2.8B scaling result, a control matching four shared-core applications, and a matched 410M hardware latency benchmark.

3.3 Scaling Effects

As shown in Figure 2 (left), both SPIRALFORMER-B and SPIRALFORMER-L define a consistently better loss–compute frontier than full-resolution LOOPEDFORMER (Anchor/MeSH) and Pythia. Besides, at matched parameter count, SpiralFormer also matches or exceeds both Pythia and LOOPEDFORMER across scales, with the gap generally widening at larger sizes (Figure 2, right). Together, the scaling results suggest that multi-resolution recursion yields improvements that persist under scaling, strengthening the overall scaling frontier in both compute- and parameter-constrained regimes.

3.4 Impact of Recurrence Ratio

To better understand the architectural trade-offs, we study the effect of the **Recurrence Ratio** in SPIRALFORMER-L. We define the recurrence ratio as the fraction % of layers assigned to the shared core, $N_{\text{loop}}/N_{\text{total}}$, where N_{total} counts the total number of Transformer layers. A ratio of 0% corresponds to a non-recursive MeSH baseline (“No loop”), while higher ratios indicate more aggressive parameter sharing inside the multi-resolution recursive core. For all settings we keep $N_{\text{pre}} = N_{\text{post}}$ in the layer allocation. We instantiate this study on the 410M-scale model with $N_{\text{total}} = 24$, and train each configuration at two compute budgets (8e19 and 16e19 FLOPs). Figure 3 plots validation loss

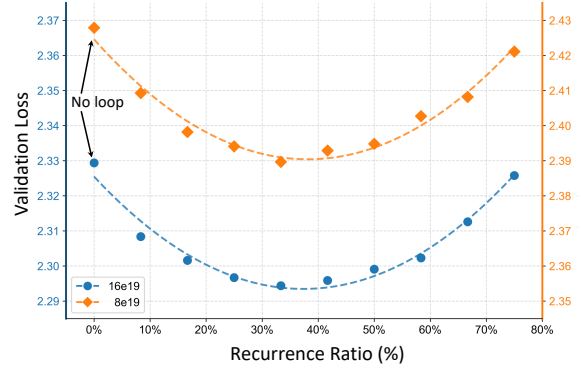


Figure 3: **Impact of recurrence ratio on validation loss.** We vary the recurrence ratio, defined as the fraction of layers placed in the shared core, $N_{\text{loop}}/N_{\text{total}}$, while keeping the total parameter fixed. Results at 410M scale are shown under two compute budgets (8e19 and 16e19 FLOPs).

as a function of Recurrence Ratio.

We observe a clear U-shaped curve. Moving from the non-looped baseline (0% ratio) to an architecture with even a small amount of looping (10%) already yields a substantial improvement in validation loss. As the Recurrence Ratio increases, the loss continues to decrease and reaches its minimum around 30–40%. Beyond this range, further increasing the ratio leads to degradation, indicating that excessive parameter sharing starts to hurt capacity. This pattern is consistent across both compute budgets, suggesting that the optimal balance between looped and non-looped layers is largely insensitive to training FLOPs. It supports the choice of the balanced configuration used for SPIRALFORMER-L in our main experiments, and indicates that this balance remains stable as we scale under different compute regimes.

3.5 Ablation study.

We conduct an ablation study on the Pythia-410M model to validate our core design choices (Table 2). Additional ablations on down/up-scaling operator variants are reported in Appendix H.

1) Topology. Replacing MeSH with the Anchor topology reduces the absolute result, indicating that MeSH-style state management provides an additional gain. However, the topology-matched Anchor comparison in Table 1 still favors SpiralFormer-B over LoopedFormer at 410M (Pile: 9.13 vs. 9.19; 5-shot: 46.30 vs. 45.68), separating the multi-resolution improvement from MeSH’s multi-slot memory.

2) Schedule. Switching the resolution schedule from coarse-to-fine to fine-to-coarse also hurts performance, supporting coarse-to-fine as the rec-

Table 2: Ablation study of SpiralFormer core design choices on the 410M model. The first row (highlighted) denotes the **Main Configuration**. Each subsequent row evaluates the impact of changing **one** specific design choice from its default setting to a variant. Best results in each column are in **bold**, and second-best are underlined.

Design Axis	Ablation (Default $\rightarrow$ Variant)	Perplexity $\downarrow$				Task Acc $\uparrow$	
		Pile	Wiki	LD-O	LD-S	0-shot	5-shot
Main Configuration (SpiralFormer-B)		9.00	21.48	<u>19.11</u>	39.78	<u>44.31</u>	46.75
(1) Topology	MeSH $\rightarrow$ Anchor	9.13	22.04	21.96	47.33	43.87	46.30
(2) Schedule	Coarse-to-Fine $\rightarrow$ Fine-to-Coarse ($\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$)	9.24	22.38	21.56	55.36	43.61	46.01
(3) Causality	Overlap ($s_t = g_t - 1$) $\rightarrow$ Parallel ($s_t = g_t$)	9.10	21.86	20.91	52.37	43.53	45.70
(4) Offset	Half-chunk ($\omega_t = \lfloor g_t/2 \rfloor$) $\rightarrow$ Zero ($\omega_t = 0$)	<u>9.01</u>	<u>21.57</u>	18.69	<u>41.29</u>	44.63	<u>46.66</u>

ommended design.

3) Causality. Using the no-overlap “parallel” regime ($s_t \geq g_t$) further reduces quality, but it introduces deployment-time pipelining opportunities (Appendix E); closing this quality gap while retaining parallelism remains an open direction.

4) Chunk Offset. Changing the chunk offset to $\omega_t = 0$ performs comparably, suggesting it is a viable option; however, offsets affect the periodic triggering pattern and can induce non-uniform per-token compute (Appendix D), so we use the half-chunk offset $\omega_t = \lfloor g_t/2 \rfloor$ as a robust default to avoid confounding comparisons.

Matched core applications. Appendix G.2 compares the coarse-to-fine schedule against a uniform-resolution control with the same four shared-core applications. The coarse-to-fine model improves every reported metric, so the gain cannot be attributed solely to using four rather than two passes.

4 Analyzing Hierarchical Dependencies in Multi-Resolution Recursion

Multi-resolution recursion is designed to encourage a coarse-to-fine computation pattern: early loop iterations operate on compressed sequences to capture global structure, while later iterations refine token-level details at higher resolution. If the model leverages this structure, we expect the *effective dependency pattern* (as reflected by attention statistics) to shift systematically across loops as the resolution increases. In this section we test this hypothesis using two complementary attention-based probes.

4.1 Measurements for Cross-Loop Dependency Shifts

Let head identity (ℓ, h) denote an attention head inside the loop-shared block at layer ℓ and head index h within that layer. We analyze attention matrices as a function of loop index t and head identity (ℓ, h) . To capture cross-loop changes in

dependency patterns, we use two metrics that emphasize different aspects of attention behavior.

Key-marginal Entropy. For each head (ℓ, h) at loop t , let $A_{\ell,h}^{(t)} \in \mathbb{R}^{L_t \times L_t}$ denote the causal attention matrix at the loop-native length $L_t = \lfloor r_t L \rfloor$. We average over queries to obtain a key-marginal distribution

$$p_{\ell,h}^{(t)}(k) \propto \frac{1}{L_t} \sum_{q=0}^{L_t-1} A_{\ell,h}^{(t)}(q, k), \quad k = 0, \dots, L_t-1,$$

and define the normalized entropy

$$H_{\ell,h}^{(t)} = - \sum_{k=0}^{L_t-1} p_{\ell,h}^{(t)}(k) \log p_{\ell,h}^{(t)}(k) / \log L_t \in [0, 1].$$

Lower entropy indicates more concentrated key usage (selective attention), while higher entropy indicates more diffuse key usage (broad/global attention).

Local Attention Mass (LAM). We quantify localness by measuring how much attention probability mass falls into a resolution-aligned causal local backward window. Let $A_{\ell,h}^{(t)} \in \mathbb{R}^{L_t \times L_t}$ be the attention matrix at loop t . We set the window size

$$M_t = \lceil \gamma r_t \rceil, \quad \gamma = 32, \quad (21)$$

and define the causal local key window for each query q as

$$W_t(q) = \{k : q - M_t \leq k < q\} \cap \{0, \dots, L_t-1\}.$$

The LAM score is

$$\text{LAM}_{\ell,h}^{(t)} = \frac{1}{L_t} \sum_{q=0}^{L_t-1} \sum_{k \in W_t(q)} A_{\ell,h}^{(t)}(q, k).$$

Higher LAM indicates stronger local refinement behavior, while lower LAM indicates weaker locality.

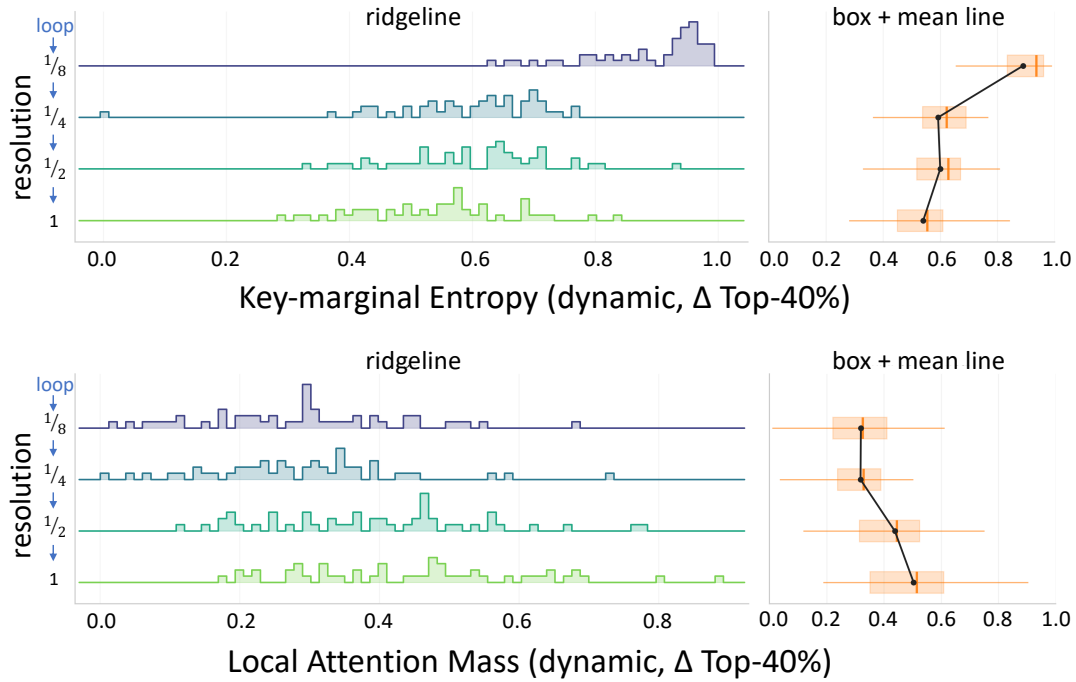


Figure 4: **Cross-loop distribution shifts of attention statistics on dynamic heads.** We visualize the distributions of (top) key-marginal entropy and (bottom) Local Attention Mass (LAM) across loop iterations (resolutions) for the *dynamic* heads (top 40% by cross-loop range for each metric) of the 410M SPIRALFORMER-B[†] model. Statistics are computed by averaging each head’s response over 500 samples from the Pile validation set. For each loop (from coarse 1/8 to full resolution 1), the left subplots show ridgeline histograms of head-wise values, and the right subplots show box plots with the loop-wise mean indicated by a black dot and connected across loops. For completeness, we further include the distributions over all heads in Appendix J.3.

Dynamic heads. To focus on heads that exhibit the clearest cross-loop adaptation, we report results on *dynamic heads* for each metric: the top 40% heads ranked by cross-loop range (ΔH for entropy and ΔLAM for LAM). (We include a head-level specialization analysis as supplementary material in Appendix J.1.)

4.2 Systematic Cross-Loop Shifts Under Coarse-to-Fine Recursion

Figure 4 visualizes the distributions of key-marginal entropy (top) and LAM (bottom) across loop iterations for the 410M SPIRALFORMER-B[†] model under a coarse-to-fine schedule. Statistics are computed on 500 sequences sampled from the Pile validation set: for each head, we average the metric over the 500 sequences to obtain one head-wise value per loop, and then plot the distribution across heads.

Observation 1 (attention becomes more selective at higher resolution). As resolution increases, the distribution of key-marginal entropy shifts downward (Figure 4, top). This indicates that higher-resolution loops allocate attention mass to a smaller subset of keys, consistent with a refine-

ment stage that selectively uses a few salient dependencies rather than maintaining a broad/global interaction pattern.

Observation 2 (local refinement strengthens at higher resolution). LAM exhibits a consistent cross-loop shift across resolutions (Figure 4, bottom). Interpreting LAM as the fraction of probability mass assigned to local neighborhoods, this provides evidence that later iterations increasingly emphasize local token-level dependencies, aligning with the intended role of fine-resolution loops as local refinement steps.

Hierarchical dependencies in the loop. Together, these two signatures show that the attention behavior changes systematically across loops as a function of resolution: coarse loops are more diffuse (higher entropy, weaker locality), whereas fine loops are more selective and more local (lower entropy, higher locality). This systematic shift supports the claim that multi-resolution recursion induces a hierarchical dependency pattern across iterations—global aggregation at coarse resolution followed by local refinement at fine resolution—even though the core parameters are shared across loops. We also apply the same probes to a full-

resolution LOOPEDFORMER[†] baseline and observe much weaker and less systematic cross-loop shifts (Appendix J.2, Figure 7), suggesting that the hierarchical pattern is tied to multi-resolution recursion rather than looping alone.

5 Conclusion

In this paper, we propose SpiralFormer, a recursive (looped) Transformer architecture that executes loop-shared layers under a multi-resolution recursion schedule. Empirically, SpiralFormer establishes a superior frontier for both parameter and compute efficiency, consistently outperforming both loop and non-loop baselines across model scales from 160M to 1.4B. Besides, through attention-based probing, we showed that multi-resolution recursion enables looped Transformers to learn hierarchical dependencies, opening sequence resolution as a potential axis for scaling recursive architectures.

Limitations

While this work establishes multi-resolution recursion as a promising architectural principle for looped Transformers, we recognize several limitations that open up avenues for future research. Our experiments validate SpiralFormer on decoder-only language models up to the Pythia-1.4B scale, trained from scratch on a deduplicated subset of the Pile. It remains unclear whether the parameter- and compute-efficiency gains persist at larger scales, under different training regimes, or after instruction tuning. In addition, our analysis relies on attention-based probes, which provide useful evidence of cross-loop specialization but do not fully characterize all internal computations induced by the proposed architecture. Finally, because SpiralFormer targets more efficient language-model training, its potential societal risks are primarily those associated with improving language-model capabilities more broadly; we do not identify additional risks specific to the proposed architectural mechanism beyond this general setting.

References

Preslav Aleksandrov, Meghdad Kurmanji, Fernando Garcia Redondo, David O’Shea, William Shen, Alex Iacob, Lorenzo Sani, Xinchu Qiu, Nicola Cancedda, and Nicholas D Lane. 2025. Abbie: Autoregressive block-based iterative encoder

for efficient sequence modeling. *arXiv preprint arXiv:2507.08567*.

Sangmin Bae, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Seungyeon Kim, and Tal Schuster. 2024. Relaxed recursive transformers: Effective parameter sharing with layer-wise lora. *arXiv preprint arXiv:2410.20672*.

Sangmin Bae, Yujin Kim, Reza Bayat, Sungnyun Kim, Jiyouon Ha, Tal Schuster, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Aaron Courville, and 1 others. 2025. Mixture-of-recursions: Learning dynamic recursive depths for adaptive token-level computation. *arXiv preprint arXiv:2507.10524*.

Loïc Barrault, Paul-Ambroise Duquenne, Maha El-bayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, and 1 others. 2024. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Nasim Borazjanizadeh and James McClelland. 2025. Modeling language as a sequence of thoughts. *arXiv preprint arXiv:2512.25026*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. 2025a. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *arXiv preprint arXiv:2505.16782*.

Yilong Chen, Junyuan Shang, Zhenyu Zhang, Yanxi Xie, Jiawei Sheng, Tingwen Liu, Shuohuan Wang, Yu Sun, Hua Wu, and Haifeng Wang. 2025b. Inner thinking transformer: Leveraging dynamic depth scaling to foster adaptive internal thinking. *arXiv preprint arXiv:2502.13842*.

Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning

- through dense representations. *arXiv preprint arXiv:2412.13171*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Beiya Dai, Yuliang Liu, Daozheng Xue, Qipeng Guo, Kai Chen, Xinbing Wang, Bowen Zhou, and Zhouhan Lin. 2025. Context-level language modeling by learning predictive context embeddings. *arXiv preprint arXiv:2510.20280*.
- Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.
- Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2018. Universal transformers. *arXiv preprint arXiv:1807.03819*.
- Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*.
- Eric Egli, Matteo Manica, and Jannis Born. 2025. Multiscale byte language models—a hierarchical architecture for causal million-length sequence modeling. *arXiv preprint arXiv:2502.14553*.
- Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. 2019. Depth-adaptive transformer. *arXiv preprint arXiv:1910.10073*.
- Ying Fan, Yilun Du, Kannan Ramchandran, and Kangwook Lee. 2024. Looped transformers for length generalization. *arXiv preprint arXiv:2409.15647*.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*.
- Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. 2023. Looped transformers as programmable computers. In *International Conference on Machine Learning*, pages 11398–11442. PMLR.
- Zixuan Gong, Yong Liu, and Jiaye Teng. 2025. What makes looped transformers perform better than non-recursive ones. *arXiv preprint arXiv:2510.10089*.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.
- Tamir David Hay and Lior Wolf. 2024. Dynamic layer tying for parameter-efficient transformers. *arXiv preprint arXiv:2401.12819*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Jaemu Heo, Eldor Fozilov, Hyunmin Song, and Taehwan Kim. 2025. Ringformer: Rethinking recurrent transformer with adaptive level signals. *arXiv preprint arXiv:2502.13181*.
- Namgyu Ho, Sangmin Bae, Taehyeon Kim, Hyunjik Jo, Yireun Kim, Tal Schuster, Adam Fisch, James Thorne, and Se-Young Yun. 2024. Block transformer: Global-to-local language modeling for fast inference. *Advances in Neural Information Processing Systems*, 37:48740–48783.
- Xiang Hu, Qingyang Zhu, Kewei Tu, and Wei Wu. 2023. Augmenting transformers with recursively composed multi-grained representations. *arXiv preprint arXiv:2309.16319*.
- Sukjun Hwang, Brandon Wang, and Albert Gu. 2025. Dynamic chunking for end-to-end hierarchical sequence modeling. *arXiv preprint arXiv:2507.07955*.
- Ahmadreza Jeddi, Marco Ciccone, and Babak Taati. 2026. Loopformer: Elastic-depth looped transformers for latent reasoning via shortcut modulation. *arXiv preprint arXiv:2602.11451*.
- Guanghao Li, Wenhao Jiang, Li Shen, Ming Tang, and Chun Yuan. 2025a. Zero token-driven deep thinking in llms: Unlocking the full potential of existing parameters via cyclic refinement. *arXiv preprint arXiv:2502.12214*.
- GuoLiang Li and Yiyang Li. 2021. Recurrent multiple shared layers in depth for neural machine translation. *arXiv preprint arXiv:2108.10417*.

- Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. 2025b. Implicit reasoning in large language models: A comprehensive survey. *arXiv preprint arXiv:2509.02350*.
- Pit Neitemeier, Björn Deiseroth, Constantin Eichenberg, and Lukas Balles. 2025. Hierarchical autoregressive transformers: Combining byte-and word-level processing for robust, adaptable language models. *arXiv preprint arXiv:2501.10322*.
- Anthony Nguyen and Wenjun Lin. 2025. Intra-layer recurrence in transformers for language modeling. *arXiv preprint arXiv:2505.01855*.
- Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambda dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.
- David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.
- Xingwei Qu, Shaowen Wang, Zihao Huang, Kai Hua, Fan Yin, Rui-Jie Zhu, Jundong Zhou, Qiyang Min, Zihao Wang, Yizhi Li, and 1 others. 2025. Dynamic large concept models: Latent reasoning in an adaptive semantic space. *arXiv preprint arXiv:2512.24617*.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.
- Chenze Shao, Darren Li, Fandong Meng, and Jie Zhou. 2025. Continuous autoregressive language models. *arXiv preprint arXiv:2510.27688*.
- Chenze Shao, Fandong Meng, and Jie Zhou. 2024. Beyond next token prediction: Patch-level training for large language models. *arXiv preprint arXiv:2407.12665*.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*.
- Zhiqiang Shen, Zechun Liu, and Eric Xing. 2022. Sliced recursive transformer. In *European Conference on Computer Vision*, pages 727–744. Springer.
- Dachuan Shi, Abedelkadir Asi, Keying Li, Xiangchi Yuan, Leyan Pan, Wenke Lee, and Wen Xiao. 2025. Swireasoning: Switch-thinking in latent and explicit for pareto-superior reasoning llms. *arXiv preprint arXiv:2510.05069*.
- Sho Takase and Shun Kiyono. 2021. Lessons on parameter sharing across layers in transformers. *arXiv preprint arXiv:2104.06022*.
- Shawn Tan, Yikang Shen, Zhenfang Chen, Aaron Courville, and Chuang Gan. 2023. Sparse universal transformer. *arXiv preprint arXiv:2310.07096*.
- Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2022. Will we run out of data? limits of llm scaling based on human-generated data. *arXiv preprint arXiv:2211.04325*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.
- Bohong Wu, Mengzhao Chen, Xiang Luo, Shen Yan, Qifan Yu, Fan Xia, Tianqi Zhang, Hongrui Zhan, Zheng Zhong, Xun Zhou, and 1 others. 2025. Parallel loop transformer for efficient test-time computation scaling. *arXiv preprint arXiv:2510.24824*.
- Liu Yang, Kangwook Lee, Robert Nowak, and Dimitris Papailiopoulos. 2023. Looped transformers are better at learning learning algorithms. *arXiv preprint arXiv:2311.12424*.
- Chengting Yu, Xiaobo Shu, Yadao Wang, Yizhen Zhang, Haoyi Wu, Jiaang Li, Rujiao Long, Ziheng Chen, Yuchi Xu, Wenbo Su, and 1 others. 2025. Mesh: Memory-as-state-highways for recursive transformers. *arXiv preprint arXiv:2510.07739*.
- Lili Yu, Dániel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis. 2023. Megabyte: Predicting million-byte sequences with multiscale transformers. *Advances in Neural Information Processing Systems*, 36:78808–78823.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.
- Boyi Zeng, He Li, Shixiang Song, Yixuan Wang, Ziwei He, Xinbing Wang, and Zhouhan Lin. 2025a. Ponderlm-2: Pretraining llm with latent thoughts in continuous space. *arXiv preprint arXiv:2509.23184*.
- Boyi Zeng, Shixiang Song, Siyuan Huang, Yixuan Wang, He Li, Ziwei He, Xinbing Wang, Zhiyu Li, and Zhouhan Lin. 2025b. Pretraining language models to ponder in continuous space. *arXiv preprint arXiv:2505.20674*.
- Xiang Zhang, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2024. Autoregressive+ chain of thought= recurrent: Recurrence’s role in language models’ computability and a revisit of recurrent transformer. *arXiv preprint arXiv:2409.09239*.

Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, and 1 others. 2025a. A survey on latent reasoning. *arXiv preprint arXiv:2507.06203*.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, and 1 others. 2025b. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*.

A Related Work

Looped and Recursive Transformers. The idea of increasing computational depth by iterating a Transformer block in a loop traces back to the Universal Transformer (UT) (Dehghani et al., 2018), which applies a weight-shared layer recurrently and thereby decouples computational depth from parameter depth. Following UT, a wide array of architectures have explored recursive/looped recurrence in different settings and training regimes (Elbayad et al., 2019; Li and Li, 2021; Takase and Kiyono, 2021; Shen et al., 2022; Tan et al., 2023; Giannou et al., 2023; Yang et al., 2023; Zhang et al., 2024; Fan et al., 2024; Hay and Wolf, 2024; Chen et al., 2025b; Nguyen and Lin, 2025; Li et al., 2025a; Aleksandrov et al., 2025; Gong et al., 2025). Recent scaling efforts including Huginn (Geiping et al., 2025) and Ouro (Zhu et al., 2025b) demonstrate that weight-shared recurrence can match frontier-level performance, while specialized recursion mechanisms such as MeSH (Yu et al., 2025), Mixture-of-Recursions (Bae et al., 2025), Relaxed recursive transformers (Bae et al., 2024), and RingFormer (Heo et al., 2025) further address limitations of vanilla looping (e.g., representational bottlenecks). LoopFormer (Jeddi et al., 2026) enables elastic-depth inference by varying trajectory length and step size along normalized recurrent depth, using shortcut conditioning and consistency training. SpiralFormer provides a complementary perspective by executing the same shared core at varying sequence resolutions, offering sequence resolution as an additional axis for improving looped Transformers.

Latent Reasoning and Implicit CoT. Latent reasoning internalizes the multi-step deliberation of Chain-of-Thought (CoT) within hidden states rather than through explicit text generation (Wei et al., 2022). Frameworks Continuous Thoughts (Hao et al., 2024; Shen et al., 2025), and Implicit CoT (Deng et al., 2024) enable models to refine internal hypotheses in continuous space before token emission. CCot (Cheng and Van Durme, 2024) utilizes pretrained models for distillation such that the model only generates important tokens in a continuous fashion. SwiReasoning (Shi et al., 2025) saves a large proportion of reasoning tokens by dynamically switching between explicit and implicit decoding mode. A recurring observation in this literature is the **token efficiency**: complex logic can be compacted into a limited set of high-capacity la-

tent slots to match the performance of much longer explicit reasoning chains. While existing methods primarily focus on distilling latent thoughts into fixed-depth models, SpiralFormer operationalizes this compression morphology as an explicit architectural primitive. By integrating multi-resolution computation into the recursive loop, we formalize the process of global conceptual planning followed by local refinement.

Hierarchical Architectures and Sequence Compression. Hierarchical modeling has been extensively explored to alleviate the $O(L^2)$ bottleneck of Transformers by introducing compressed or multi-scale representations. ReCAT (Hu et al., 2023) augments Transformers with recursively composed multi-grained span representations via contextual inside–outside passes. MegaByte (Yu et al., 2023), Hierarchical Autoregressive Transformer (Neitemeier et al., 2025), Thought Gestalt (Borazjanizadeh and McClelland, 2025) and Block Transformer (Ho et al., 2024) implement global-to-local processing to reduce inference overhead. Patch-level training (Shao et al., 2024), MBLM (Egli et al., 2025) and H-Net (Hwang et al., 2025) pursue training-time efficiency and context-length scaling through token aggregation and learned segmentation. CALM (Shao et al., 2025) uses an energy-based generative head to predict divergent patch representations. ContextLM (Dai et al., 2025) augments next-token prediction with objectives that learn predictive multi-token context embeddings. LCM (Barrault et al., 2024) and DLCM (Qu et al., 2025) attempt to process text sequence in the concept level. In contrast, SpiralFormer explores a single shared core across a coarse-to-fine resolution schedule within a loop, turning compression into an internal computation primitive rather than a fixed architectural partition.

B MeSH topology: full update equations

This appendix provides the explicit write–read equations for the MeSH topology used as an instantiation of the unified topology update in Eq. (4).

Memory buffer. MeSH maintains a B -slot memory buffer

$$M^{(t)} = \{\mathbf{m}_b^{(t)}\}_{b=0}^{B-1}, \\ \mathbf{m}_b^{(t)} \in \mathbb{R}^{L \times d}.$$

Write/read routers output token-wise distributions over the B slots (softmax over the slot dimension).

Pre-loop initialization ($t = -1$). Before the main loop starts, we initialize a designated anchor slot with the input embeddings $\mathbf{x}$:

$$\mathbf{m}_0^{(-1)} = \mathbf{x}, \quad \mathbf{m}_{b>0}^{(-1)} = \mathbf{0}. \quad (22)$$

Transitional write-read (reorganization before the main loop). We follow Yu et al. (2025) and perform a transitional write-read cycle before entering the main recursion. Let $\mathbf{v} = f_{\text{pre}}(\mathbf{x})$ denote the pre-loop output. Transitional routers take $\mathbf{x}$ as input and produce token-wise write/read weights

$$\begin{aligned} \mathbf{w}_{\text{write}}^{(-1)} &= \text{Softmax}_b \left(\text{Linear}_{\text{write}}^{(-1)}(\mathbf{x}) \right), \\ \mathbf{w}_{\text{read}}^{(-1)} &= \text{Softmax}_b \left(\text{Linear}_{\text{read}}^{(-1)}(\mathbf{x}) \right). \end{aligned} \quad (23)$$

where $\mathbf{w}_{\text{write}}^{(-1)}, \mathbf{w}_{\text{read}}^{(-1)} \in \mathbb{R}^{L \times B}$ and the softmax is applied over the slot dimension b for each token position. We write the pre-loop output into the buffer:

$$\begin{aligned} \mathbf{m}_b^{(0)} &= \mathbf{m}_b^{(-1)} + \mathbf{v} \odot \mathbf{w}_{\text{write},b}^{(-1)}, \\ b &= 0, \dots, B-1. \end{aligned} \quad (24)$$

and synthesize the initial loop state by reading from the updated buffer:

$$\mathbf{h}^{(0)} = \sum_{b=0}^{B-1} \mathbf{m}_b^{(0)} \odot \mathbf{w}_{\text{read},b}^{(-1)}. \quad (25)$$

Main-loop write-read update. At iteration t , given the current state $\mathbf{h}^{(t)}$, the causality-corrected update $\tilde{\mathbf{u}}^{(t)}$ (Eq. (20)), and the current buffer $\mathbf{M}^{(t)}$, MeSH computes step-wise routing weights

$$\begin{aligned} \mathbf{w}_{\text{write}}^{(t)} &= \text{Softmax}_b \left(\text{Linear}_{\text{write}}^{(t)}(\mathbf{h}^{(t)}) \right), \\ \mathbf{w}_{\text{read}}^{(t)} &= \text{Softmax}_b \left(\text{Linear}_{\text{read}}^{(t)}(\mathbf{h}^{(t)}) \right). \end{aligned} \quad (26)$$

updates the buffer by a distributed write

$$\begin{aligned} \mathbf{m}_b^{(t+1)} &= \mathbf{m}_b^{(t)} + \tilde{\mathbf{u}}^{(t)} \odot \mathbf{w}_{\text{write},b}^{(t)}, \\ b &= 0, \dots, B-1. \end{aligned} \quad (27)$$

and synthesizes the next loop state by reading from the updated buffer:

$$\mathbf{h}^{(t+1)} = \sum_{b=0}^{B-1} \mathbf{m}_b^{(t+1)} \odot \mathbf{w}_{\text{read},b}^{(t)}. \quad (28)$$

Equations (26)–(28) define the MeSH instantiation of the unified topology update in Eq. (4) by setting $\mathcal{H}^{(t)} \equiv \mathbf{M}^{(t)}$.

C Right-shift for strict causality and overlap regimes

Proposition 1 (Causality threshold and overlap).

Assume at iteration t , we partition positions into non-overlapping chunks of size g_t . Let Agg be any within-chunk aggregation operator that maps chunk tokens $\{\mathbf{h}_i^{(t)} : i \in \mathcal{I}_{t,j}\}$ to a single coarse vector, and let Exp be any within-chunk expansion operator that maps the coarse vector back to all positions in the same chunk. Denote the pre-shift expanded update by $\mathbf{u}^{(t)} \in \mathbb{R}^{L \times d}$. We form a shifted update $\tilde{\mathbf{u}}^{(t)}$ by right shifting $\mathbf{u}^{(t)}$ by s_t positions:

$$\tilde{\mathbf{u}}^{(t)}[i] = \begin{cases} 0, & i < s_t, \\ \mathbf{u}^{(t)}[i - s_t], & i \geq s_t. \end{cases}$$

Then: (1) If $s_t \leq g_t - 2$, $\tilde{\mathbf{u}}^{(t)}$ is not strictly causal in general. (2) If $s_t = g_t - 1$, $\tilde{\mathbf{u}}^{(t)}$ is strictly causal and induces a single-token overlap between the chunk generating the update and the chunk receiving it. (3) If $s_t \geq g_t$, $\tilde{\mathbf{u}}^{(t)}$ is strictly causal and induces no overlap (chunk-wise parallel regime).

Proof. Let chunks be indexed by $j = 0, \dots, L/g_t - 1$. Define the chunk index function $\pi_t(i) = \lfloor i/g_t \rfloor$. By construction, the pre-shift expansion is chunk-local: for any position k ,

$$\begin{aligned} \mathbf{u}^{(t)}[k] &\text{ depends only on } \\ &\{\mathbf{h}_u^{(t)} : u \in \mathcal{I}_{t,\pi_t(k)}\}. \end{aligned} \quad (\text{A.1})$$

After shifting, for any $i \geq s_t$, we have $\tilde{\mathbf{u}}^{(t)}[i] = \mathbf{u}^{(t)}[i - s_t]$ and therefore by (A.1),

$$\begin{aligned} \tilde{\mathbf{u}}^{(t)}[i] &\text{ depends only on } \\ &\{\mathbf{h}_u^{(t)} : u \in \mathcal{I}_{t,\pi_t(i-s_t)}\}. \end{aligned} \quad (\text{A.2})$$

(1) Non-causality when $s_t \leq g_t - 2$. Choose any chunk $j \geq 1$ and consider position $i = jg_t + 1$ (the second token of chunk j). Since $s_t \leq g_t - 2$, we have $i - s_t \geq jg_t - (g_t - 3)$ and there exist valid indices where $\pi_t(i - s_t) = j$. Then by (A.2), $\tilde{\mathbf{u}}^{(t)}[i]$ may depend on tokens in chunk j , including the last position $jg_t + (g_t - 1)$, which is a future token relative to i . Hence strict causality can be violated in general.

(2) Causality and single-token overlap when $s_t = g_t - 1$. Fix any $i \geq g_t - 1$ and write $i = jg_t + p$ with $p \in \{0, \dots, g_t - 1\}$. Then

$$i - s_t = jg_t + p - (g_t - 1) = (j - 1)g_t + (p + 1).$$

If $p \leq g_t - 2$, then $\pi_t(i - s_t) = j - 1$, so $\tilde{\mathbf{u}}^{(t)}[i]$ depends only on tokens in the previous chunk, all with indices $\leq jg_t - 1 < i$, hence causal. If $p = g_t - 1$, then $i - s_t = jg_t$ and $\pi_t(i - s_t) = j$. This is the unique position in chunk j whose update depends on chunk j itself. Moreover, $\mathbf{u}^{(t)}[jg_t]$ depends only on tokens within chunk j , whose maximum index is $jg_t + (g_t - 1) = i$, hence still causal. Therefore strict causality holds and the only overlap is at this single boundary position.

(3) Causality and no-overlap when $s_t \geq g_t$. For any $i = jg_t + p$, we have $i - s_t \leq jg_t + p - g_t \leq jg_t - 1$, hence $\pi_t(i - s_t) \leq j - 1$. Therefore $\tilde{\mathbf{u}}^{(t)}[i]$ depends only on tokens in chunks strictly before j , all with indices $< jg_t \leq i$, implying strict causality. Since $\pi_t(i - s_t) \neq j$ for all positions in chunk j , there is no overlap between the chunk receiving updates and the chunk generating them.

D Inference-Time Motivation for Chunk Offset

This appendix clarifies an inference-time phenomenon induced by blockwise multi-resolution recursion, and motivates our use of a chunk offset ω_t in Eq. (12) (default $\omega_t = \lfloor g_t/2 \rfloor$). We refer to Appendix Algorithm 2 for pseudocode showing how ω_t affects chunk indexing and the resulting chunk-triggered recomputation pattern in decoding.

Phenomenon: periodic triggering leads to non-uniform per-token compute. In our multi-resolution recursion, iteration t uses chunk size g_t and a causal right-shift of size $s_t = g_t - 1$ (Eq. (20)). Under autoregressive decoding, cached intermediate states can be reused. As a result, low-resolution computations (compress $\rightarrow$ core $\rightarrow$ decompress at chunk size g_t) are not necessarily triggered at every token position. Instead, they are triggered at a periodic subset of positions determined by the chunk structure, which implies that the number of resolution-level computations attributed to different tokens can vary. Intuitively, tokens that coincide with these triggering positions (one per chunk at each resolution) can incur more work than tokens that do not. Concretely, in Algorithm 2, low-resolution recomputation at iteration t is triggered only when the in-chunk index satisfies $\rho = g_t - 1$.

Key observation (under $s_t = g_t - 1$): one triggering position per chunk. Define the offset chunk

maps (Eq. (12))

$$\begin{aligned}\pi_t(i) &= \left\lfloor \frac{i + \omega_t}{g_t} \right\rfloor, \\ \rho_t(i) &= (i + \omega_t) - g_t \pi_t(i).\end{aligned}$$

Let $\mathbf{u}^{(t)}$ be the pre-shift upscaled update and $\tilde{\mathbf{u}}^{(t)}$ the shifted update. With $s_t = g_t - 1$, we have

$$\tilde{\mathbf{u}}^{(t)}[i] = \mathbf{u}^{(t)}[i - (g_t - 1)].$$

For positions with $\rho_t(i) \leq g_t - 2$, the source index $i - (g_t - 1)$ lies in the previous chunk, i.e., $\pi_t(i - (g_t - 1)) = \pi_t(i) - 1$. Therefore, their shifted updates can be formed using already-available summaries from previous chunks. In contrast, for the unique position with $\rho_t(i) = g_t - 1$ in each chunk, the source index satisfies $\pi_t(i - (g_t - 1)) = \pi_t(i)$, meaning the shifted update can depend on the current chunk summary. Consequently, at chunk size g_t , there is exactly one position per chunk that is responsible for triggering the new low-resolution computation. This corresponds to the unique “chunk trigger” branch in Algorithm 2 (i.e., the $\rho = g_t - 1$ condition), which updates the cached latent $\hat{\mathbf{z}}_{\text{last}}^{(t)}$ once per chunk.

Example: non-uniform compute under coarse-to-fine schedules ($\omega_t = 0$). Consider a coarse-to-fine schedule $[1/8, 1/4, 1/2, 1]$ with chunk sizes $g \in \{8, 4, 2, 1\}$ and $\omega_t = 0$ for all iterations. Then the triggering sets are nested residue classes:

$$\begin{aligned}g = 2 : \quad i &\equiv 1 \pmod{2}, \\ g = 4 : \quad i &\equiv 3 \pmod{4}, \\ g = 8 : \quad i &\equiv 7 \pmod{8},\end{aligned}$$

while $g = 1$ triggers at every position. Thus, some tokens (e.g., $i = 7$) can trigger multiple resolutions simultaneously, whereas others (e.g., $i = 0$) trigger only the full-resolution iteration, yielding non-uniform per-token compute.

How a half-chunk offset changes the triggering pattern. With a non-zero offset, the triggering positions shift to a different residue class. Since $\rho_t(i) = (i + \omega_t) \bmod g_t$, the trigger condition $\rho_t(i) = g_t - 1$ is equivalent to

$$i \equiv g_t - 1 - \omega_t \pmod{g_t}. \quad (29)$$

Choosing $\omega_t = \lfloor g_t/2 \rfloor$ shifts the triggering positions by half a chunk. Across multiple resolutions (especially when g_t are powers of two), this tends

to interleave triggering positions across resolutions, making the per-token compute allocation uniform in practice. Algorithm 2 implements this effect by computing (π, ρ) from $(i + \omega_t)$, so changing ω_t directly shifts the set of token positions that satisfy the trigger condition.

Remark. Chunk offsets only change how tokens are grouped within each iteration. They do not affect model parameters, and strict causality is still enforced by the same right-shift operator in Eq. (20).

E Inference-Time Parallelism in the Parallel (no-overlap) Regime

This appendix discusses an inference-time property that arises when using a causal right-shift size $s_t \geq g_t$ at iteration t , i.e., the *no-overlap / parallel regime* in Proposition 1. While our main experiments set $s_t = g_t - 1$ as default, the no-overlap setting reveals an additional form of decoding parallelism that can be exploited in implementations.

Chunk-level time lag induced by $s_t \geq g_t$. Consider iteration t with chunk size g_t and shifted update $\tilde{\mathbf{u}}^{(t)}[i] = \mathbf{u}^{(t)}[i - s_t]$ for $i \geq s_t$. From Proposition 1(3), when $s_t \geq g_t$, for any position i that lies in chunk j , the source index $i - s_t$ necessarily lies in a chunk strictly earlier than j . Equivalently, the update written into the current chunk cannot depend on any token within the current chunk. We can summarize this as a **chunk-level time lag**:

$$\begin{aligned} \tilde{\mathbf{u}}^{(t)}[i] \text{ depends only on } \\ \{\mathbf{h}_k^{(t)} : \pi_t(k) \leq \pi_t(i) - 1\}, \quad (30) \\ \text{when } s_t \geq g_t. \end{aligned}$$

where $\pi_t(i) = \lfloor (i + \omega_t) / g_t \rfloor$ is the (offset) chunk index map defined in Eq. (12). In particular, the update written into chunk $\pi_t(i)$ cannot depend on any token inside the same chunk, i.e., the multi-resolution recursion is always at least one chunk behind the full-resolution stream.

Decoding implication: overlapping full-resolution for current token with multi-resolution for future tokens. In autoregressive decoding, next-token logits at time step i are computed from the model state at position i . Under the lag property in Eq. (30), the multi-resolution branch (with $s_t \geq g_t$) does not require the newly arrived token(s) of the current chunk in order to produce updates for positions in that chunk;

instead it only depends on already-finalized earlier chunks. This opens a pipeline opportunity:

- **Critical path (current token):** compute the full-resolution pathway needed to produce logits for position i (including pre-/post-loop and any iterations operating at $g = 1$).
- **Background work (future tokens):** concurrently run the low-resolution recursion (downscale $\rightarrow$ shared core $\rightarrow$ upscale) on earlier-chunk summaries to precompute $\hat{\mathbf{z}}^{(t)}$ (and/or $\tilde{\mathbf{u}}^{(t)}$) that will be consumed by future positions.

Concretely, when $s_t \geq g_t$, the low-resolution latent $\hat{\mathbf{z}}_j^{(t)}$ for chunk j only depends on tokens within that chunk and earlier ones. Once chunk j is finalized (i.e., all its fine-grained tokens have been processed at full resolution), its downscaling and core pass can be launched immediately and asynchronously, and the resulting $\hat{\mathbf{z}}_j^{(t)}$ (or the corresponding shifted updates $\tilde{\mathbf{u}}^{(t)}$) can be cached. Later tokens that fall into chunk j or subsequent chunks then only need to read these precomputed chunk-level latents on their own critical path. Thus, the bulk of the multi-resolution computation is moved off the per-token latency path and amortized over time. In other words, the extra test-time compute introduced by multi-resolution recursion can be scheduled *off the per-token latency-critical path*, improving hardware utilization while preserving strict causality.

Connection to cross-loop parallelism in Parallel Loop Transformer. The benefit above is closely related in spirit to cross-loop parallelism (CLP) proposed in the Parallel Loop Transformer (PLT) (Wu et al., 2025), which overlaps loop- ℓ computation for the current token with loop- $(\ell + 1)$ computation for earlier tokens to collapse sequential loop latency. Here, the no-overlap regime plays an analogous role: the right-shift $s_t \geq g_t$ removes within-chunk dependencies, enabling the multi-resolution recursion at iteration t to be advanced on earlier chunks while the model processes the current token at full resolution.

We view this as a promising systems optimization direction for SpiralFormer deployments, complementary to the architectural gains from multi-resolution recursion itself.

F Implementation Details

Pretraining data. All models are pretrained from scratch on the same deduplicated subset of the Pile (Gao et al., 2020) for one epoch (250B tokens), following the Pythia recipe (Biderman et al., 2023). The comparisons therefore hold token exposure fixed rather than total training FLOPs. We use the original GPT-NeoX tokenizer with a vocabulary size of 50,257.

Model architecture. We use decoder-only GPT-NeoX blocks with RMSNorm and rotary positional embeddings (RoPE). SpiralFormer is instantiated on the Middle-cycle backbone. We denote layer allocations as $N_{\text{pre}} + N_{\text{loop}} \times \{r_0, \dots, r_{T-1}\} + N_{\text{post}}$, where N_{loop} layers are shared across T loop iterations and $\{r_t\}$ is the multi-resolution schedule. Unless stated otherwise, SpiralFormer uses a coarse-to-fine schedule ($r_t = 2r_{t-1}$) with chunk sizes $g_t = \lfloor \frac{1}{r_t} \rfloor$, causal right-shift $s_t = g_t - 1$ (single-token overlap), and half-chunk offset $\omega_t = \lfloor g_t/2 \rfloor$ (Appendix D). Down/Up-scaling use learnable self-aggregation for downsampling and output-dependent allocation for upsampling (§2.3). We evaluate both Anchor and MeSH topologies; for MeSH, we follow Yu et al. (2025) and treat the causality-corrected update $\tilde{\mathbf{u}}^{(t)}$ as the produced state written into the MeSH buffer.

Training hyperparameters. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and weight decay 0.01. The learning rate follows a cosine decay schedule with 1% warmup and decays to 10% of the peak learning rate. The global batch size is 512 and the sequence length is 4096. All runs use BF16 mixed precision and FlashAttention-2 (Dao, 2023). Distributed training uses DeepSpeed ZeRO Stage 0.

Evaluation protocol. We report validation perplexity on the Pile validation slice, WikiText, and Lambada (OpenAI and Standard) (Paperno et al., 2016). For downstream evaluation, we use the LM Evaluation Harness (Gao et al., 2024) and evaluate 0-shot and 5-shot accuracy on 9 tasks: Lambada (Paperno et al., 2016) in both OpenAI (LD-O) and Standard (LD-S) versions, HellaSwag (HS) (Zellers et al., 2019), PIQA (Bisk et al., 2020), WinoGrande (WG) (Sakaguchi et al., 2021), ARC-Easy (ARC-E) and ARC-Challenge (ARC-C) (Clark et al., 2018), SciQ (Welbl et al., 2017), and continuation-MMLU (cMMLU) (Hendrycks et al., 2020). We use length-normalized accuracy

Table 3: Additional 2.8B result. Lower perplexity is better; higher average accuracy is better. † denotes MeSH topology.

Model	Pile	Wiki	LD-O	LD-S	0-shot	5-shot
LoopedFormer	6.90	14.18	8.41	16.87	52.49	54.92
LoopedFormer †	6.70	13.60	7.30	11.36	54.71	56.85
SpiralFormer-B†	6.62	13.36	7.11	11.76	55.02	57.28

Table 4: 410M control matching the number of shared-core applications. All models use MeSH topology.

Model / schedule	Loop	Pile	Wiki	LD-O	LD-S	0-shot	5-shot
LoopedFormer, full resolution	2	9.09	21.84	19.63	42.51	44.12	45.56
LoopedFormer, uniform $\frac{1}{2}$	4	9.35	22.93	21.78	49.49	43.45	45.83
SpiralFormer-B, coarse-to-fine	4	9.00	21.48	19.11	39.78	44.31	46.75

for PIQA, HellaSwag, ARC-E, ARC-C, and SciQ and standard accuracy for Lambada, WinoGrande, and cMMLU. All 5-shot results use the default LM Evaluation Harness configuration consistently across models, and each reported pretraining result is based on a single training run. Detailed 0-shot and 5-shot results appear in Table 7.

G Additional Scaling, Controls, and Hardware Efficiency

G.1 Scaling to 2.8B Parameters

We additionally evaluate a completed 2.8B experiment under the same 250B-token, one-epoch training and evaluation protocol. Table 3 extends the controlled trend within the tested Pythia family. This result extends the scaling evidence beyond the main 160M–1.4B suite without establishing generalization to other model families or training regimes.

G.2 Matched Number of Shared-Core Applications

To test whether SpiralFormer benefits merely from using four rather than two shared-core applications, Table 4 adds a 410M MeSH control with four uniform-resolution applications, $\{\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\}$. With the number of applications matched, the coarse-to-fine schedule improves every reported metric. This comparison isolates the organization of computation across resolutions from the number of core passes.

G.3 Matched Hardware Latency

We benchmark the 410M Pythia baseline and parameter-matched SPIRALFORMER-L † using a PyTorch implementation adapted from Hugging

Table 5: Matched 410M latency on one NVIDIA H20 (BF16, batch size 8).

Stage	Pythia	SpiralFormer-L [†]	Change
Prefill (4096 tokens)	0.321 s	0.329 s	+2.5%
Decode (128 tokens)	2.658 s	2.354 s	-11.4%

Table 6: Additional ablations of down/up-scaling operators on the 410M model.

Axis	Variant	Pile	Wiki	LD-O	LD-S	0-shot	5-shot
Main	Learnable	9.00	21.48	<u>19.11</u>	39.78	<u>44.31</u>	46.75
Downscale	Mean pooling	<u>9.01</u>	21.65	20.37	48.45	44.33	46.44
Upscale	Uniform broadcast	9.03	21.73	19.88	44.25	44.22	46.14
Both	Pooling+broadcast	9.03	<u>21.64</u>	20.13	48.86	43.97	45.40

Face GPTNeoX on one NVIDIA H20 in BF16, with batch size 8, a 4096-token prefill followed by 128 decoded tokens, standard attention/KV caching where applicable, and KV-cache sharing. The SpiralFormer timing instantiates the chunk-triggered multi-resolution decoder, so its end-to-end decode time includes periodic recomputation. As shown in Table 5, prefill is 2.5% slower in this setup, whereas decoding is 11.4% faster.

H Operator Variants and Ablations

Mean-pooling down-scaling. As a non-learnable alternative to self-aggregation, we consider mean pooling within each chunk:

$$\mathbf{z}_j^{(t)} = \frac{1}{g_t} \sum_{i \in \mathcal{I}_{t,j}} \mathbf{h}_i^{(t)}. \quad (31)$$

Uniform-broadcast up-scaling. As a non-learnable alternative to output-dependent allocation, we broadcast within each chunk by setting $\beta_{j,\rho}^{(t)} = 1/g_t$, yielding

$$\mathbf{u}_i^{(t)} = \frac{\lambda_t}{g_t} \hat{\mathbf{z}}_{\pi_t(i)}^{(t)}. \quad (32)$$

These ablations show that self-aggregation outperforms mean pooling and output-dependent allocation outperforms uniform broadcast; combining both learnable operators gives the strongest overall result.

I Pseudocode

Here we summarize detailed pseudocode for the multi-resolution recursive architecture discussed in the main paper. Algorithm 1 outlines the training-time implementation of SpiralFormer, where a single loop-shared core is executed under a multi-resolution recursion schedule. Algorithm 2 details the corresponding autoregressive decoding

Algorithm 1 SpiralFormer: Multi-Resolution Recursion

Require: token embeddings $\mathbf{x} \in \mathbb{R}^{L \times d}$; pre-loop f_{pre} ; loop-shared core f_{loop} ; post-loop f_{post} ; schedule $\{r_t\}_{t=0}^{T-1}$; shift sizes $\{s_t\}_{t=0}^{T-1}$; topology update $\mathcal{U}$

- 1: $\mathbf{v} \leftarrow f_{\text{pre}}(\mathbf{x})$
- 2: $(\mathbf{h}^{(0)}, \mathcal{H}^{(0)}) \leftarrow \text{INITTOPO}(\mathbf{x}, \mathbf{v})$
- 3: **for** $t = 0$ to $T - 1$ **do**
- 4: $\mathbf{z}^{(t)} \leftarrow \mathcal{S}_{\downarrow}^{(t)}(\mathbf{h}^{(t)}; r_t)$ ▷ Down-scaling
- 5: $\hat{\mathbf{z}}^{(t)} \leftarrow f_{\text{loop}}(\mathbf{z}^{(t)})$ ▷ Shared operator
- 6: $\mathbf{u}^{(t)} \leftarrow \mathcal{S}_{\uparrow}^{(t)}(\hat{\mathbf{z}}^{(t)}; \mathbf{h}^{(t)}, r_t)$ ▷ Up-scaling
- 7: $\tilde{\mathbf{u}}^{(t)} \leftarrow \text{CAUSALSHIFT}(\mathbf{u}^{(t)}, s_t)$ ▷ Causal shift
- 8: $(\mathbf{h}^{(t+1)}, \mathcal{H}^{(t+1)}) \leftarrow \mathcal{U}(\tilde{\mathbf{u}}^{(t)}, \mathbf{h}^{(t)}, \mathcal{H}^{(t)}; t)$
- 9: **end for**
- 10: $\mathbf{h}^{\text{out}} \leftarrow f_{\text{post}}(\mathbf{h}^{(T)})$
- 11: **return** $\mathbf{h}^{\text{out}}$

procedure and shows how chunk-triggered multi-resolution updates are combined with caching to realize the same recursion pattern at inference time.

J More Discussion

J.1 Additional Analysis: Where Cross-Loop Changes Concentrate

Definition of dynamic heads (per metric). Let (ℓ, h) denote an attention head at layer ℓ and head index h within that layer. For a metric $m \in \{H, \text{LAM}\}$, let $m_{\ell,h}^{(t)}$ be the head’s metric value at loop t (computed as in §4.1, and averaged over 500 Pile validation sequences as described above). We define the cross-loop range

$$\Delta m_{\ell,h} = \max_t m_{\ell,h}^{(t)} - \min_t m_{\ell,h}^{(t)}.$$

A head (ℓ, h) is called *dynamic* under metric m if its $\Delta m_{\ell,h}$ is within the top 40% among all heads in the loop-shared block (ranked separately for each metric).

Observation 3 (Cross-loop variability concentrates in a subset of heads). Figure 5 shows that cross-loop changes are not uniformly distributed across attention heads: large variations in both ΔH and ΔLAM concentrate in specific layer-head regions. Moreover, the highlighted dynamic heads differ between ΔH and ΔLAM , suggesting heterogeneity in how heads respond to the coarse-to-fine resolution schedule. This observation complements

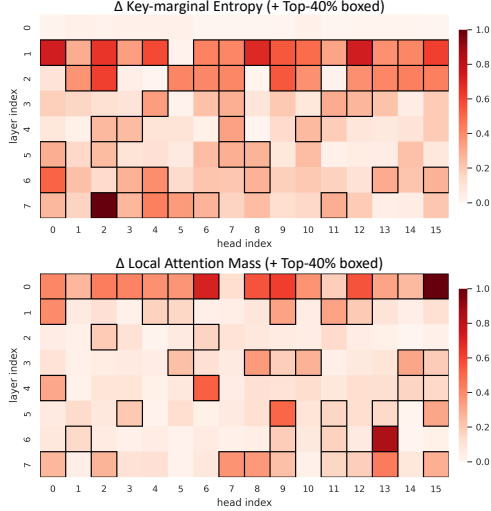


Figure 5: **Head-wise localization of cross-loop variability.** We visualize the cross-loop ranges of (top) key-marginal entropy (ΔH) and (bottom) Local Attention Mass (ΔLAM) as layer-head heatmaps, where each cell corresponds to one attention head indexed by its layer and head indices (ℓ, h) . To compare patterns across layers/heads, we normalize each heatmap to $[0, 1]$ using min-max normalization over all cells of that heatmap: $x' = (x - \min(x)) / (\max(x) - \min(x))$; if $\max(x) - \min(x)$ is numerically negligible (or values are non-finite), we set all entries to 0. Black boxes mark *dynamic heads*, defined as the top 40% heads ranked by cross-loop range for the corresponding metric (defined separately per metric). Statistics are computed on 500 sequences sampled from the Pile validation set: for each loop t and head (ℓ, h) , we first average the metric over the 500 sequences, and then compute cross-loop ranges over t .

the main-text result that attention statistics shift systematically across loops.

J.2 Additional Analysis: LoopedFormer Controls (Distribution & Head-wise)

Distribution-level trends. To contextualize the cross-loop dependency shifts observed in SpiralFormer, we repeat the same attention-based probes (key-marginal entropy and LAM) on a *full-resolution* LoopedFormer baseline, where all loop iterations operate at token resolution (i.e., no multi-resolution compression). We use the same evaluation setup as in §4.2, 500 sequences from the Pile validation set, and report distributions over *dynamic heads* (top 40% by cross-loop range, defined per metric).

Head-wise localization control. Beyond distribution-level trends, we additionally examine *where* cross-loop variability concentrates in the network for a full-resolution LOOPEDFORMER[†] control. We follow the same procedure as in Ap-

pendix J.1: for each head (ℓ, h) , we compute the cross-loop ranges $\Delta H_{\ell, h}$ and $\Delta \text{LAM}_{\ell, h}$ (after averaging metrics over 500 Pile validation sequences at each loop), min-max normalize each heatmap to $[0, 1]$, and box dynamic heads (top 40% by range, per metric). Figure 6 shows that LOOPEDFORMER exhibits a qualitatively different and less structured variability pattern compared to SpiralFormer, supporting that the iteration-wise specialization observed in the main paper is tied to *multi-resolution* recursion rather than looping alone.

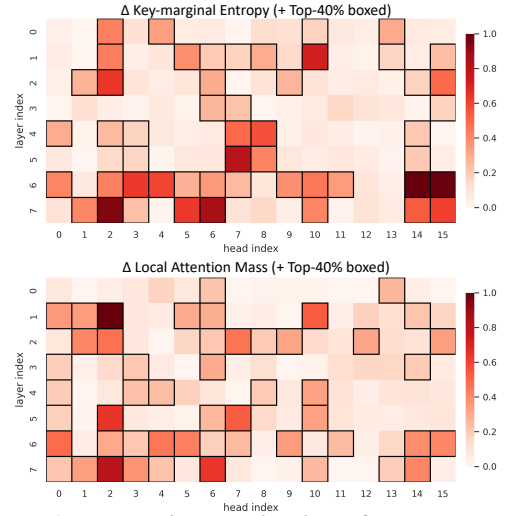


Figure 6: **Head-wise localization of cross-loop variability in full-resolution LOOPEDFORMER[†] (control).** We visualize the cross-loop ranges of (top) key-marginal entropy (ΔH) and (bottom) Local Attention Mass (ΔLAM) as layer-head heatmaps, with per-heatmap min-max normalization to $[0, 1]$. Black boxes mark dynamic heads (top 40% by cross-loop range for each metric). Compared to Figure 5 (SpiralFormer), the control exhibits weaker and less structured head-wise specialization.

J.3 Additional Analysis: Attention Statistics over All Heads (100%)

In §4.2, to emphasize heads that exhibit the clearest iteration-wise adaptation under multi-resolution recursion, we reported attention-statistic trends on *dynamic heads* (top 40% by cross-loop range; defined separately for key-marginal entropy and Local Attention Mass). A potential concern is that restricting to dynamic heads may introduce selection bias and overstate cross-loop specialization. To address this, we provide the same distributional visualizations computed over **all heads (100%)** in the loop-shared block, using the identical evaluation protocol: 500 sequences sampled from the Pile validation set, and per-head metrics averaged

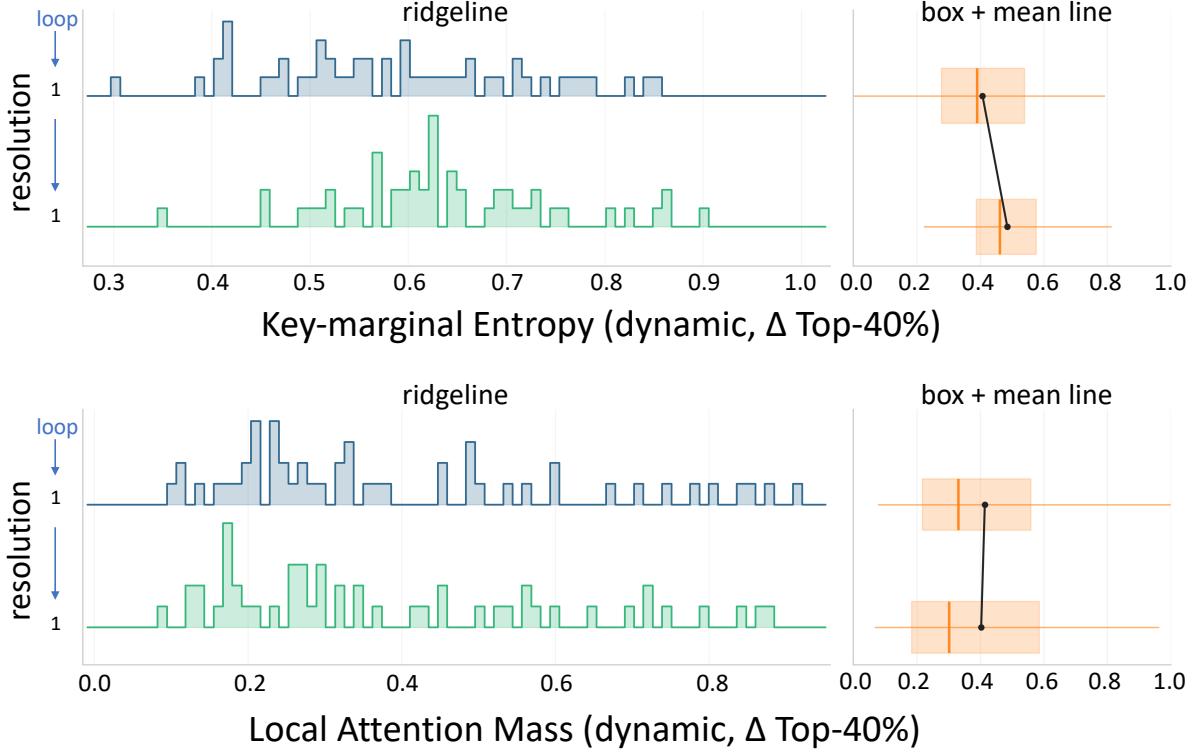


Figure 7: **LoopedFormer control visualization (full-resolution recursion)**. Distributions of key-marginal entropy (top) and Local Attention Mass (bottom) across loop iterations for the 410M full-resolution LOOPEDFORMER[†] baseline, under the same probe protocol as Figure 4. We observe qualitatively different cross-loop behaviors compared to SpiralFormer, indicating that resolution changes play an important role in shaping iteration-wise specialization.

over sequences before forming loop-wise head distributions. Figure 8 shows that the same qualitative cross-loop shifts remain visible without any head filtering: as resolution increases, (i) key-marginal entropy shifts downward (attention becomes more selective), and (ii) Local Attention Mass increases (local refinement strengthens). While the effect sizes are naturally diluted when including heads with weak cross-loop variation, the persistence of these trends over all heads supports the claim that the observed hierarchical dependency pattern is a global property induced by multi-resolution recursion rather than an artifact of selecting a subset of heads.

K Detailed Downstream Results

Table 7 provides the task-specific 0-shot and 5-shot accuracy for all models across the four scaling points. The average (Avg) column corresponds to the task accuracy reported in the main results (Table 1).

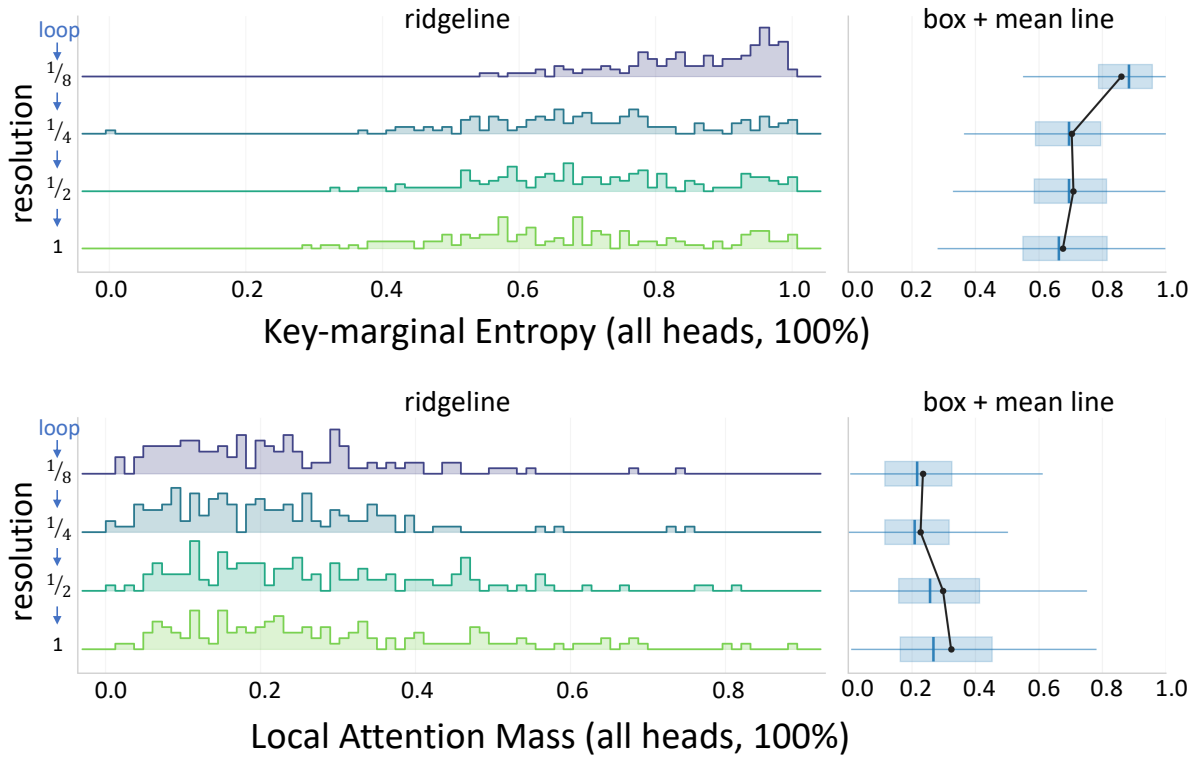


Figure 8: **Cross-loop distribution shifts of attention statistics on all heads (100%).** Distributions of (top) key-marginal entropy and (bottom) Local Attention Mass (LAM) across loop iterations (resolutions) for the SPIRALFORMER-B[†] model at 410M scale. Unlike Figure 4, which focuses on dynamic heads (top 40% by cross-loop range per metric), this figure includes **all** attention heads in the loop-shared block. Statistics are computed by averaging each head’s metric over 500 samples from the Pile validation set, then plotting the head-wise distributions at each loop. The same coarse-to-fine trend remains visible at the population level: entropy decreases and LAM increases as resolution grows.

Table 7: Detailed downstream evaluation results on 9 tasks. For each model size, we report 0-shot and 5-shot accuracy. The average accuracy (‘Avg’) is reported in the final column, corresponding to the values in Table 1. We use notation from the main paper for model configurations, where ‘*’ denotes Anchor and ‘†’ denotes MeSH topology. Best result in each column (within a model size and shot setting) is in **bold**, and second-best is underlined.

MODEL	CONFIG	SHOT	LD-O	LD-S	HS	PQ	WG	ARC-E	ARC-C	SciQ	MMLU	AVG
Pythia-160M												
BASELINE (PYTHIA)	12 Layers	0-SHOT	32.31	23.64	31.14	62.46	50.59	39.56	23.21	70.30	25.69	39.88
		5-SHOT	27.11	24.22	31.38	62.95	50.67	42.21	22.53	78.20	25.55	40.54
LOOPEDFORMER *	2+4×{1,1}+2	0-SHOT	30.04	21.11	30.93	60.39	51.14	38.13	23.81	68.30	25.40	38.81
		5-SHOT	26.16	21.23	31.44	61.15	50.75	41.71	23.04	80.20	25.70	40.15
LOOPEDFORMER †	2+4×{1,1}+2	0-SHOT	31.32	21.48	31.02	60.66	53.43	39.06	22.27	69.70	25.73	39.41
		5-SHOT	26.43	21.00	31.48	60.72	51.93	42.93	23.04	81.90	26.00	40.60
SPIRALFORMER-B †	2+4×{1/8, 1/4, 1/2, 1}+2	0-SHOT	32.93	23.50	31.51	61.32	50.99	39.02	23.98	69.10	25.27	39.73
		5-SHOT	27.07	22.65	31.87	62.19	52.25	42.51	23.46	81.10	26.07	41.02
SPIRALFORMER-L †	4+4×{1/16, 1/8, 1/4, 1/2}+4	0-SHOT	32.78	22.76	32.65	60.72	48.93	40.28	23.04	67.10	25.42	39.30
		5-SHOT	27.52	22.34	32.81	61.70	51.07	44.36	23.72	82.70	26.09	41.37
Pythia-410M												
BASELINE (PYTHIA)	24 Layers	0-SHOT	41.74	29.65	37.65	64.80	51.93	43.60	25.68	73.10	26.68	43.87
		5-SHOT	35.59	28.92	38.01	67.19	50.36	50.08	25.43	85.20	27.03	45.31
LOOPEDFORMER *	4+8×{1,1}+4	0-SHOT	41.45	31.24	36.82	64.09	52.80	43.14	23.72	73.60	26.39	43.70
		5-SHOT	36.56	30.06	37.13	65.62	51.30	49.24	25.43	88.80	26.94	45.68
LOOPEDFORMER †	4+8×{1,1}+4	0-SHOT	41.92	32.33	37.27	64.20	53.83	42.30	25.09	73.50	26.66	44.12
		5-SHOT	36.25	31.71	38.00	65.40	51.22	49.37	24.49	87.00	26.93	45.56
SPIRALFORMER-B *	4+8×{1/8, 1/4, 1/2, 1}+4	0-SHOT	41.03	33.55	37.33	64.09	53.04	42.51	24.23	72.40	26.64	43.87
		5-SHOT	36.13	32.74	38.07	65.72	52.09	50.46	26.11	88.10	27.30	46.30
SPIRALFORMER-B †	4+8×{1/8, 1/4, 1/2, 1}+4	0-SHOT	43.02	33.69	37.98	65.83	52.49	42.72	23.64	72.70	26.72	44.31
		5-SHOT	38.95	35.94	39.00	65.56	52.96	49.50	23.46	88.10	27.29	46.75
SPIRALFORMER-L †	8+8×{1/16, 1/8, 1/4, 1/2}+8	0-SHOT	41.61	32.45	39.63	66.59	53.91	44.91	24.57	74.00	27.10	44.97
		5-SHOT	37.28	32.89	40.00	66.05	51.14	52.65	26.11	89.80	27.63	47.06
Pythia-1B												
BASELINE (PYTHIA)	16 Layers	0-SHOT	46.73	34.02	43.61	66.87	52.01	48.53	26.28	76.60	27.86	46.95
		5-SHOT	40.60	34.41	43.98	68.44	52.33	54.46	28.75	89.90	28.71	49.07
LOOPEDFORMER *	3+5×{1,1}+3	0-SHOT	46.17	34.68	42.62	67.68	53.51	46.80	25.26	75.90	27.99	46.73
		5-SHOT	39.92	32.89	43.26	69.15	53.12	55.05	27.22	90.00	28.83	48.83
LOOPEDFORMER †	3+5×{1,1}+3	0-SHOT	48.40	36.95	44.36	67.03	52.01	46.93	26.54	77.60	27.91	47.53
		5-SHOT	42.62	34.87	44.68	67.95	52.96	55.14	27.22	91.40	28.71	49.51
SPIRALFORMER-B †	3+5×{1/8, 1/4, 1/2, 1}+3	0-SHOT	49.51	37.86	44.85	68.01	55.80	47.48	23.46	78.10	28.22	48.14
		5-SHOT	44.32	37.09	45.71	68.83	53.83	55.18	26.71	91.60	28.96	50.25
Pythia-1.4B												
BASELINE (PYTHIA)	24 Layers	0-SHOT	51.08	39.82	47.74	68.83	55.41	50.04	26.11	77.30	29.18	49.50
		5-SHOT	46.17	39.69	48.01	69.64	54.22	59.22	29.27	91.20	29.95	51.93
BASELINE †	24 Layers	0-SHOT	53.35	42.50	49.52	69.80	53.67	51.77	27.90	78.50	29.05	50.67
		5-SHOT	47.99	40.37	49.74	69.86	56.51	61.28	30.29	93.00	30.02	53.23
LOOPEDFORMER *	4+8×{1,1}+4	0-SHOT	50.75	40.93	47.65	69.75	53.75	48.40	26.62	78.00	28.62	49.39
		5-SHOT	45.99	40.95	47.85	69.37	52.96	56.82	26.79	91.00	29.65	51.27
LOOPEDFORMER †	4+8×{1,1}+4	0-SHOT	53.46	41.84	48.58	69.53	54.85	49.75	27.82	80.30	28.89	50.56
		5-SHOT	49.14	42.69	49.21	69.70	54.78	57.79	29.35	92.70	29.76	52.79
SPIRALFORMER-B †	4+8×{1/8, 1/4, 1/2, 1}+4	0-SHOT	54.03	45.29	49.57	70.35	56.35	50.08	29.10	79.50	29.01	51.48
		5-SHOT	46.61	43.82	50.51	70.08	54.30	59.68	31.49	92.30	30.15	53.22
SPIRALFORMER-L †	8+8×{1/16, 1/8, 1/4, 1/2}+8	0-SHOT	52.86	46.69	50.88	69.64	56.43	50.59	28.75	80.70	29.24	51.75
		5-SHOT	46.32	45.29	51.66	70.51	59.51	60.61	31.91	93.20	30.35	54.37

Algorithm 2 Autoregressive decoding with SpiralFormer (chunk-triggered multi-resolution recursion)

Require: prompt tokens $\{x_0, \dots, x_{n-1}\}$; max new tokens M ; chunk sizes $\{g_t\}_{t=0}^{T-1}$; offsets $\{\omega_t\}$; shifts $\{s_t\}$ (default $s_t = g_t - 1$); modules $f_{\text{pre}}, f_{\text{loop}}, f_{\text{post}}$; routers $\{\mathcal{A}^{(t)}, \mathcal{B}^{(t)}\}$; topology update $\mathcal{U}$ (default: MeSH).

```

1: Init KV caches for  $f_{\text{pre}}, f_{\text{loop}}, f_{\text{post}}$ 
2: for  $t = 0$  to  $T - 1$  do
3:   Init chunk buffer  $\mathbf{B}^{(t)} \in \mathbb{R}^{g_t \times d}$ , current chunk id  $c^{(t)} \leftarrow -1$ 
4:   Init last latent  $\hat{\mathbf{z}}_{\text{last}}^{(t)} \leftarrow \mathbf{0}$ 
5: end for
6: PREFILL( $\{x_0, \dots, x_{n-1}\}$ ) ▷ can be batched over the prompt
7:  $S \leftarrow n$ 
8: for  $m = 1$  to  $M$  do ▷ Generate (sequential)
9:    $\mathbf{h}^{\text{out}} \leftarrow \text{DECODESTEP}(S - 1, x_{S-1})$ 
10:   $x_S \leftarrow \text{SAMPLE}(\text{LMHEAD}(\mathbf{h}^{\text{out}}))$ 
11:   $S \leftarrow S + 1$ 
12: end for
13: function PREFILL( $\{x_0, \dots, x_{n-1}\}$ )
14:   for  $i = 0$  to  $n - 1$  do
15:     DECODESTEP( $i, x_i$ )
16:   end for
17: end function
18: function DECODESTEP( $i, x_i$ )
19:   $\mathbf{v}_i \leftarrow f_{\text{pre}}(\text{Embed}(x_i); \text{KV}_{\text{pre}})$ 
20:   $(\mathbf{h}_i^{(0)}, \mathcal{H}_i^{(0)}) \leftarrow \text{TOPOINITAT}(i, \text{Embed}(x_i), \mathbf{v}_i)$  ▷ e.g., MeSH transitional init
21:  for  $t = 0$  to  $T - 1$  do
22:     $(\mathbf{h}_i^{(t+1)}, \mathcal{H}_i^{(t+1)}) \leftarrow \text{LOOPUPDATEAT}(t, i, \mathbf{h}_i^{(t)}, \mathcal{H}_i^{(t)})$ 
23:  end for
24:   $\mathbf{h}_i^{\text{out}} \leftarrow f_{\text{post}}(\mathbf{h}_i^{(T)}; \text{KV}_{\text{post}})$ 
25:  return  $\mathbf{h}_i^{\text{out}}$ 
26: end function
27: function LOOPUPDATEAT( $t, i, \mathbf{h}_i^{(t)}, \mathcal{H}_i^{(t)}$ )
28:   $(\pi, \rho) \leftarrow \left( \left\lfloor \frac{i + \omega_t}{g_t} \right\rfloor, (i + \omega_t) - g_t \pi \right)$  ▷ chunk index and in-chunk offset, Eq. (12)
29:  if  $\pi \neq c^{(t)}$  then ▷ entered a new chunk
30:     $c^{(t)} \leftarrow \pi; \mathbf{B}^{(t)} \leftarrow \mathbf{0}$ 
31:  end if
32:   $\mathbf{B}^{(t)}[\rho] \leftarrow \mathbf{h}_i^{(t)}$  ▷ write fine-grained token into chunk buffer
33:  if  $\rho = g_t - 1$  then ▷ chunk boundary triggers coarse computation
34:     $\alpha \leftarrow \text{Softmax}_k(\mathcal{A}^{(t)}(\mathbf{B}^{(t)}[k]))$  ▷ self-aggregation weights, Eq. (16)
35:     $\mathbf{z} \leftarrow \sum_{k=0}^{g_t-1} \alpha_k \mathbf{B}^{(t)}[k]$  ▷ Down-scaling  $\mathcal{S}_{\downarrow}^{(t)}$ , Eq. (17)
36:     $\hat{\mathbf{z}}_{\text{last}}^{(t)} \leftarrow f_{\text{loop}}(\mathbf{z}; \text{KV}_{\text{loop}, t})$  ▷ shared loop core on compressed sequence, Eq. (9)
37:  end if
38:   $\tilde{\mathbf{u}} \leftarrow \mathbf{0}$ 
39:  if  $i \geq s_t$  then ▷ causal right-shift, Eq. (20); default  $s_t = g_t - 1$  (§2.3.4)
40:     $j \leftarrow i - s_t$ 
41:     $(\pi_s, \rho_s) \leftarrow \left( \left\lfloor \frac{j + \omega_t}{g_t} \right\rfloor, (j + \omega_t) - g_t \pi_s \right)$ 
42:     $\hat{\mathbf{z}}_s \leftarrow \hat{\mathbf{z}}_{\text{last}}^{(t)}$  ▷ latest finalized chunk latent
43:     $\beta_s \leftarrow \text{Softmax}(\mathcal{B}^{(t)}(\hat{\mathbf{z}}_s)) \in \mathbb{R}^{g_t}$  ▷ output-dependent allocation, Eq. (19)
44:     $\tilde{\mathbf{u}} \leftarrow \lambda_t (\beta_s)_{\rho_s} \cdot \hat{\mathbf{z}}_s$  ▷ Up-scaling  $\mathcal{S}_{\uparrow}^{(t)}$  with gain  $\lambda_t = \sqrt{g_t}$ , Eq. (18)
45:  end if
46:   $(\mathbf{h}_i^{(t+1)}, \mathcal{H}_i^{(t+1)}) \leftarrow \mathcal{U}(\tilde{\mathbf{u}}, \mathbf{h}_i^{(t)}, \mathcal{H}_i^{(t)}; t)$  ▷ default: MeSH (Eq. (6))
47:  return  $(\mathbf{h}_i^{(t+1)}, \mathcal{H}_i^{(t+1)})$ 
48: end function

```
