



Josep Marques¹, Hyunyoung A. Cho¹, Youngeun Kim², Youjia Zhang¹, Aecheon Jung¹, Kyuhong Shim¹, Sungeun Hong^{1†}

¹ Sungkyunkwan University ² Korea University

Research Gap

Previous Work

Layer as Reuse Unit

Sequence sharing

Network as Reuse Unit

Cycle sharing

Problem

Repeating each layer makes representations homogeneous across depth

One set of shared weights cannot specialize to different depths

Ours

Block as Reuse Unit

Grouping adjacent layers into blocks enables local refinement while preserving depth specialization

Motivation

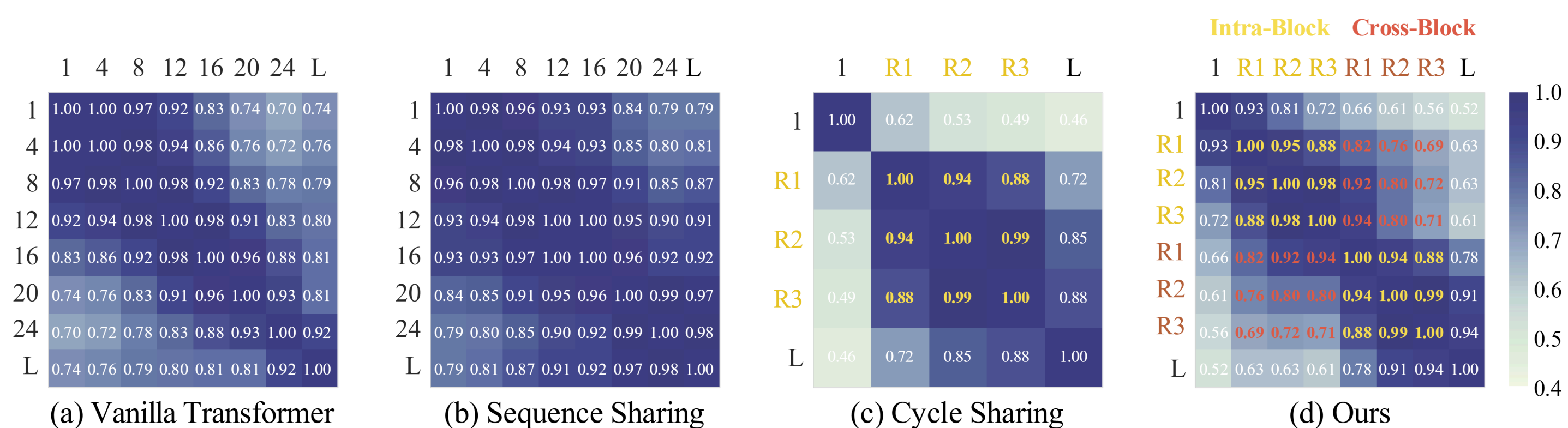
Reuse Topology Is an Overlooked Design Axis

- Sequence and Cycle sharing are two boundary cases: overly fine-grained vs. overly global reuse
- Layers at different depths play specialized roles; lower layers keep local and lexical information, upper layers integrate higher-level features

How should shared parameters be organized across depth?

Observation

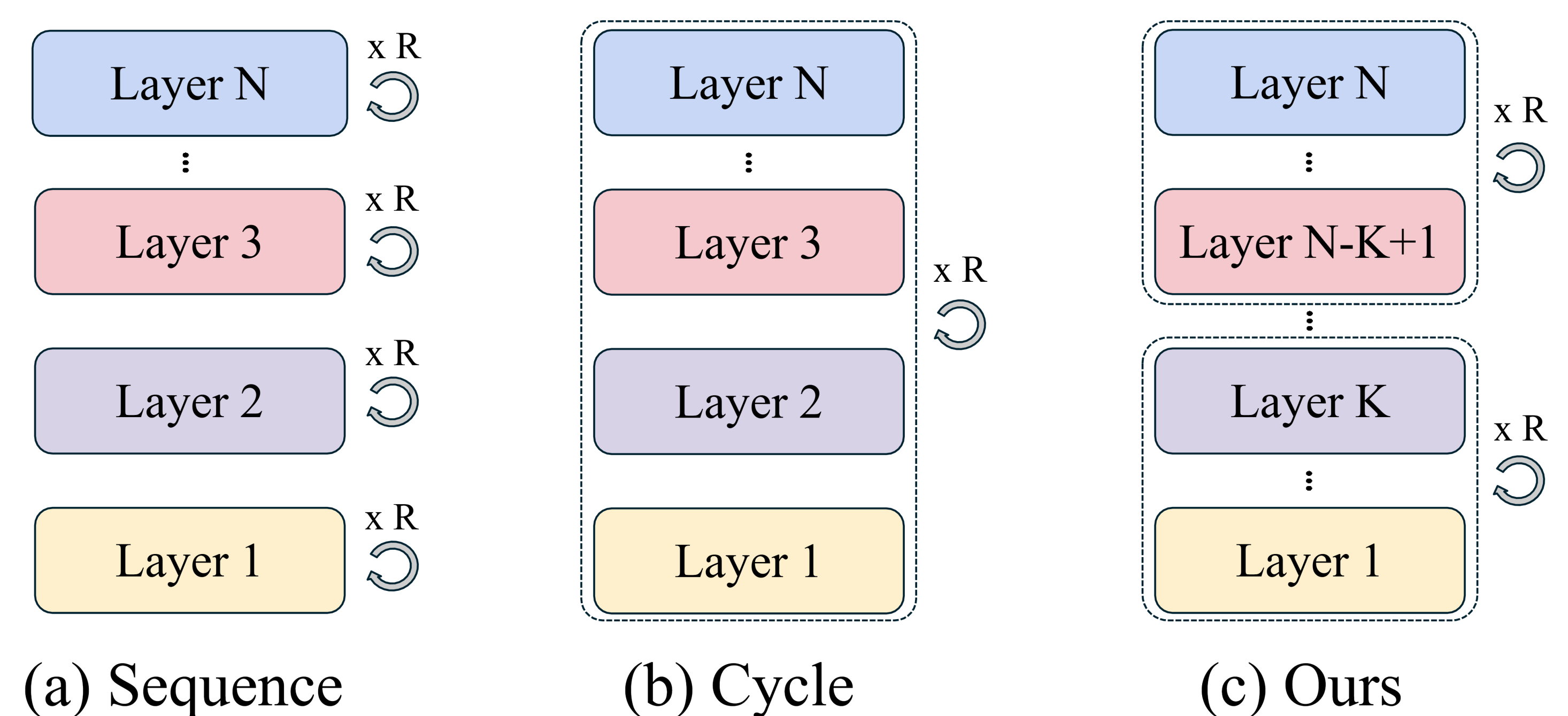
BRT Refines Locally, Specializes Across Depth



Functional specialization patterns support block reuse. BRT shows local refinement with clear differentiation across depth.

Method

Block Recursive Transformers (BRT)



(a) Sequence

(b) Cycle

(c) Ours

$$F_s = B(\cdot; \theta_{s,K}) \circ \dots \circ B(\cdot; \theta_{s,1})$$

$$h_{\text{out}} = B(F_s^R \dots F_1^R(B(h_0; \theta_1)); \theta_L)$$

Unifying Static Sharing Topologies

$S = I$ Global cyclic reuse (Cycle / Middle-Cycle)

$S = P$ Layer-wise repetition (Sequence / Middle-Sequence)

$I < S < P$ BRT: block-level recursive reuse

Experimental Results

Comparison across SmolLM model sizes: 135M, 360M and 1.7B parameters. Results are mean $\pm$ std over three seeds.

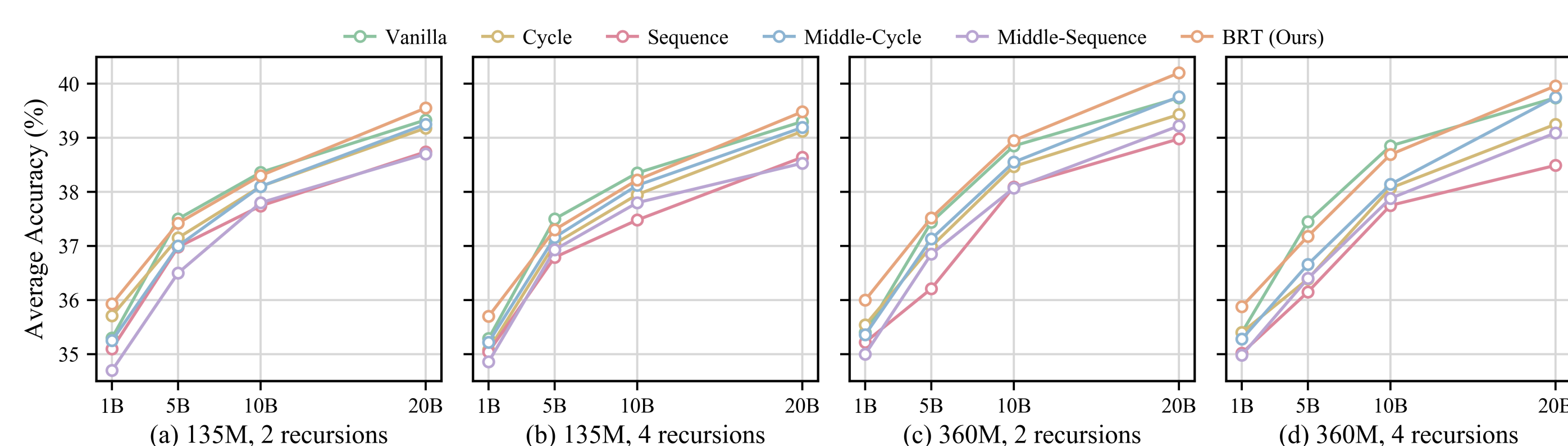
Main Results

Model	Recursion		Pretrain		Few-shot Accuracy $\uparrow$						
	N_r	Share	Param	N_{tok}	ARC	WG	MMLU	HS	PIQA	Avg	
Vanilla	-	-	315M	20B	34.87 $\pm$ 0.31	50.92 $\pm$ 0.41	25.12 $\pm$ 0.17	28.32 $\pm$ 0.16	59.45 $\pm$ 0.32	39.74 $\pm$ 0.14	
Recursive	2	Seq	157M	20B	32.90 $\pm$ 0.40	50.03 $\pm$ 0.42	24.88 $\pm$ 0.34	27.51 $\pm$ 0.22	59.61 $\pm$ 0.31	38.98 $\pm$ 0.30	
		Cyc	157M	20B	34.12 $\pm$ 0.50	50.37 $\pm$ 1.13	24.39 $\pm$ 0.26	28.37 $\pm$ 0.10	59.90 $\pm$ 0.43	39.43 $\pm$ 0.20	
		M-Seq	167M	20B	32.96 $\pm$ 0.11	50.76 $\pm$ 0.34	25.21 $\pm$ 0.41	28.02 $\pm$ 0.34	59.17 $\pm$ 0.50	39.22 $\pm$ 0.23	
		M-Cyc	167M	20B	34.73 $\pm$ 0.21	51.12 $\pm$ 0.63	25.12 $\pm$ 0.19	28.59 $\pm$ 0.12	59.23 $\pm$ 0.10	39.76 $\pm$ 0.12	
		Ours	157M	20B	35.48 $\pm$ 0.35	51.32 $\pm$ 0.64	25.46 $\pm$ 0.22	28.98 $\pm$ 0.07	59.74 $\pm$ 0.50	40.20 $\pm$ 0.23	
	4	Seq	78M	20B	32.03 $\pm$ 0.33	50.83 $\pm$ 1.08	24.38 $\pm$ 0.21	27.34 $\pm$ 0.22	57.88 $\pm$ 0.28	38.49 $\pm$ 0.26	
		Cyc	78M	20B	33.48 $\pm$ 0.23	50.22 $\pm$ 1.04	25.51 $\pm$ 0.18	28.46 $\pm$ 0.10	58.60 $\pm$ 0.35	39.25 $\pm$ 0.13	
		M-Seq	98M	20B	32.56 $\pm$ 0.54	51.79 $\pm$ 0.52	24.90 $\pm$ 0.32	27.62 $\pm$ 0.46	58.61 $\pm$ 0.53	39.09 $\pm$ 0.38	
		M-Cyc	98M	20B	33.67 $\pm$ 0.23	51.89 $\pm$ 0.59	25.18 $\pm$ 0.23	28.47 $\pm$ 0.23	59.55 $\pm$ 0.24	39.75 $\pm$ 0.13	
		Ours	78M	20B	34.43 $\pm$ 0.09	51.91 $\pm$ 0.41	25.32 $\pm$ 0.16	28.73 $\pm$ 0.35	59.43 $\pm$ 0.21	39.96 $\pm$ 0.10	

Ablation on Block Granularity

S	Sharing		Pretrain		Few-shot Accuracy $\uparrow$						
	Untied	Share	Param	N_{tok}	ARC	WG	MMLU	HS	PIQA	Avg	
1	No	BRT	157M	10B	31.35 $\pm$ 0.39	50.75 $\pm$ 0.59	24.93 $\pm$ 0.26	27.33 $\pm$ 0.13	58.01 $\pm$ 0.16	38.47 $\pm$ 0.24	
2	No	BRT	157M	10B	32.10 $\pm$ 0.35	50.31 $\pm$ 0.26	24.74 $\pm$ 0.22	27.70 $\pm$ 0.11	59.05 $\pm$ 0.75	38.77 $\pm$ 0.24	
3	No	BRT	147M	10B	32.08 $\pm$ 0.36	50.31 $\pm$ 0.36	24.53 $\pm$ 0.31	27.54 $\pm$ 0.13	57.92 $\pm$ 0.31	38.47 $\pm$ 0.04	
4	No	BRT	157M	10B	32.07 $\pm$ 0.17	49.40 $\pm$ 0.53	24.35 $\pm$ 0.29	27.49 $\pm$ 0.11	58.48 $\pm$ 0.32	38.34 $\pm$ 0.07	
16	No	BRT	157M	10B	30.42 $\pm$ 0.50	51.22 $\pm$ 1.17	24.71 $\pm$ 0.19	26.79 $\pm$ 0.20	57.20 $\pm$ 0.20	38.09 $\pm$ 0.32	
1	Yes	BRT	167M	10B	31.16 $\pm$ 0.20	51.85 $\pm$ 1.09	24.70 $\pm$ 0.19	27.39 $\pm$ 0.16	57.65 $\pm$ 0.44	38.55 $\pm$ 0.28	
2	Yes	BRT	157M	10B	31.95 $\pm$ 0.45	51.57 $\pm$ 0.16	24.84 $\pm$ 0.21	27.47 $\pm$ 0.08	58.94 $\pm$ 0.53	38.95 $\pm$ 0.07	
3	Yes	BRT	167M	10B	31.92 $\pm$ 0.38	50.76 $\pm$ 0.37	24.96 $\pm$ 0.36	27.78 $\pm$ 0.08	58.44 $\pm$ 0.13	38.76 $\pm$ 0.08	
4	Yes	BRT	138M	10B	31.47 $\pm$ 0.32	50.60 $\pm$ 0.71	24.68 $\pm$ 0.27	27.23 $\pm$ 0.13	58.13 $\pm$ 0.44	38.42 $\pm$ 0.17	
15	Yes	BRT	167M	10B	30.31 $\pm$ 0.58	51.28 $\pm$ 1.16	24.23 $\pm$ 0.17	27.37 $\pm$ 0.04	57.18 $\pm$ 0.68	38.07 $\pm$ 0.29	

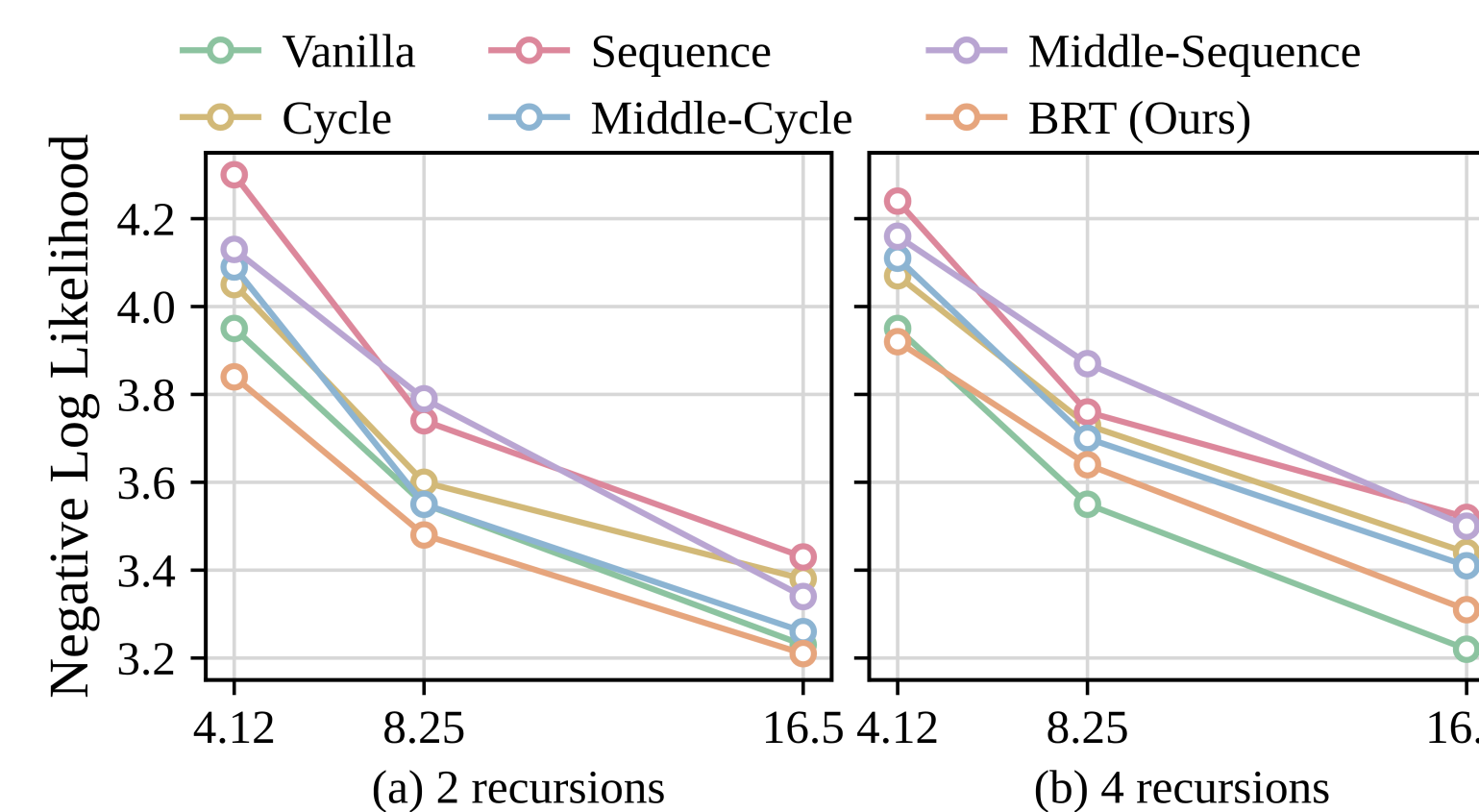
Performance across Token Budgets



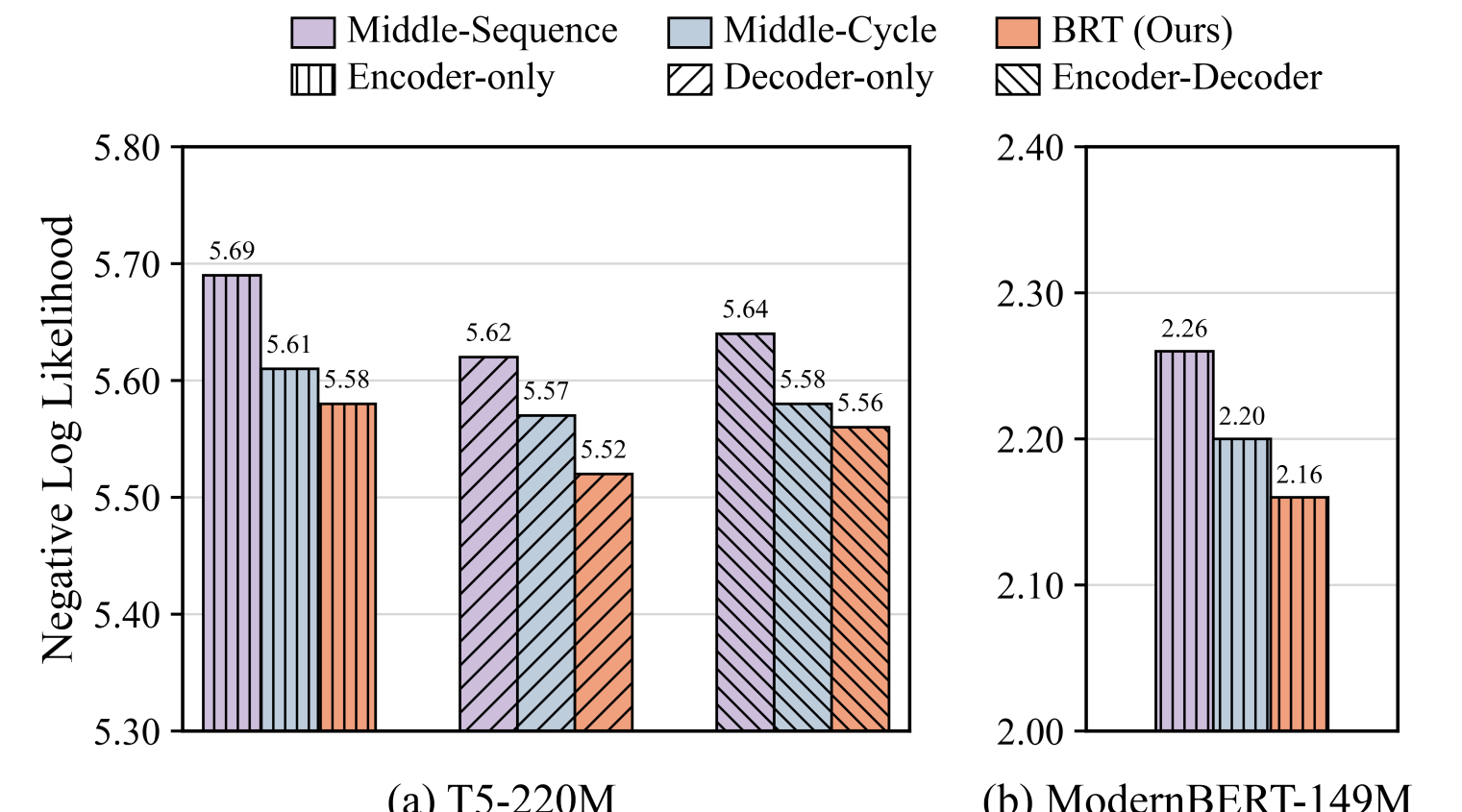
Results at 1.7B Scale

Model	Recursion		Pretrain		Few-shot Accuracy $\uparrow$						
	N_r	Share	Param	N_{tok}	ARC	WG	MMLU	HS	PIQA	Avg	
Vanilla	-	-	1.61B	20B	38.12 $\pm$ 0.22	51.55 $\pm$ 0.72	25.97 $\pm$ 0.26	31.76 $\pm$ 0.17	61.03 $\pm$ 0.16	41.68 $\pm$ 0.25	
Recursive	2	M-Cyc	870M	20B	37.65 $\pm$ 0.26	50.93 $\pm$ 0.89	25.64 $\pm$ 0.21	31.31 $\pm$ 0.32	61.13 $\pm$ 0.12	41.33 $\pm$ 0.20	
	2	Ours	805M	20B	38.25 $\pm$ 0.29	51.30 $\pm$ 0.61	26.10 $\pm$ 0.18	31.69 $\pm$ 0.22	60.82 $\pm$ 0.27	41.63 $\pm$ 0.22	

NLL across Compute Budgets



Generalization across Architectures



Key Takeaways

- Reuse topology is a key design factor for recursive architectures
- Vanilla-level accuracy with 50% of parameters
- Gains hold up to 1.7B model size

Future Work

- Adaptive or learned block partitions
- Input-dependent topology
- Scaling architecture size
- Broader number of tasks
- Multimodal and vision models