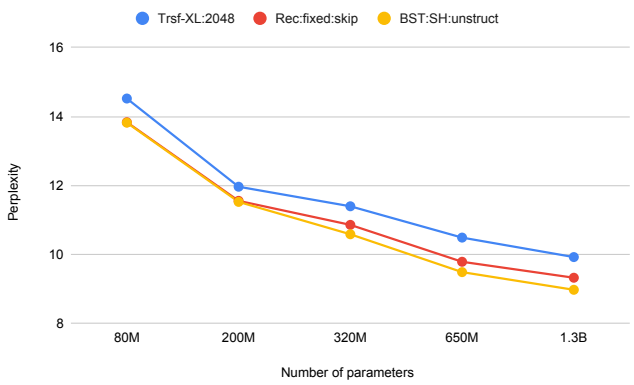

Figure 1: Scaling properties, BST vs BRT vs Trsf-XL on PG-19. Red: 12-layer Block-Recurrent Transformer (`Rec:fixed:skip`) Yellow: 12-layer Block-State Transformer (`BST:SH:unstruct`) Blue: 13-layer Transformer-XL (`Trsf-XL-2048`)

MODEL (∼150M)	PERPLEXITY
`Hyena`	14.6
`Hybrid-H3`	16.2
`BRecT:fixed:skip`	13.5
`BST:SH:S4`	13.4
`BST:SH:unstruct`	**13.2**

Table 1: Perplexity on PG-19. For a fair comparison, we use GPT2 tokenizer with vocabulary size of 50257 as in Hyena and H3. Our `BST` and `BRT` models are 12 layer models with an embedding dimension of 512, 8 attention heads, 128 head dimension and an intermediate layer dimension of 4096 and an SSM dimension of 128.

MODEL	LISTOPTS	TEXT	RETRIEVAL	IMAGE	PATHFINDER	PATH-X	AVG
Transformer	36.37	64.27	57.46	42.44	71.40	✗	53.66
Linear Trans.	16.13	65.90	53.09	42.34	75.30	✗	50.46
Reformer	37.27	56.10	53.40	38.07	68.50	✗	50.56
Performer	18.01	65.40	53.82	42.77	77.05	✗	51.18
BigBird	36.05	64.02	59.29	40.83	74.87	✗	54.17
Mega	**63.14**	**90.43**	91.25	**90.44**	**96.01**	97.98	**88.21**
S4D	60.47	86.18	89.46	88.19	93.06	91.95	84.89
S4	59.60	86.82	90.90	88.65	94.20	96.35	86.09
S5	62.15	89.32	**91.40**	88.00	95.33	**98.58**	87.46
Methods with chunked input sequences							
`BRecT:fixed:skip`	37.29	66.14	58.76	50.41	76.33	75.89	60.80
`Mega-chunk`	58.76	**90.19**	**90.97**	85.80	94.41	93.81	85.66
`BST:SH:S4` (ours)	**61.49**	87.63	90.51	**91.07**	**95.75**	**95.28**	**86.96**

Table 2: Performance on Long-Range Arena (LRA). For fair comparison, we adjust the number of layers and model dimensions on each task so that `BST` and `BRT` have similar number of parameters with `S4` and `Mega-chunk`. `BST:SH:S4` is composed of six BST layers (no BRT layers are interleaved). We use the same standard block length of 512 for `BST` and `BRT`. We use AdamW as our optimizer (Loshchilov and Hutter, 2017) with a warmup for the learning rate, where we start from a value of 1e7 and increase the learning rate linearly up a specified value $\in \{1e-3, 2e3, 4e3\}$ for the first 10% of training. This is followed by cosine annealing for the rest of training down to a value of 1e7. Except for Path-X experiments, we use weight decays $\in \{0.03, 0.05, 0.07\}$ for all parameters except S4 matrices A and B.