

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=qVyeW-grC2k>.

## A TASK EXAMPLES

Task	Description	$x$	$f(x)$
base	Reproduce string unchanged	JUeQoLmUdF	JUeQoLmUdF
rev	Reverse string	JUeQoLmUdF	FdUmLoQeUJ
even	If string length is even, reproduce string	JUeQoLmUdF	JUeQoLmUdF
	If it is odd, reverse string	JUeQoLmUdFa	aFdUmLoQeUJ
odd	Identify any character that appears an odd number of times	fQbfhQMb	hM
sort	Sort characters alphabetically	JUeQoLmUdF	FJLQUUdemo
pal	Add character to make palindrome	CoIBnnBoC	CoIBnnBIoC
dyck	Add character to make valid Dyck string	dfBaaccBeed	dfBaaccBfeed or dffBaaccBeed
case	Switch case of each character	JUeQoLmUdF	juEqOlMuDf
period	Identify repeated substring	mIoHmIoHmIoH	mIoH

Table 2: Description of tasks

## B FURTHER RESULTS

### B.1 EFFECT OF MODEL SIZE

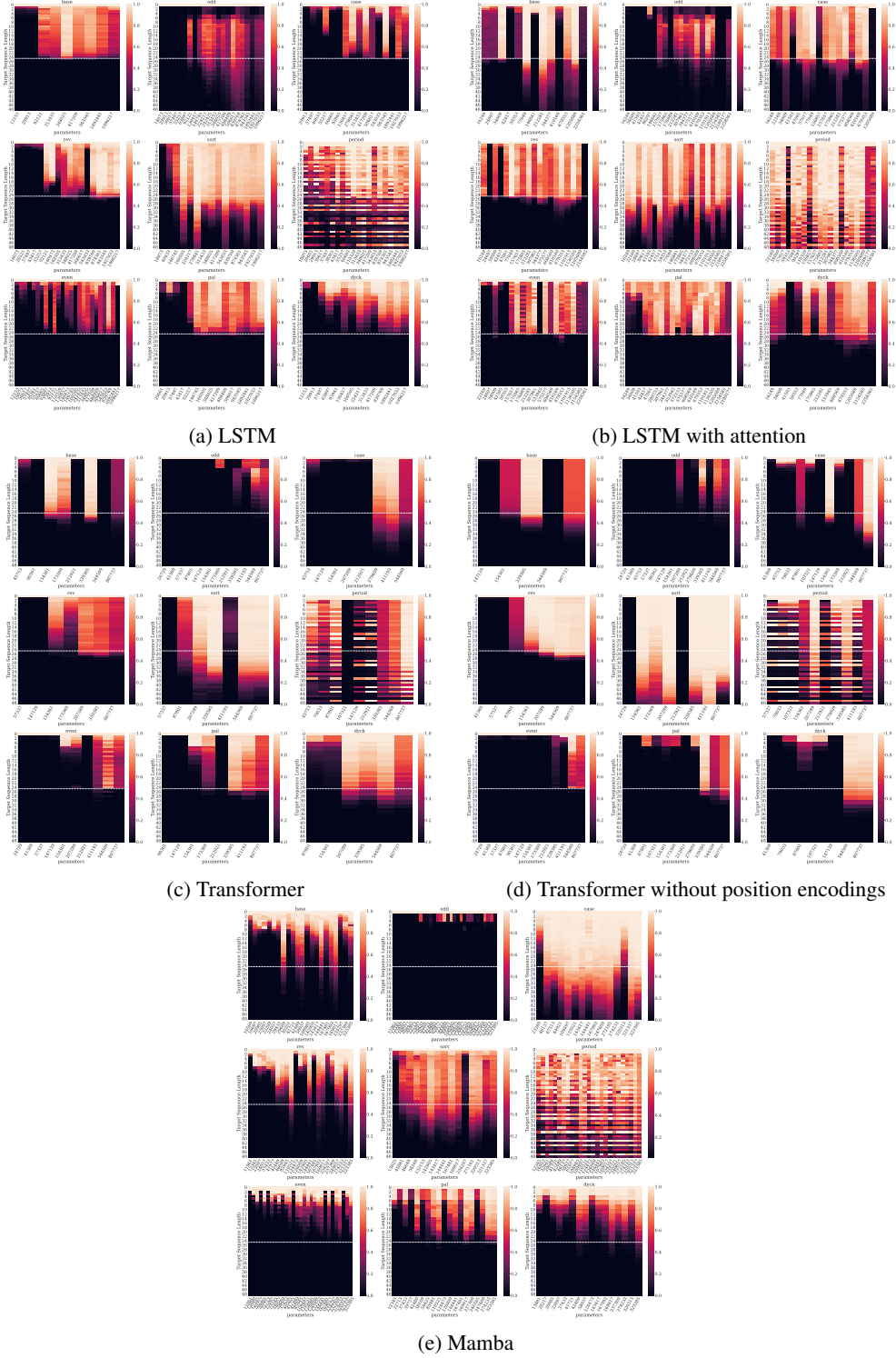


Figure 5: Accuracy broken down by sequence length across models with different numbers of parameters

## B.2 CROSS-TASK COMPARISON

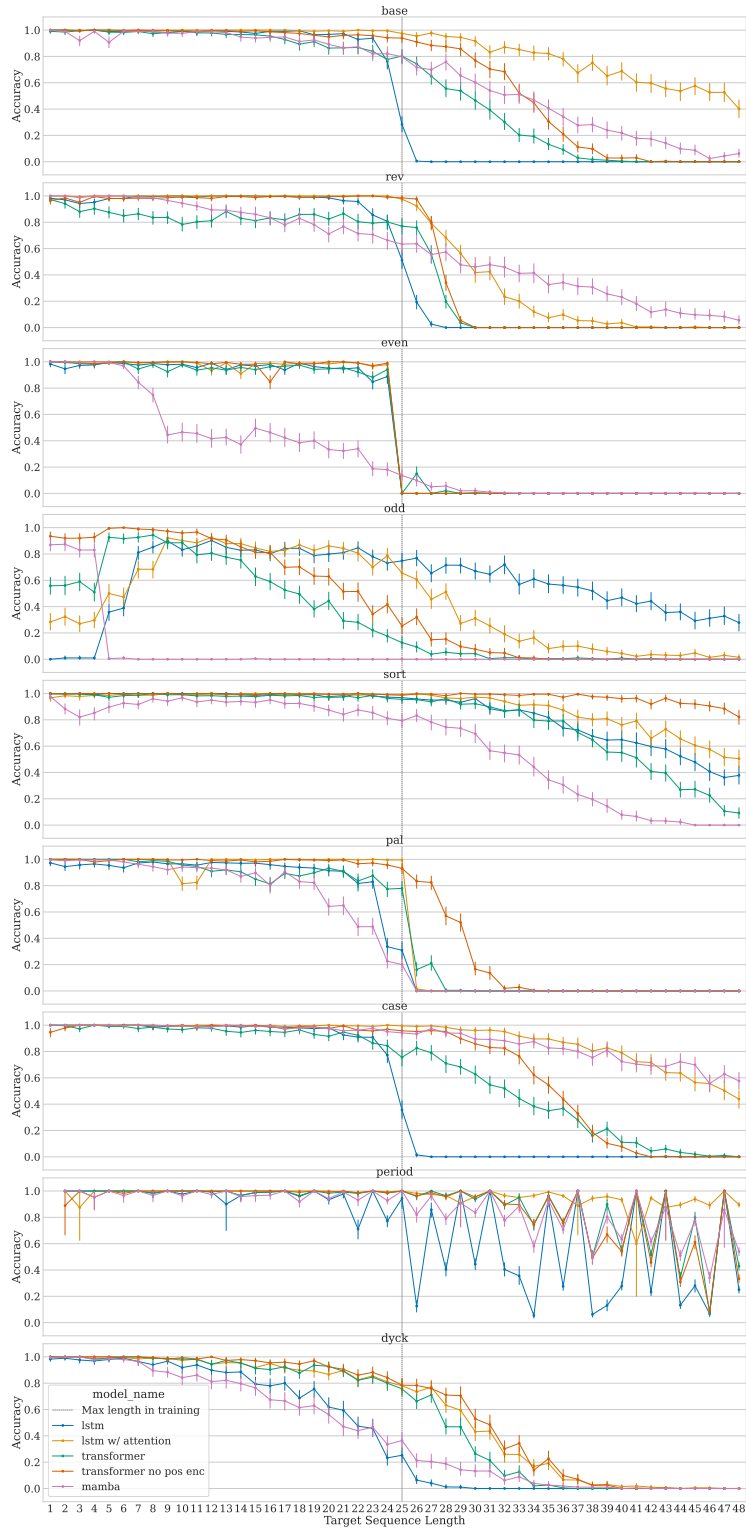


Figure 6: Best model performance on each task shown across target sequence length

## C HYPERPARAMETER DETAILS

		Number of Parameters	Embedding Size	Number of Layers	Hidden Size	Learning Rate	Batch Size
LSTM	base	313,433	32	1	256	0.00045	256
	rev	839,769	32	2	256	0.00010	32
	even	506,617	64	4	128	0.00021	128
	odd	839,769	32	2	256	0.00425	256
	sort	543,033	128	4	128	0.00425	128
	pal	1,927,033	64	4	256	0.00095	128
	dyck	1,892,441	32	4	256	0.00010	64
	case	943,545	128	2	256	0.00021	128
	period	313,433	32	1	256	0.00900	32
LSTM with at- tention	base	212,281	128	1	128	0.00425	64
	rev	679,353	128	1	256	0.00425	256
	even	610,169	64	1	256	0.00010	32
	odd	610,169	64	1	256	0.00425	256
	sort	47,513	64	4	32	0.00425	128
	pal	1,136,505	64	2	256	0.00900	256
	dyck	679,353	128	1	256	0.00201	256
	case	212,281	128	1	128	0.00425	256
	period	212,281	128	1	128	0.00900	64

Table 3: Best hyperparams for LSTMs

		Number of Parameters	Embedding Size	Number of Layers	Feedforward Dimension	Number of Heads	Learning Rate	Batch Size
Transformer	base	807,737	128	4	512	4	0.00010	32
	rev	339,385	64	4	512	4	0.00021	32
	even	544,569	128	4	256	4	0.00010	32
	odd	807,737	128	4	512	4	0.00010	256
	sort	544,569	128	4	256	4	0.00010	128
	pal	339,385	64	4	512	8	0.00045	32
	dyck	544,569	128	4	256	8	0.00045	256
	case	411,193	128	2	512	8	0.00201	256
	period	807,737	128	4	512	8	0.00201	128
Transformer without position encodings	base	807,737	128	4	512	4	0.00021	128
	rev	544,569	128	4	256	8	0.00010	128
	even	807,737	128	4	512	8	0.00021	64
	odd	807,737	128	4	512	8	0.00021	256
	sort	411,193	128	2	512	8	0.00045	128
	pal	807,737	128	4	512	8	0.00021	256
	dyck	544,569	128	4	256	4	0.00010	32
	case	807,737	128	4	512	4	0.00010	32
	period	807,737	128	4	512	4	0.00095	128

Table 4: Best hyperparams for Transformers

		Number of Parameters	Embedding Size	d.state	d.conv	expand	Learning Rate	Batch Size
	base	169,017	128	8	64	2	0.00425	32
	rev	323,385	128	8	64	4	0.00900	256
	even	169,017	128	8	64	2	0.00425	256
	odd	72,121	64	2	16	4	0.00425	256
Mamba	sort	142,905	128	2	32	2	0.00095	32
	pal	323,385	128	8	64	4	0.00201	128
	dyck	323,385	128	8	64	4	0.00095	64
	case	321,337	128	4	64	4	0.00425	256
	period	321,337	128	4	64	4	0.00010	128

Table 5: Best hyperparams for Mamba