Table 1: Baselines

Hyperparameter	XXSmall	XSmall	Small	Medium
Number of layers	3	6	12	24
d_{model}	768	768	768	1024
Attention heads	12	12	12	16
Context size	1024	1024	1024	1024
Dropout	0.1	0.1	0.1	0.1
Attention dropout	0.1	0.1	0.1	0.1
Total batch size	80	80	80	80
Optimizer	Adam	Adam	Adam	Adam
Peak learning rate	$5e-4$	$5e-4$	$5e-4$	$3e-4$
Learning rate decay	cosine	cosine	cosine	cosine
Warmup tokens	409.6M	409.6M	409.6M	409.6M
Decay tokens	32.8G	32.8G	32.8G	40.9G
Max training tokens	32G	32G	32G	32G
Weight decay	0.0	0.0	0.0	0.0
Gradient clipping	1.0	1.0	1.0	1.0

Table 2: Doped models

Hyperparameter	Doped XSmall	Doped Small	Doped Medium
Total number of layers	5	11	23
Number of trainable layers	3	6	12
Number of frozen layers	2	5	11
d_{model}	768	768	1024
Attention heads	12	12	16
Context size	1024	1024	1024
Dropout	0.1	0.1	0.1
Attention dropout	0.1	0.1	0.1
Total batch size	80	80	80
Optimizer	Adam	Adam	Adam
Peak learning rate	$5e-4$	$5e-4$	$3e-4$
Learning rate decay	cosine	cosine	cosine
Warmup tokens	409.6M	409.6M	409.6M
Decay tokens	32.8G	32.8G	40.9G
Max training tokens	32G	32G	32G
Weight decay	0.0	0.0	0.0
Gradient clipping	1.0	1.0	1.0

A Appendix

We include all the hyperparameter and architectural details for the baselines (Table 1), the doped models (Table 2, and the structured models (Table 3). The doped models are built by alternating trainable transformer and frozen MLP layers, so that the first and last layers of the models are always trainable.

Table 3: Structured Models

Hyperparameter	Structued Adaptive FastFood	Structued Block Diagonal
Number of layers	12	12
d_{model}	1024	1024
Attention heads	16	16
Context size	1024	1024
Dropout	0.1	0.1
Attention dropout	0.1	0.1
Total batch size	80	80
Optimizer	Adam	Adam
Peak learning rate	$5e-4$	$5e-4$
Learning rate decay	cosine	cosine
Warmup tokens	409.6M	409.6M
Decay tokens	32.8G	40.9G
Max training tokens	32G	32G
Weight decay	0.0	0.0
Gradient clipping	1.0	1.0