	17M	**32M**	**68M**	**150M**	**400M**	**1B**
Parameter	**(XXS)**	**(XS)**	**(Small)**	**(Base)**	**(Large)**	**(XL)**
Layers	7	10	19	22	28	28
Hidden Size	256	384	512	768	1024	1792
Intermediate Size	384	576	768	1152	2624	3840
Attention Heads	4	6	8	12	16	28
Learning Rate	3.0e-3	3e-3	3e-3	8e-4	5e-4	5e-4
Weight Decay	3.0e-4	3.0e-4	3.0e-4	1.0e-5	1.0e-5	5.0e-5
Warmup Tokens (B)	4	4	3	3	2	2
BS Warmup (B)	125	100	75	50	10	3