

Table 1: Model configurations for different parameter sizes.

Params(B)	Layers	Hidden Dim	Num Heads	L.R.	Batch Size	SeqLen
0.16	12	768	12	3E-04	32	2048
0.41	24	1024	16	3E-04	32	2048
1.45	24	2048	32	3E-04	32	2048