Hyperparameter Guide

Parameter	Base Model	Large Model
📐 d_model	512	1024
🏗️ Layers (N)	6	12+
🎯 Attention Heads (h)	8	16
🔷 d_ff (FFN hidden)	2048	4096
💧 Dropout Rate	0.1	0.1

🔥 **Warmup Steps**

4000 steps (typical)

📦 **Batch Size**

As large as GPU memory allows

⚙️ Adam Optimizer

β_1
0.9

β_2
0.98

ε
10^{-9}

💡

Base model suitable for most tasks
Large model for complex problems