

Masked Language Modeling Process

Input: "The [MASK] sat on the mat"



Step 1: Token + Position Embeddings



Step 2: 12 Transformer Encoder Layers



Step 3: Output for [MASK] position



Step 4: Project to vocabulary (30K)



Step 5: Softmax → $P(\text{cat})=0.73$