Stage 1: Pre-training with Masked Language Model (MLM)

Step 1: Mask 15% of Tokens

Original Sentence:

"*The movie was fantastic*"

Masked Input (15% masked):

The | movie | **[MASK]** | fantastic

BERT must predict masked words from context

Step 2: BERT Processes Input

Token Embeddings

Transformer Layer 1

Transformer Layer 2

Transformer Layer 3

Output Embeddings

Bidirectional Context

Step 3: Predict Masked Token

Top Predictions for [MASK]:

Loss = -log(0.85) = 0.16
(Low loss = good prediction)