

Raw Text
↓
Byte-level-encoding
↓
BPE Merge Rules

Token IDs (integer) [shape: (seq-len)]

Token IDs (token embedding lookup-wte)
Position IDs (0...T-1) (position embedding lookup-wpe)

$x_0 = \text{token-embeddings} + \text{position-embeddings}$
[shape: (seq-len, d-model)]
shapes shown omit B
(seq: (B, seq-len, d-model))

GPT-2
Transformer
Block x N
 $x_i \rightarrow \text{Block}_{i+1} \rightarrow x_{i+1}$



