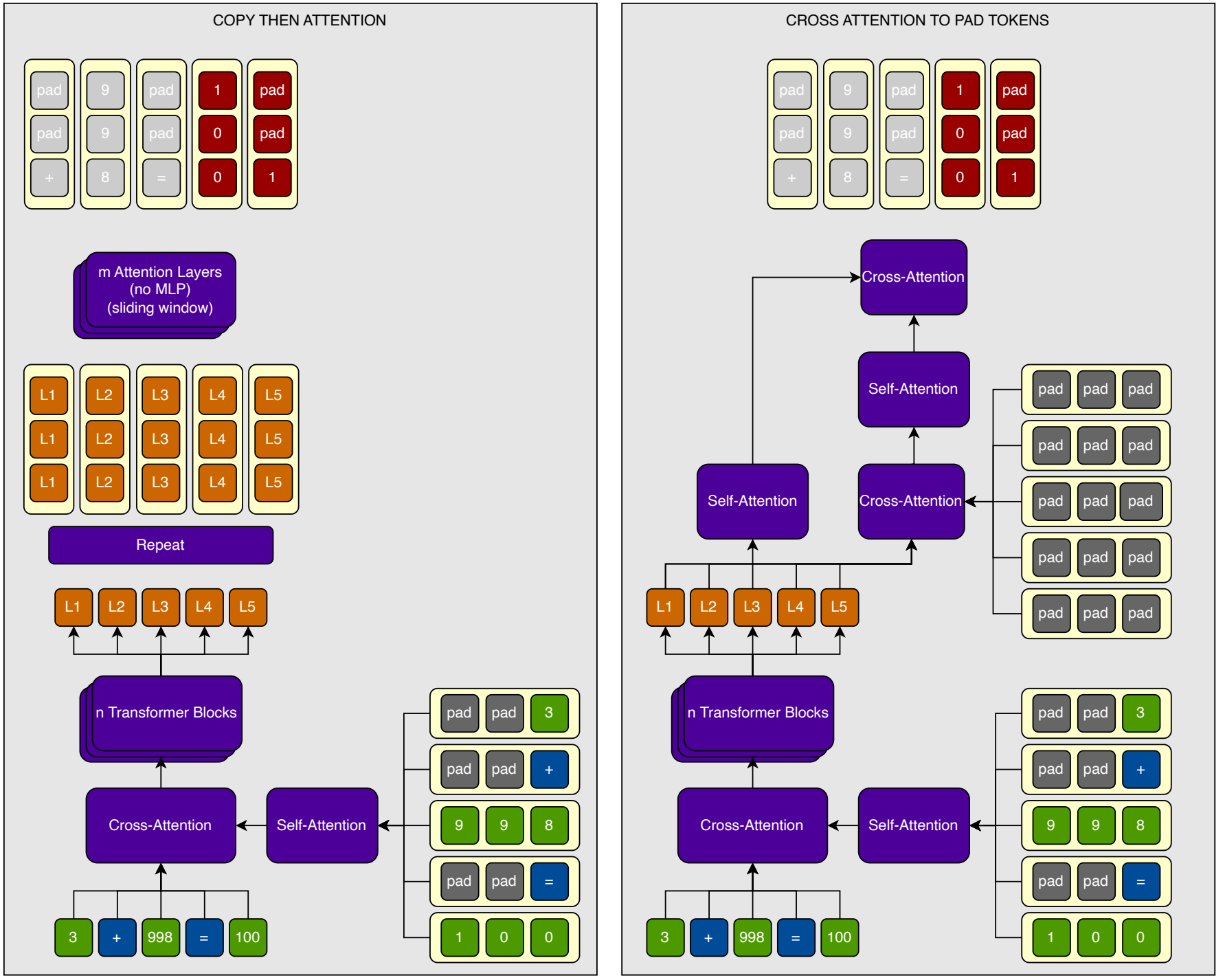
COPY THEN ATTENTION
pad pad + 9 9 8 pad pad = 1 0 0 pad pad 1
m Attention Layers (no MLP) (sliding window)
L1 L1 L1 L2 L2 L2 L3 L3 L3 L4 L4 L4 L5 L5 L5
Repeat
L1 L2 L3 L4 L5
n Transformer Blocks
Cross-Attention
Self-Attention
3 + 998 = 100
pad pad 3
pad pad +
9 9 8
pad pad =
1 0 0
CROSS ATTENTION TO PAD TOKENS
pad pad + 9 9 8 pad pad = 1 0 0 pad pad 1
Cross-Attention
Self-Attention
Self-Attention
Cross-Attention
pad pad pad
pad pad pad
pad pad pad
pad pad pad
pad pad pad
L1 L2 L3 L4 L5
n Transformer Blocks
Cross-Attention
Self-Attention
3 + 998 = 100
pad pad 3
pad pad +
9 9 8
pad pad =
1 0 0