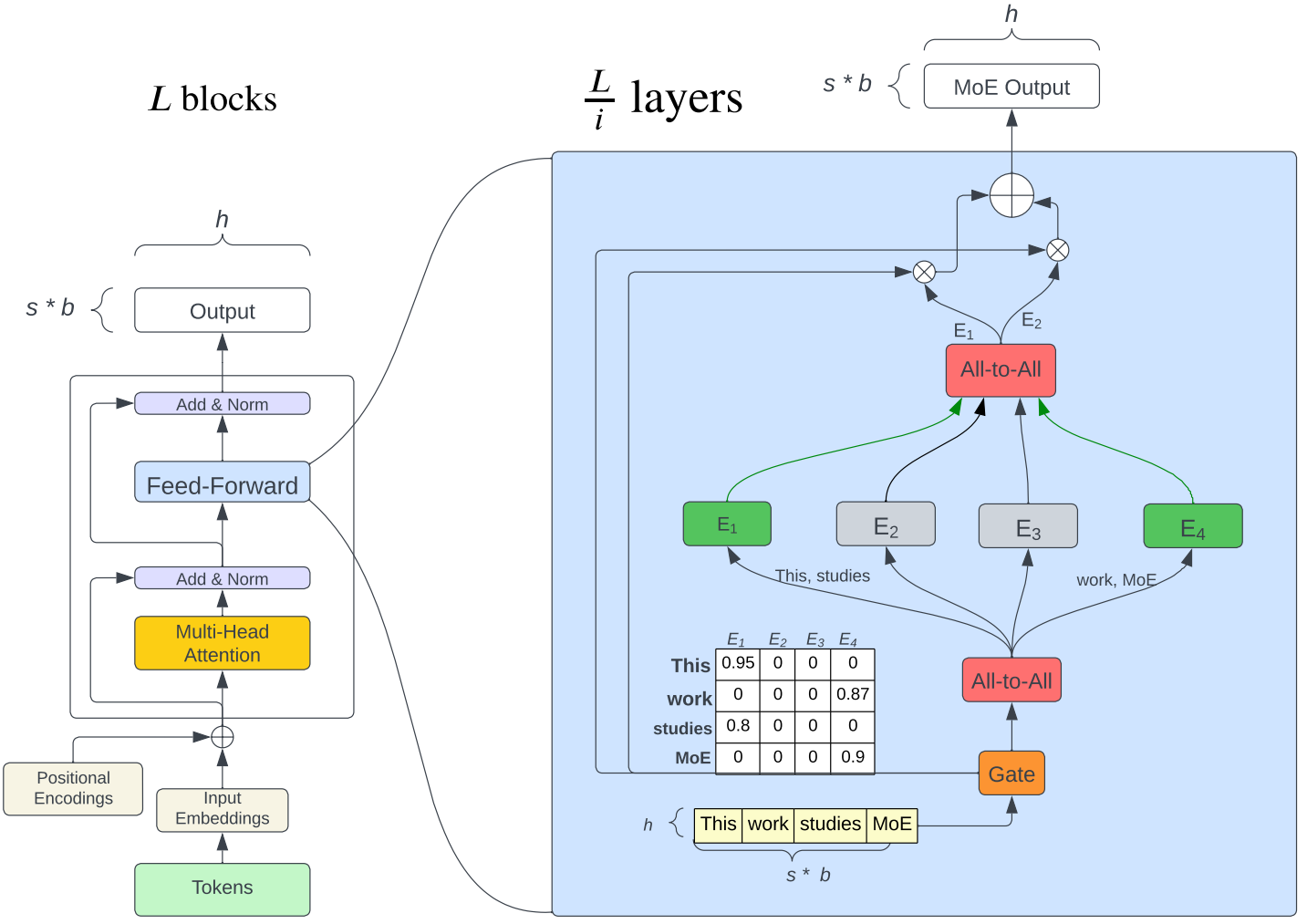
L blocks
L/i layers
h
s * b
Output
Add & Norm
Feed-Forward
Add & Norm
Multi-Head Attention
Positional Encodings
Input Embeddings
Tokens
MoE Output
E1
E2
All-to-All
E1
E2
E3
E4
This, studies
work, MoE
All-to-All
Gate
This
work
studies
MoE
E1 E2 E3 E4
This 0.95 0 0 0
work 0 0 0 0.87
studies 0.8 0 0 0
MoE 0 0 0 0.9