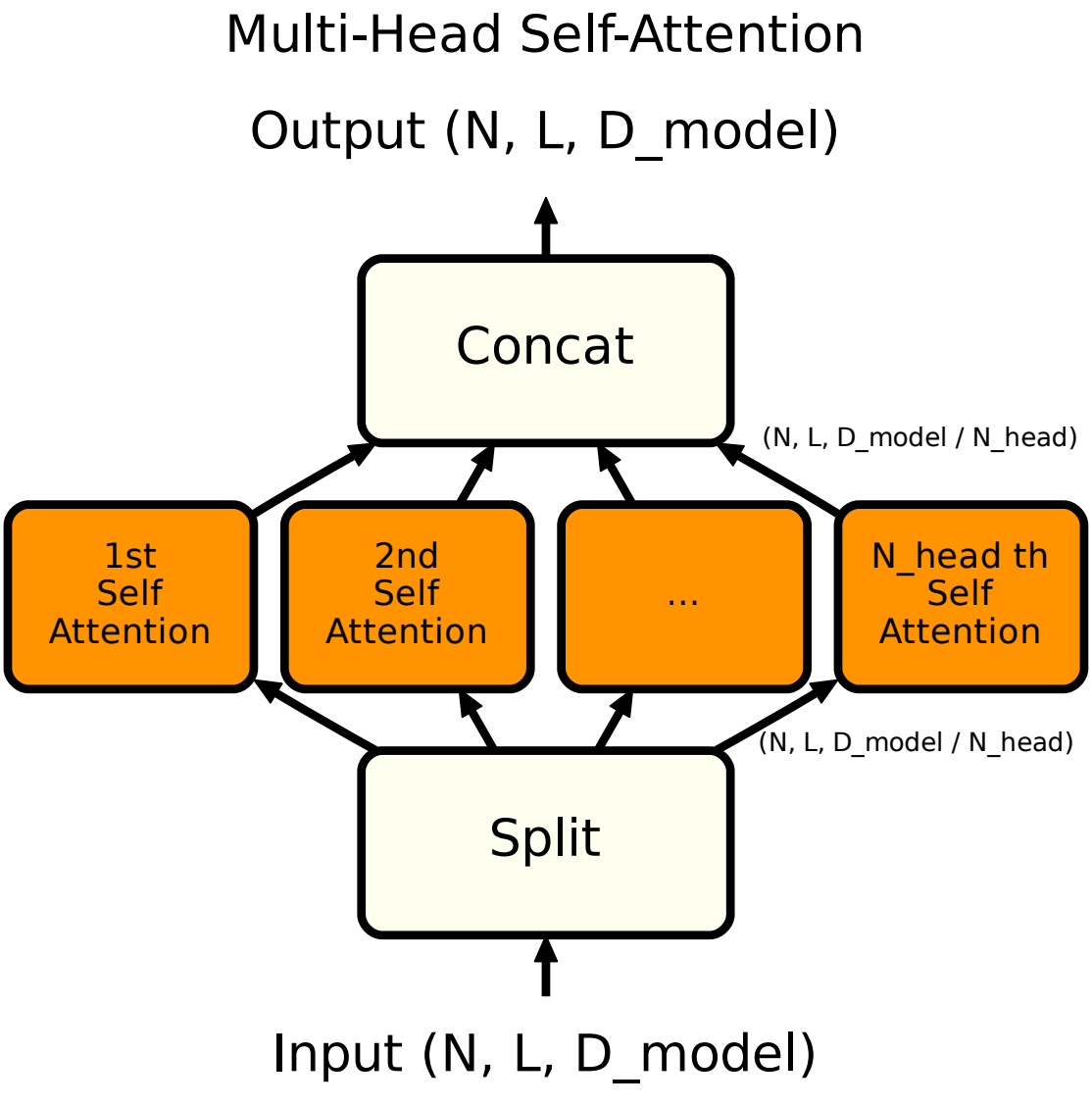
Multi-Head Self-Attention
Output (N, L, D_model)
Concat
(N, L, D_model / N_head)
1st Self Attention
2nd Self Attention
...
N_head th Self Attention
(N, L, D_model / N_head)
Split
Input (N, L, D_model)