deepseek

Multi-head Attention

Head 2

Head 1

Query

Y N

Native Sparse Attention

Y 0

Head 1

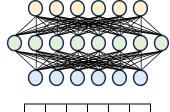
Ν

Tokens Value Key Value Recent Key Value Key Value -5 -4 -3 -5 -2 0 0 0 0 0 Ν Ν -1 -10 9 -2 -7 8 0 3 -4 -10 2 Selected 5 -6 -7 3 0 -7 2 4 -6 -10 9 -1 N 0 -2 -1 N N 3 0 -1 10 5 7 -2 1 2 -2 1 3 2 1 -3 3 -2 -3 N N -3 2 -2 2 2 -2 0 -1 -1 -9 -10 8 -1 3 Compressed



-3

0



-10

-5

7 10 -4

8

-5

-3

0