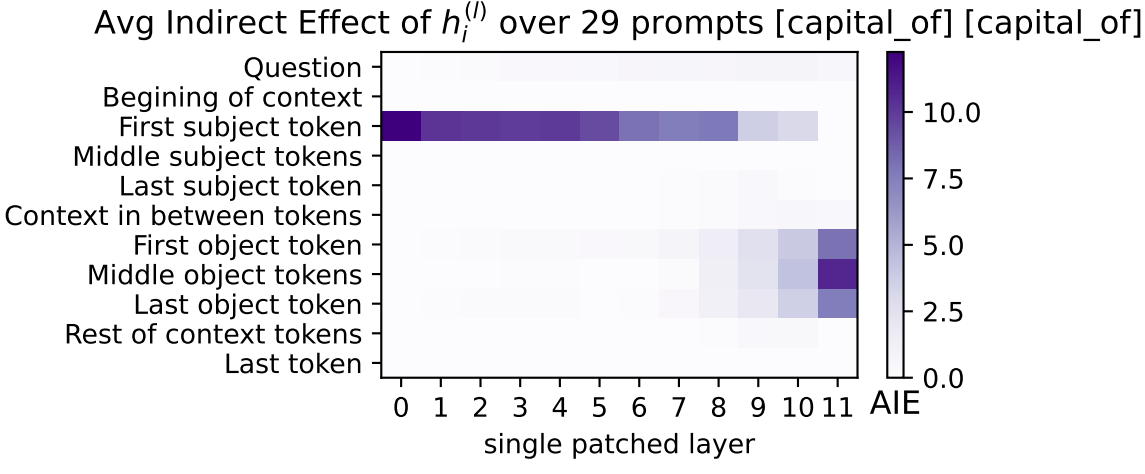
Avg Indirect Effect of $h_i^{(l)}$ over 29 prompts [capital_of] [capital_of]
Question
Begining of context
First subject token
Middle subject tokens
Last subject token
Context in between tokens
First object token
Middle object tokens
Last object token
Rest of context tokens
Last token
0 1 2 3 4 5 6 7 8 9 10 11
single patched layer
10.0
7.5
5.0
2.5
0.0
AIE