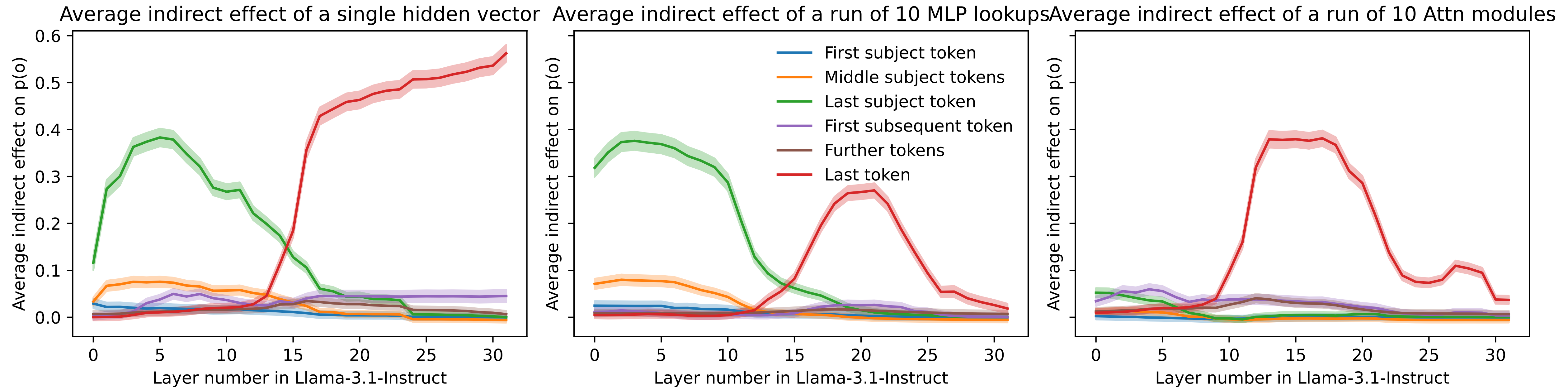
Average indirect effect of a single hidden vector
Average indirect effect of a run of 10 MLP lookups
Average indirect effect of a run of 10 Attn modules
Average indirect effect on p(o)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5
10
15
20
25
30
Layer number in Llama-3.1-Instruct
First subject token
Middle subject tokens
Last subject token
First subsequent token
Further tokens
Last token