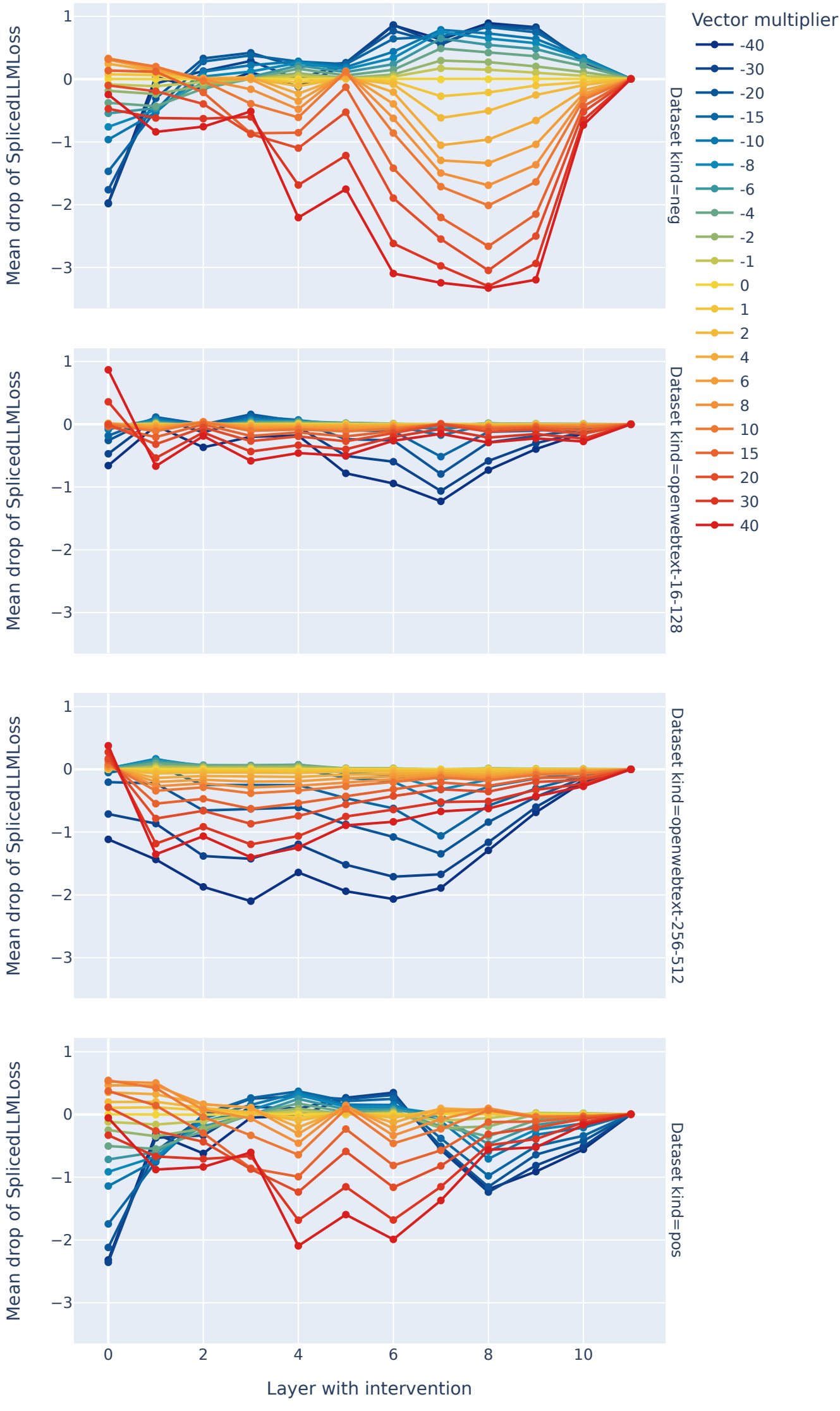
Vector multiplier
-40
-30
-20
-15
-10
-8
-6
-4
-2
-1
0
1
2
4
6
8
10
15
20
30
40
Mean drop of SplicedLLMLoss
Dataset kind=neg
Dataset kind=openwebtext-16-128
Dataset kind=openwebtext-256-512
Dataset kind=pos
1
0
−1
−2
−3
0
2
4
6
8
10
Layer with intervention