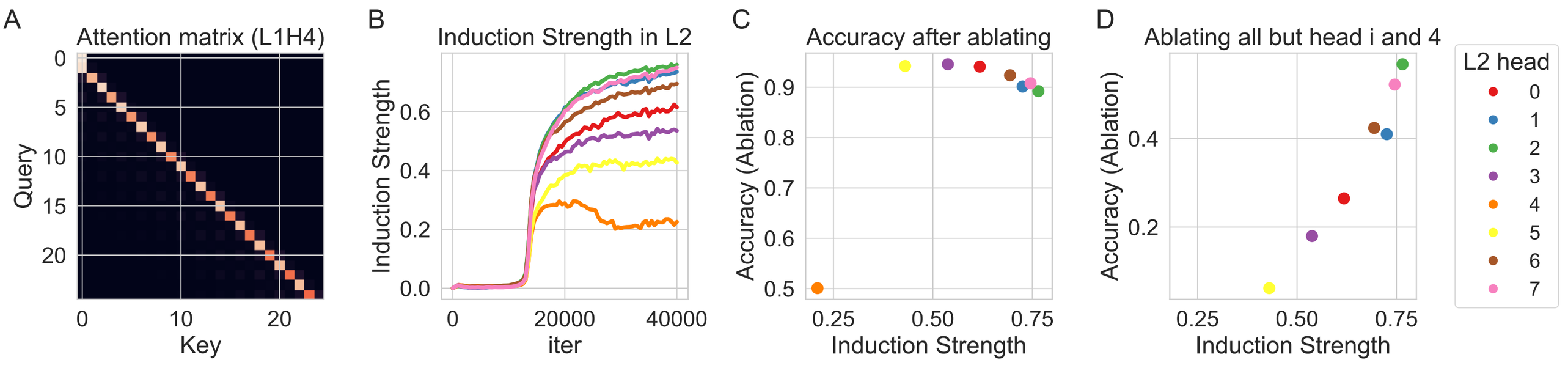
A
Attention matrix (L1H4)
Query
Key
0
5
10
15
20
0
10
20
B
Induction Strength in L2
Induction Strength
iter
0.0
0.2
0.4
0.6
0
20000
40000
C
Accuracy after ablating
Accuracy (Ablation)
Induction Strength
0.5
0.6
0.7
0.8
0.9
0.25
0.50
0.75
D
Ablating all but head i and 4
Accuracy (Ablation)
Induction Strength
0.2
0.4
0.25
0.50
0.75
L2 head
0
1
2
3
4
5
6
7