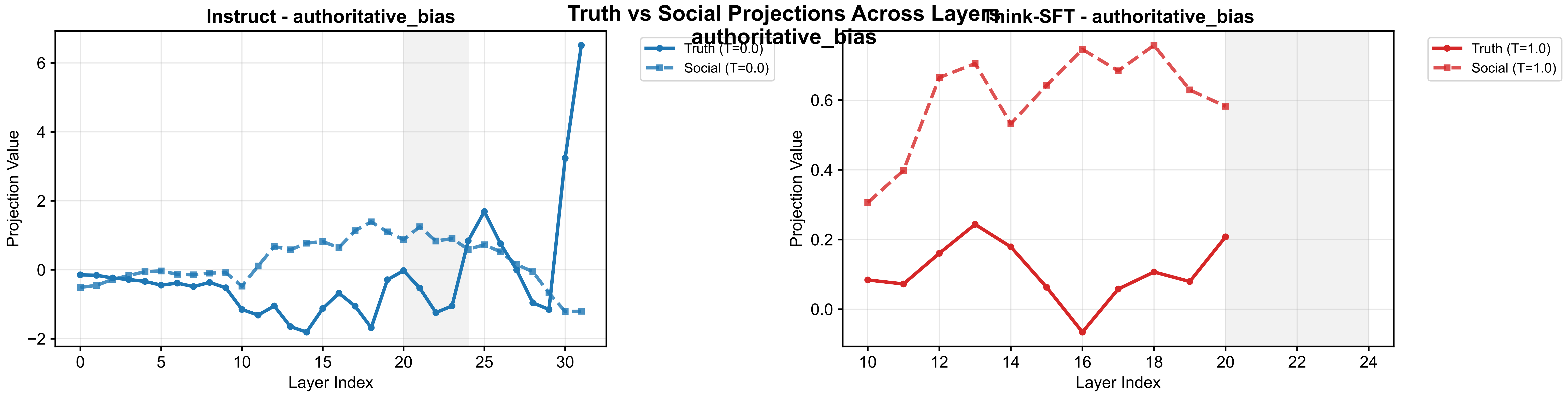
Truth vs Social Projections Across Layers
authoritative_bias
Instruct - authoritative_bias
Think-SFT - authoritative_bias
Projection Value
Layer Index
−2
0
2
4
6
0
5
10
15
20
25
30
Truth (T=0.0)
Social (T=0.0)
0.0
0.2
0.4
0.6
10
12
14
16
18
20
22
24
Truth (T=1.0)
Social (T=1.0)