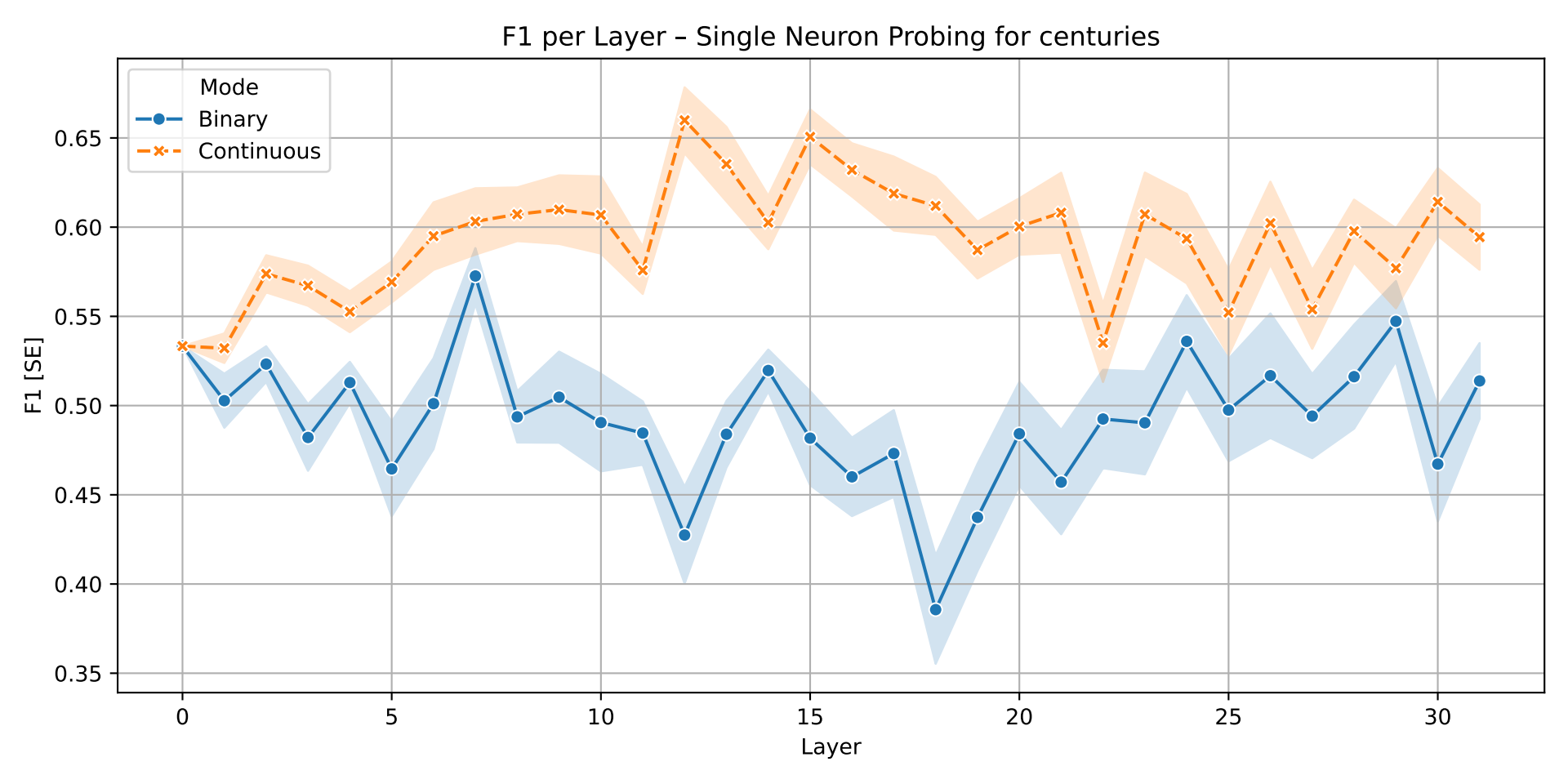
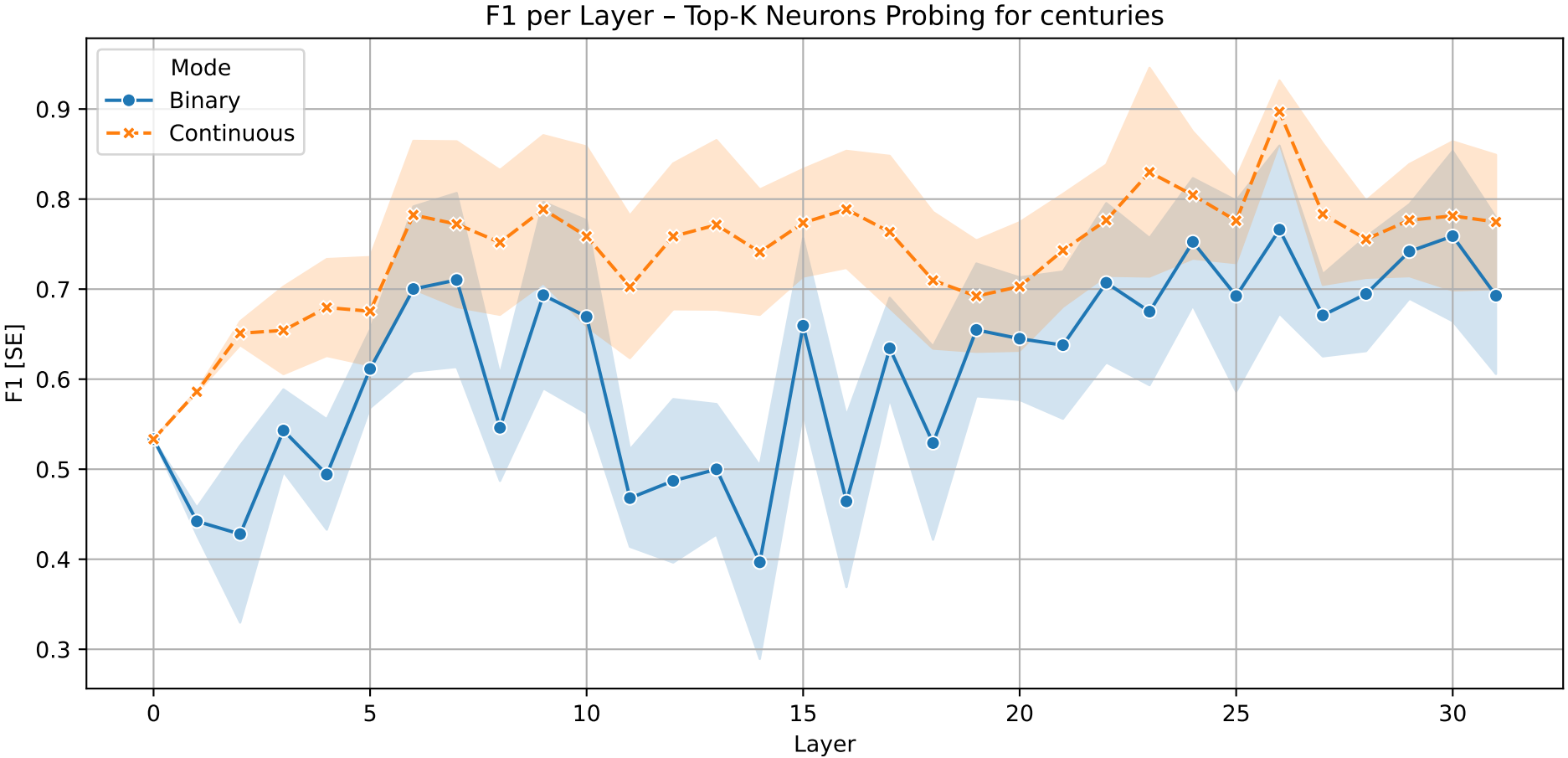
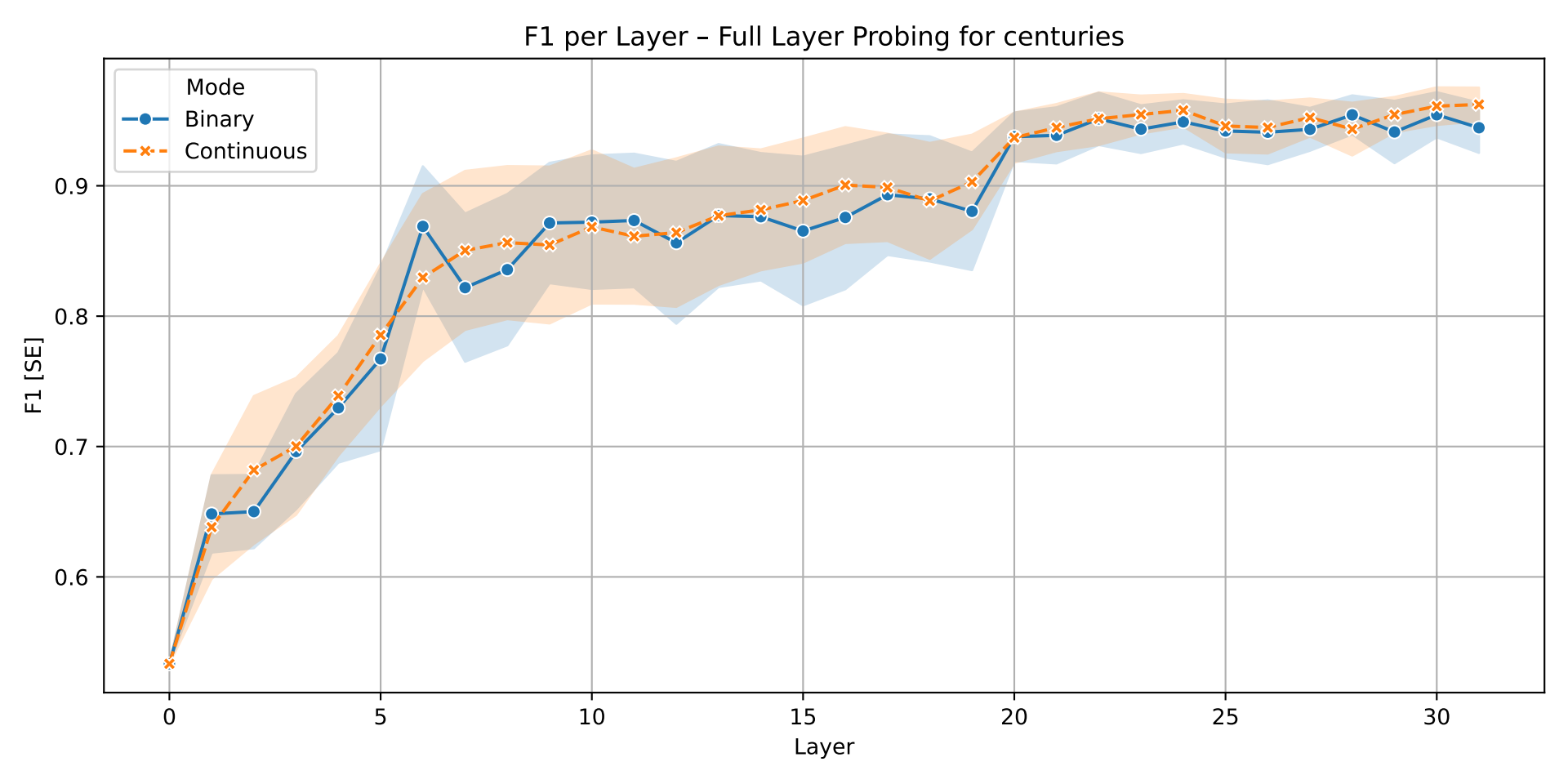
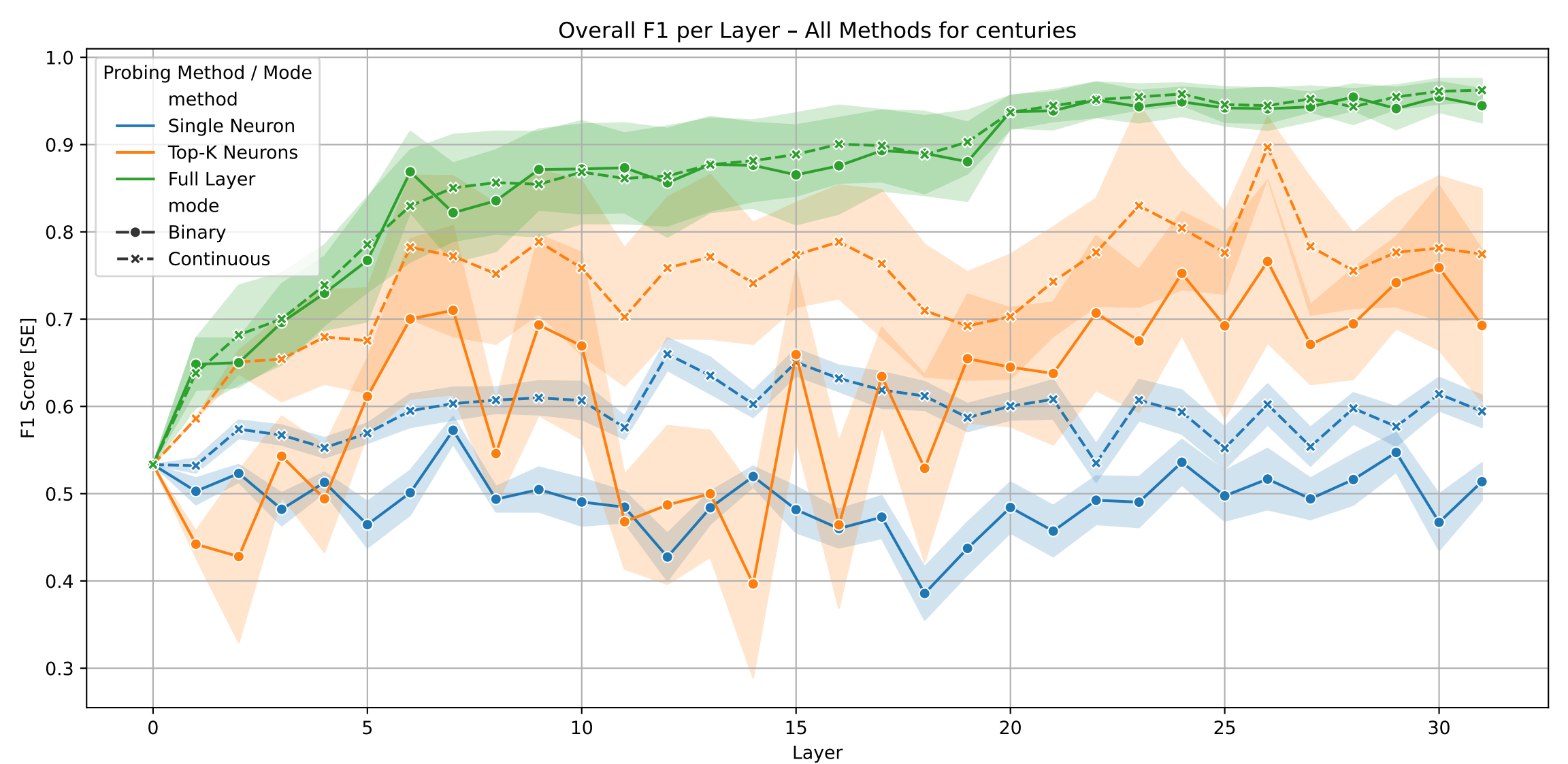

F1 per Layer – Single Neuron Probing for centuries
Mode
Binary
Continuous
F1 [SE]
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0
5
10
15
20
25
30
Layer

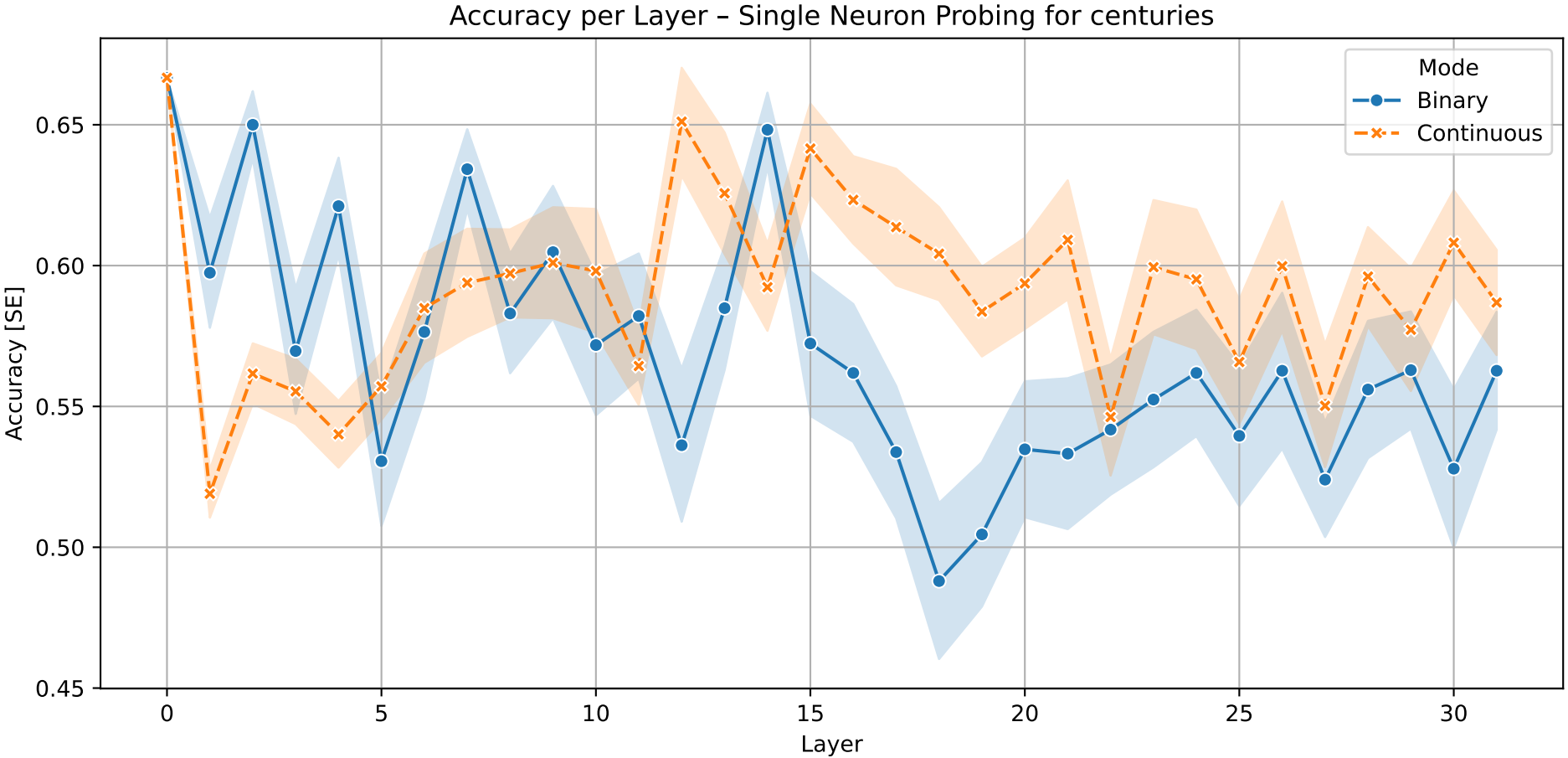

F1 per Layer – Top-K Neurons Probing for centuries
Mode
Binary
Continuous
F1 [SE]
Layer
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
5
10
15
20
25
30

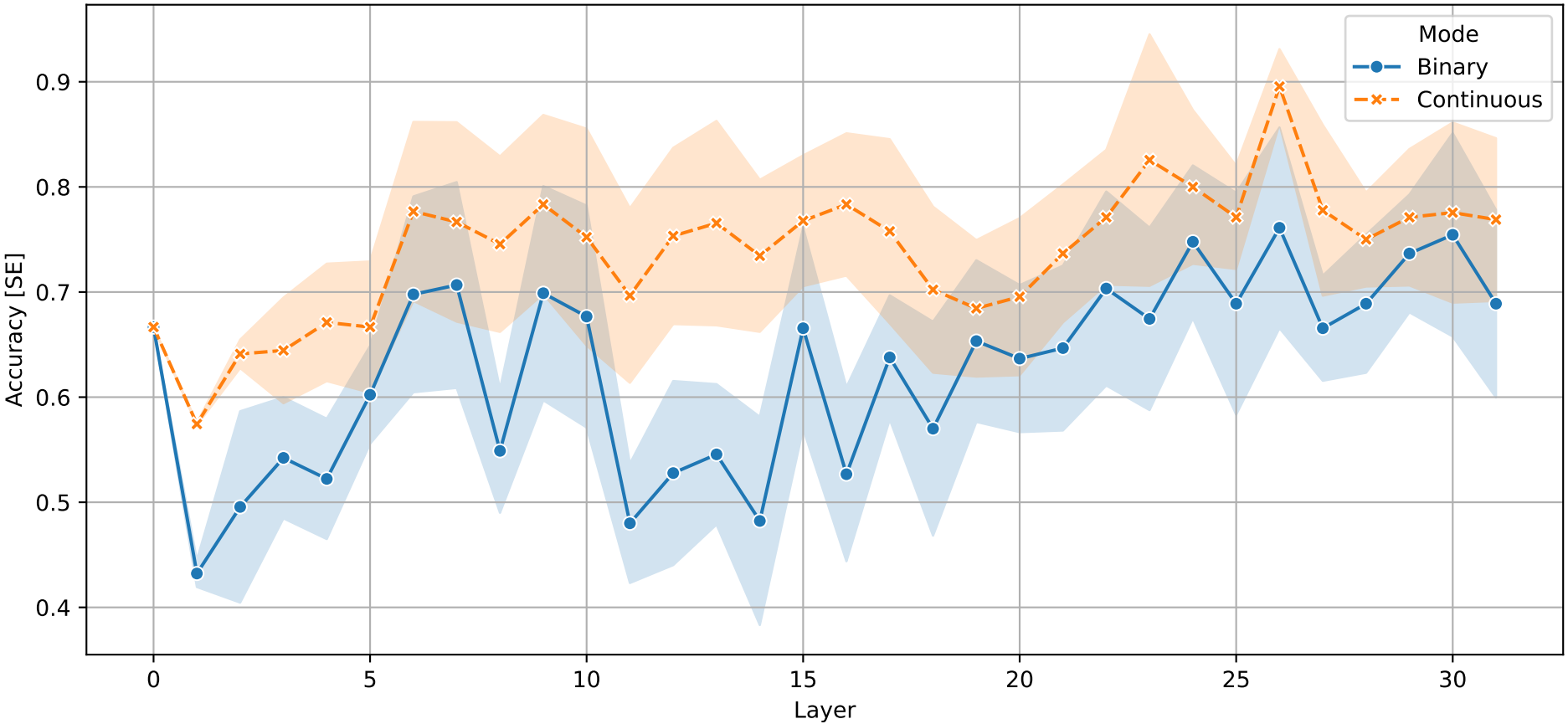

F1 per Layer – Full Layer Probing for centuries
Mode
Binary
Continuous
F1 [SE]
0.6
0.7
0.8
0.9
0
5
10
15
20
25
30
Layer

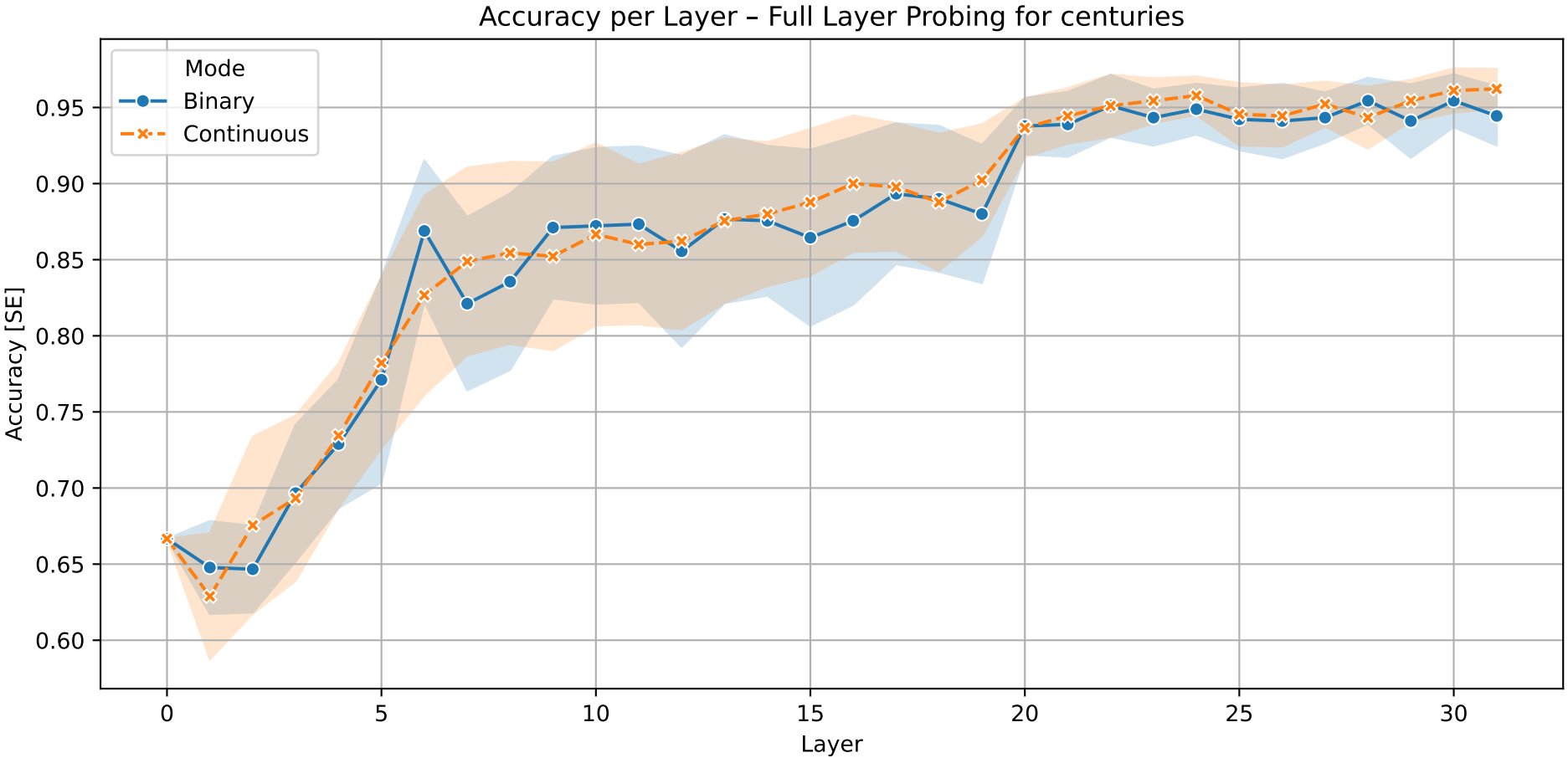

Overall F1 per Layer – All Methods for centuries
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
5
10
15
20
25
30

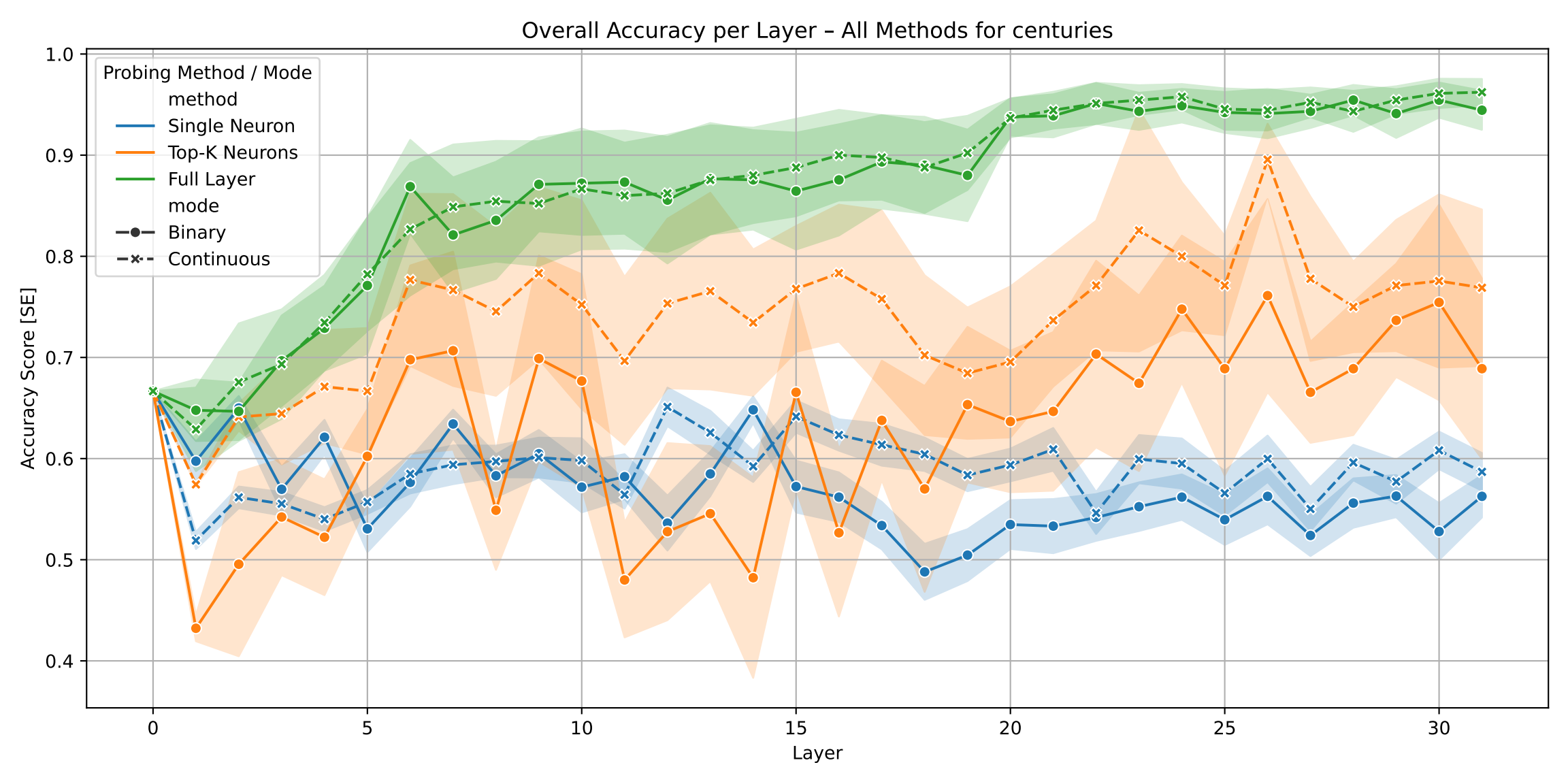
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	30.0	31.0
Full Layer	f1_max	0.9867	0.9866
Full Layer	f1_mean	0.8601	0.866
Full Layer	f1_std	0.1188	0.1191
Single Neuron	f1_best_layer	7.0	12.0
Single Neuron	f1_max	0.9086	0.9504
Single Neuron	f1_mean	0.4921	0.5925
Single Neuron	f1_std	0.1344	0.1022
Top-K Neurons	f1_best_layer	26.0	26.0
Top-K Neurons	f1_max	0.9467	0.9634
Top-K Neurons	f1_mean	0.6125	0.7417
Top-K Neurons	f1_std	0.1564	0.1215


Accuracy per Layer – Single Neuron Probing for centuries
Mode
Binary
Continuous
Accuracy [SE]
0.65
0.60
0.55
0.50
0.45
0
5
10
15
20
25
30
Layer


Accuracy per Layer – Top-K Neurons Probing for centuries
Mode
Binary
Continuous
Accuracy [SE]
Layer
0.4
0.5
0.6
0.7
0.8
0.9
0
5
10
15
20
25
30


Accuracy per Layer – Full Layer Probing for centuries
Mode
Binary
Continuous
Accuracy [SE]
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
5
10
15
20
25
30
Layer


Overall Accuracy per Layer – All Methods for centuries
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20
25
30

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	28.0	31.0
Full Layer	accuracy_max	0.9867	0.9867
Full Layer	accuracy_mean	0.8641	0.8685
Full Layer	accuracy_std	0.1093	0.1114
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.91	0.95
Single Neuron	accuracy_mean	0.568	0.5907
Single Neuron	accuracy_std	0.1286	0.1024
Top-K Neurons	accuracy_best_layer	26.0	26.0
Top-K Neurons	accuracy_max	0.9467	0.9633
Top-K Neurons	accuracy_mean	0.6273	0.7398
Top-K Neurons	accuracy_std	0.1444	0.1203