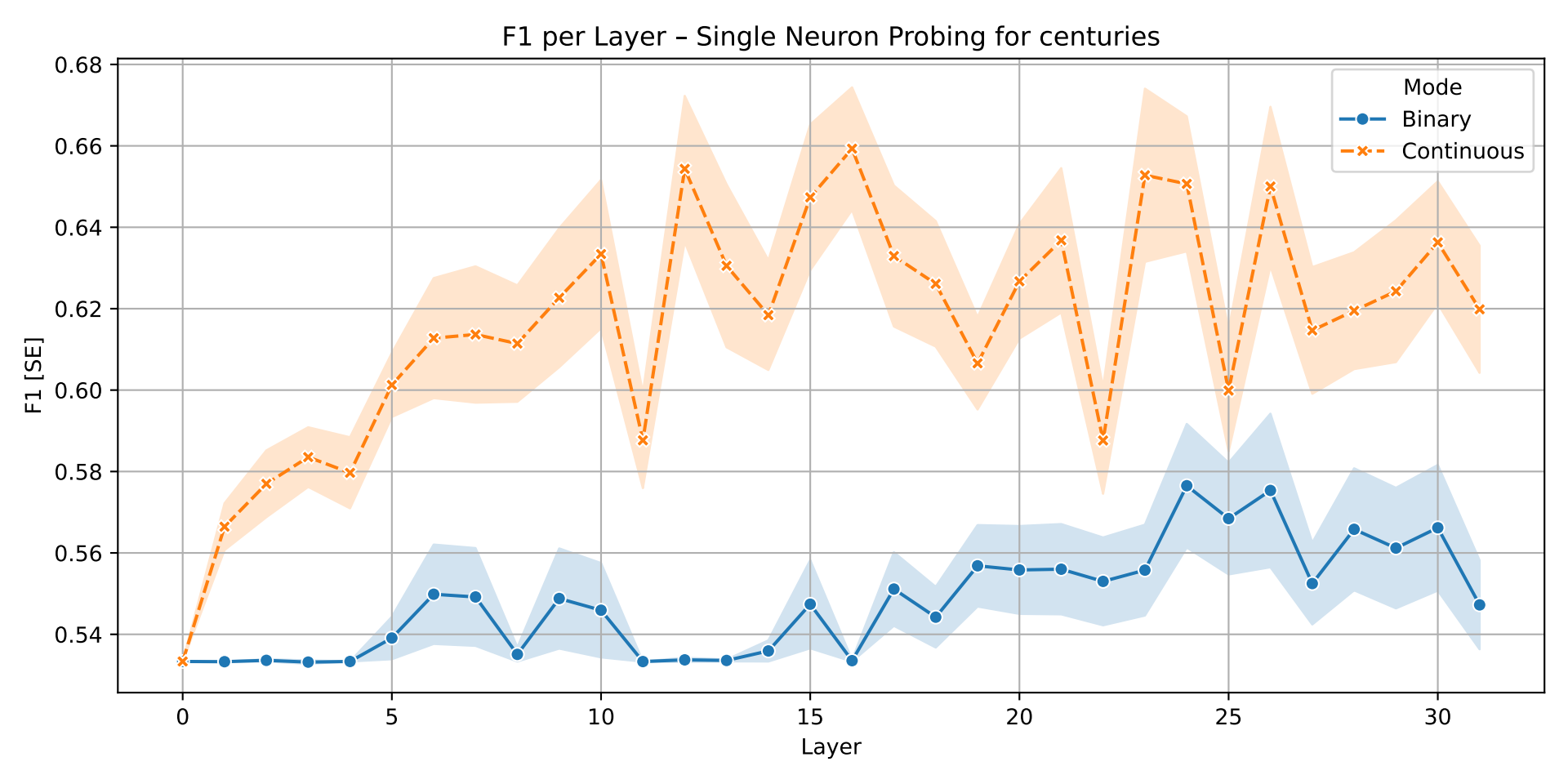
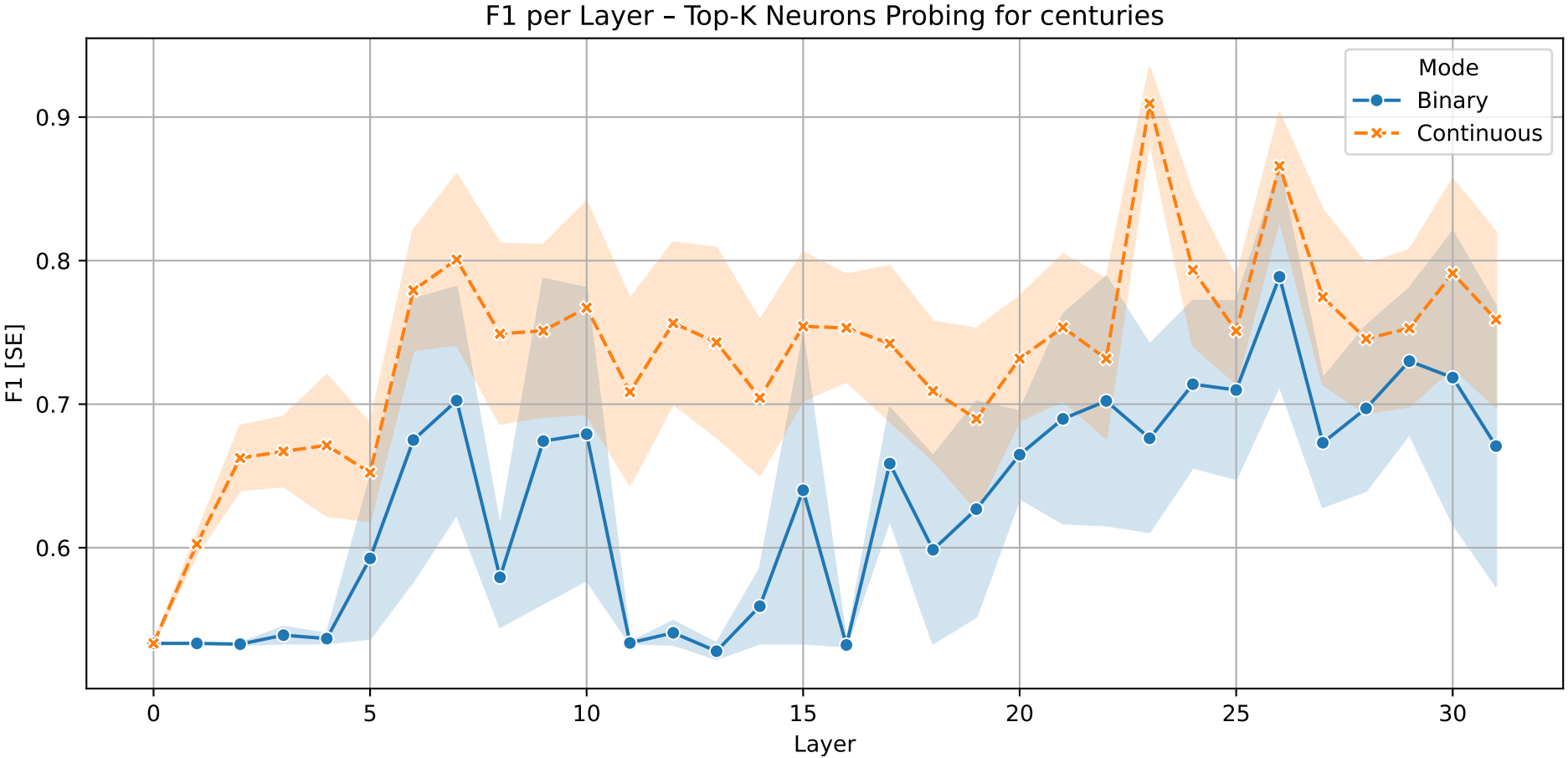
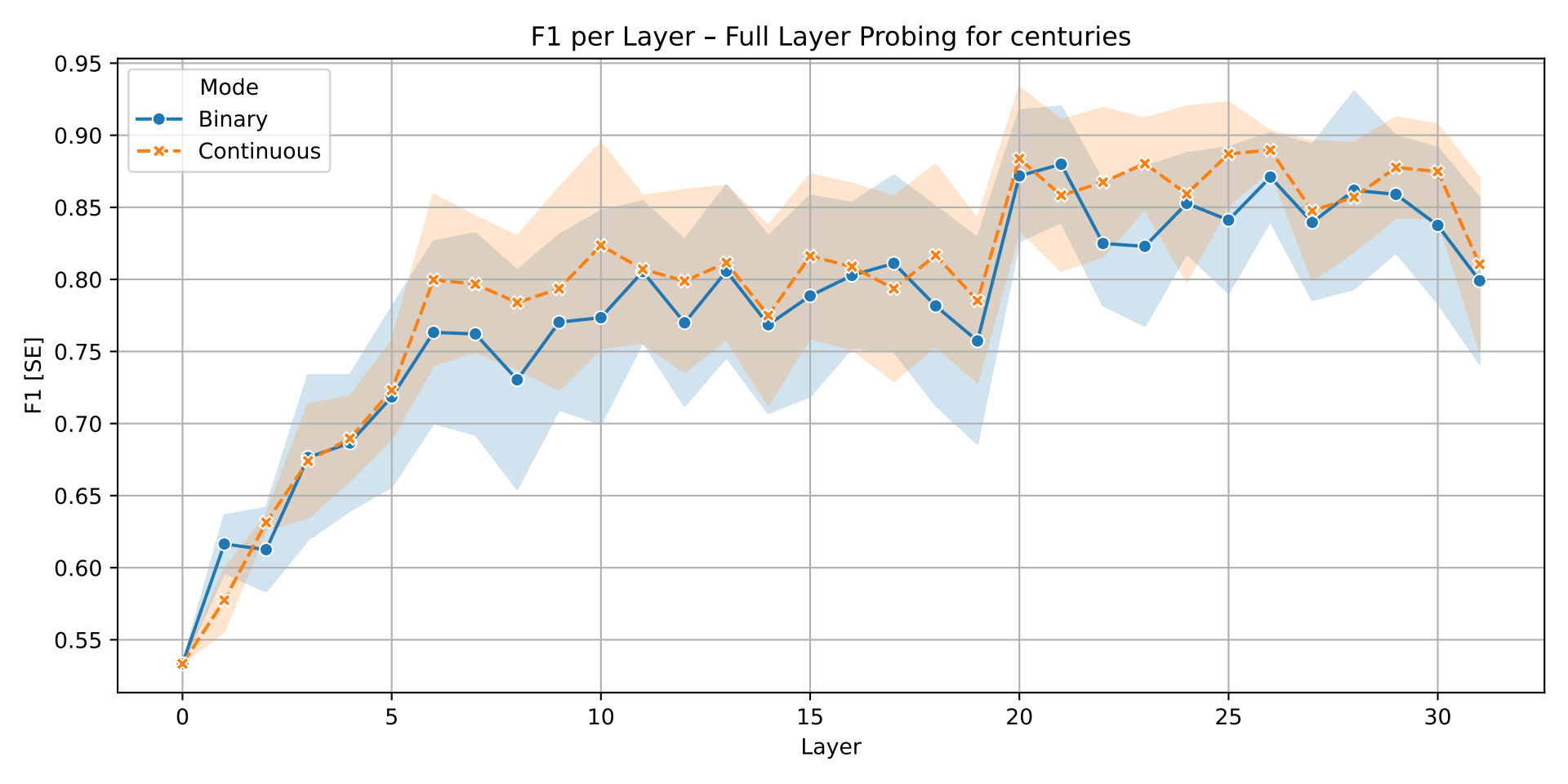
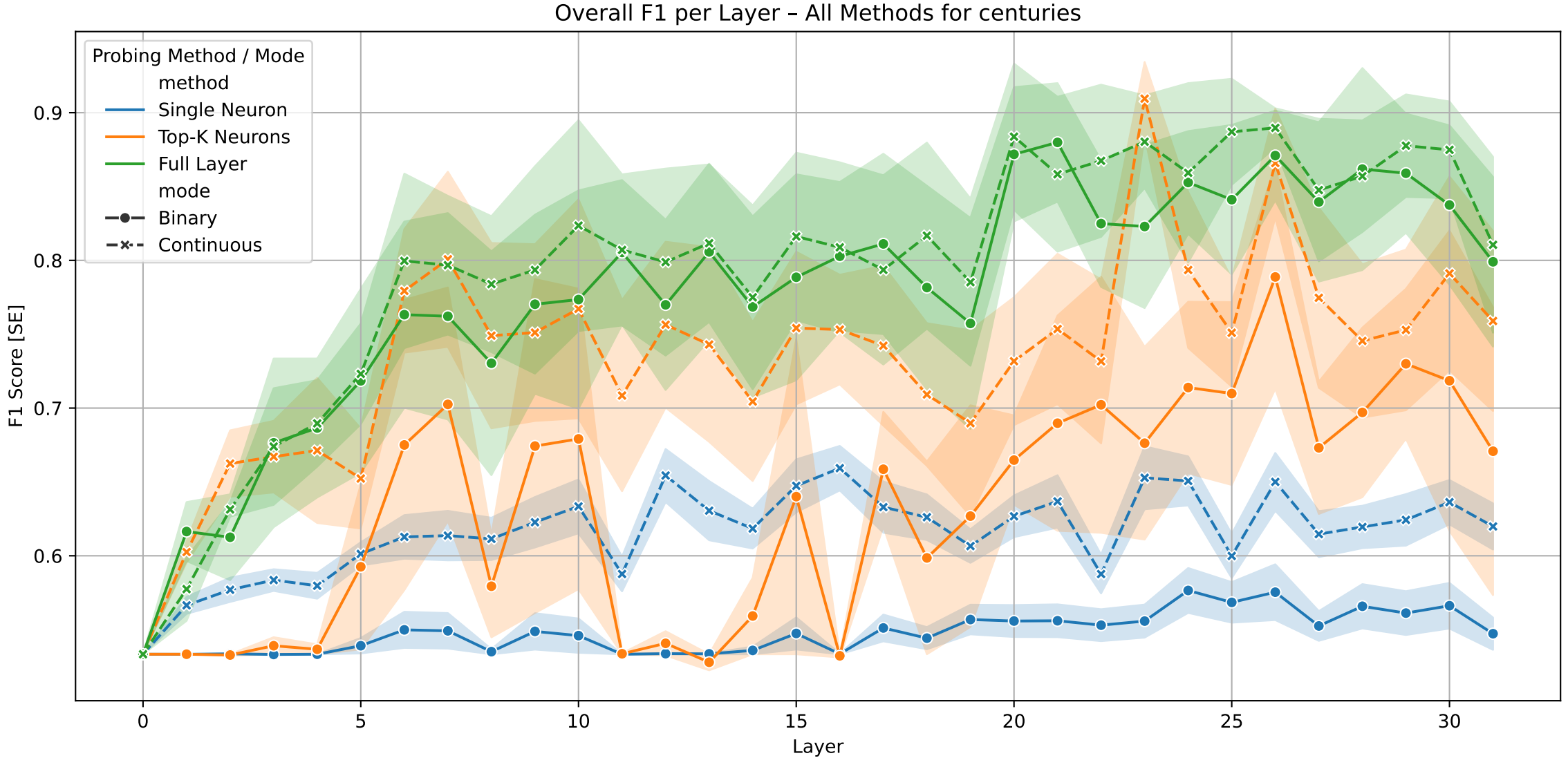

F1 per Layer – Single Neuron Probing for centuries
F1 [SE]
Layer
Mode
Binary
Continuous
0.68
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0
5
10
15
20
25
30

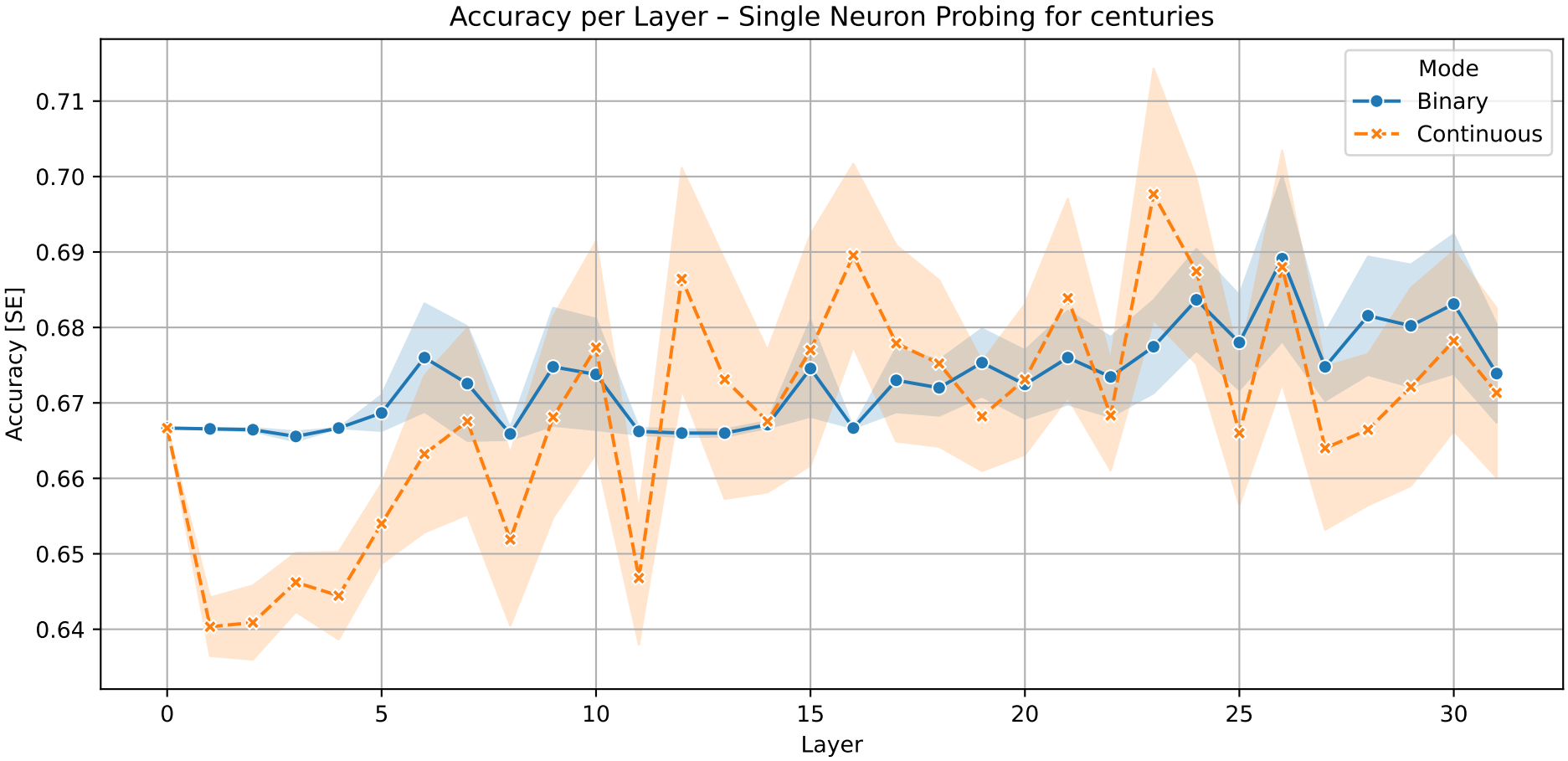

F1 per Layer – Top-K Neurons Probing for centuries
Mode
Binary
Continuous
F1 [SE]
0.6
0.7
0.8
0.9
0
5
10
15
20
25
30
Layer

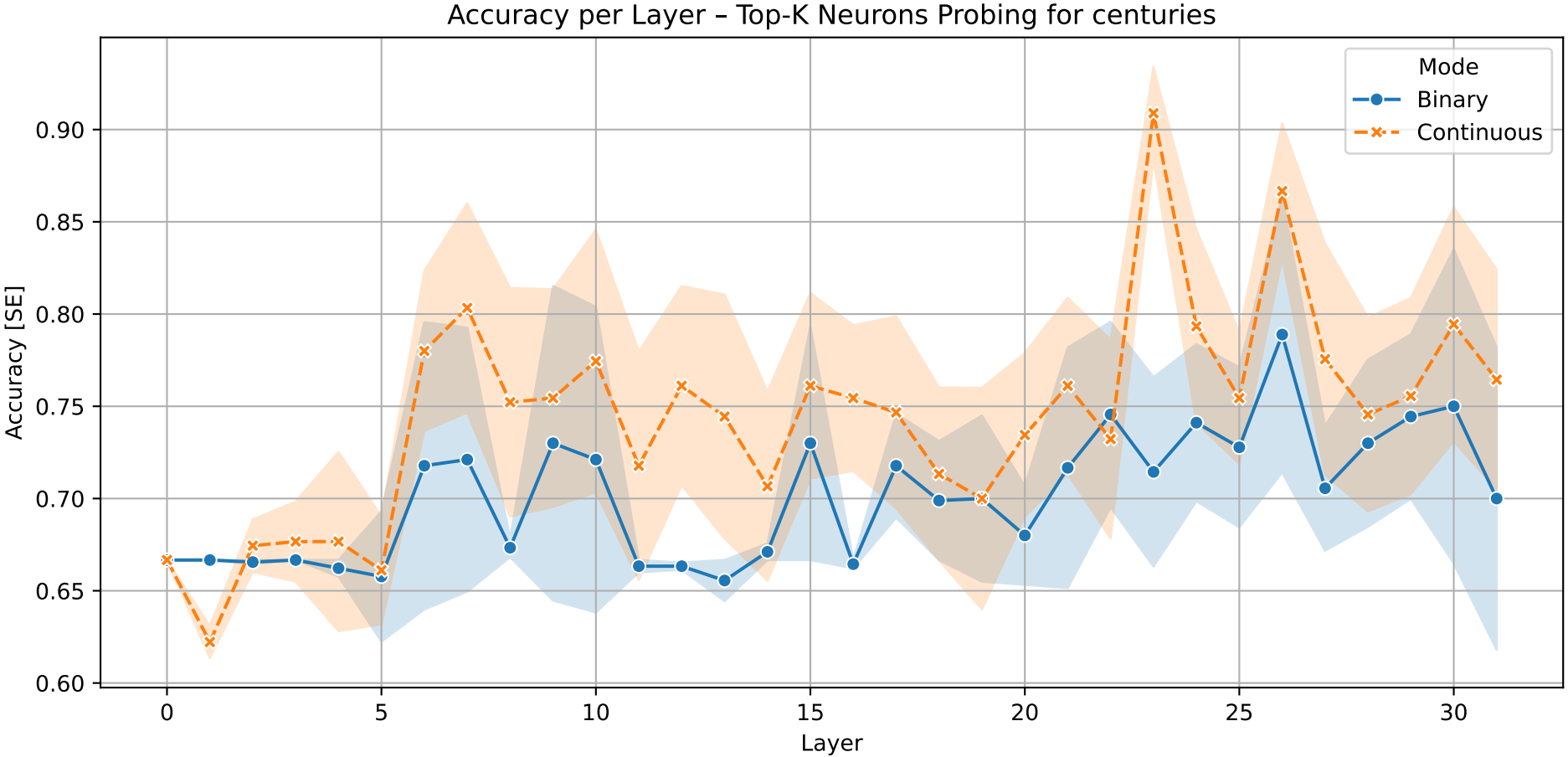

F1 per Layer – Full Layer Probing for centuries
Mode
Binary
Continuous
F1 [SE]
Layer
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
25
30

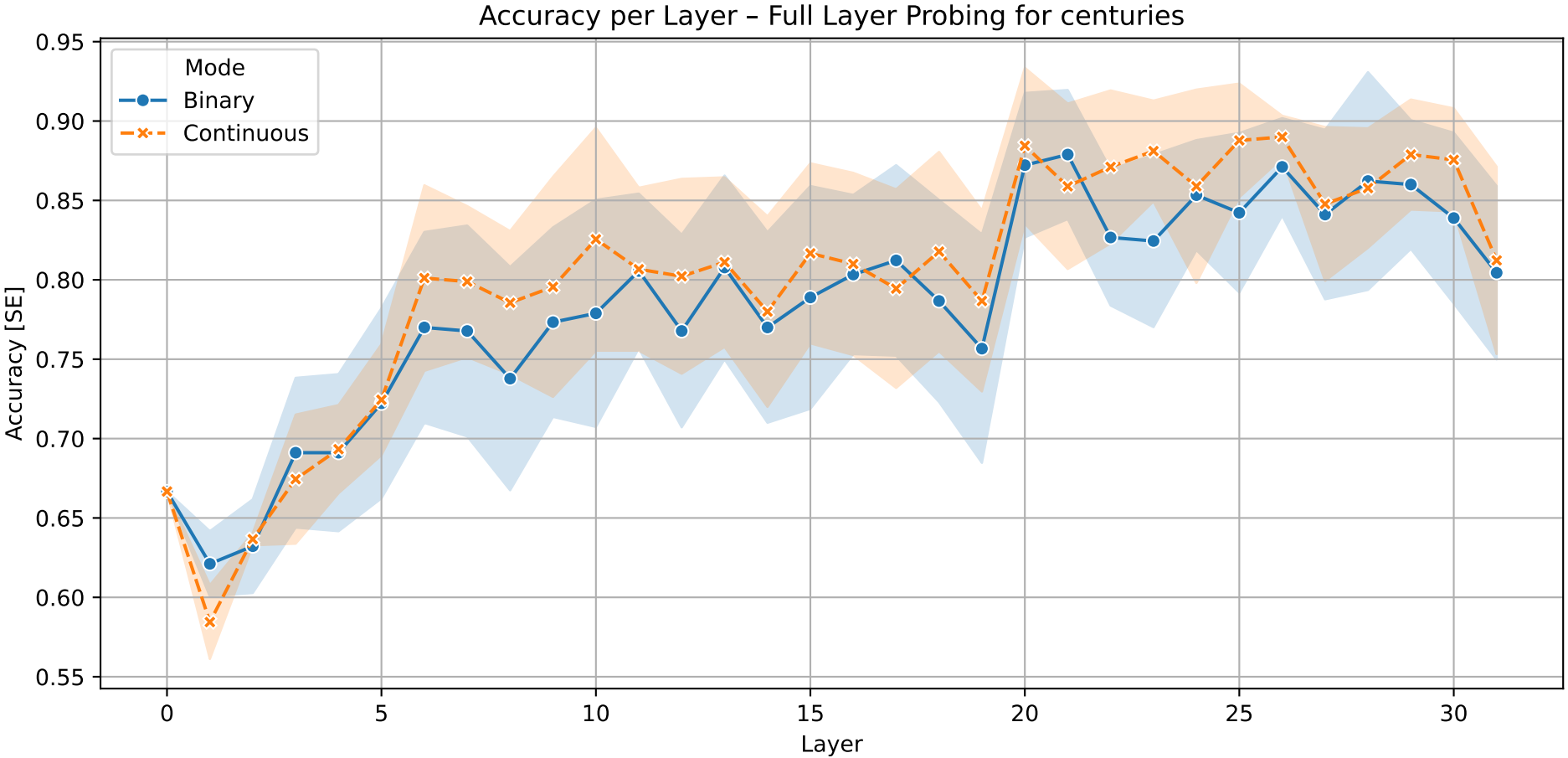

Overall F1 per Layer – All Methods for centuries
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.6
0.7
0.8
0.9
0
5
10
15
20
25
30
Layer

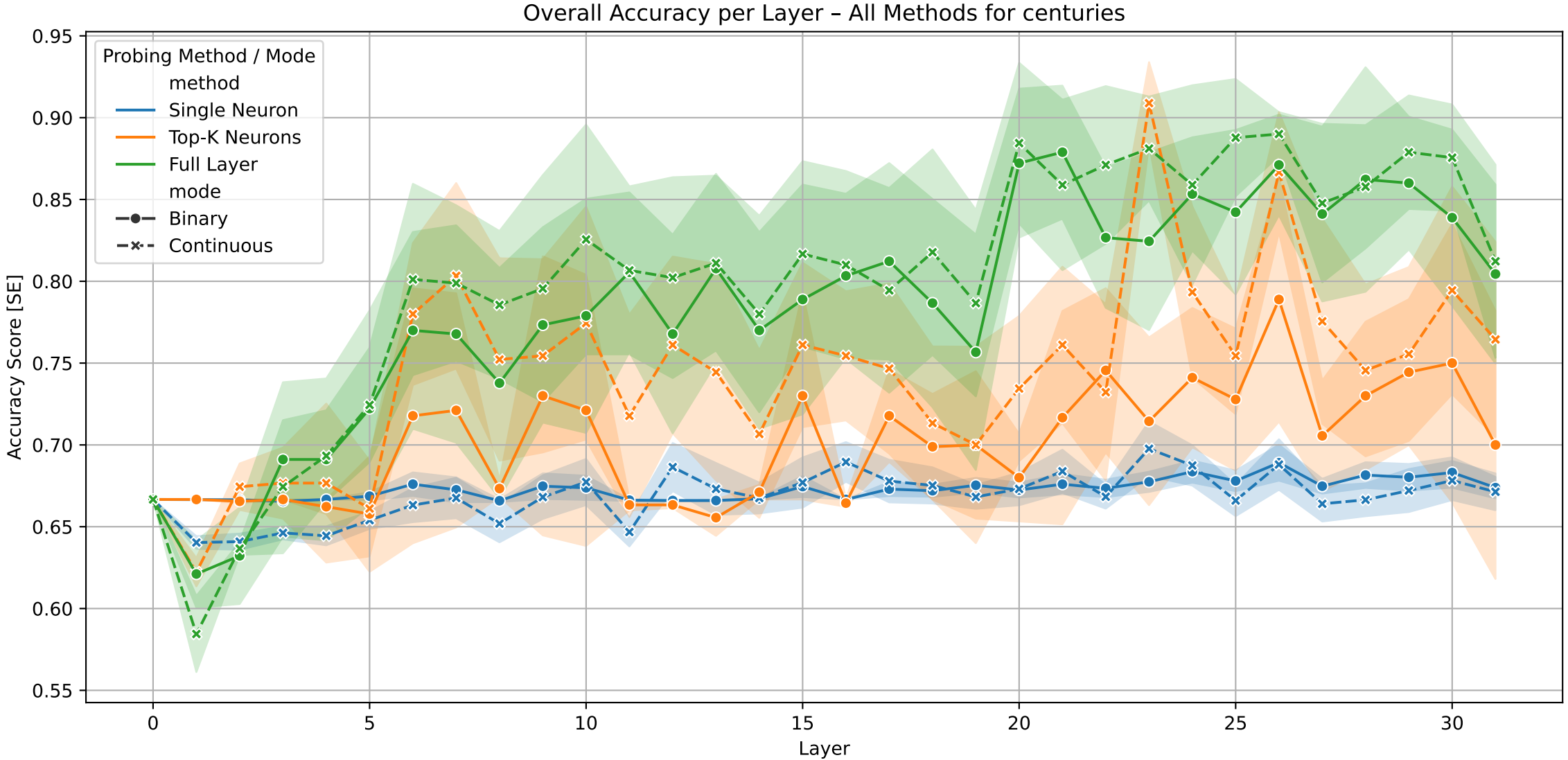
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	21.0	26.0
Full Layer	f1_max	0.9601	0.9697
Full Layer	f1_mean	0.778	0.7948
Full Layer	f1_std	0.1123	0.1118
Single Neuron	f1_best_layer	24.0	16.0
Single Neuron	f1_max	0.9086	0.9568
Single Neuron	f1_mean	0.5481	0.6162
Single Neuron	f1_std	0.0548	0.0851
Top-K Neurons	f1_best_layer	26.0	23.0
Top-K Neurons	f1_max	0.9332	0.9536
Top-K Neurons	f1_mean	0.6322	0.7362
Top-K Neurons	f1_std	0.1169	0.0995


Accuracy per Layer – Single Neuron Probing for centuries
Mode
Binary
Continuous
Accuracy [SE]
0.71
0.70
0.69
0.68
0.67
0.66
0.65
0.64
0
5
10
15
20
25
30
Layer


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


Accuracy per Layer – Full Layer Probing for centuries
Mode
Binary
Continuous
Accuracy [SE]
Layer
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
25
30


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
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
25
30

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	21.0	26.0
Full Layer	accuracy_max	0.96	0.97
Full Layer	accuracy_mean	0.7852	0.8005
Full Layer	accuracy_std	0.1022	0.103
Single Neuron	accuracy_best_layer	26.0	23.0
Single Neuron	accuracy_max	0.91	0.9567
Single Neuron	accuracy_mean	0.6729	0.6687
Single Neuron	accuracy_std	0.0296	0.0618
Top-K Neurons	accuracy_best_layer	26.0	23.0
Top-K Neurons	accuracy_max	0.9333	0.9533
Top-K Neurons	accuracy_mean	0.7018	0.7448
Top-K Neurons	accuracy_std	0.0761	0.0903