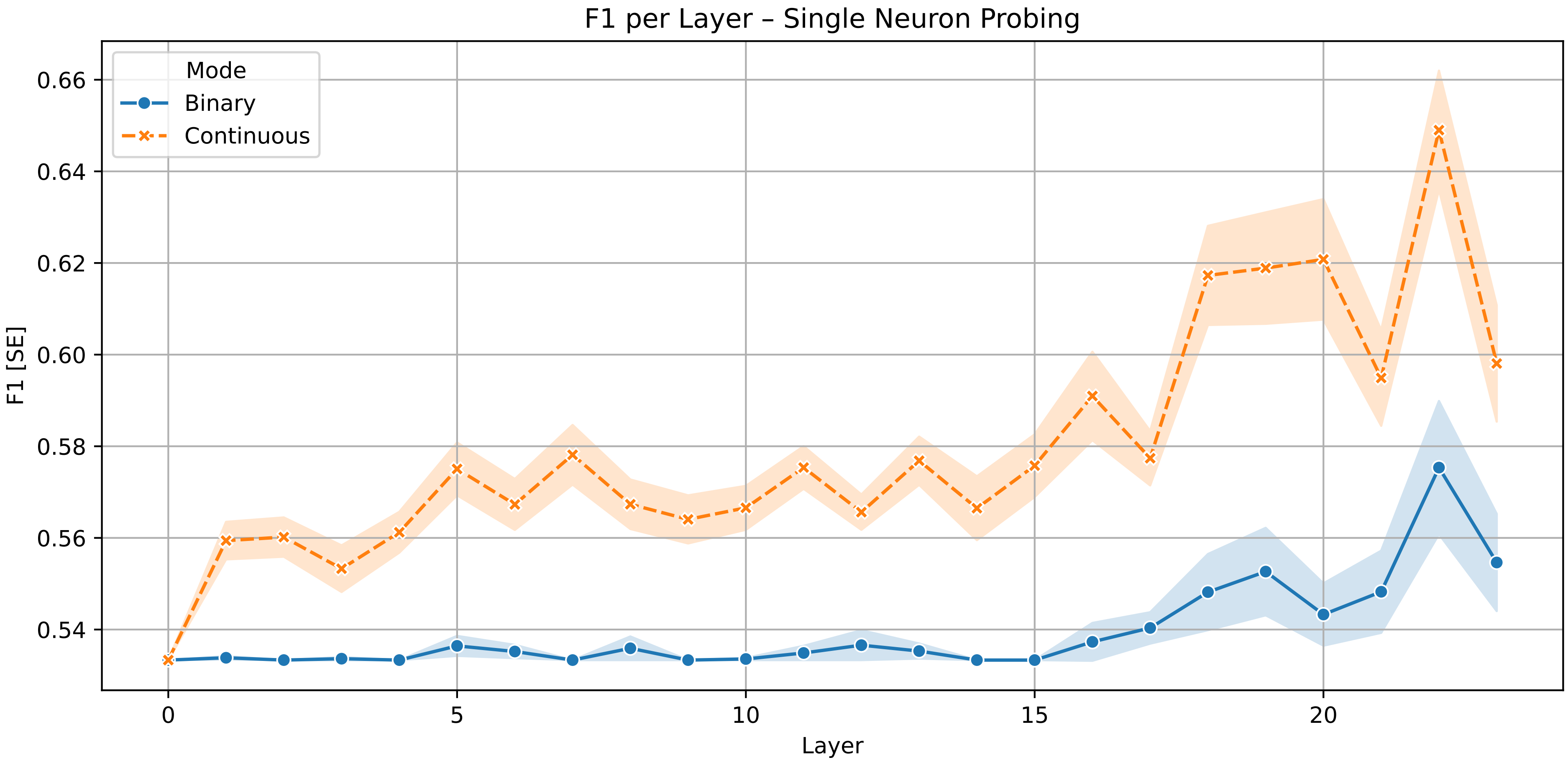
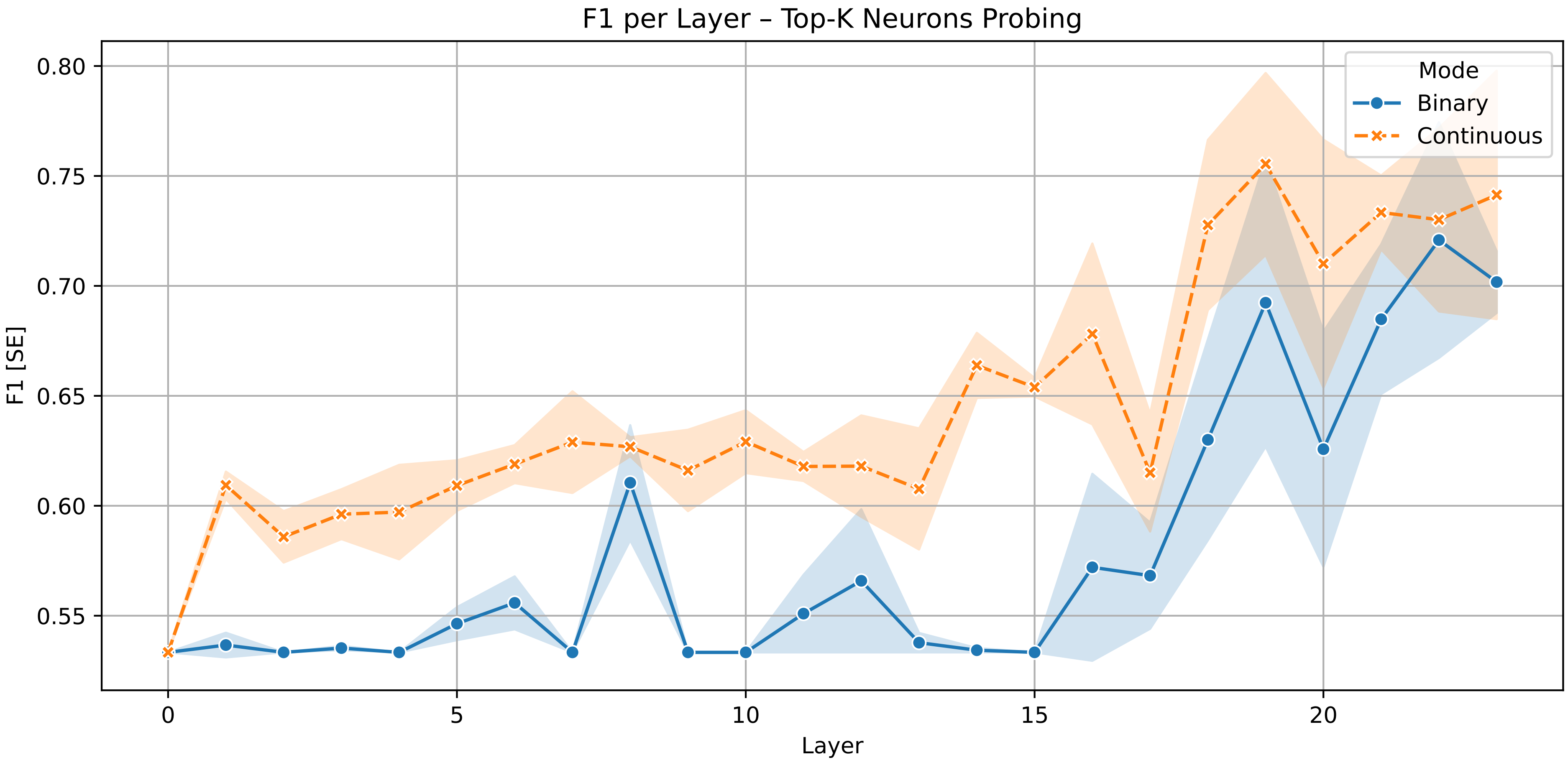
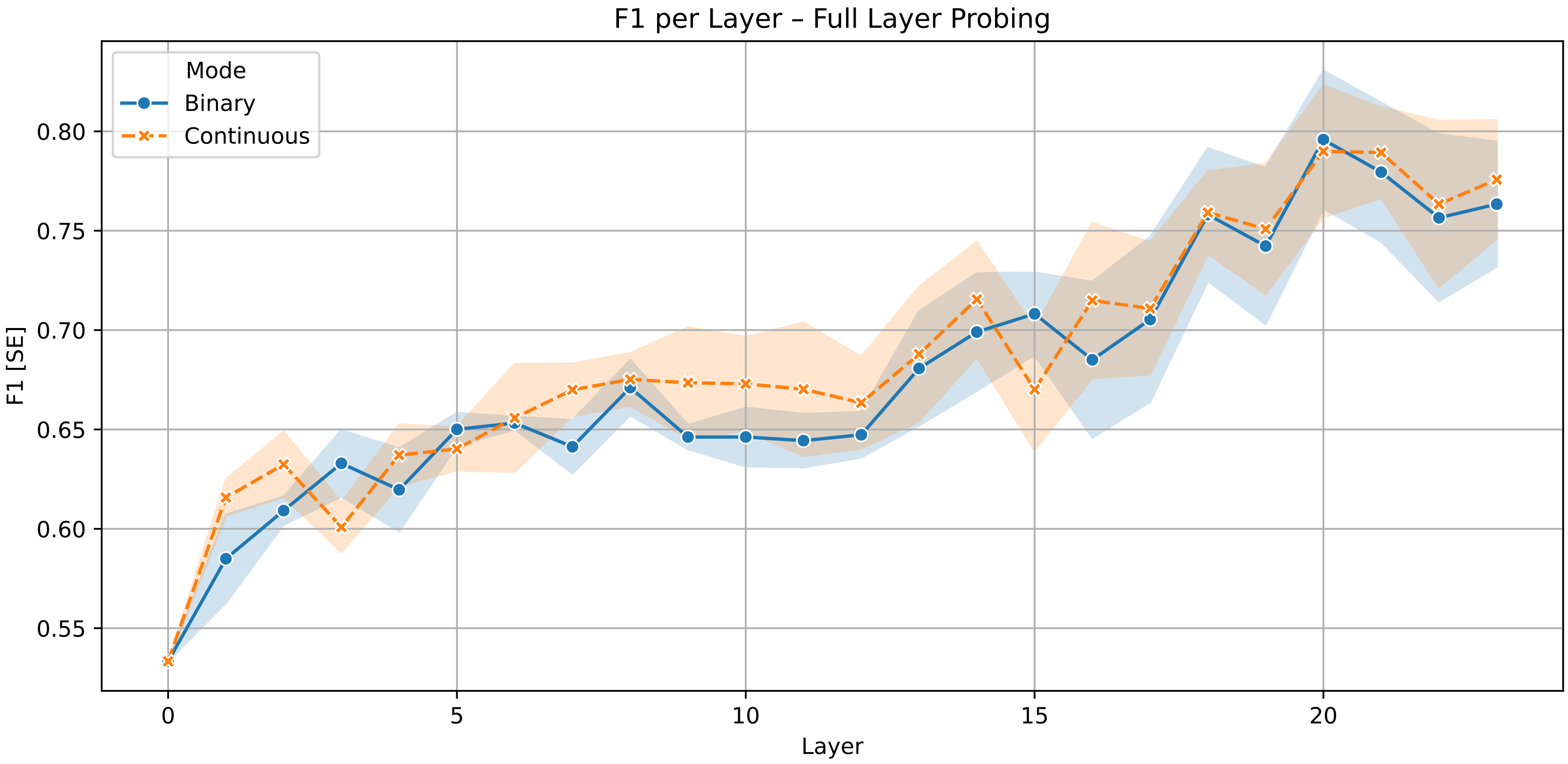
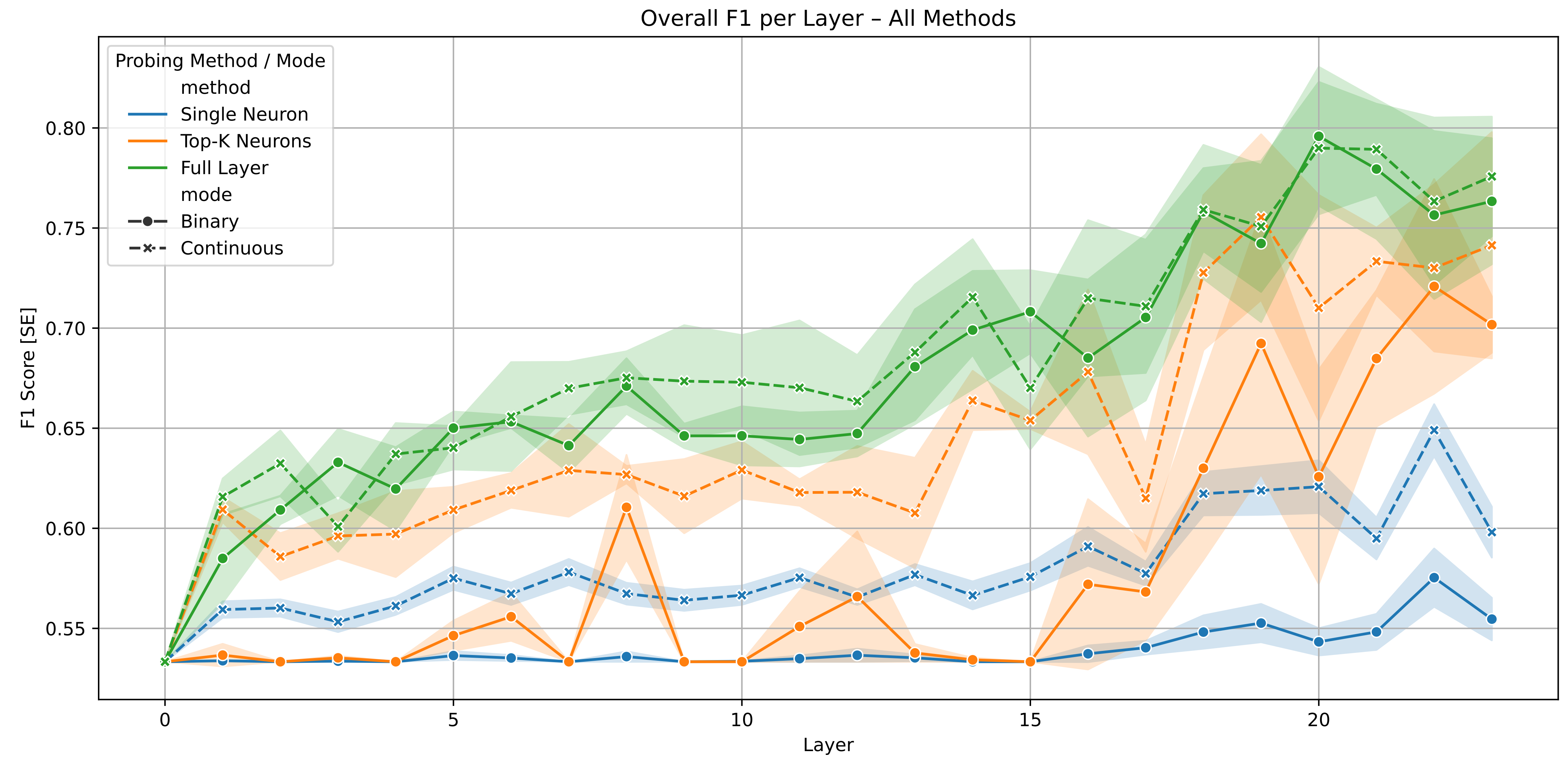
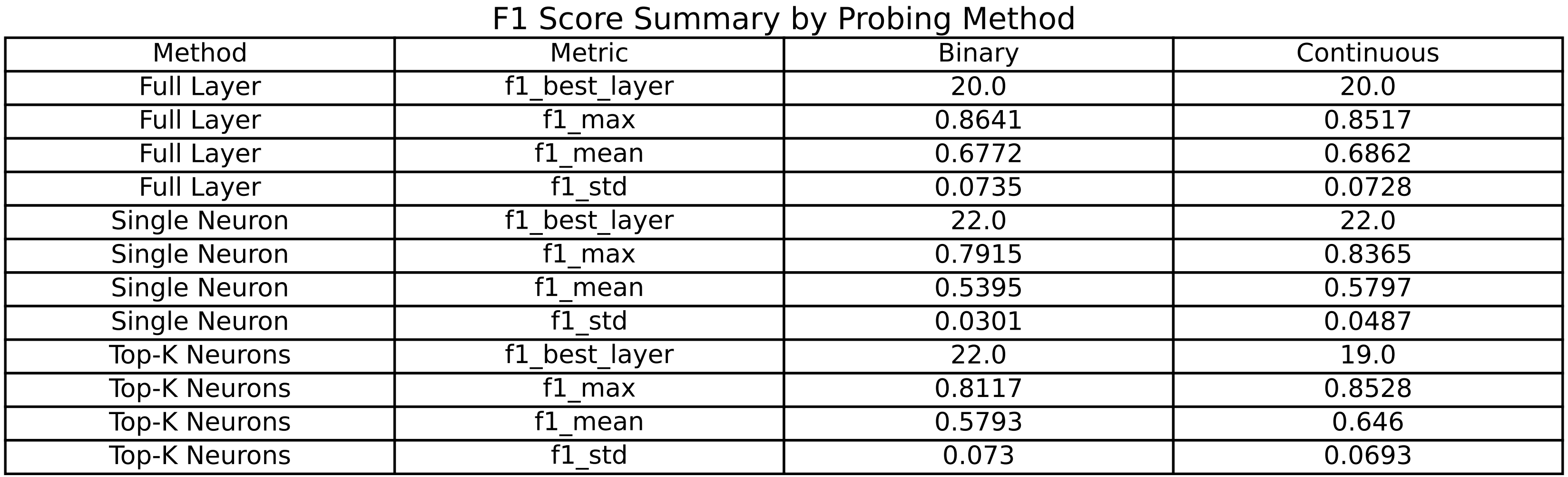

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0
5
10
15
20
Layer


F1 per Layer – Top-K Neurons Probing
F1 [SE]
0.80
0.75
0.70
0.65
0.60
0.55
Layer
0
5
10
15
20
Mode
Binary
Continuous

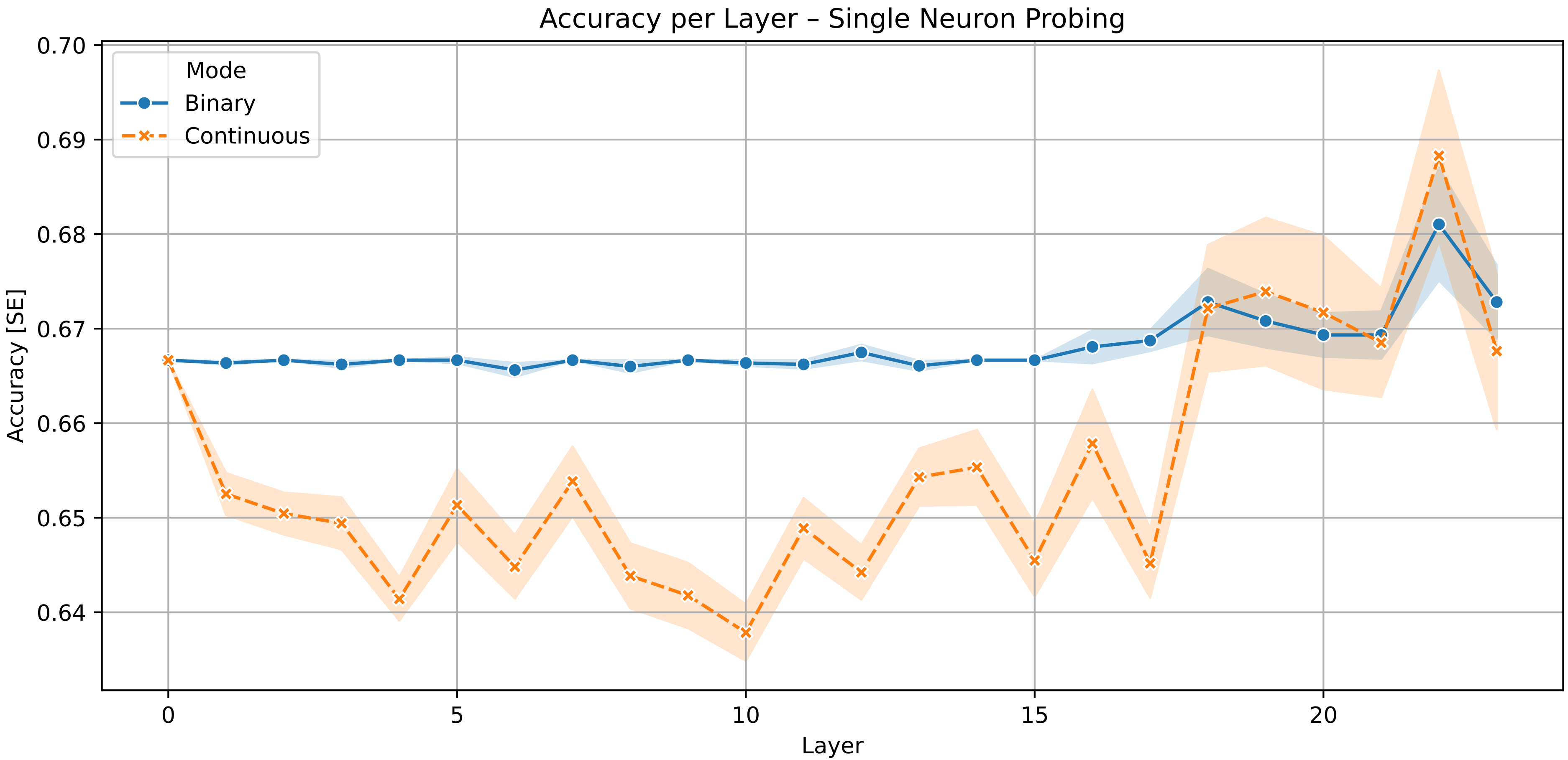

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.55
0.60
0.65
0.70
0.75
0.80
Layer
0
5
10
15
20

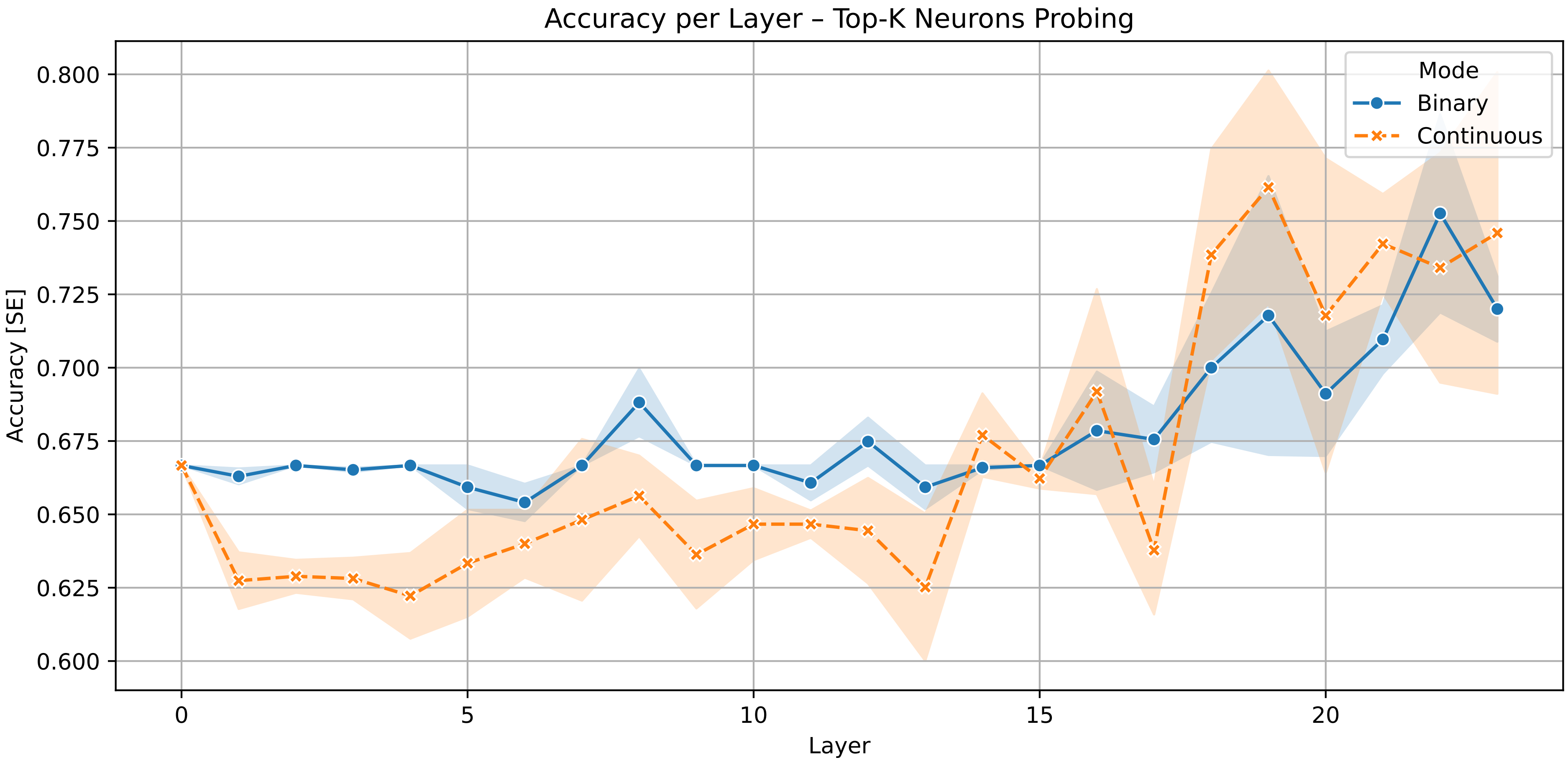

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.80
0.75
0.70
0.65
0.60
0.55
Layer
0
5
10
15
20

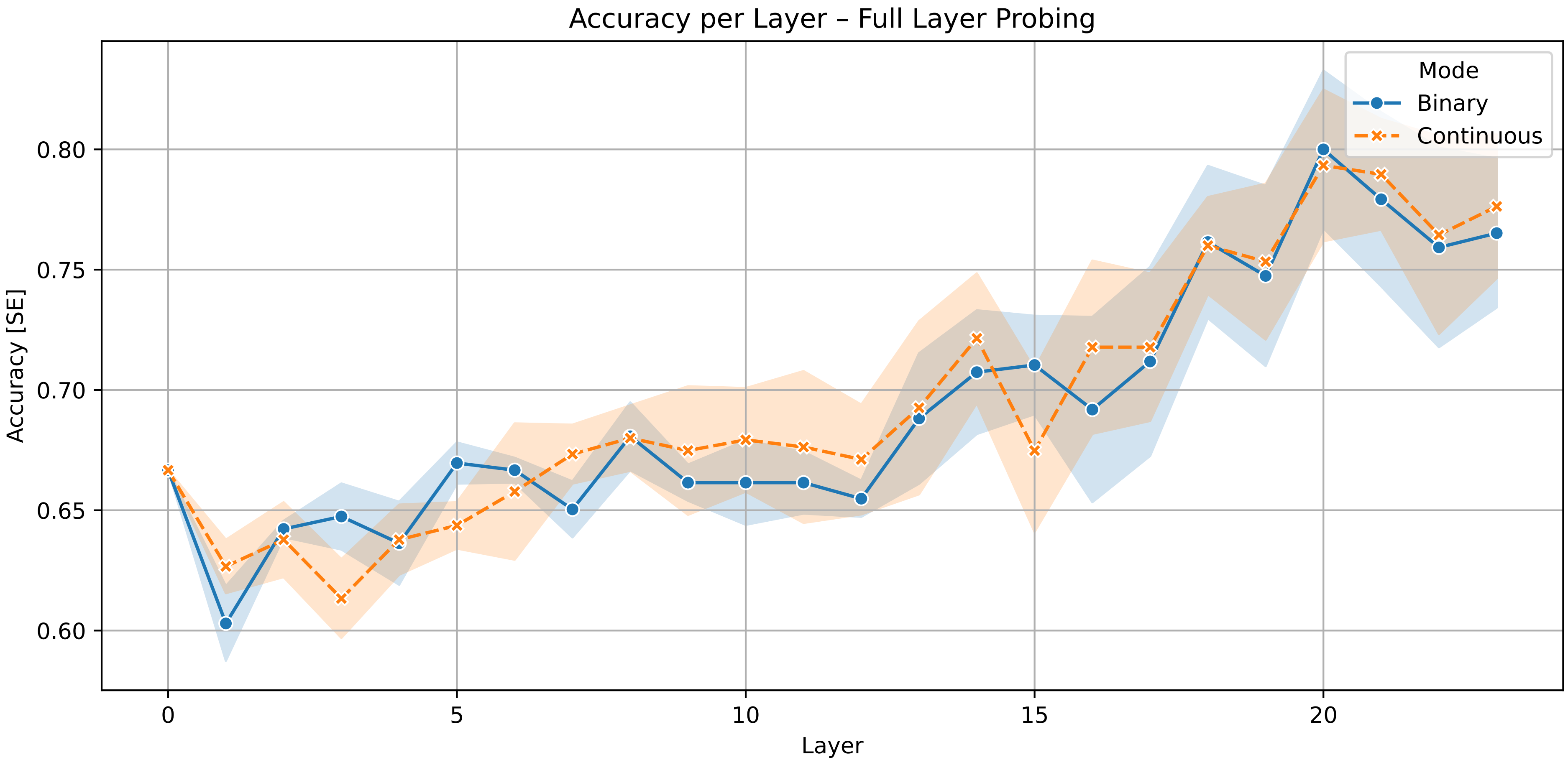


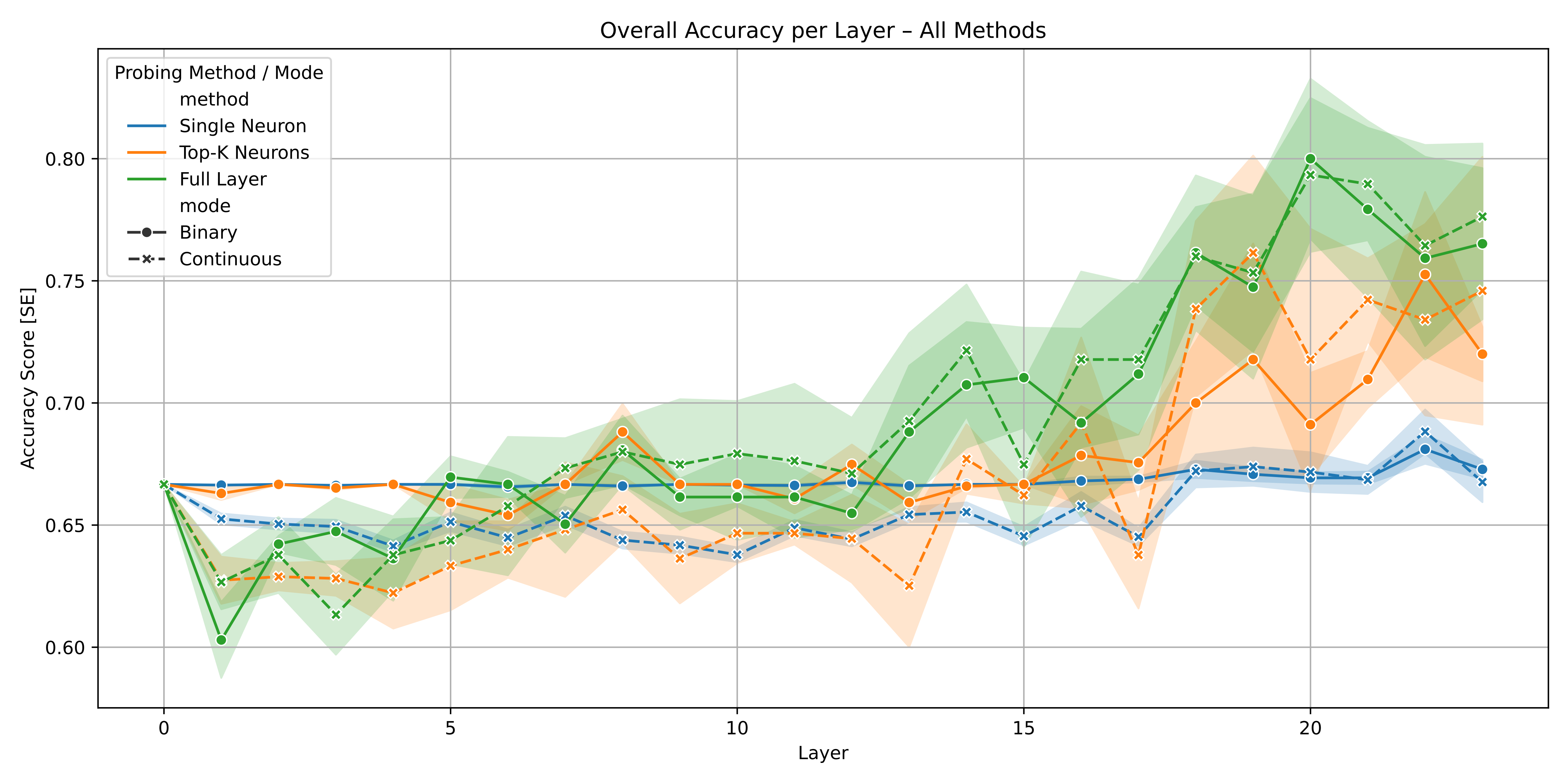
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	20.0	20.0
Full Layer	f1_max	0.8641	0.8517
Full Layer	f1_mean	0.6772	0.6862
Full Layer	f1_std	0.0735	0.0728
Single Neuron	f1_best_layer	22.0	22.0
Single Neuron	f1_max	0.7915	0.8365
Single Neuron	f1_mean	0.5395	0.5797
Single Neuron	f1_std	0.0301	0.0487
Top-K Neurons	f1_best_layer	22.0	19.0
Top-K Neurons	f1_max	0.8117	0.8528
Top-K Neurons	f1_mean	0.5793	0.646
Top-K Neurons	f1_std	0.073	0.0693


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


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.800
0.775
0.750
0.725
0.700
0.675
0.650
0.625
0.600
Layer
0
5
10
15
20


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.60
0.65
0.70
0.75
0.80
Layer
0
5
10
15
20


Overall Accuracy per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
Accuracy Score [SE]
0.60
0.65
0.70
0.75
0.80
0
5
10
15
20
Layer

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	20.0	20.0
Full Layer	accuracy_max	0.8644	0.8533
Full Layer	accuracy_mean	0.6927	0.6958
Full Layer	accuracy_std	0.0615	0.0638
Single Neuron	accuracy_best_layer	22.0	22.0
Single Neuron	accuracy_max	0.7889	0.8378
Single Neuron	accuracy_mean	0.6682	0.6553
Single Neuron	accuracy_std	0.0109	0.0288
Top-K Neurons	accuracy_best_layer	22.0	19.0
Top-K Neurons	accuracy_max	0.8156	0.8533
Top-K Neurons	accuracy_mean	0.6793	0.6691
Top-K Neurons	accuracy_std	0.0325	0.0577