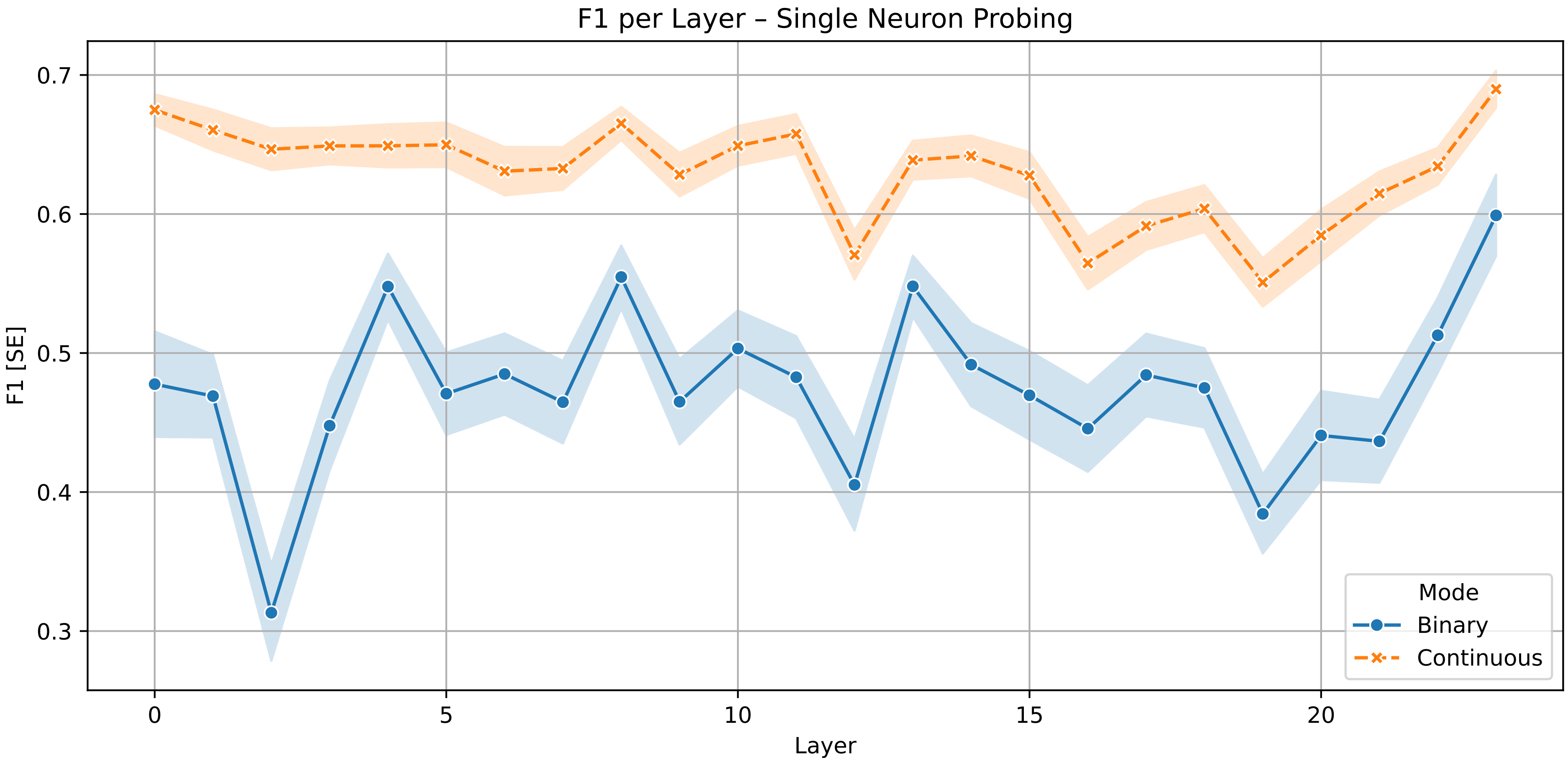
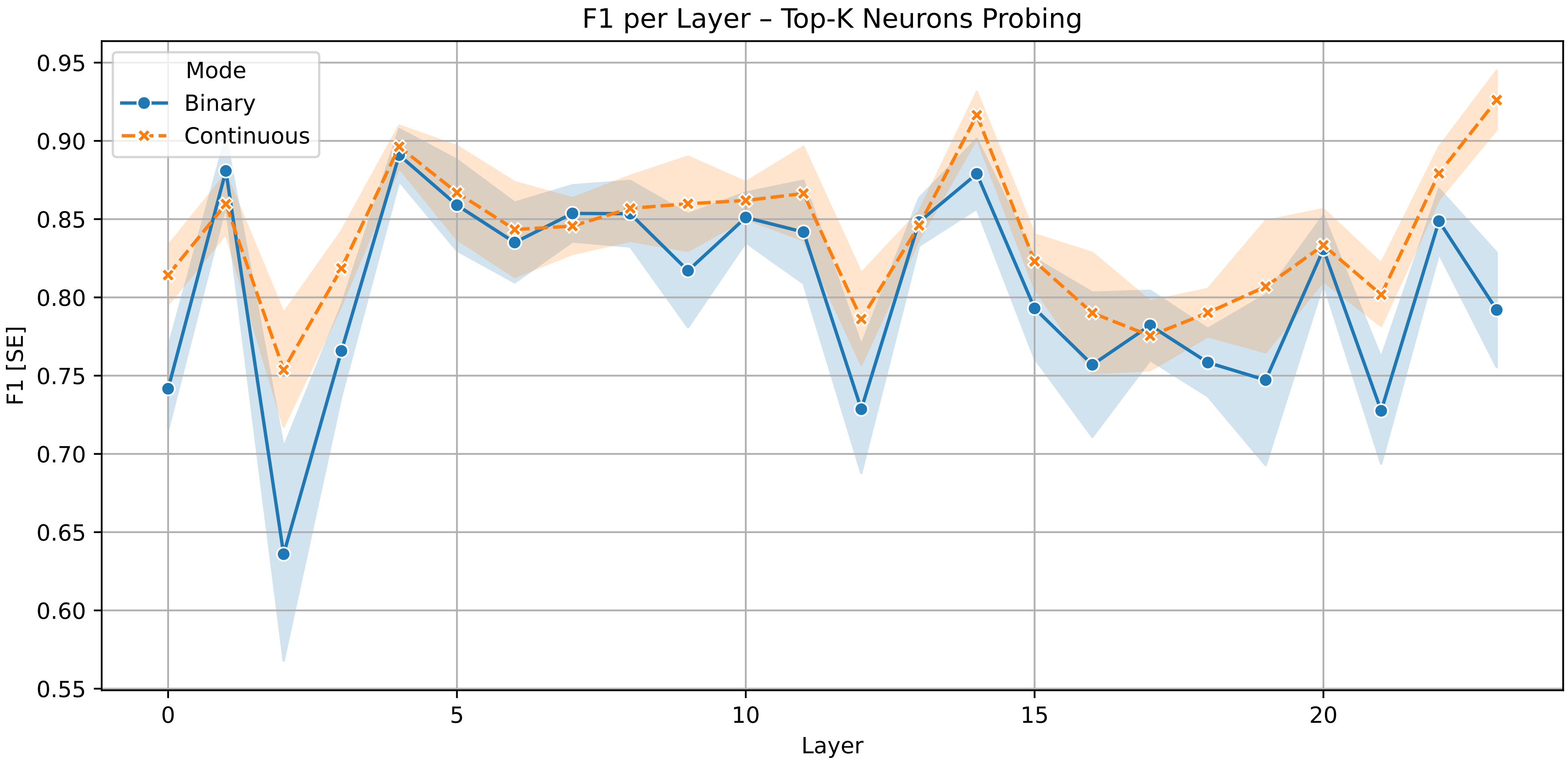
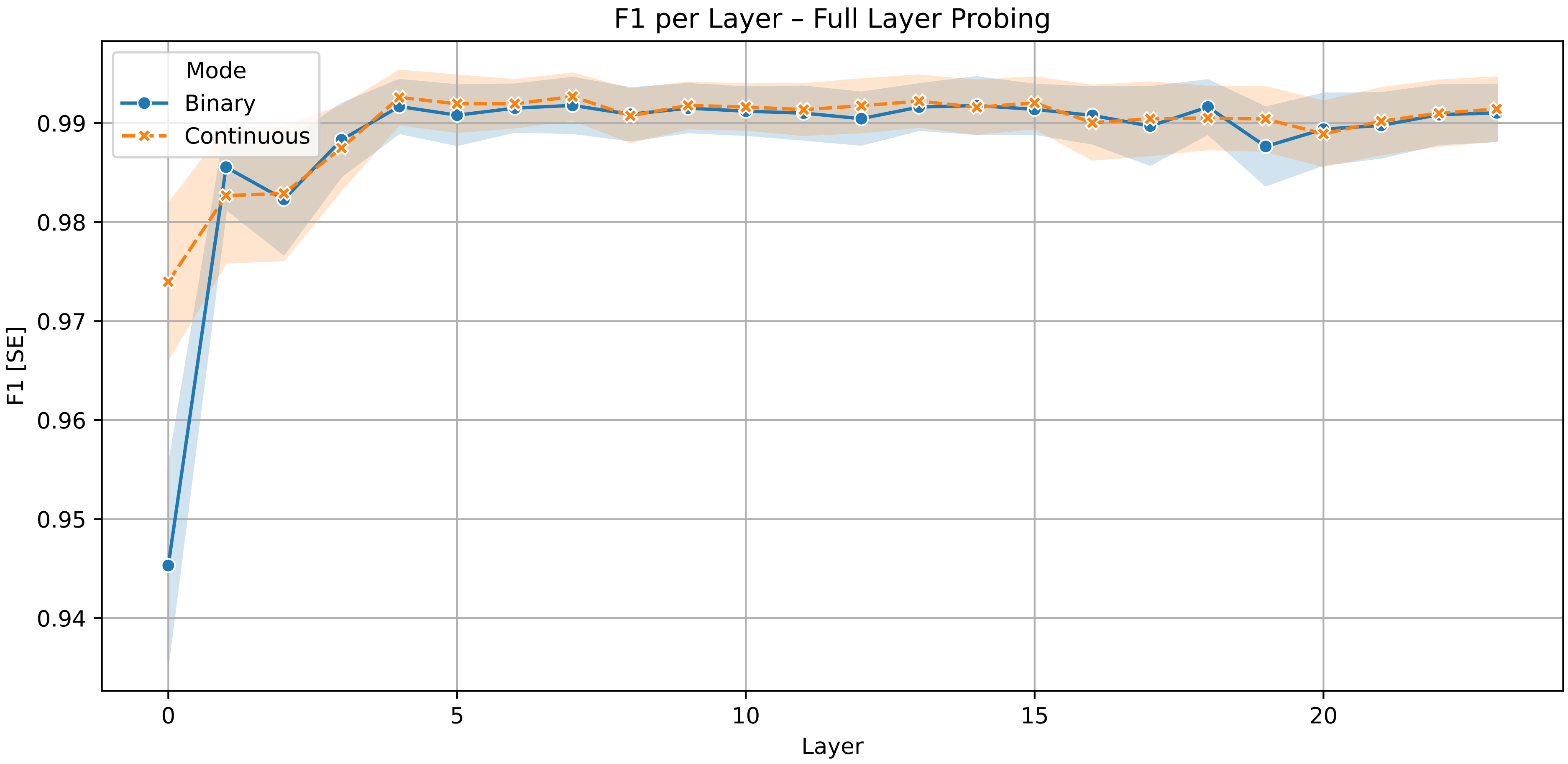
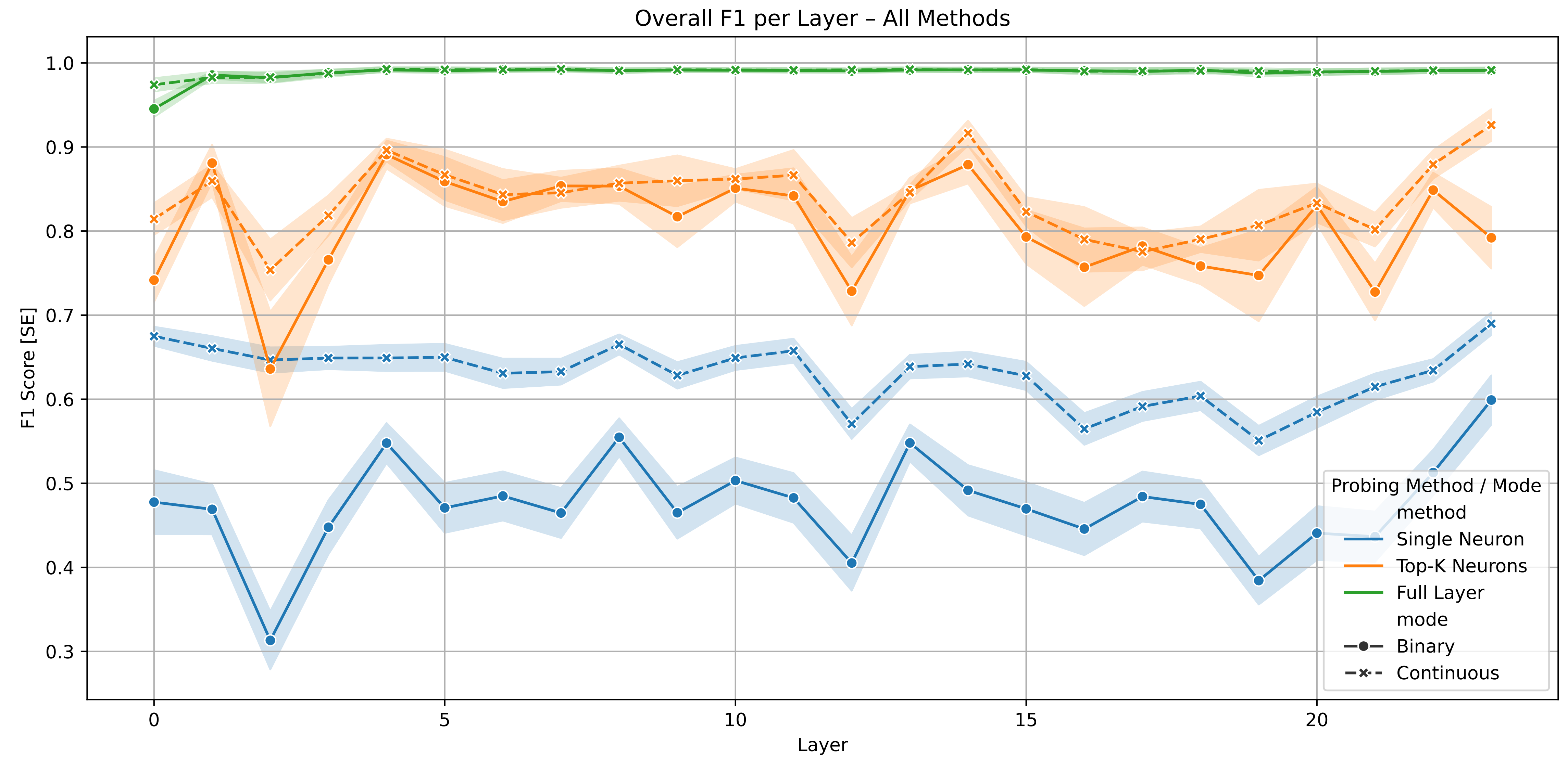
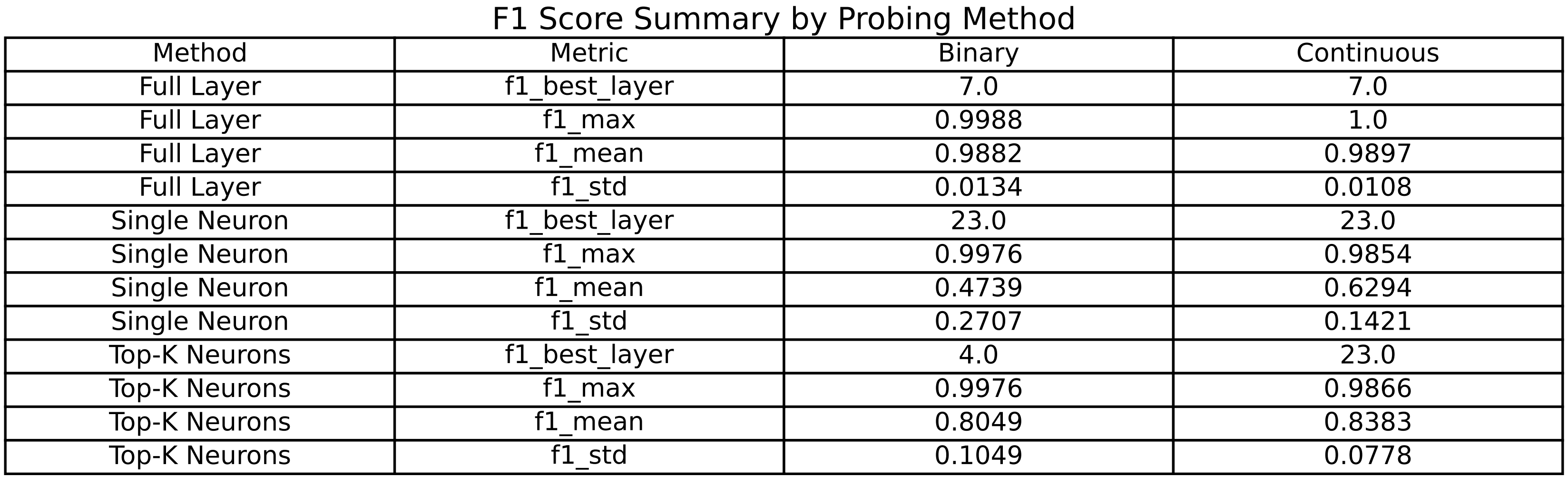

F1 per Layer – Single Neuron Probing
F1 [SE]
0.7
0.6
0.5
0.4
0.3
Layer
0
5
10
15
20
Mode
Binary
Continuous


F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
Layer
0
5
10
15
20

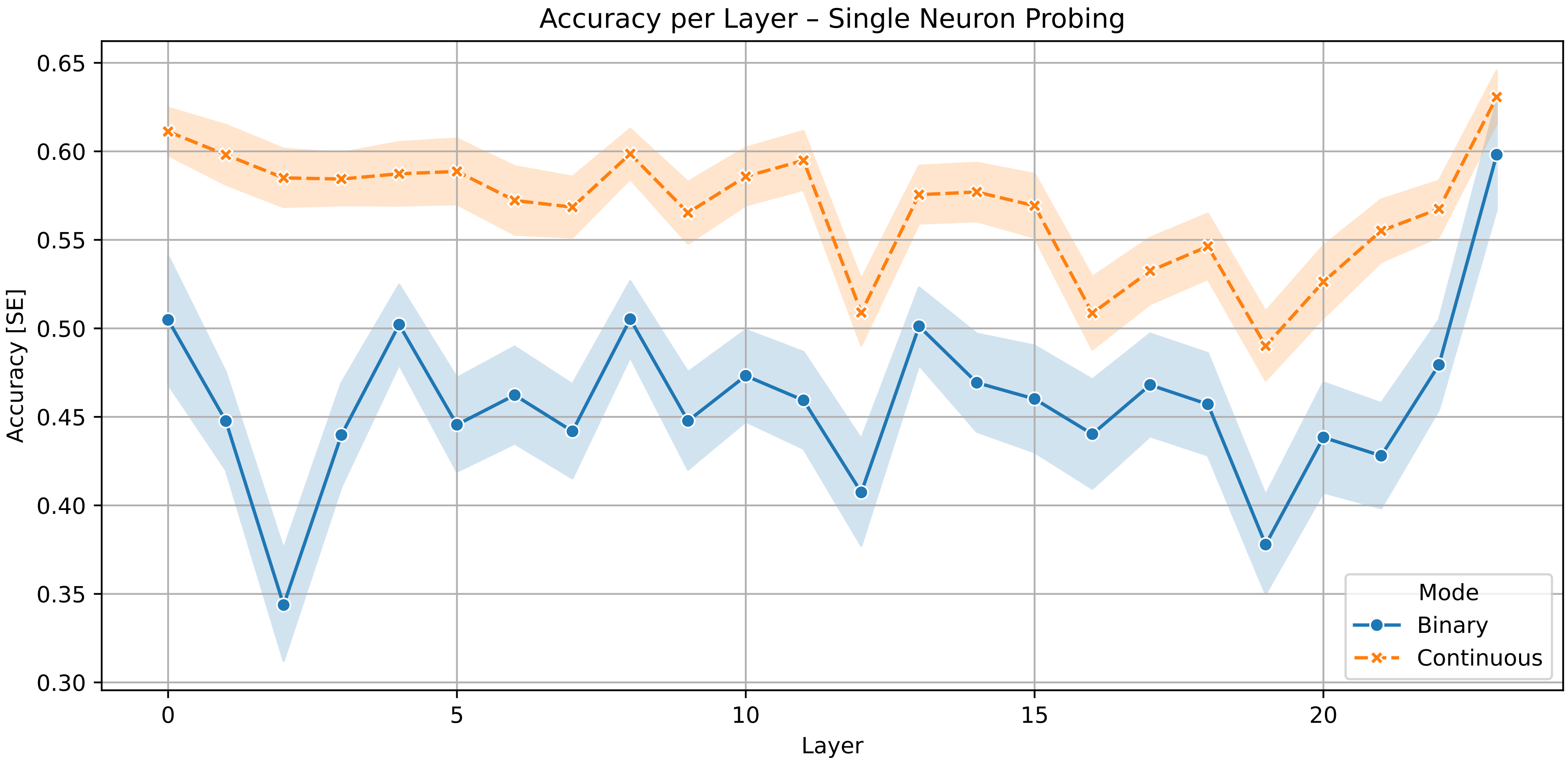

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.94
0.95
0.96
0.97
0.98
0.99
Layer
0
5
10
15
20

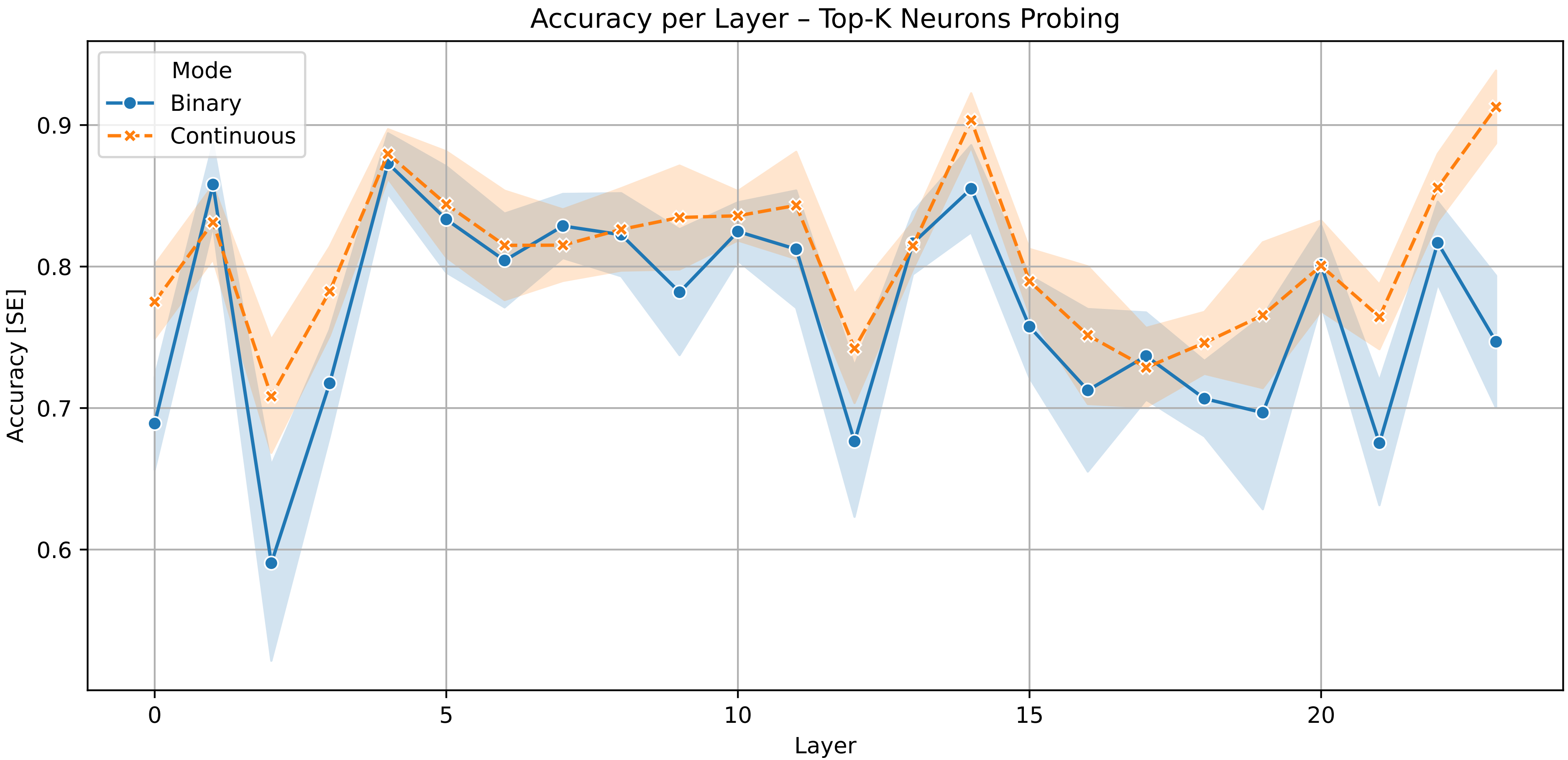

Overall F1 per Layer – All Methods
F1 Score [SE]
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Layer
0
5
10
15
20
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

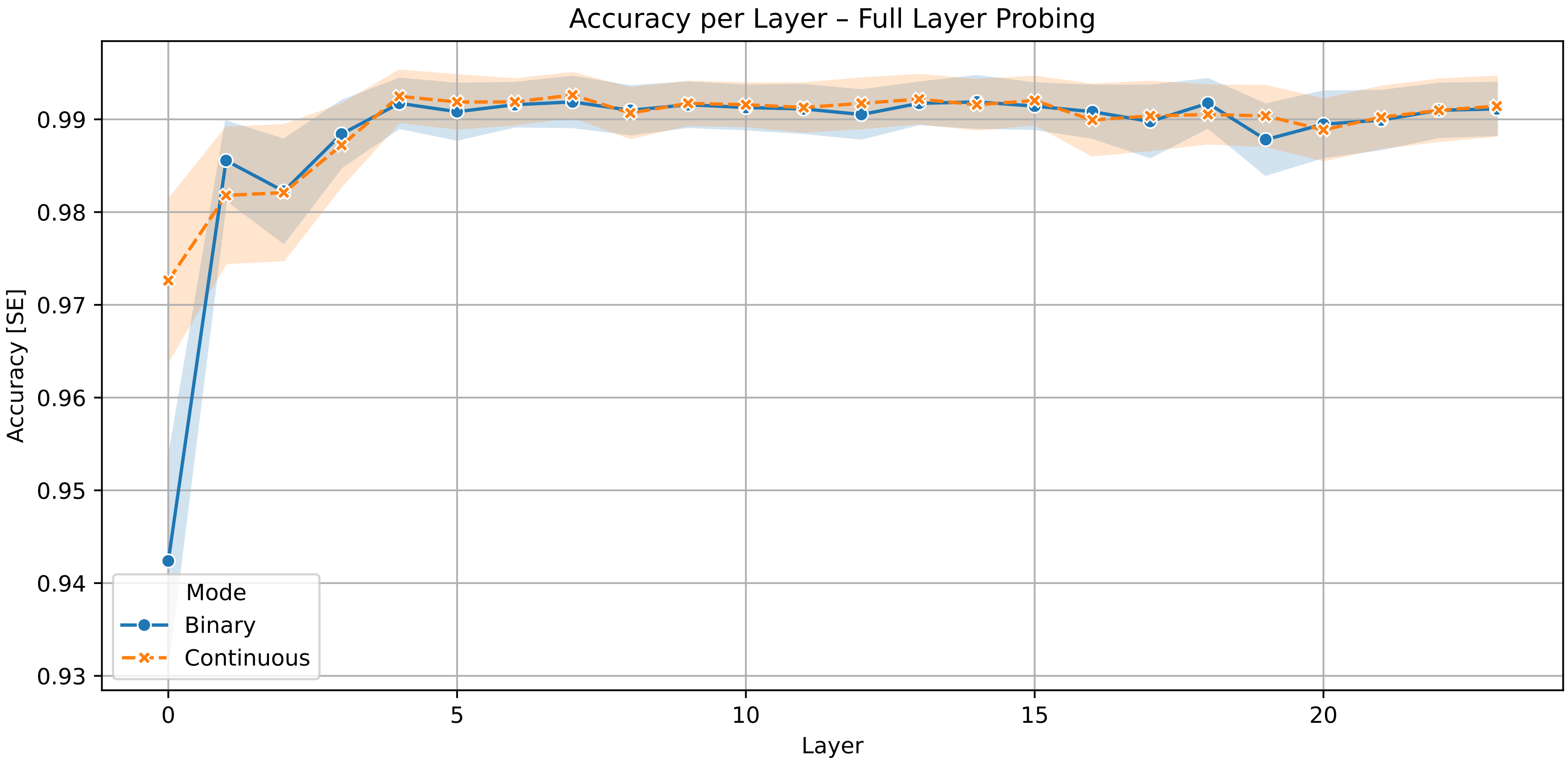


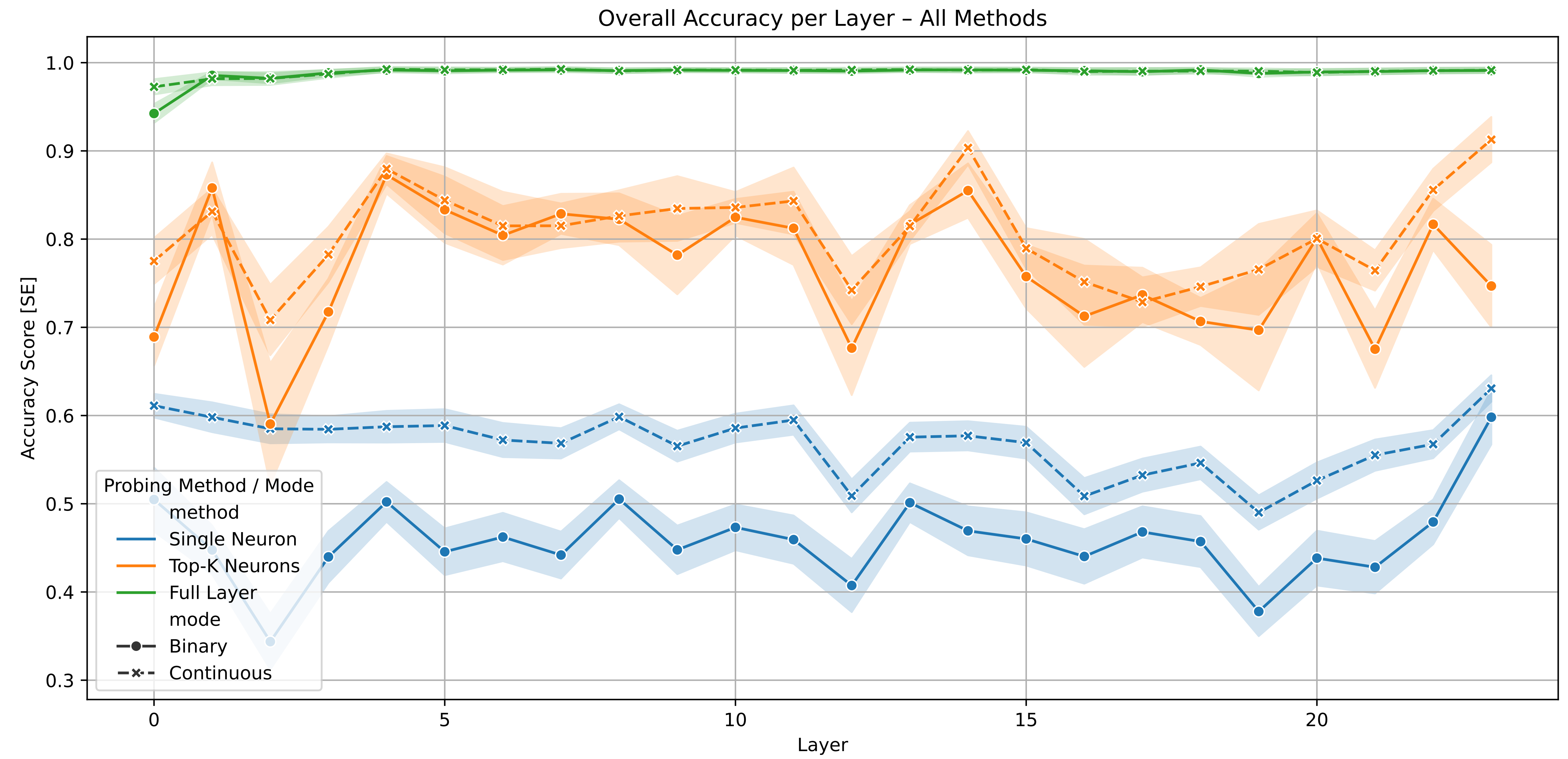
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	7.0	7.0
Full Layer	f1_max	0.9988	1.0
Full Layer	f1_mean	0.9882	0.9897
Full Layer	f1_std	0.0134	0.0108
Single Neuron	f1_best_layer	23.0	23.0
Single Neuron	f1_max	0.9976	0.9854
Single Neuron	f1_mean	0.4739	0.6294
Single Neuron	f1_std	0.2707	0.1421
Top-K Neurons	f1_best_layer	4.0	23.0
Top-K Neurons	f1_max	0.9976	0.9866
Top-K Neurons	f1_mean	0.8049	0.8383
Top-K Neurons	f1_std	0.1049	0.0778


Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
Layer
0
5
10
15
20
Mode
Binary
Continuous


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


Accuracy per Layer – Full Layer Probing
Accuracy [SE]
0.93
0.94
0.95
0.96
0.97
0.98
0.99
Layer
0
5
10
15
20
Mode
Binary
Continuous


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

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	7.0	7.0
Full Layer	accuracy_max	0.9988	1.0
Full Layer	accuracy_mean	0.9882	0.9895
Full Layer	accuracy_std	0.014	0.0113
Single Neuron	accuracy_best_layer	23.0	23.0
Single Neuron	accuracy_max	0.9976	0.9856
Single Neuron	accuracy_mean	0.4583	0.5678
Single Neuron	accuracy_std	0.2554	0.158
Top-K Neurons	accuracy_best_layer	4.0	23.0
Top-K Neurons	accuracy_max	0.9976	0.9868
Top-K Neurons	accuracy_mean	0.7681	0.8069
Top-K Neurons	accuracy_std	0.1266	0.0981