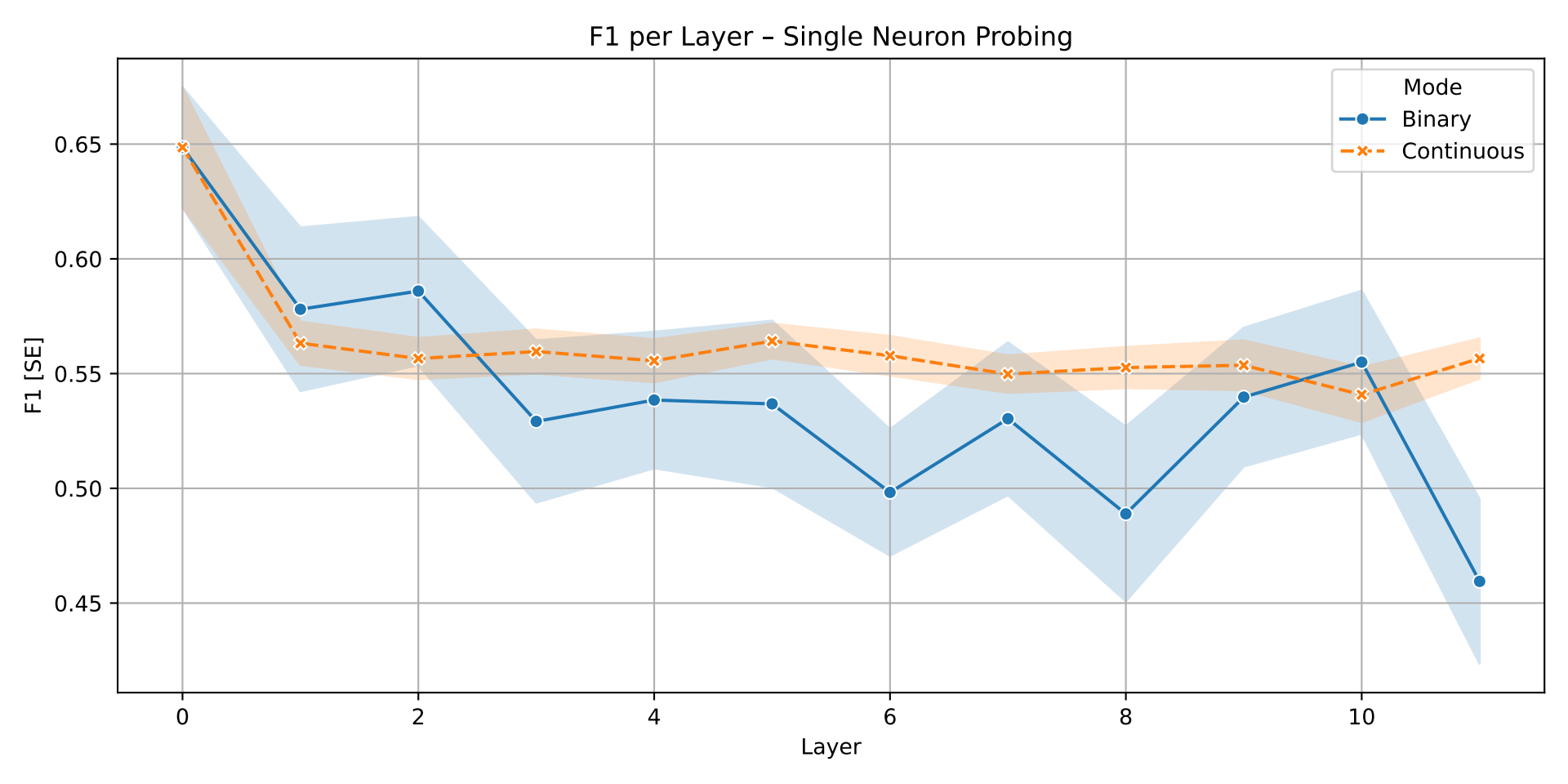
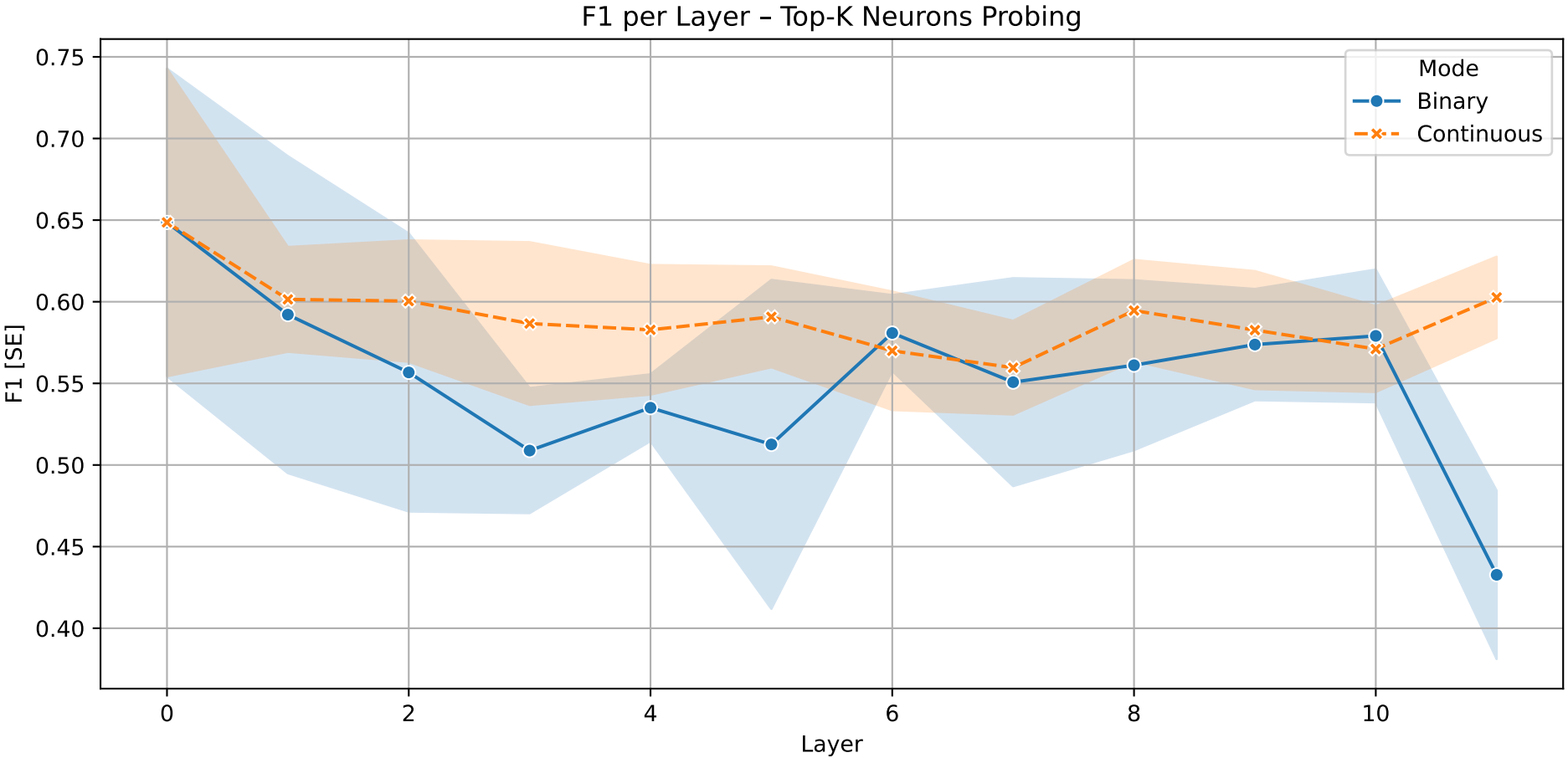
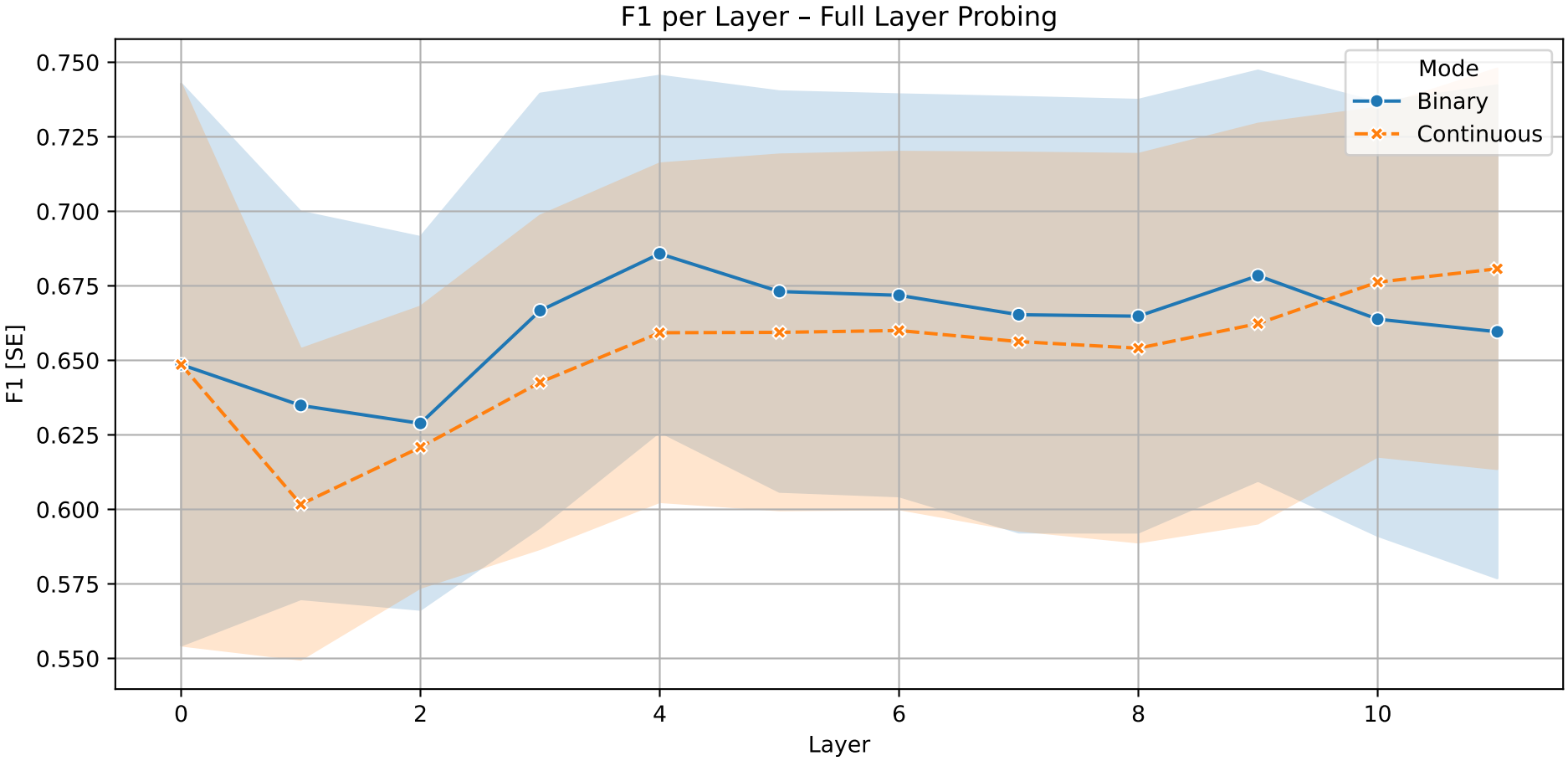
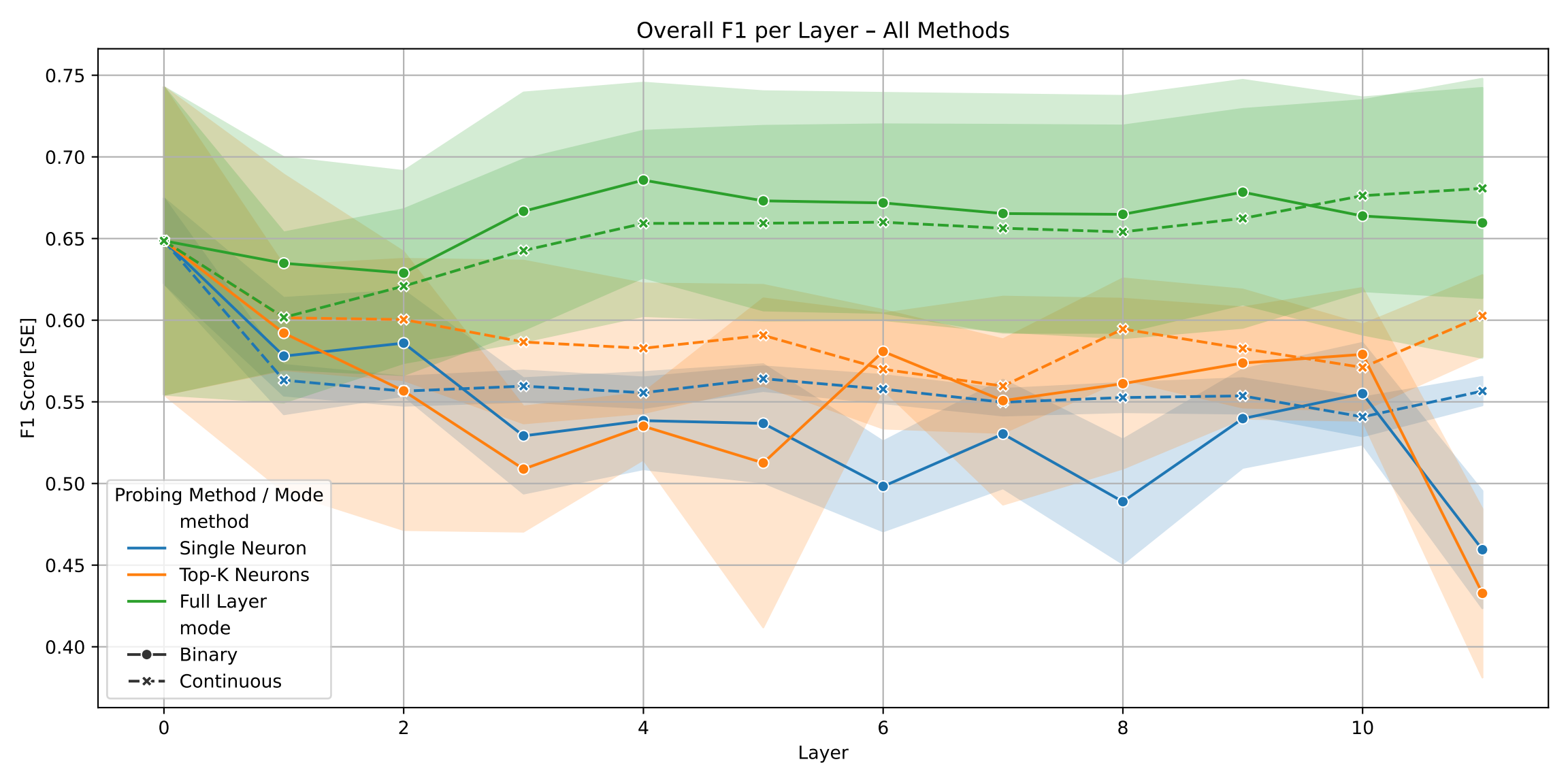

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.65
0.60
0.55
0.50
0.45
0
2
4
6
8
10
Layer

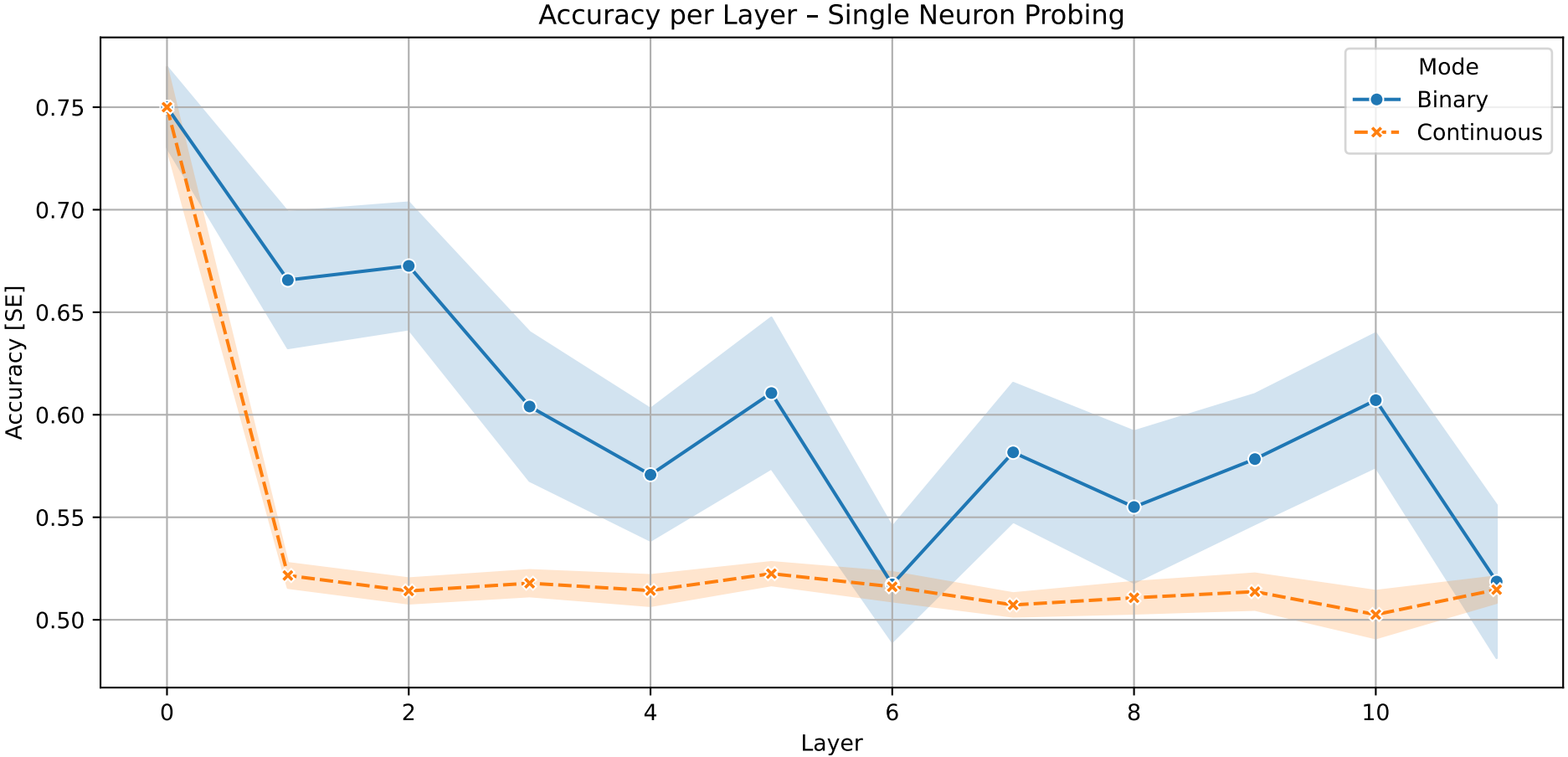

F1 per Layer – Top-K Neurons Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
2
4
6
8
10

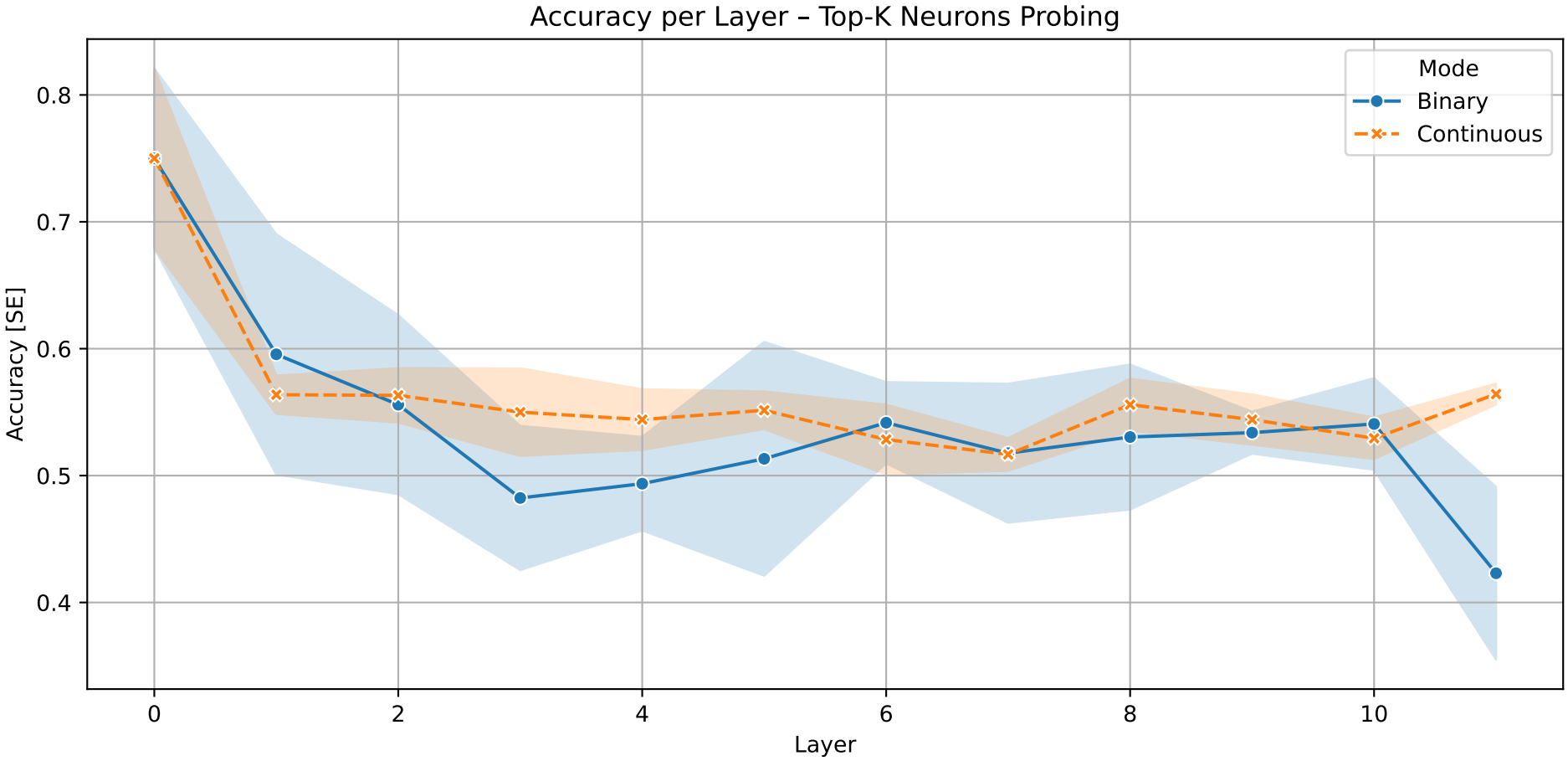

F1 per Layer – Full Layer Probing
F1 [SE]
0.750
0.725
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0
2
4
6
8
10
Layer
Mode
Binary
Continuous

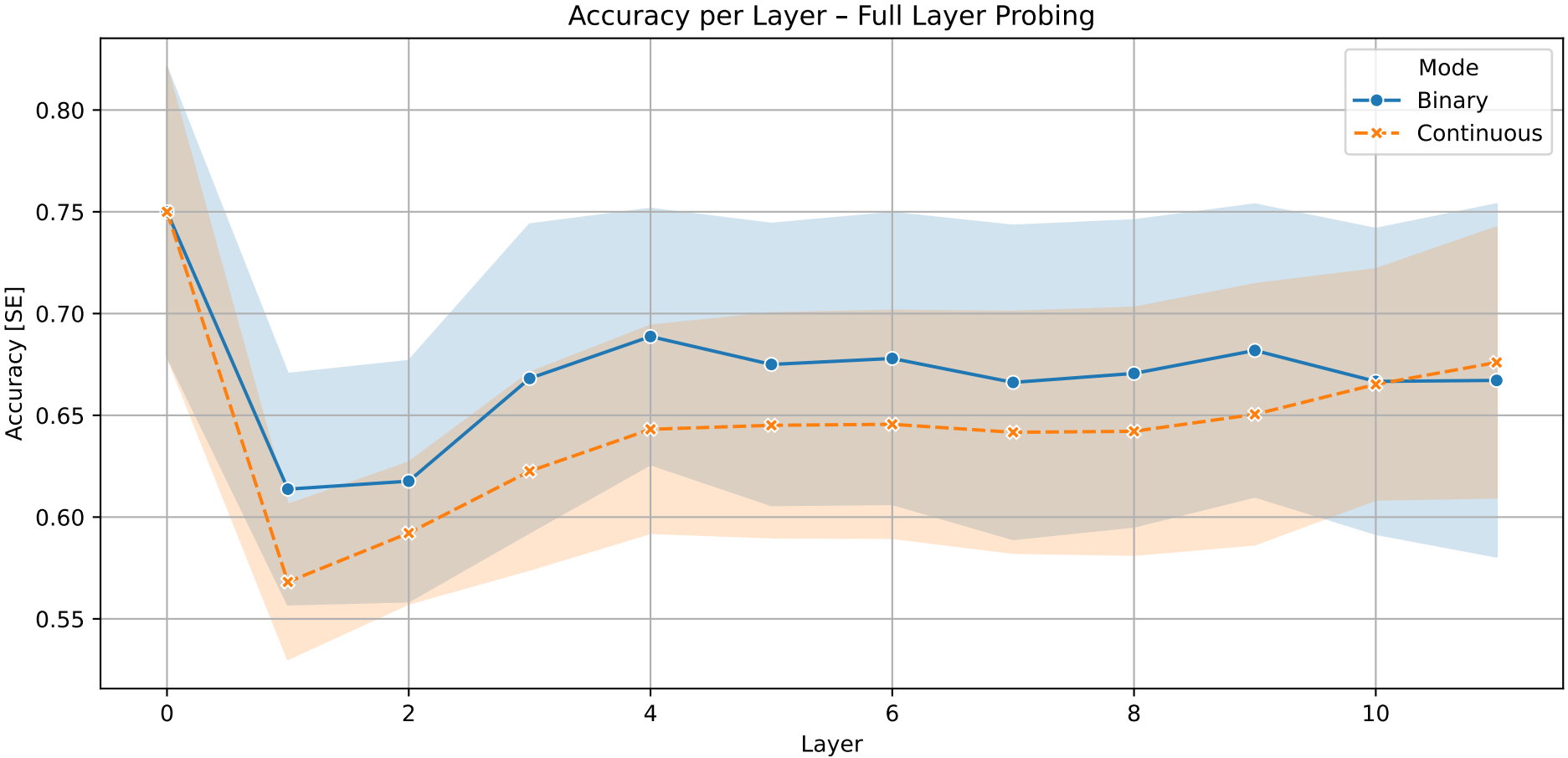

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
2
4
6
8
10
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

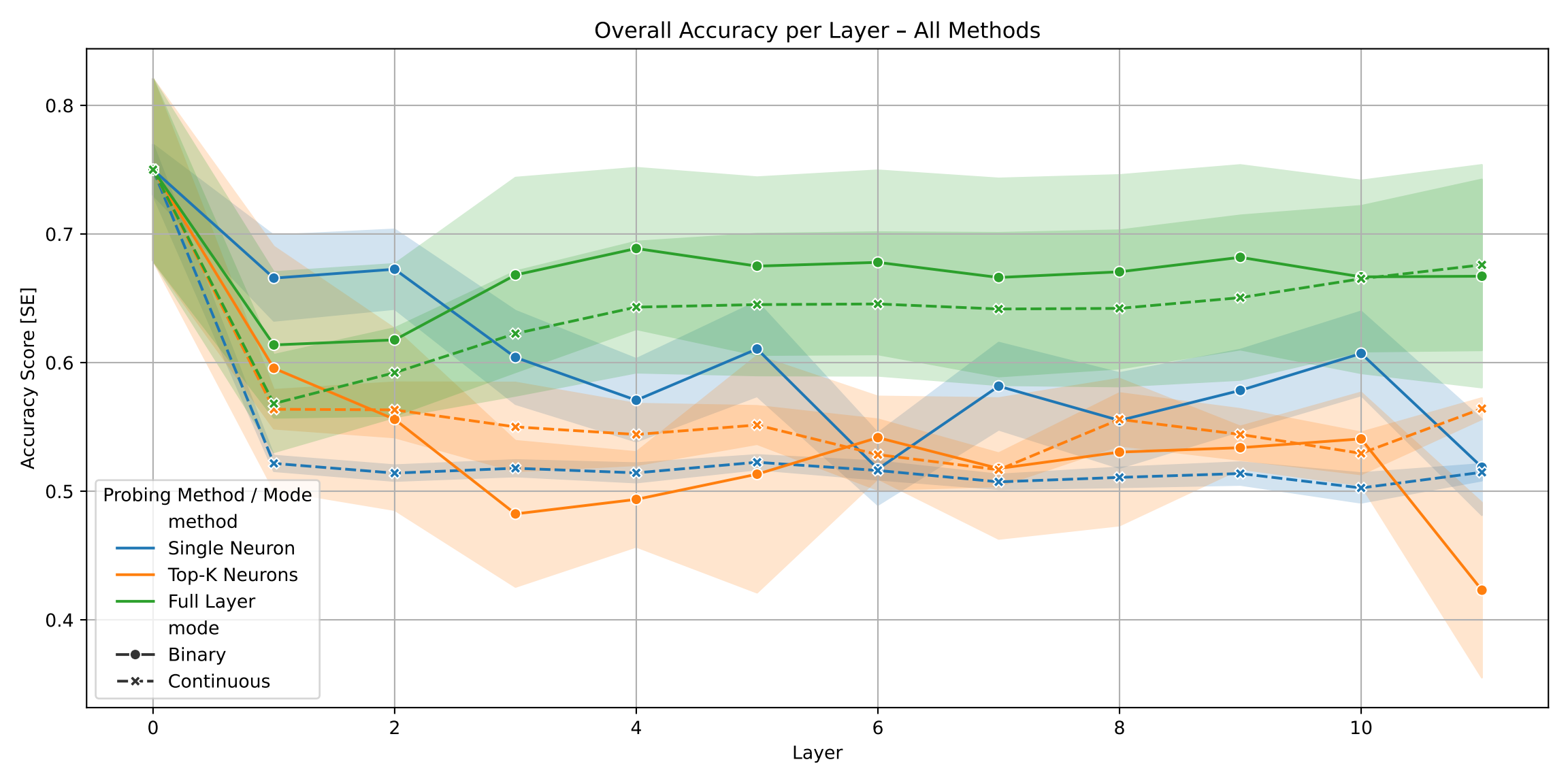
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	4.0	11.0
Full Layer	f1_max	0.8556	0.8526
Full Layer	f1_mean	0.6618	0.6518
Full Layer	f1_std	0.1272	0.1127
Single Neuron	f1_best_layer	0.0	0.0
Single Neuron	f1_max	0.8536	0.8526
Single Neuron	f1_mean	0.5407	0.5632
Single Neuron	f1_std	0.2106	0.0778
Top-K Neurons	f1_best_layer	0.0	0.0
Top-K Neurons	f1_max	0.8526	0.8526
Top-K Neurons	f1_mean	0.5527	0.5909
Top-K Neurons	f1_std	0.1245	0.0781


Accuracy per Layer – Single Neuron Probing
Mode
Binary
Continuous
Accuracy [SE]
0.75
0.70
0.65
0.60
0.55
0.50
0
2
4
6
8
10
Layer


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.8
0.7
0.6
0.5
0.4
0
2
4
6
8
10
Layer


Accuracy per Layer – Full Layer Probing
Mode
Binary
Continuous
Accuracy [SE]
0.80
0.75
0.70
0.65
0.60
0.55
0
2
4
6
8
10
Layer


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
0.8
0.7
0.6
0.5
0.4
0
2
4
6
8
10
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	0.0	0.0
Full Layer	accuracy_max	0.9	0.9
Full Layer	accuracy_mean	0.6703	0.6452
Full Layer	accuracy_std	0.1292	0.107
Single Neuron	accuracy_best_layer	0.0	0.0
Single Neuron	accuracy_max	0.9	0.9
Single Neuron	accuracy_mean	0.6026	0.5338
Single Neuron	accuracy_std	0.2147	0.0865
Top-K Neurons	accuracy_best_layer	0.0	0.0
Top-K Neurons	accuracy_max	0.9	0.9
Top-K Neurons	accuracy_mean	0.5398	0.5634
Top-K Neurons	accuracy_std	0.1319	0.077