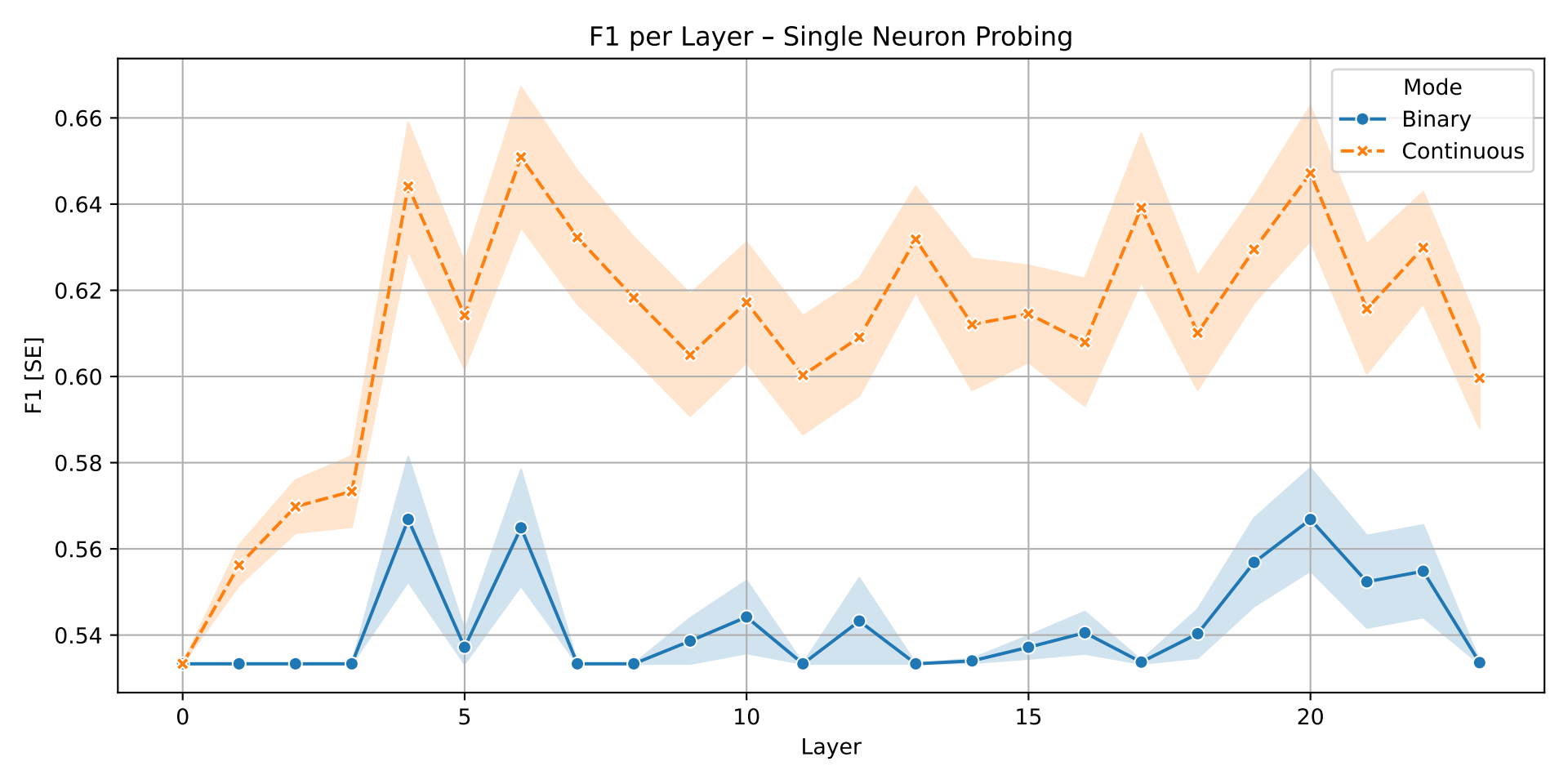
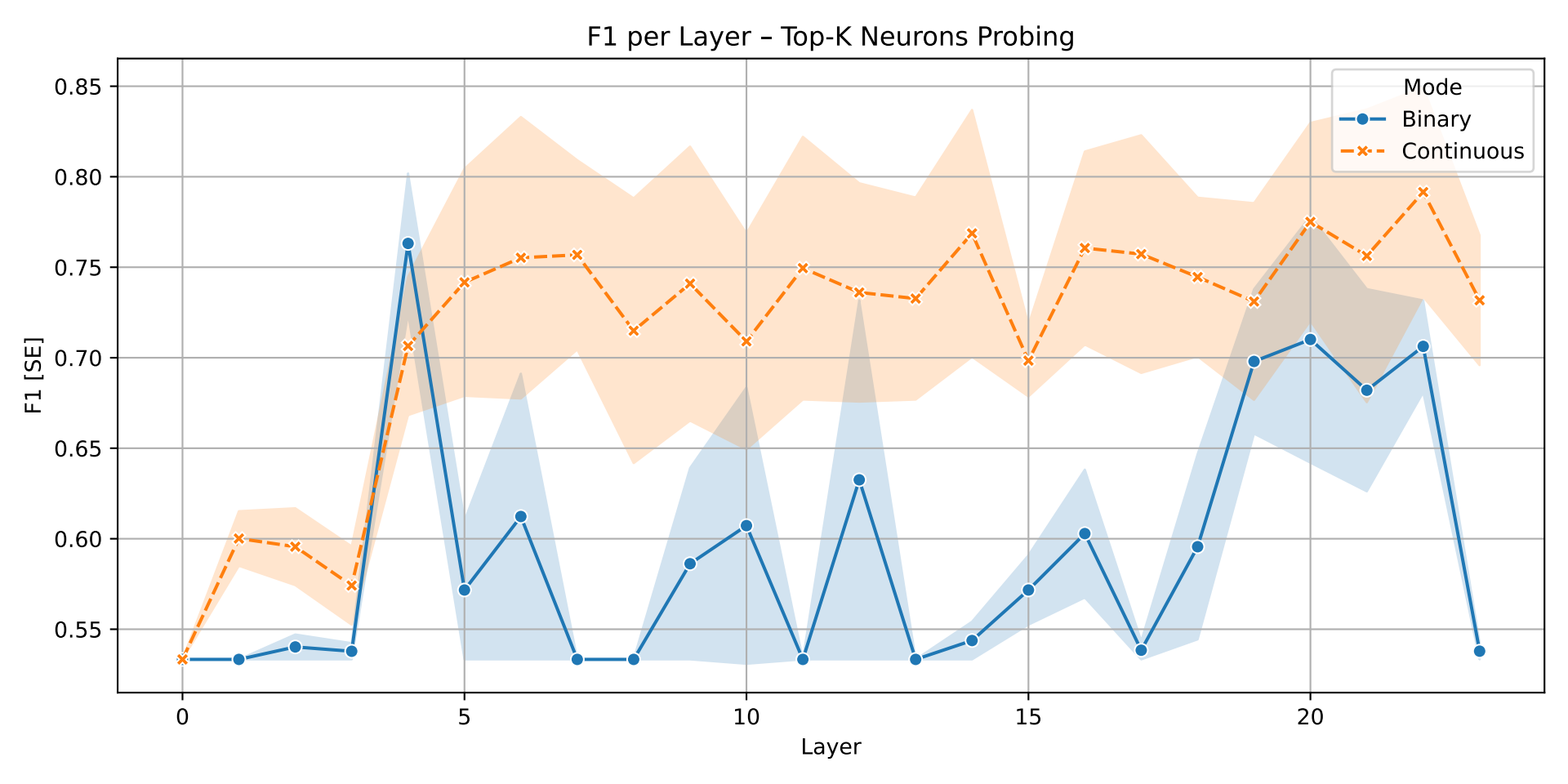
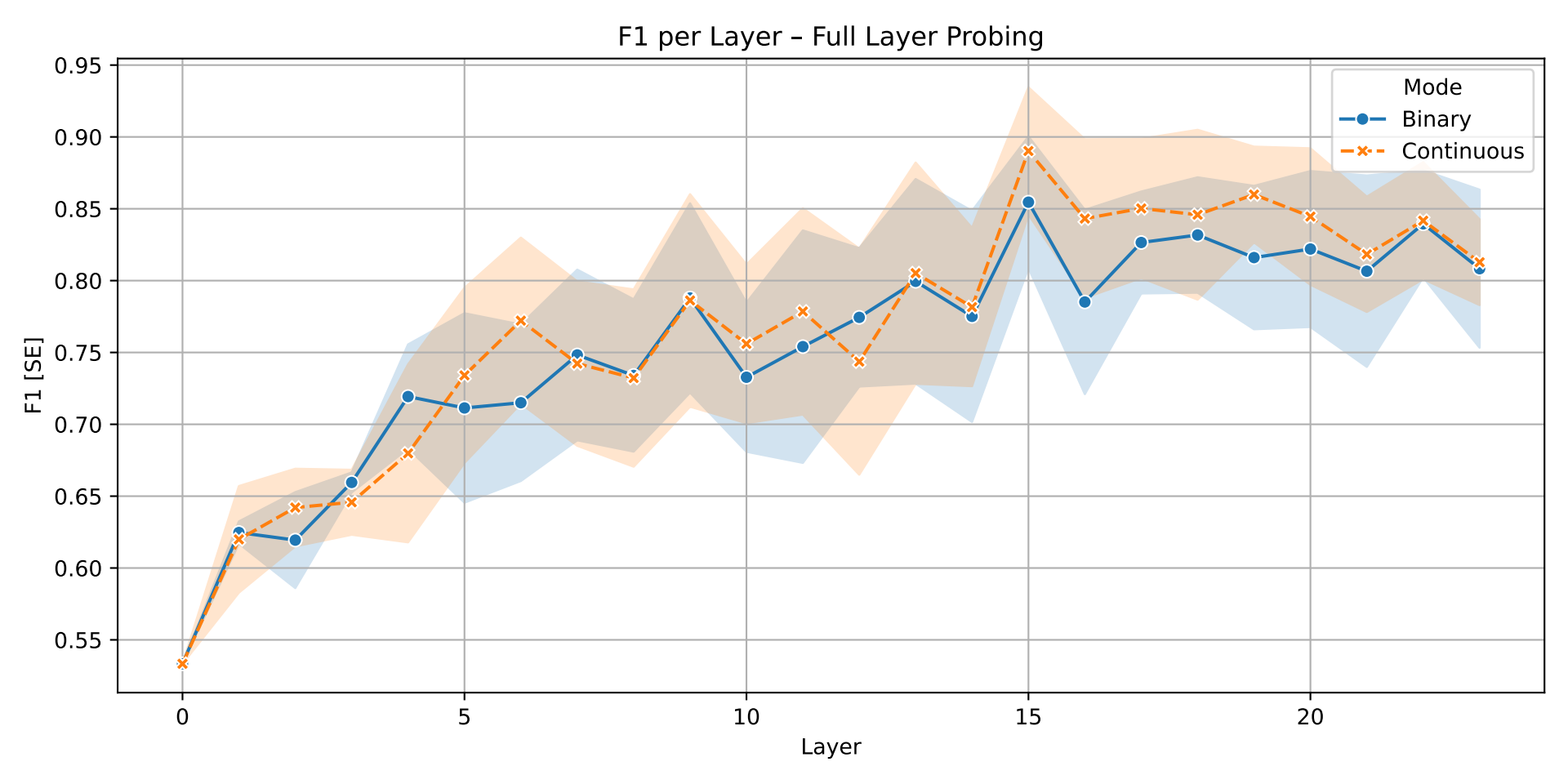
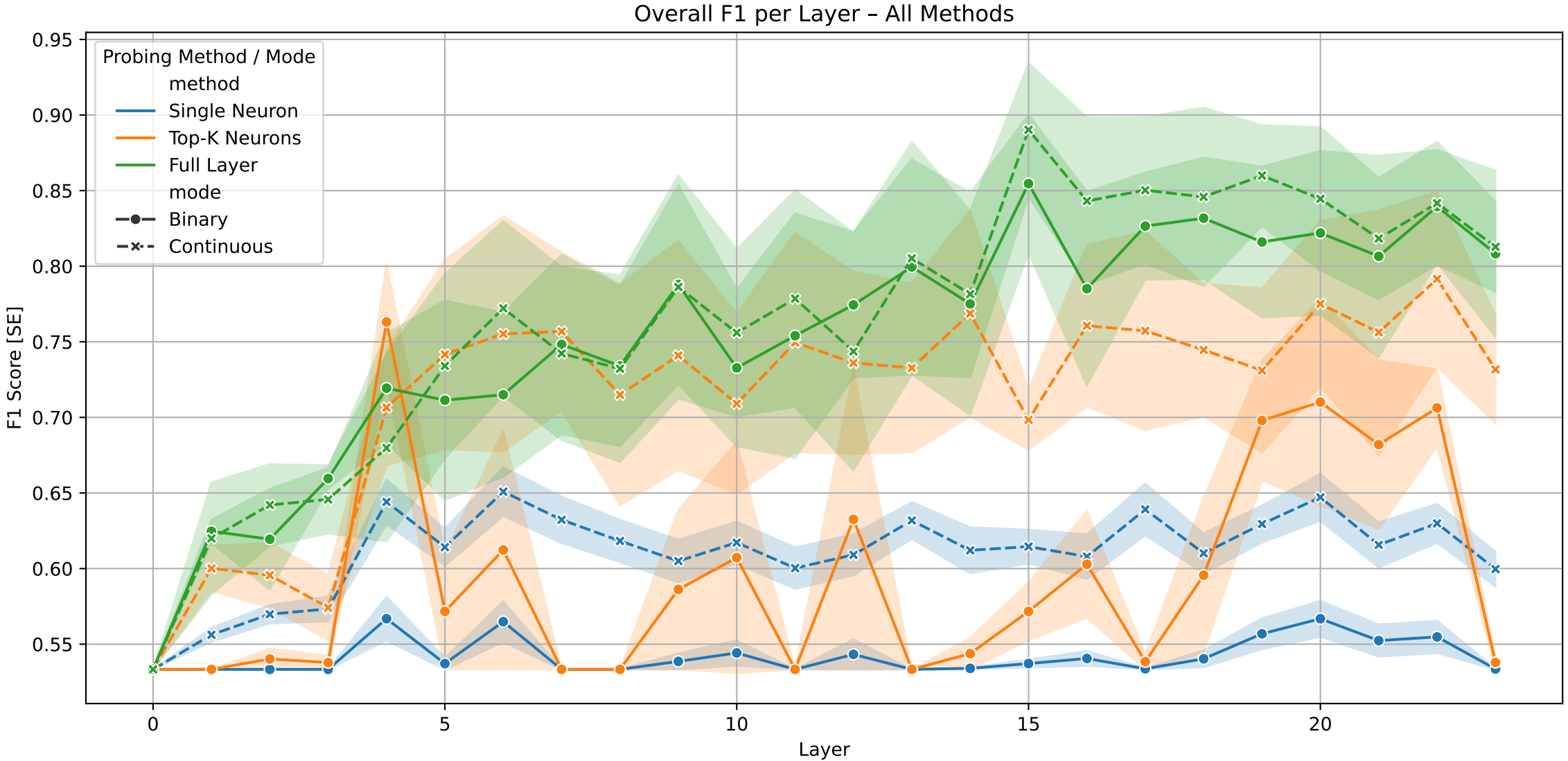

F1 per Layer – Single Neuron Probing
Mode
Binary
Continuous
F1 [SE]
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0
5
10
15
20
Layer

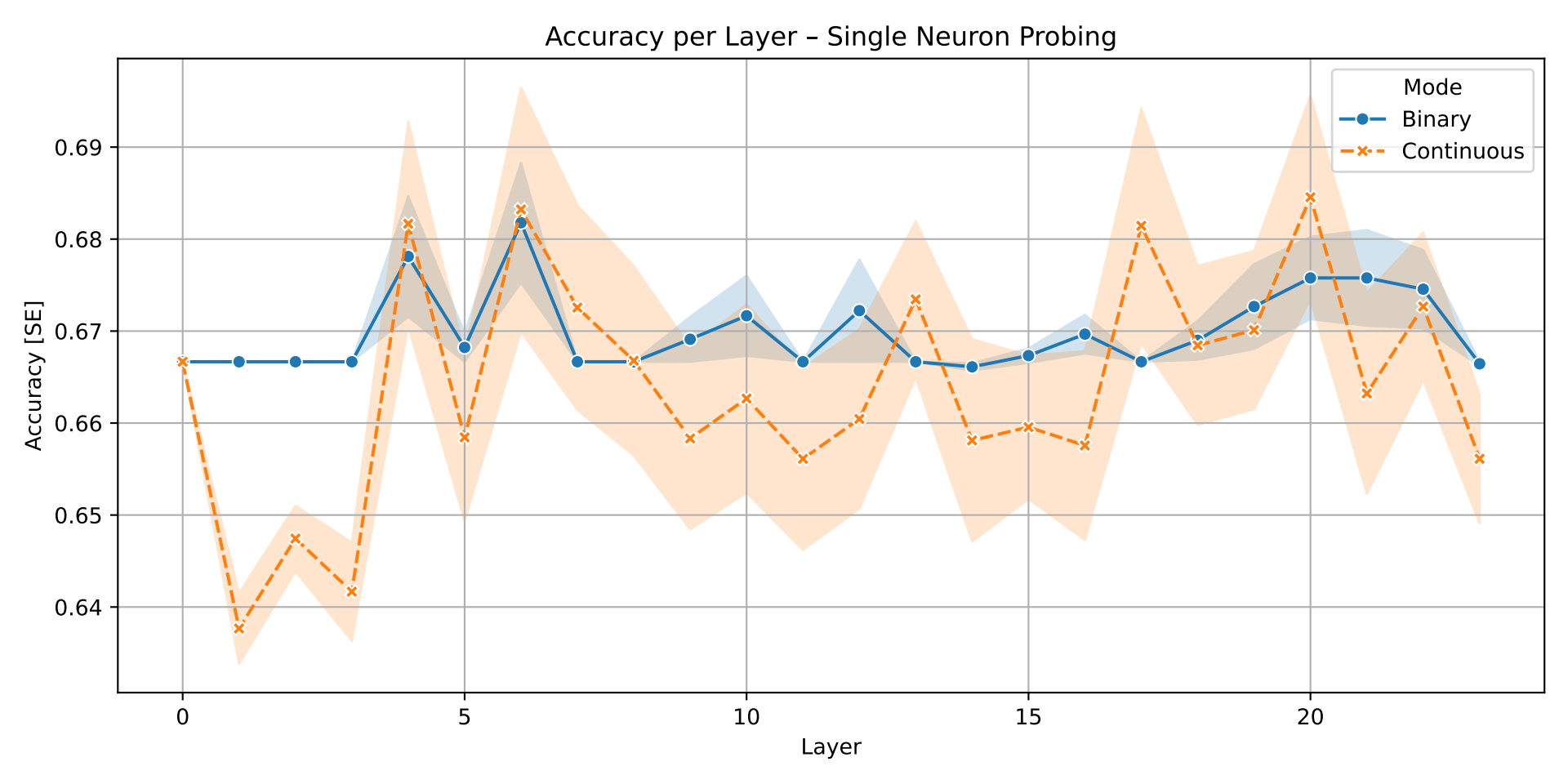

F1 per Layer – Top-K Neurons Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20

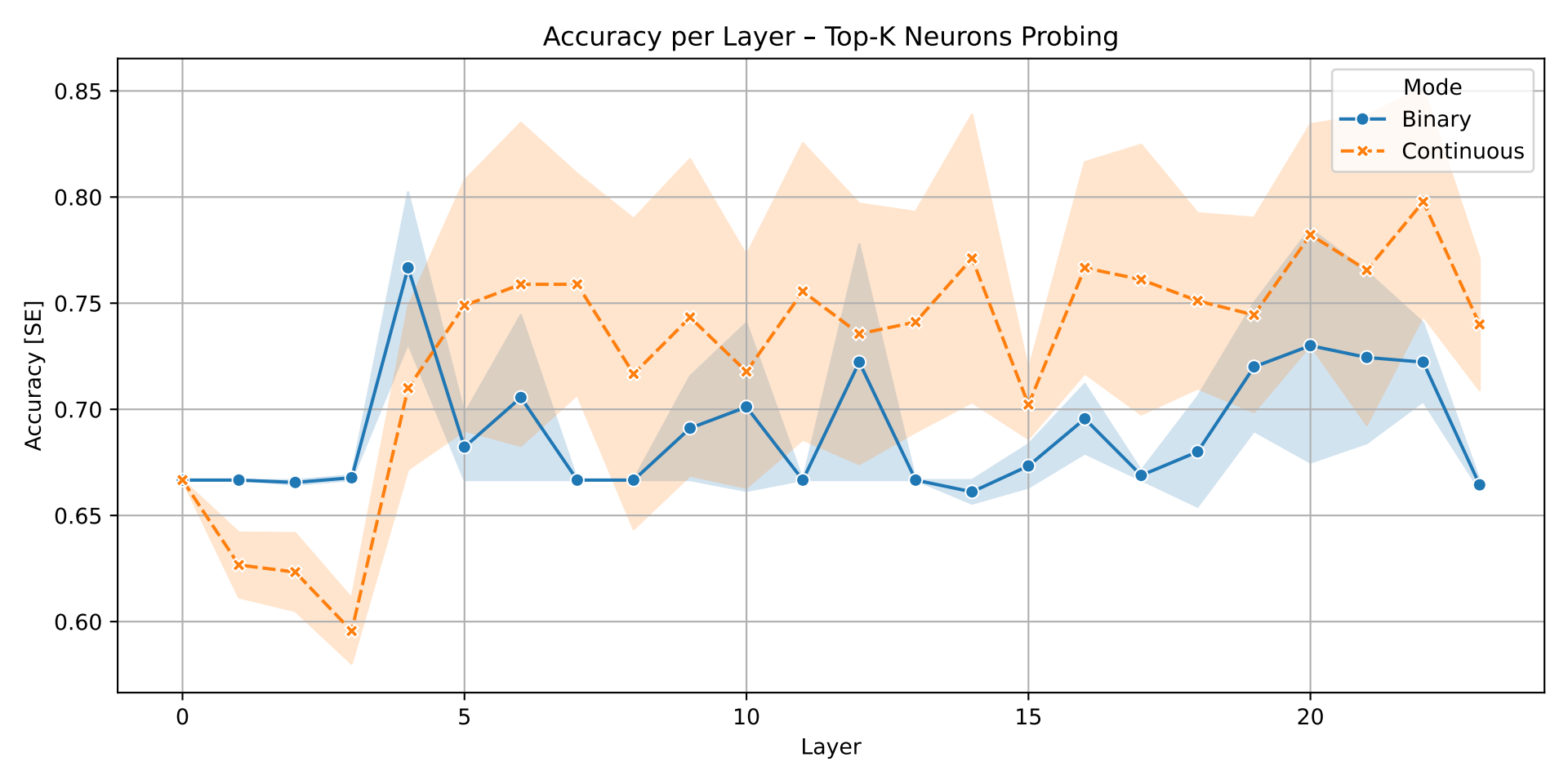

F1 per Layer – Full Layer Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20

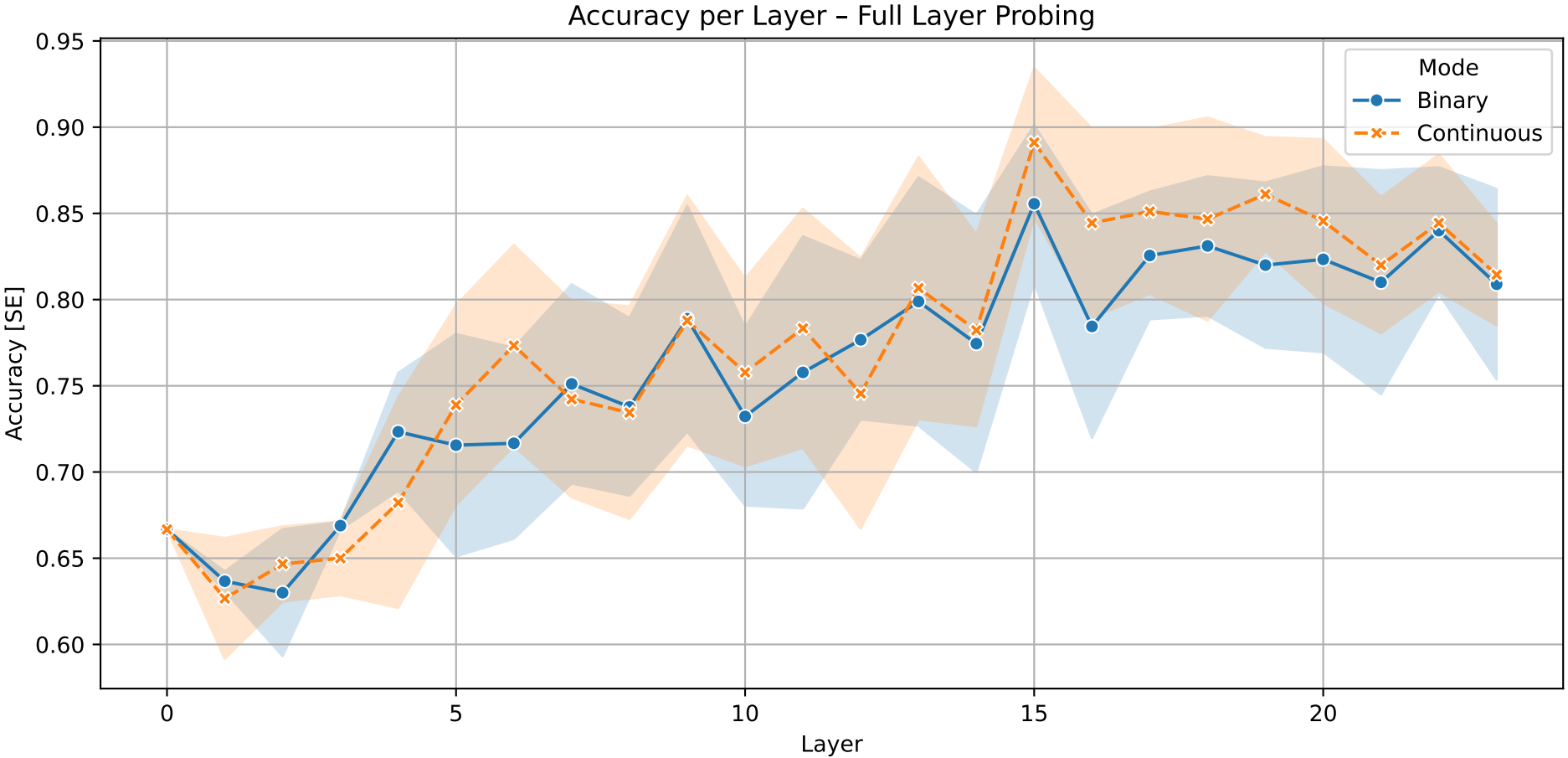

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
Layer
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20

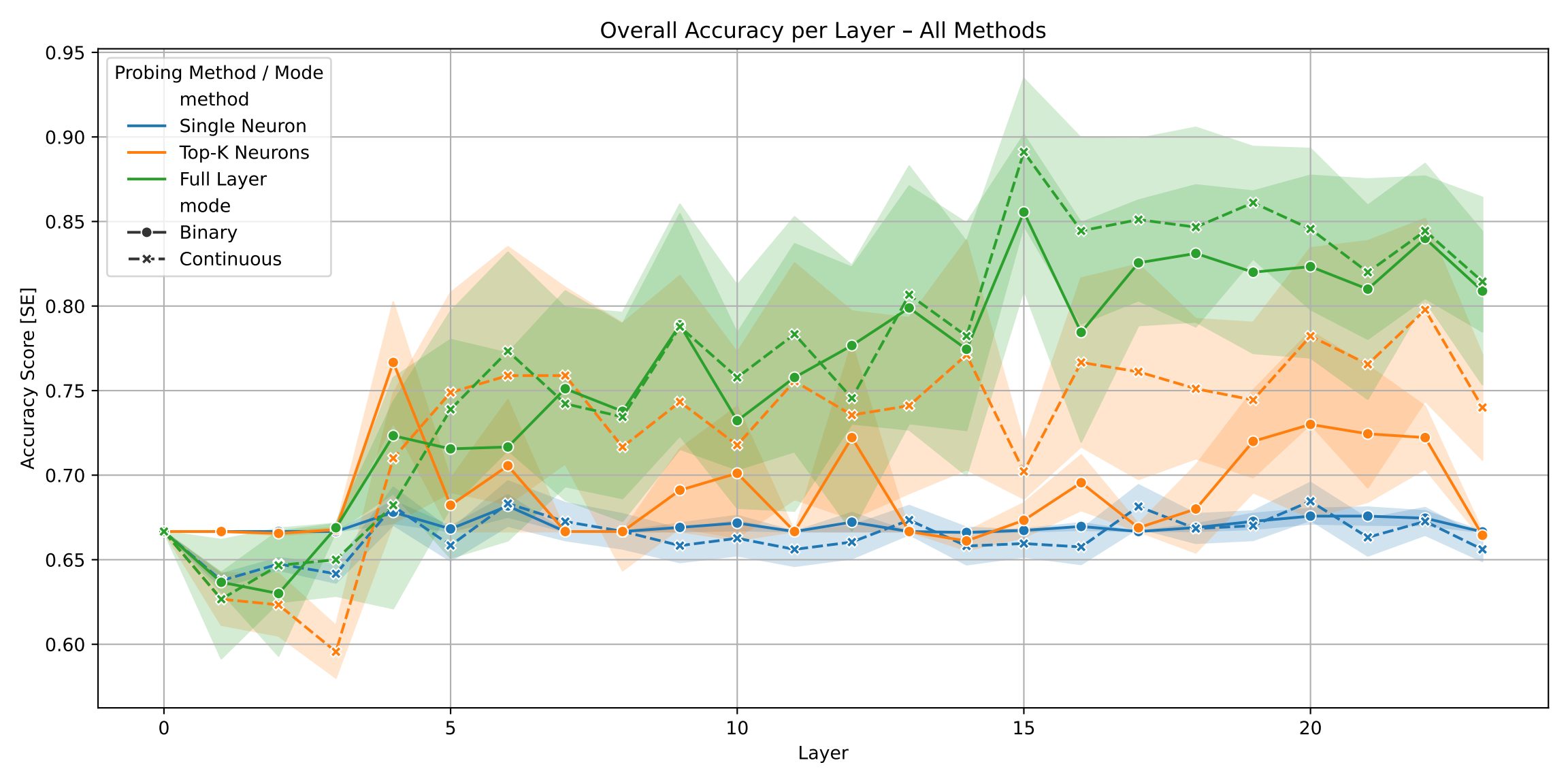
F1 Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	f1_best_layer	15.0	15.0
Full Layer	f1_max	0.9459	0.9666
Full Layer	f1_mean	0.7533	0.765
Full Layer	f1_std	0.1083	0.1153
Single Neuron	f1_best_layer	4.0	6.0
Single Neuron	f1_max	0.831	0.908
Single Neuron	f1_mean	0.5422	0.6109
Single Neuron	f1_std	0.0388	0.0757
Top-K Neurons	f1_best_layer	4.0	22.0
Top-K Neurons	f1_max	0.831	0.9062
Top-K Neurons	f1_mean	0.5932	0.7151
Top-K Neurons	f1_std	0.0914	0.1034


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


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


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

Accuracy Score Summary by Probing Method

Method	Metric	Binary	Continuous
Full Layer	accuracy_best_layer	15.0	15.0
Full Layer	accuracy_max	0.9467	0.9667
Full Layer	accuracy_mean	0.7614	0.7726
Full Layer	accuracy_std	0.0978	0.1053
Single Neuron	accuracy_best_layer	6.0	20.0
Single Neuron	accuracy_max	0.8333	0.91
Single Neuron	accuracy_mean	0.6699	0.6641
Single Neuron	accuracy_std	0.0176	0.0517
Top-K Neurons	accuracy_best_layer	4.0	22.0
Top-K Neurons	accuracy_max	0.8333	0.9067
Top-K Neurons	accuracy_mean	0.6893	0.7284
Top-K Neurons	accuracy_std	0.0456	0.0908