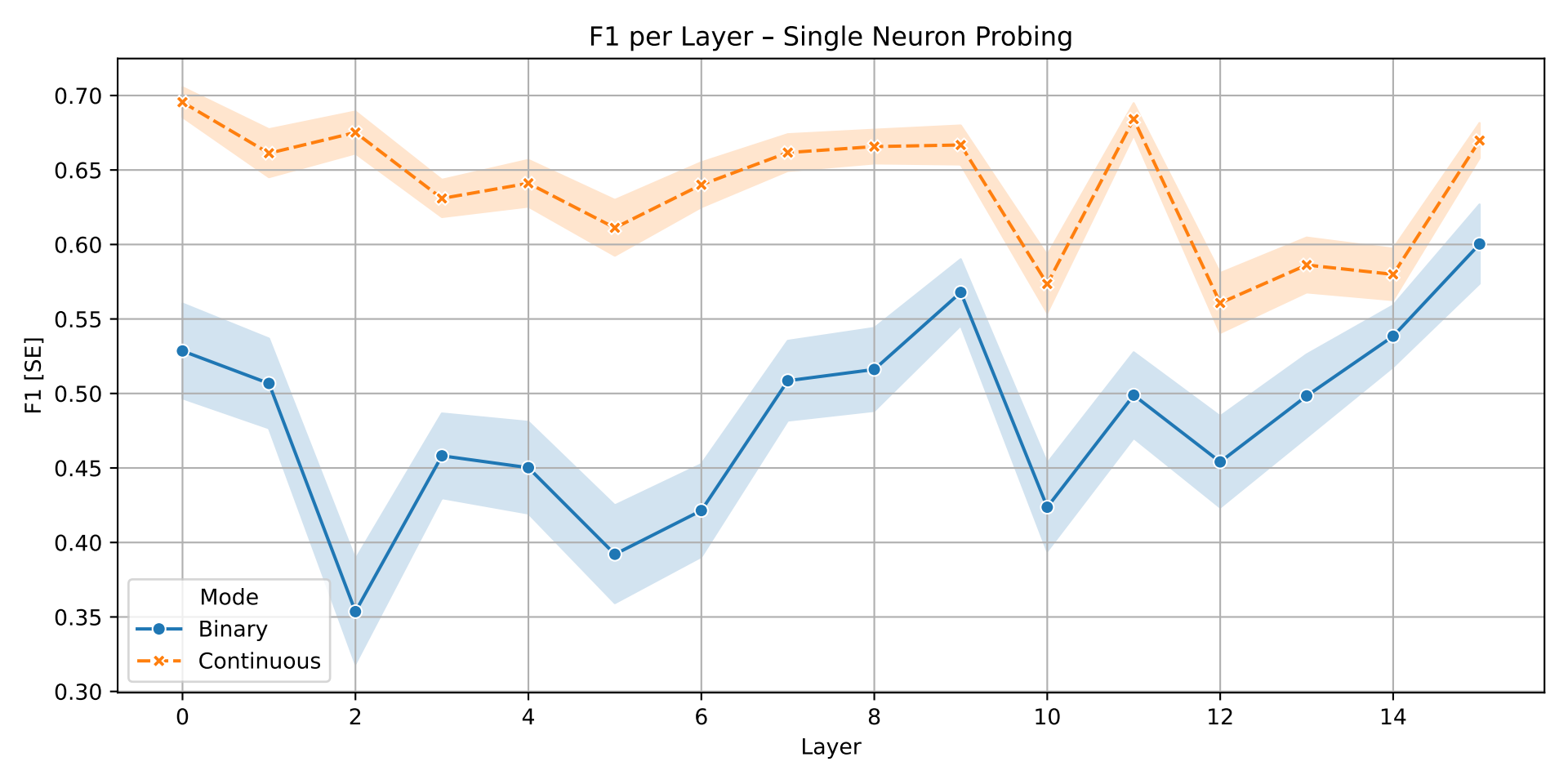
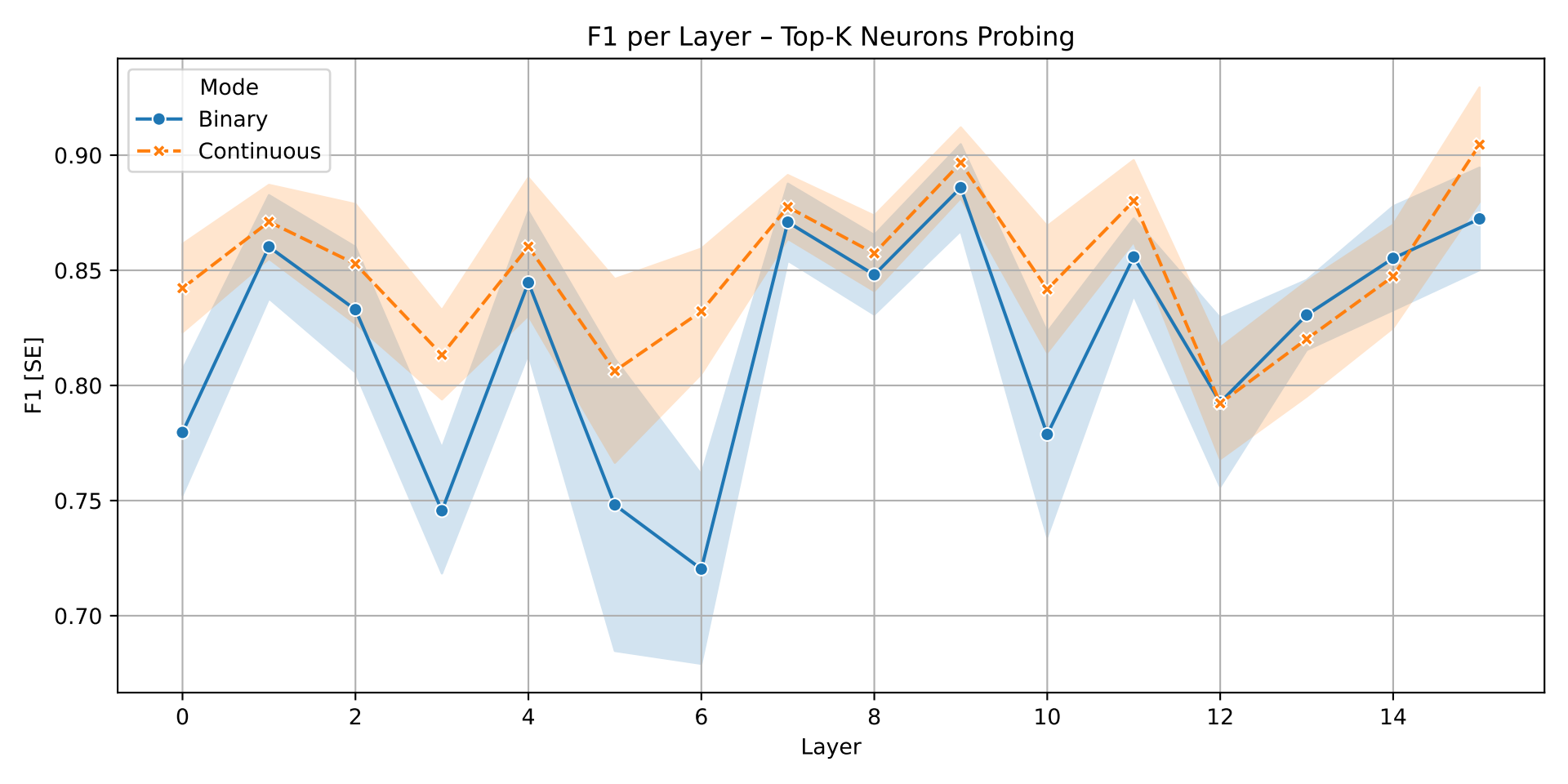
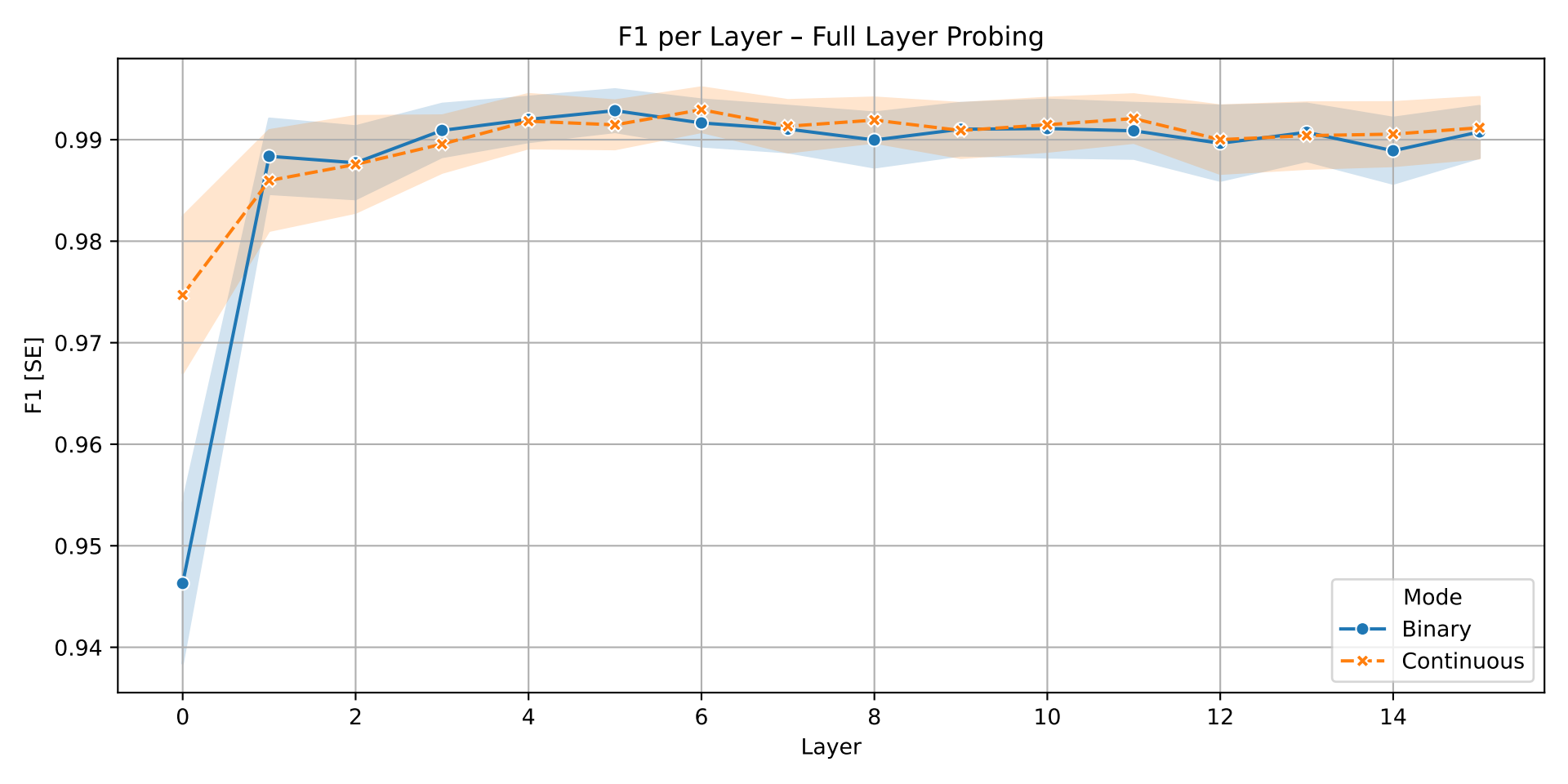
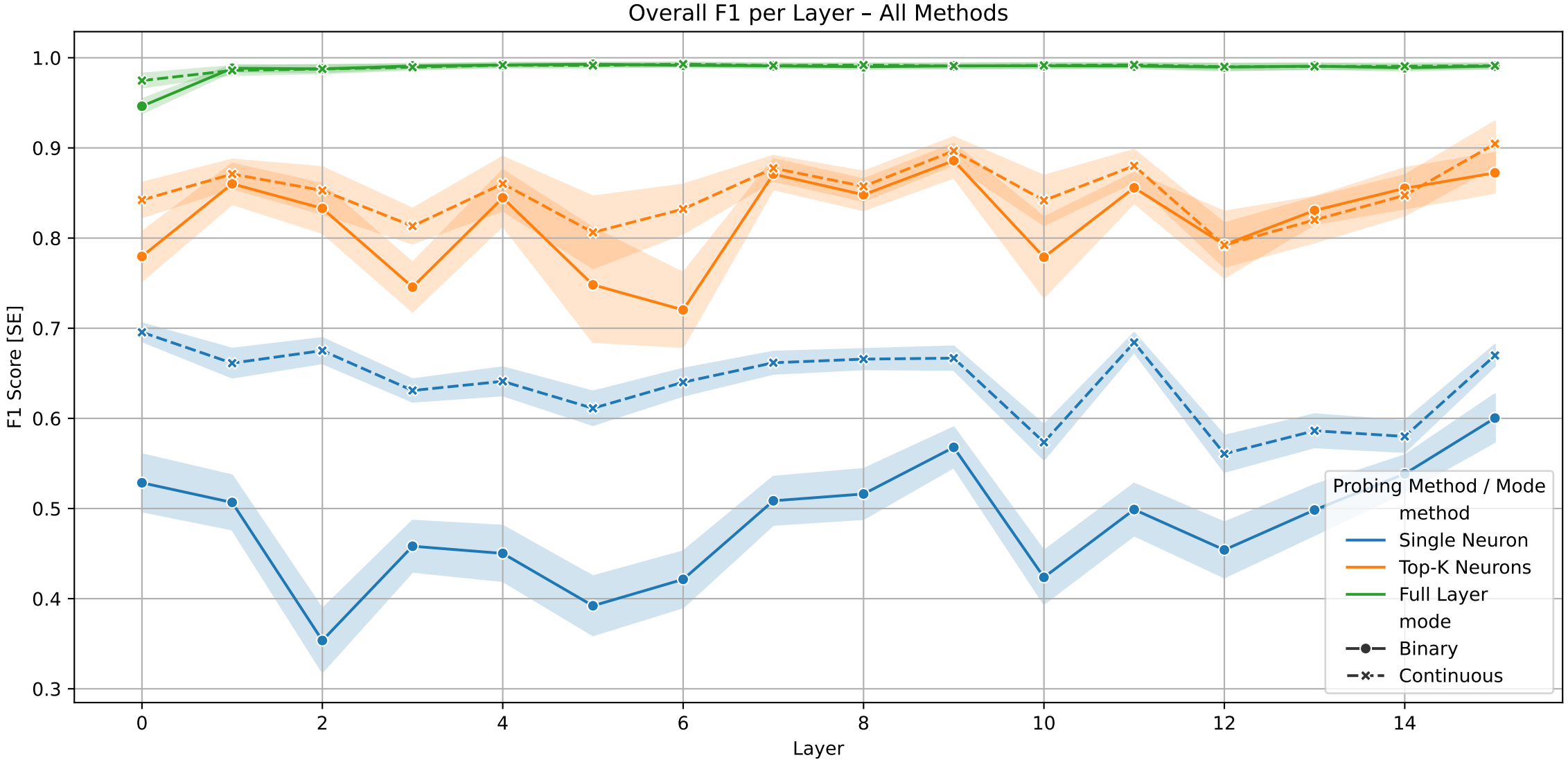

F1 per Layer – Single Neuron Probing
F1 [SE]
Layer
Mode
Binary
Continuous
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0
2
4
6
8
10
12
14

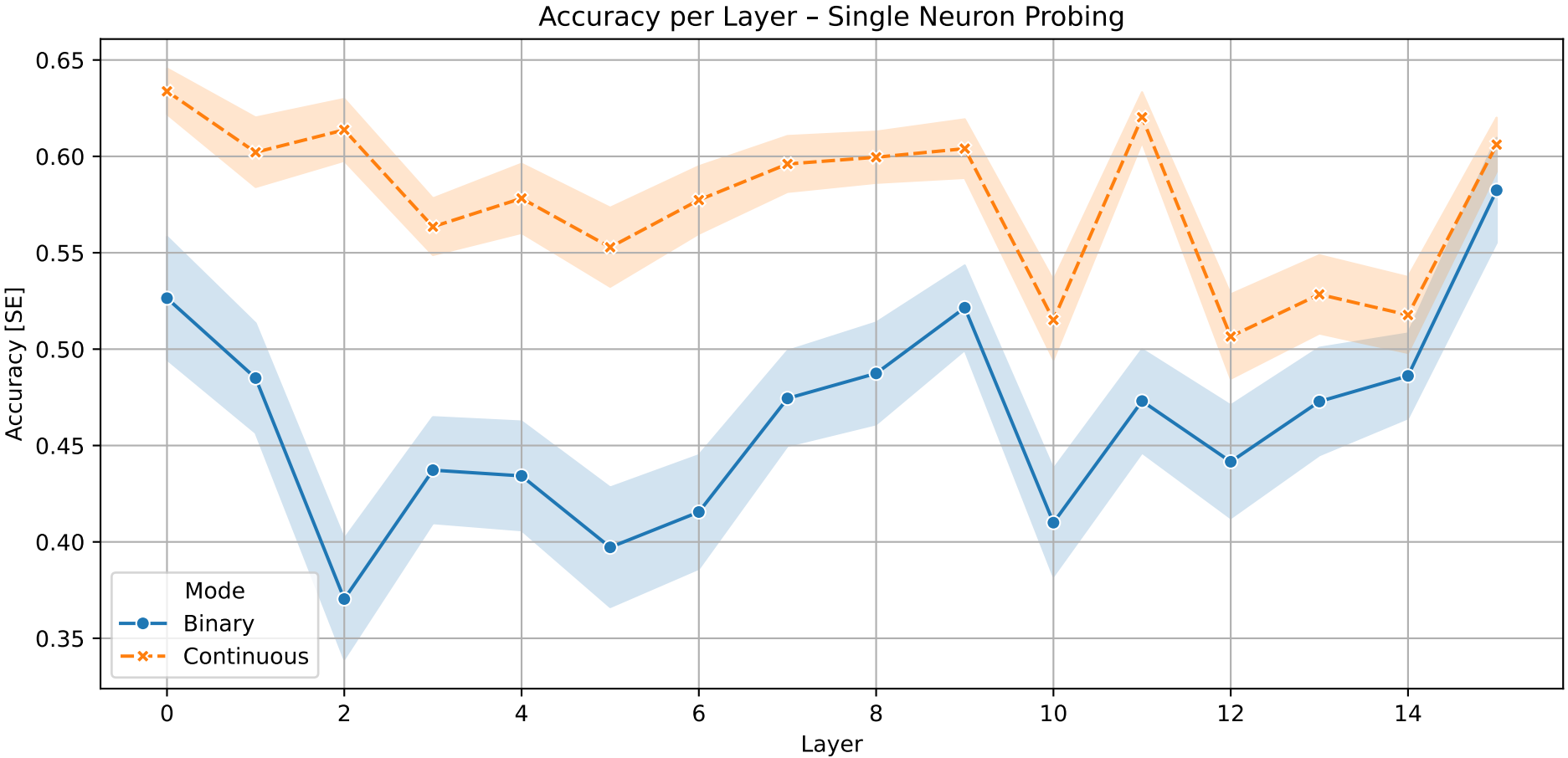

F1 per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
F1 [SE]
0.90
0.85
0.80
0.75
0.70
0
2
4
6
8
10
12
14
Layer

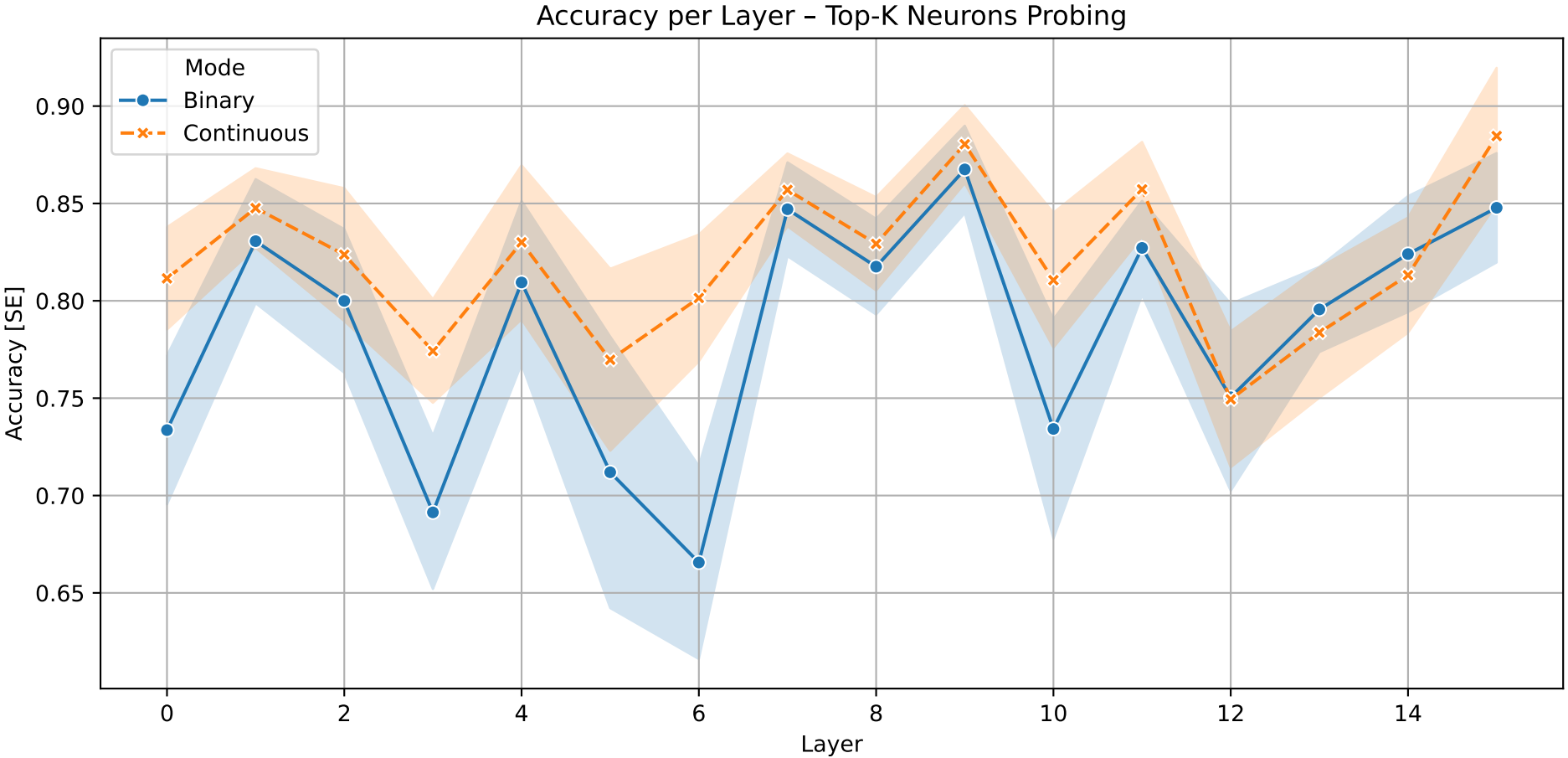

F1 per Layer – Full Layer Probing
F1 [SE]
0.99
0.98
0.97
0.96
0.95
0.94
0
2
4
6
8
10
12
14
Layer
Mode
Binary
Continuous

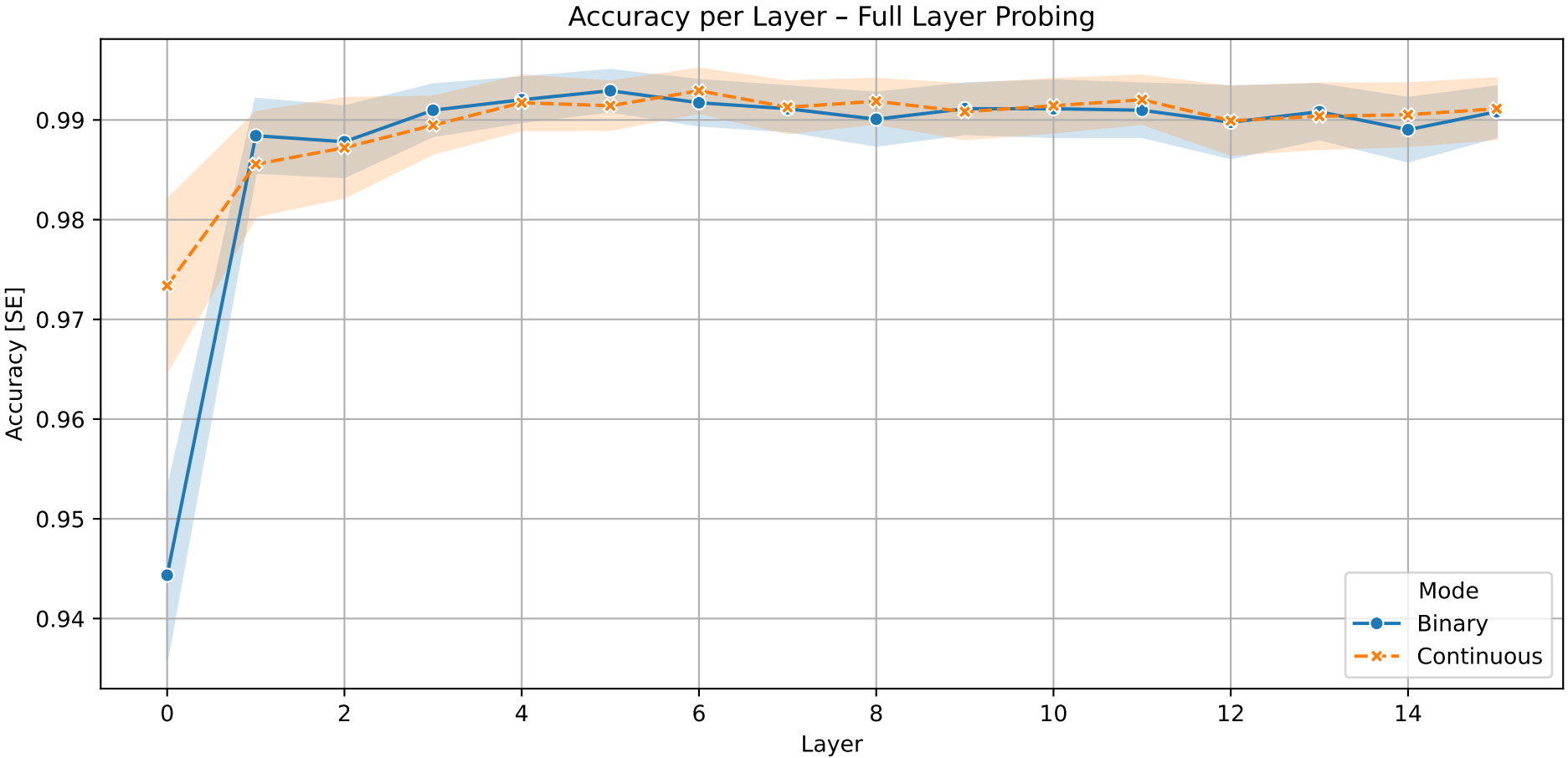

Overall F1 per Layer – All Methods
F1 Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
2
4
6
8
10
12
14
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous

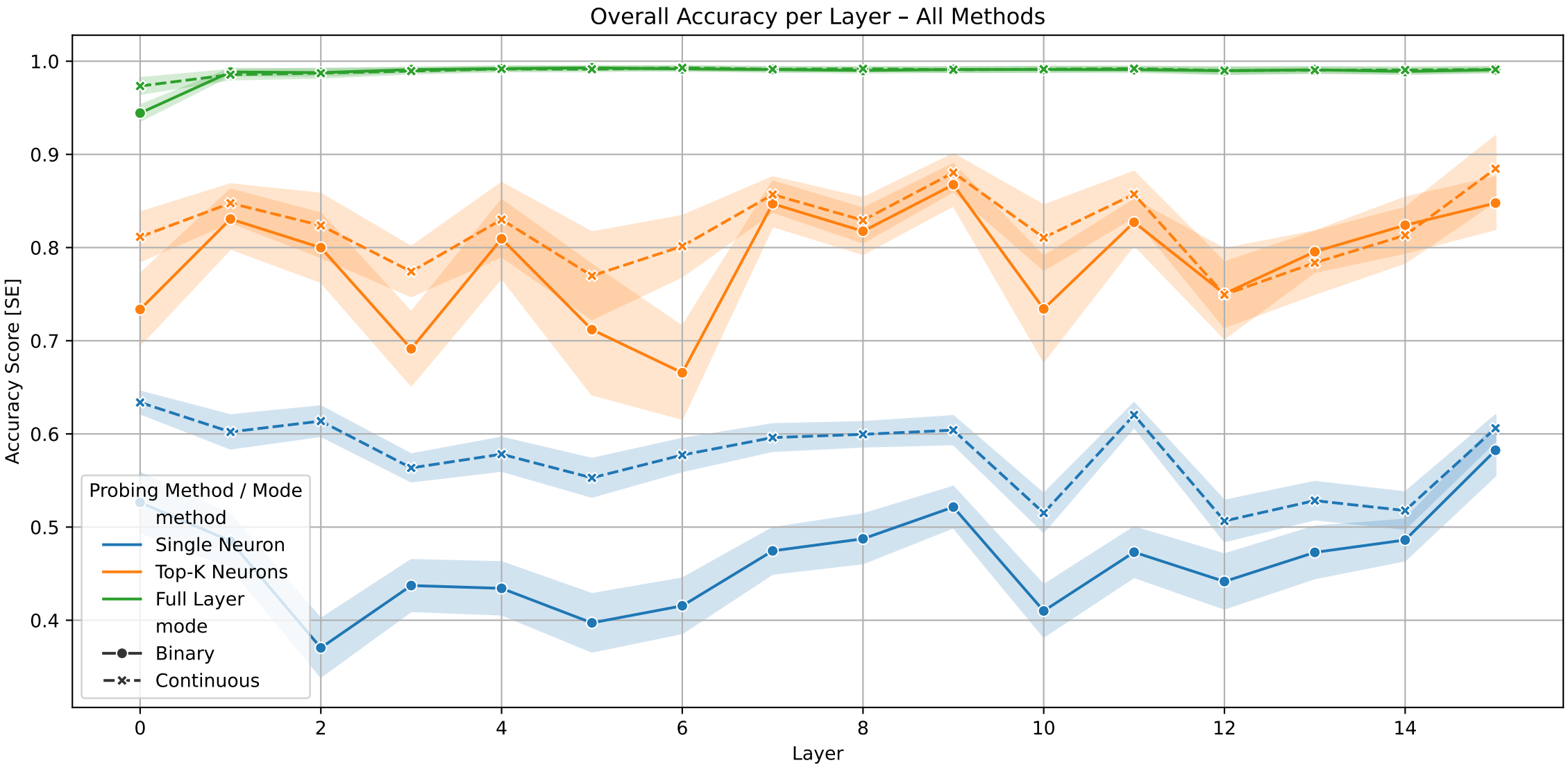
F1 Score Summary by Probing Method

| Method | Metric | Binary | Continuous |
|---|---|---|---|
| Full Layer | f1_best_layer | 5.0 | 6.0 |
| Full Layer | f1_max | 0.9988 | 1.0 |
| Full Layer | f1_mean | 0.9877 | 0.9896 |
| Full Layer | f1_std | 0.0141 | 0.0104 |
| Single Neuron | f1_best_layer | 15.0 | 0.0 |
| Single Neuron | f1_max | 0.9976 | 0.9903 |
| Single Neuron | f1_mean | 0.4823 | 0.6377 |
| Single Neuron | f1_std | 0.2653 | 0.1402 |
| Top-K Neurons | f1_best_layer | 9.0 | 15.0 |
| Top-K Neurons | f1_max | 0.9976 | 0.994 |
| Top-K Neurons | f1_mean | 0.8201 | 0.8497 |
| Top-K Neurons | f1_std | 0.096 | 0.0703 |


Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
Layer
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0
2
4
6
8
10
12
14
Mode
Binary
Continuous


Accuracy per Layer – Top-K Neurons Probing
Mode
Binary
Continuous
Accuracy [SE]
0.90
0.85
0.80
0.75
0.70
0.65
0
2
4
6
8
10
12
14
Layer


Accuracy per Layer – Full Layer Probing
Accuracy [SE]
0.99
0.98
0.97
0.96
0.95
0.94
0
2
4
6
8
10
12
14
Layer
Mode
Binary
Continuous


Overall Accuracy per Layer – All Methods
Accuracy Score [SE]
Layer
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
2
4
6
8
10
12
14
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 | 5.0 | 6.0 |
| Full Layer | accuracy_max | 0.9988 | 1.0 |
| Full Layer | accuracy_mean | 0.9877 | 0.9894 |
| Full Layer | accuracy_std | 0.0145 | 0.011 |
| Single Neuron | accuracy_best_layer | 15.0 | 0.0 |
| Single Neuron | accuracy_max | 0.9976 | 0.9904 |
| Single Neuron | accuracy_mean | 0.4635 | 0.576 |
| Single Neuron | accuracy_std | 0.2518 | 0.1567 |
| Top-K Neurons | accuracy_best_layer | 9.0 | 15.0 |
| Top-K Neurons | accuracy_max | 0.9976 | 0.994 |
| Top-K Neurons | accuracy_mean | 0.7846 | 0.8202 |
| Top-K Neurons | accuracy_std | 0.1197 | 0.0907 |