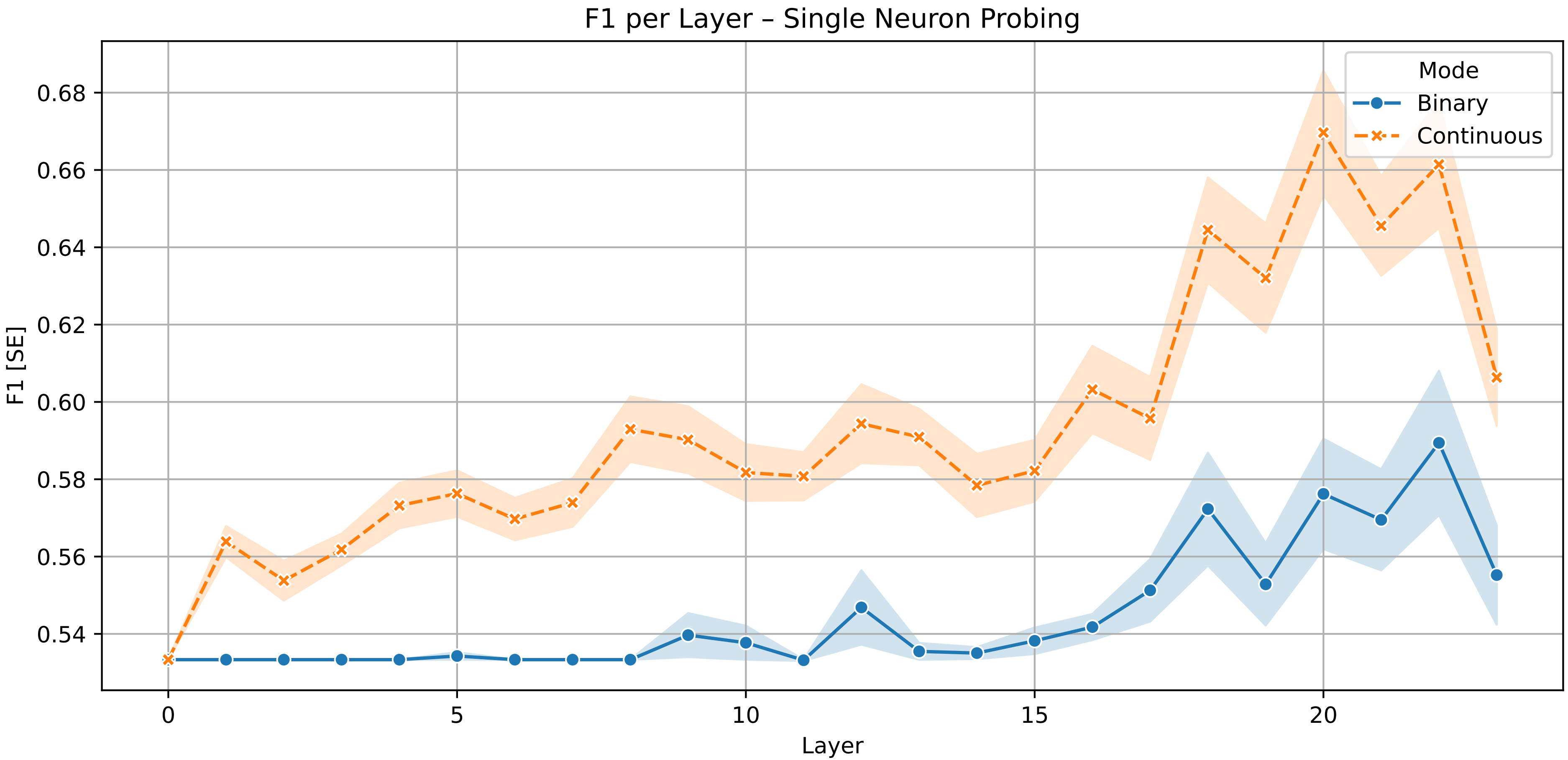
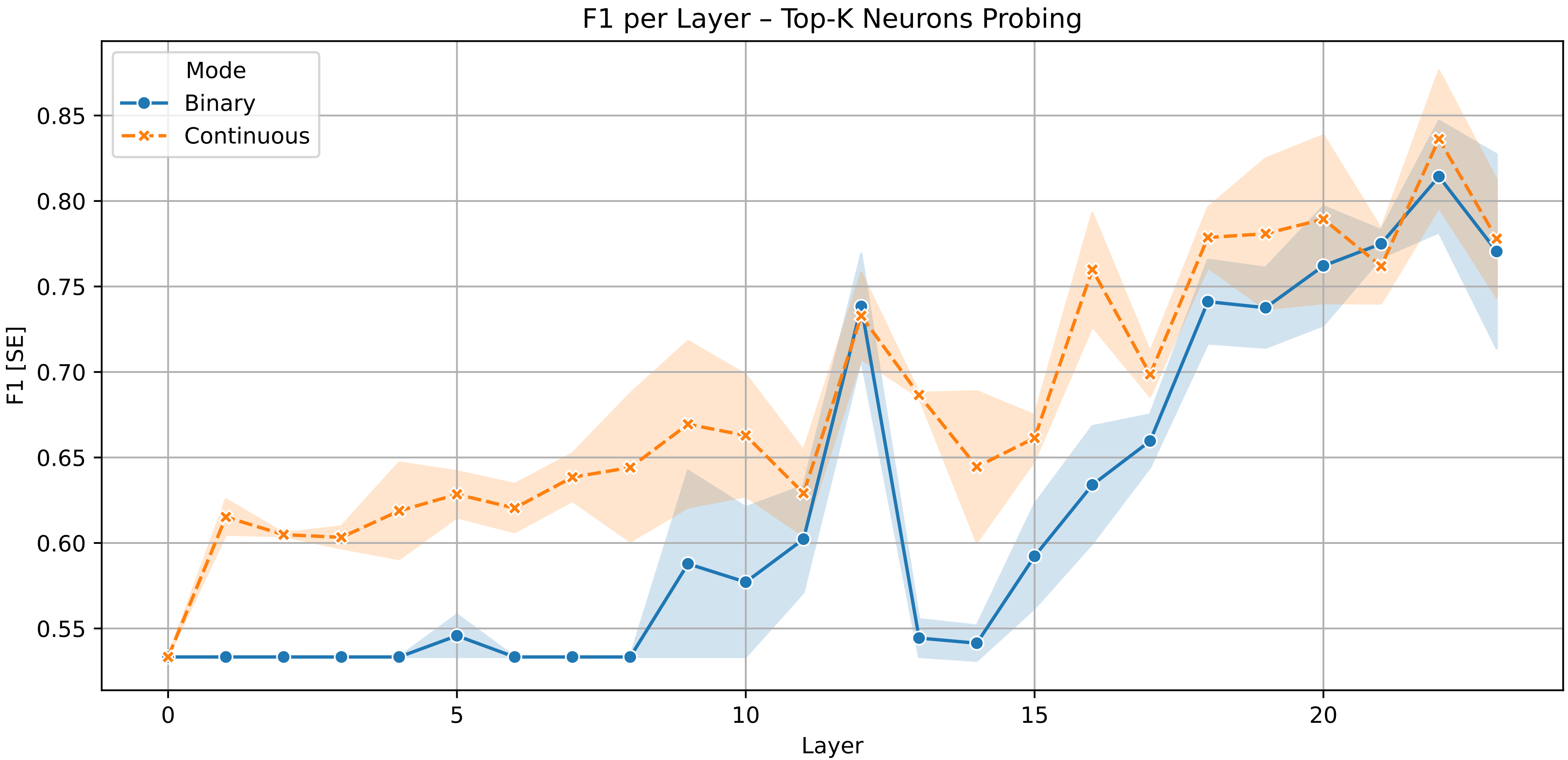
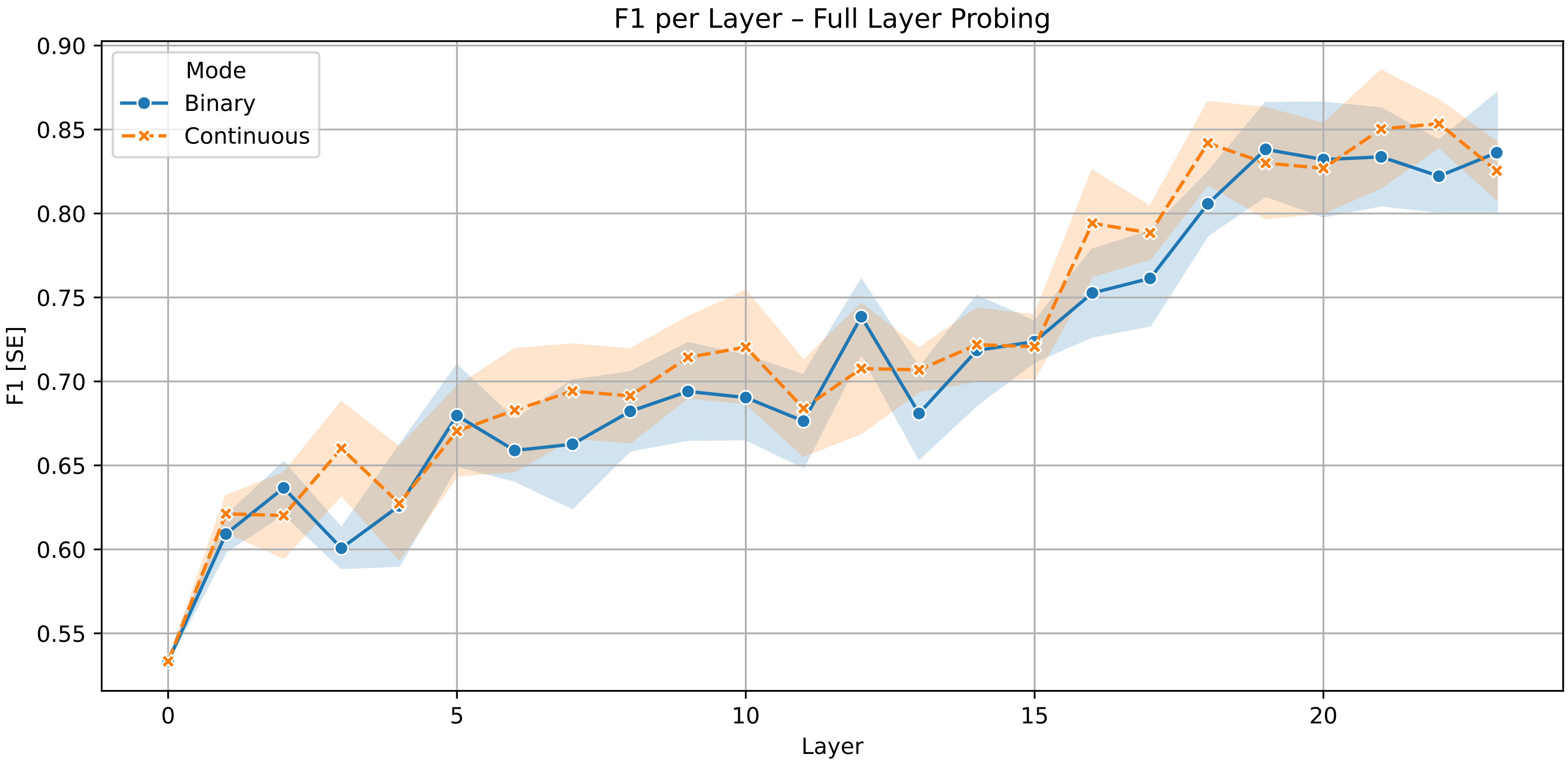
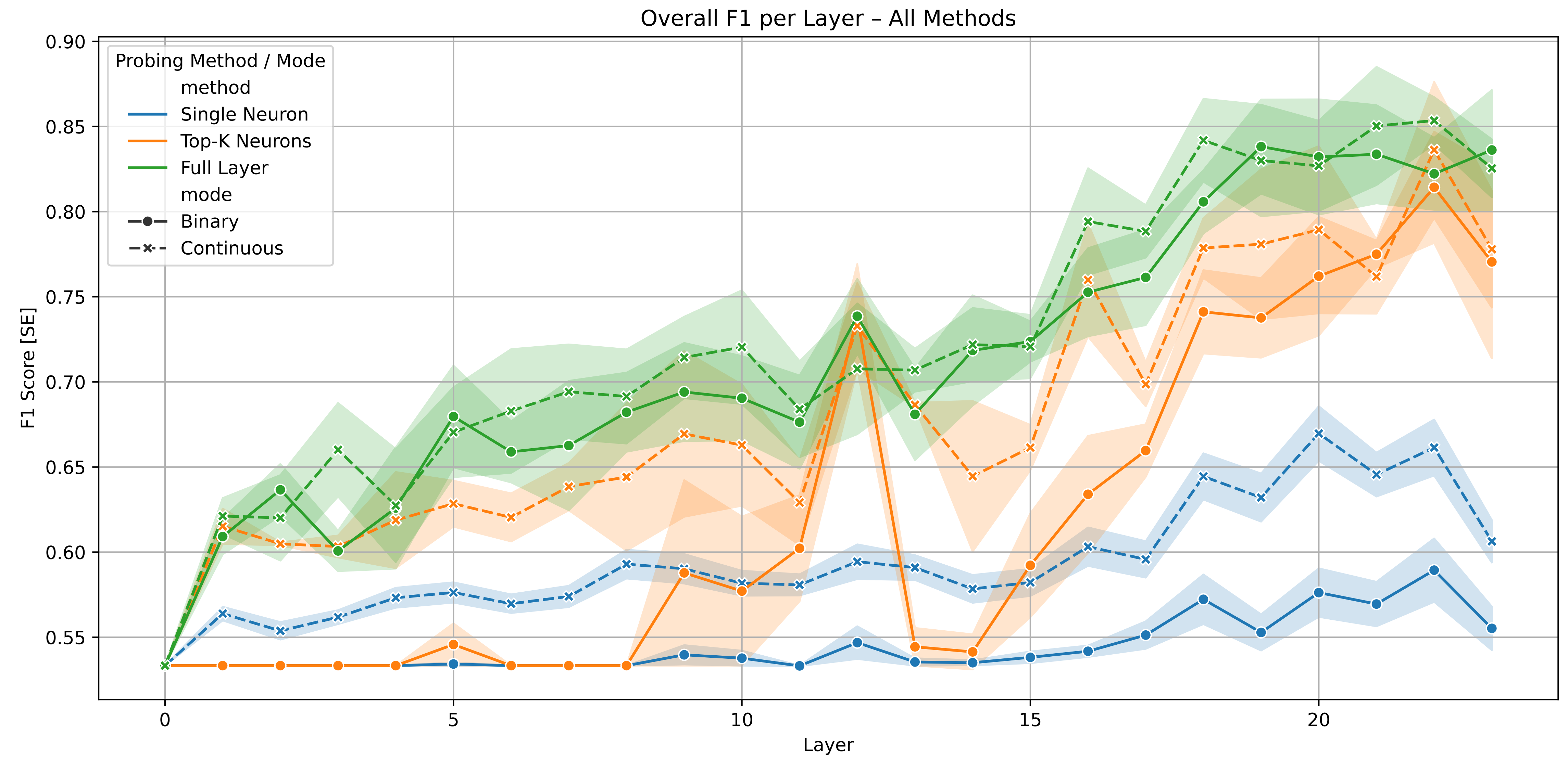
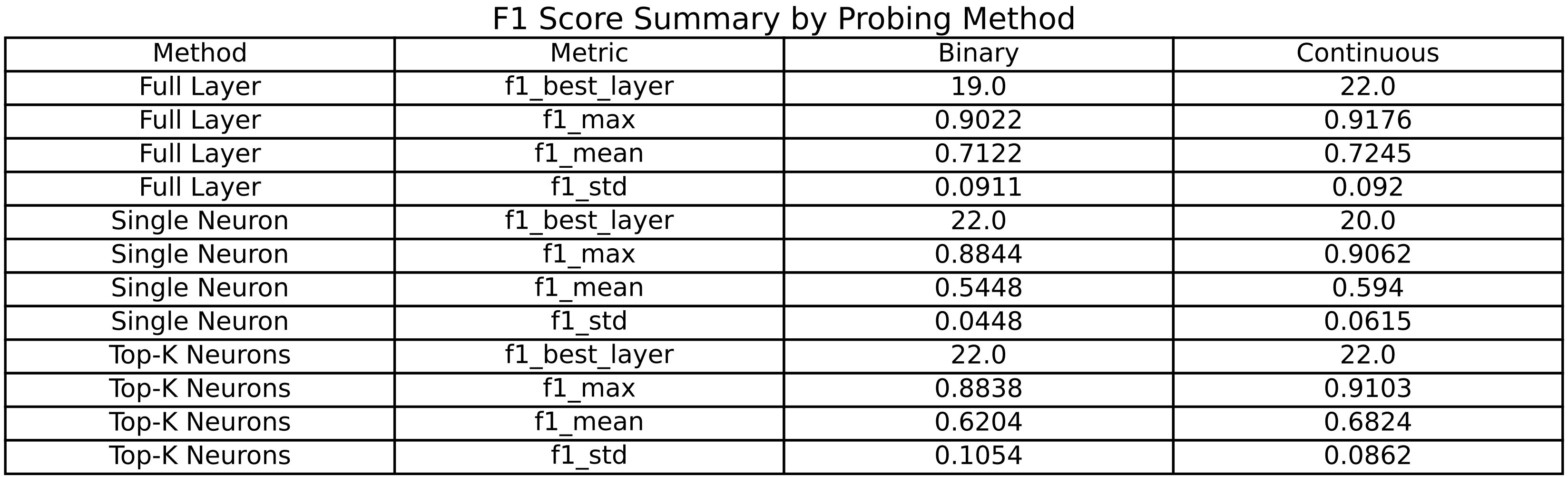


F1 per Layer – Single Neuron Probing
F1 [SE]
0.54
0.56
0.58
0.60
0.62
0.64
0.66
0.68
Layer
0
5
10
15
20
Mode
Binary
Continuous


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

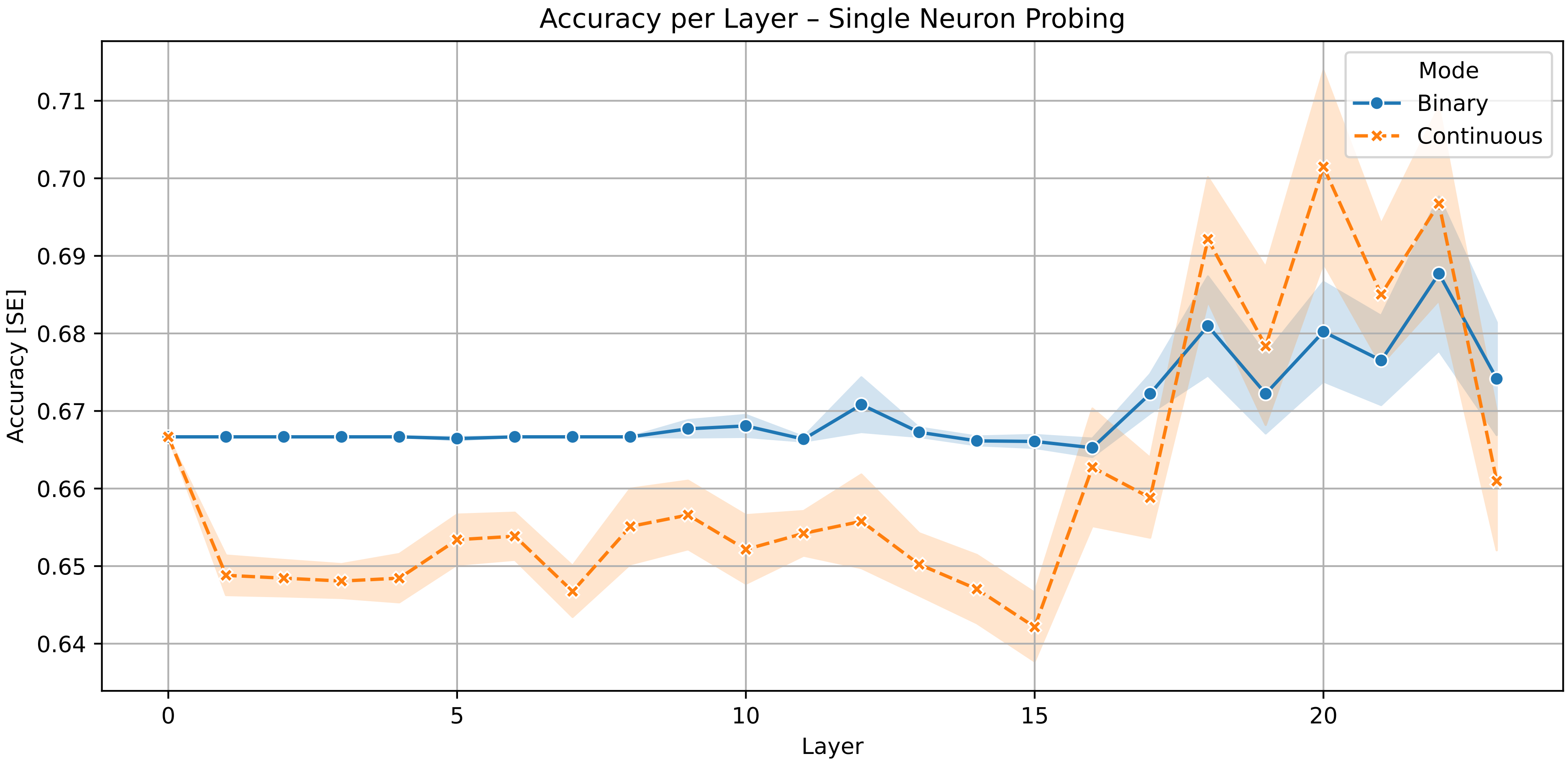

F1 per Layer – Full Layer Probing
Mode
Binary
Continuous
F1 [SE]
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
Layer

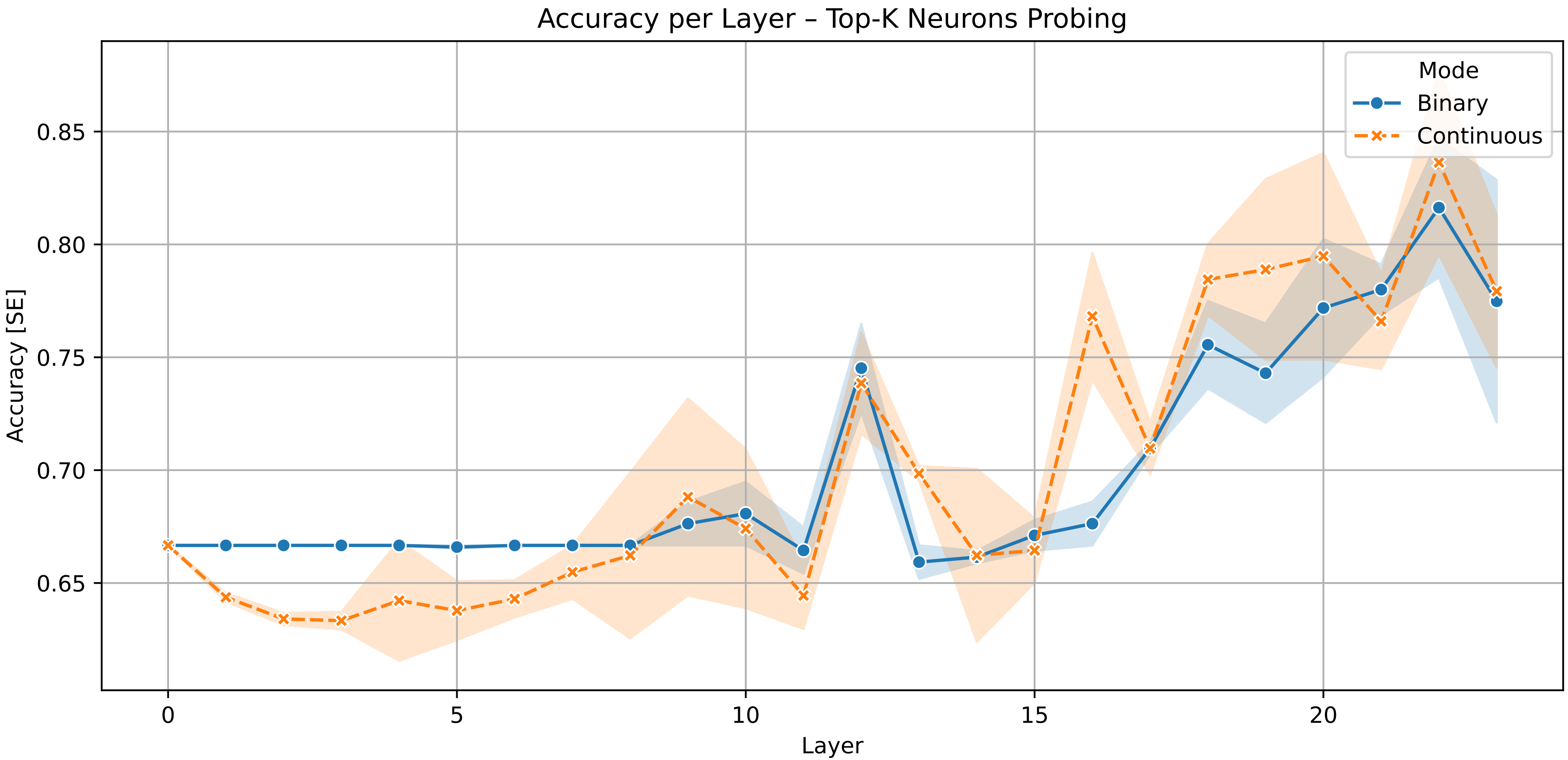

Overall F1 per Layer – All Methods
Probing Method / Mode
method
Single Neuron
Top-K Neurons
Full Layer
mode
Binary
Continuous
F1 Score [SE]
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
Layer
0
5
10
15
20

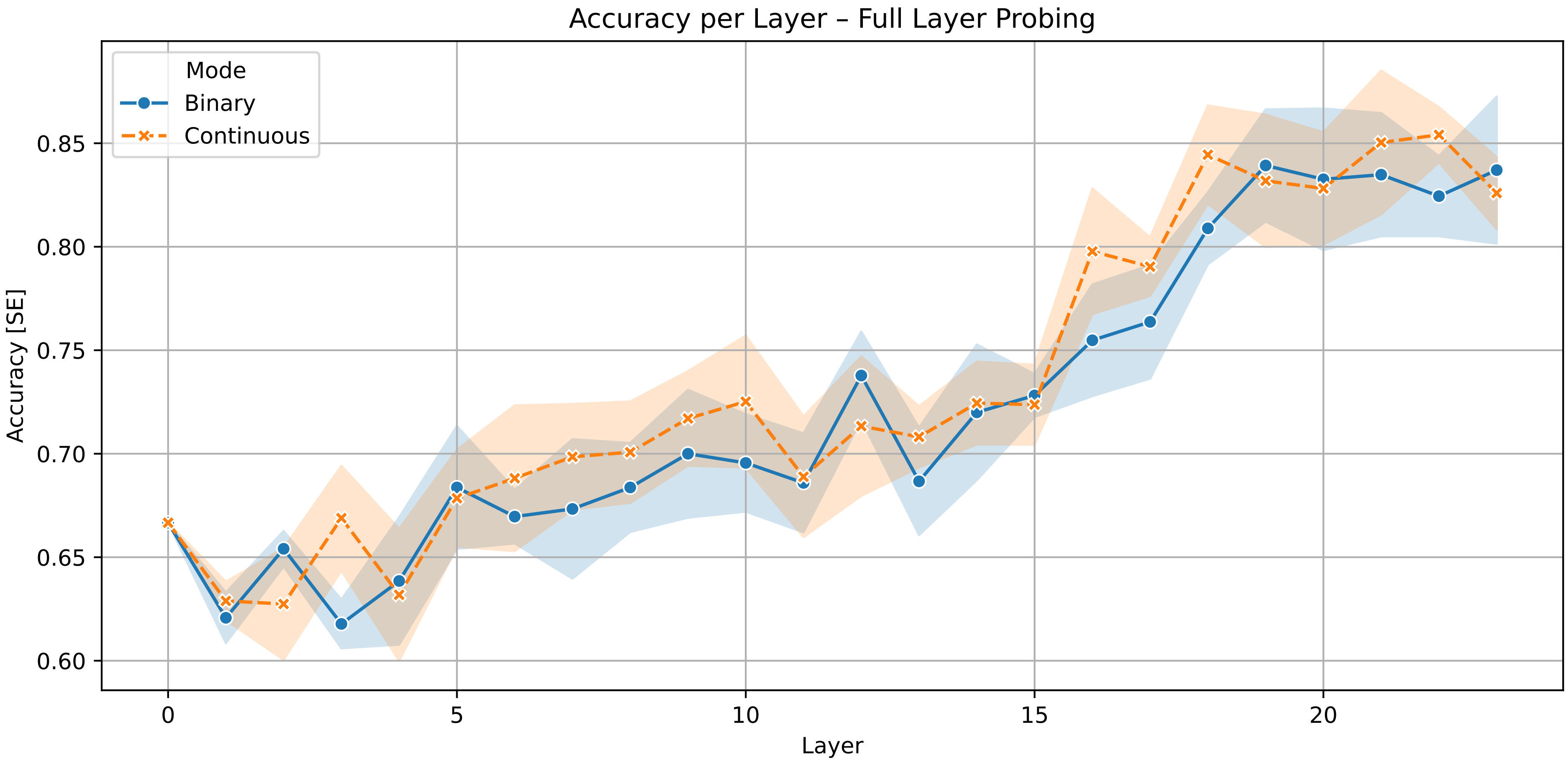


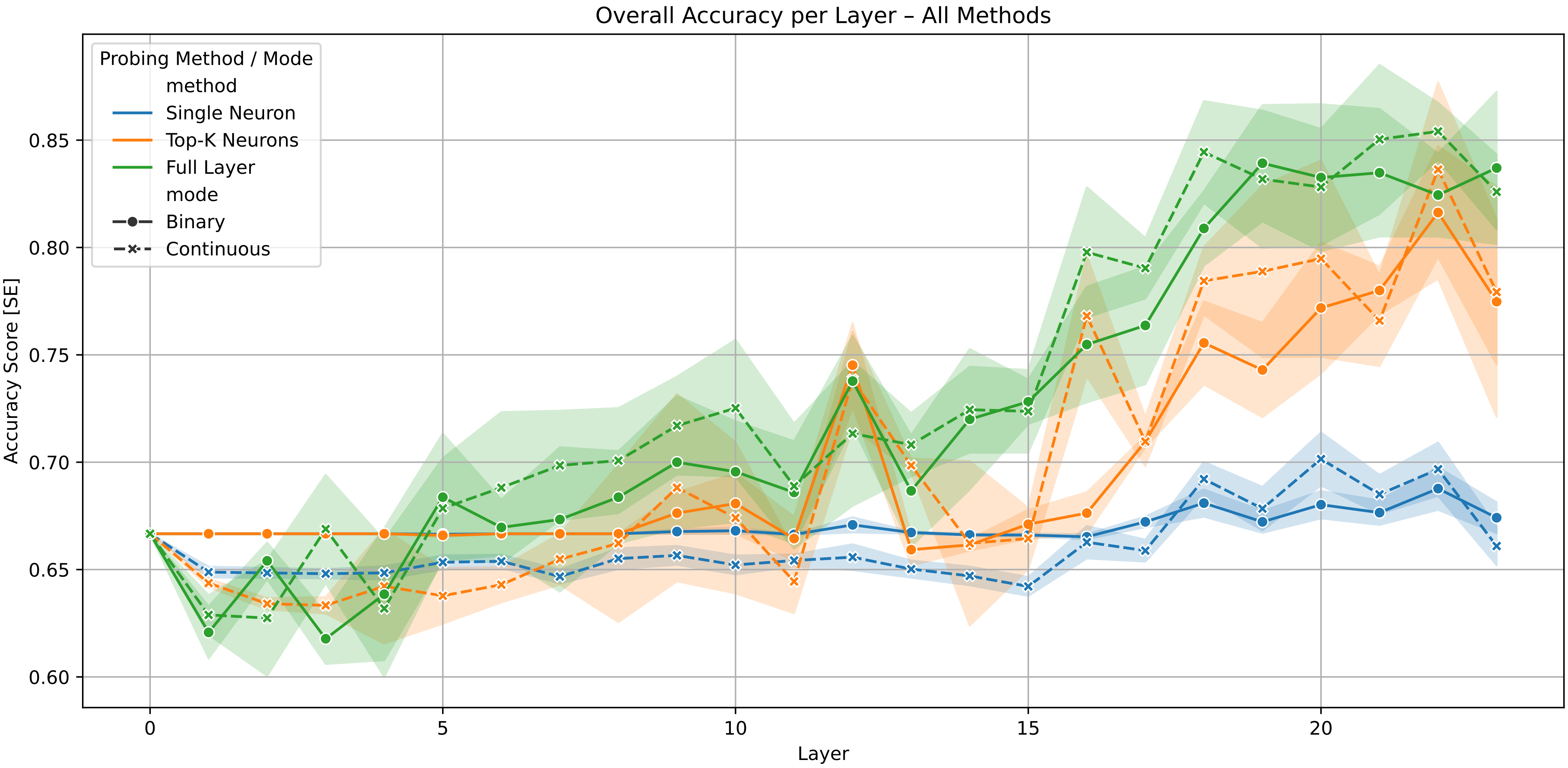
F1 Score Summary by Probing Method

| Method | Metric | Binary | Continuous |
|---|---|---|---|
| Full Layer | f1_best_layer | 19.0 | 22.0 |
| Full Layer | f1_max | 0.9022 | 0.9176 |
| Full Layer | f1_mean | 0.7122 | 0.7245 |
| Full Layer | f1_std | 0.0911 | 0.092 |
| Single Neuron | f1_best_layer | 22.0 | 20.0 |
| Single Neuron | f1_max | 0.8844 | 0.9062 |
| Single Neuron | f1_mean | 0.5448 | 0.594 |
| Single Neuron | f1_std | 0.0448 | 0.0615 |
| Top-K Neurons | f1_best_layer | 22.0 | 22.0 |
| Top-K Neurons | f1_max | 0.8838 | 0.9103 |
| Top-K Neurons | f1_mean | 0.6204 | 0.6824 |
| Top-K Neurons | f1_std | 0.1054 | 0.0862 |

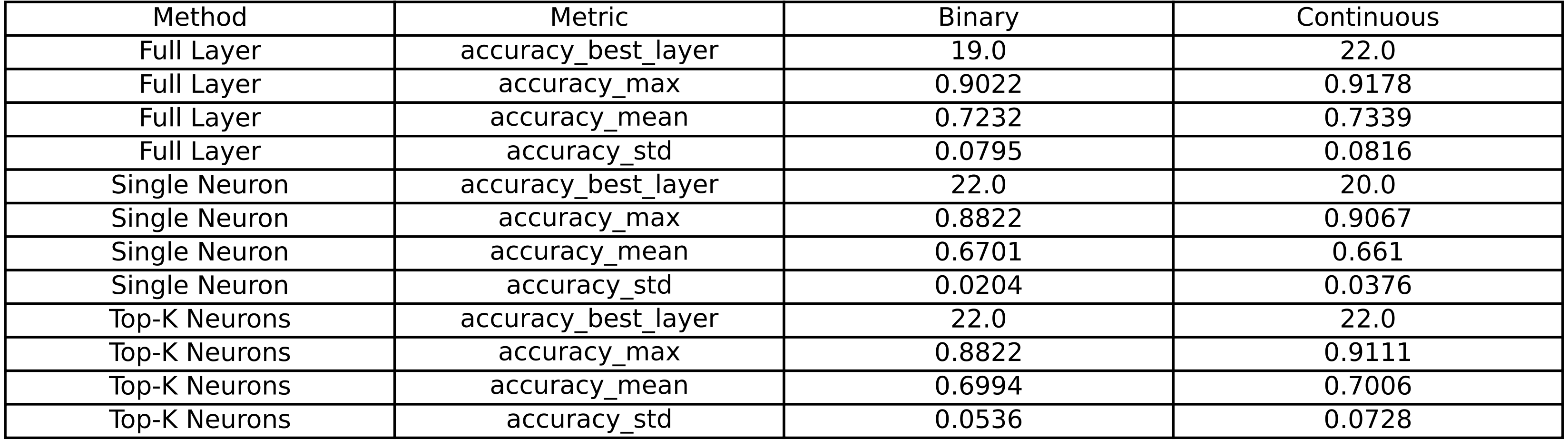

Accuracy per Layer – Single Neuron Probing
Accuracy [SE]
0.64
0.65
0.66
0.67
0.68
0.69
0.70
0.71
Layer
0
5
10
15
20
Mode
Binary
Continuous


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


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
0.85
Layer
0
5
10
15
20

## Accuracy Score Summary by Probing Method



| Method | Metric | Binary | Continuous |
|---|---|---|---|
| Full Layer | accuracy_best_layer | 19.0 | 22.0 |
| Full Layer | accuracy_max | 0.9022 | 0.9178 |
| Full Layer | accuracy_mean | 0.7232 | 0.7339 |
| Full Layer | accuracy_std | 0.0795 | 0.0816 |
| Single Neuron | accuracy_best_layer | 22.0 | 20.0 |
| Single Neuron | accuracy_max | 0.8822 | 0.9067 |
| Single Neuron | accuracy_mean | 0.6701 | 0.661 |
| Single Neuron | accuracy_std | 0.0204 | 0.0376 |
| Top-K Neurons | accuracy_best_layer | 22.0 | 22.0 |
| Top-K Neurons | accuracy_max | 0.8822 | 0.9111 |
| Top-K Neurons | accuracy_mean | 0.6994 | 0.7006 |
| Top-K Neurons | accuracy_std | 0.0536 | 0.0728 |