F1 per Layer - Single Neuron Probing for landmarks Mode Binary 0.60 -Continuous 0.55 EJ 0.50 0.45 0.40 10 20 25 30 15 Layer

F1 per Layer - Top-K Neurons Probing for landmarks Mode Binary 0.8 Continuous 0.7 0.6 F1 [SE] 0.4 0.3 0.2

15

Layer

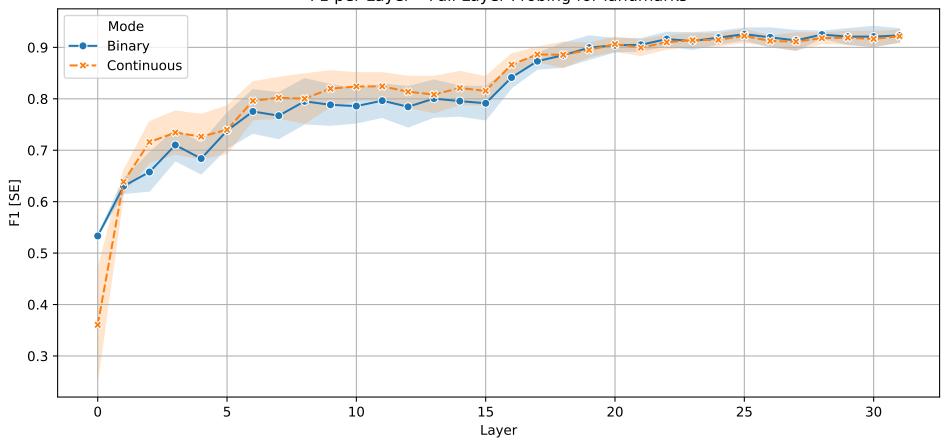
20

25

10

30

F1 per Layer - Full Layer Probing for landmarks



Overall F1 per Layer - All Methods for landmarks 0.9 8.0 0.7 Score [SE] Probing Method / Mode 0.4 method Single Neuron Top-K Neurons 0.3 — Full Layer mode -●- Binary 0.2 -x-- Continuous 10 15 20 25 30 Layer

F1 Score Summary by Probing Method Method Metric Binary f1 best layer

f1 max

f1 mean

f1_std

Full Layer

Full Laver

Top-K Neurons

Top-K Neurons

Tun Luyer	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	0.5555	0.5400
Full Layer	f1_mean	0.823	0.8293
Full Layer	f1_std	0.1075	0.1214
Single Neuron	f1_best_layer	26.0	27.0
Single Neuron	f1_max	0.7987	0.8296
Single Neuron	f1_mean	0.4865	0.5716
Single Neuron	f1_std	0.1316	0.0686
Top-K Neurons	f1_best_layer	26.0	28.0
Top-K Neurons	f1_max	0.8452	0.8709

25.0

0 9599

0.5851

0.1477

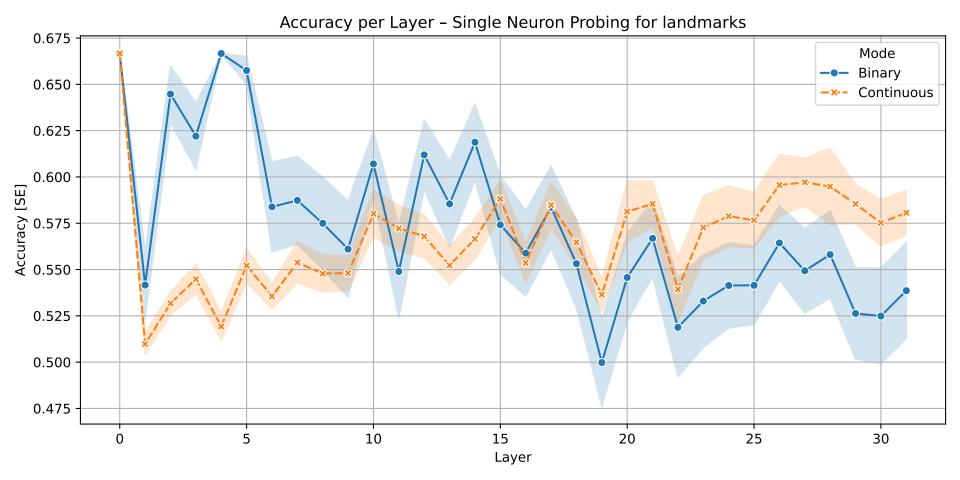
Continuous

25.0

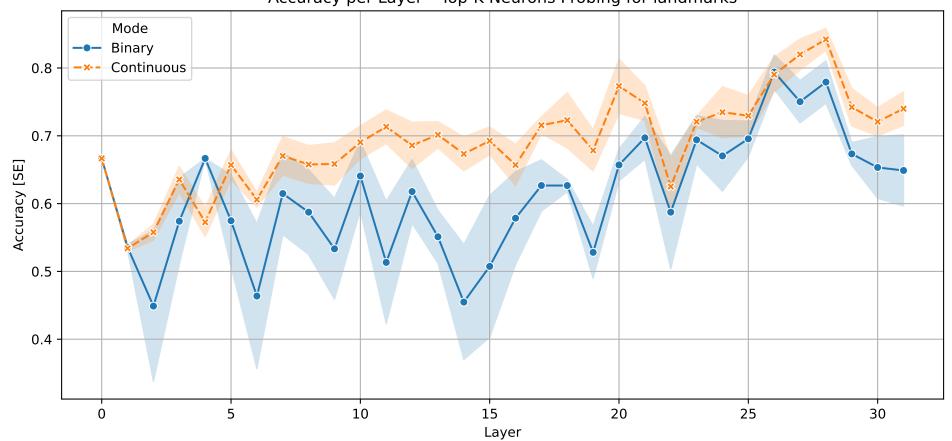
0.9468

0.6945

0.0802



Accuracy per Layer – Top-K Neurons Probing for landmarks



Accuracy per Layer - Full Layer Probing for landmarks Mode Binary 0.9 Continuous 8.0 Accuracy [SE] 0.0 990 0.5 0.4 10 20 25 30 15 Layer

Overall Accuracy per Layer – All Methods for landmarks Probing Method / Mode method 0.9 -Single Neuron Top-K Neurons Full Layer mode 0.8 -**−•−** Binary -x-- Continuous Accuracy Score [SE] 0.5 0.4 10 15 20 25 30 Layer

	ary by Probing Method	
Method	Metric	Binary

accuracy best layer

accuracy_std

Full Layer

Top-K Neurons

Full Layer	accuracy_max	0.96	0.9467	
Full Layer	accuracy_mean	0.8278	0.8316	
Full Layer	accuracy_std	0.0985	0.1099	
Single Neuron	accuracy_best_layer	0.0	0.0	
Single Neuron	accuracy_max	0.8044	0.8289	
Single Neuron	accuracy_mean	0.5737	0.5669	
Single Neuron	accuracy_std	0.1253	0.0714	
Top-K Neurons	accuracy_best_layer	26.0	28.0	
Top-K Neurons	accuracy_max	0.8422	0.8689	
Top-K Neurons	accuracy_mean	0.6129	0.6917	

25.0

0.1201

Continuous

25.0

0.0777