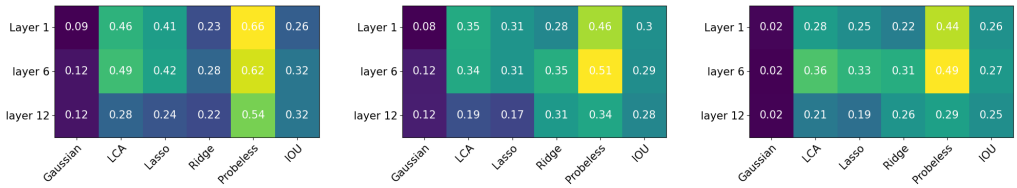
6 Appendix



(a) BERT (b) RoBERTa (c) XLMR

Figure 2: *Task: POS,* Pairwise comparison of MeanSelect with other methods using 10–50 neurons

6.1 Case Study

Mean Select: We propose a new corpus-based method, `MeanSelect`, for neuron ranking as a case-study to illustrate how researchers can use our proposed methodology. Kádár et al. (2017b) generated explanations for neurons by extracting top-N 5-gram context for each neuron based on the magnitude of their activations, followed by human annotation. Na et al. (2019) removed the human-in-the-loop by extracting concepts of various granularities from a parsed tree and aligned highly activating neurons to the concepts. Inspired by these works, we propose a novel method, `MeanSelect`, that generates a ranking of neurons with respect to a concept. The intuition is that a neuron learning a concept will have consistently high activations across different contexts where that concept appears. However, there may be a neuron that always activates with high value irrespective of the concept. A difference in the mean value of the neuron activating the concept and all other concepts will provide the true behavior of the neuron for the concept. Following our notation in Section 2, the score of a given neuron n is defined as follows:

$$R(n, \mathcal{C}) = \frac{\mu(\mathcal{C}) - \mu(\hat{\mathcal{C}})}{n_{max} - n_{min}} \tag{10}$$

where $\mu(\mathcal{C})$ is the average, n_{max} is the max and n_{min} is the min of activations $z(n, w)$ where $w \in \mathcal{C}$, and $\mu(\hat{\mathcal{C}})$ is the average of activations over the random concept set.

Compatibility Metrics Table 5 shows the compatibility score of `MeanSelect` using the representation of 3 different layers and compared it with the Random selection of neurons. The high compatibility scores show that `MeanSelect` discovers neurons that are endorsed by other neuron interpretation methods. This serves as a measure of confidence in the proposed method. Moreover, one may observe that the method has a relatively lower score for the last layer compared to other layers, giving insight into potential improvements that can be made to the behavior of this method.

Pairwise Comparison The pairwise comparison of methods further provides insights into how the new method relates to other methods in terms of the resulting neurons. Figure 2 shows the heatmaps of three pre-trained models. MeanSelect has the highest overlap with the Probeless method and except Gaussian, it shows an overlap of at least 0.23 points with other methods. While the high overlap of MeanSelect with Probeless is not surprising given both are based on similar intuitions, the overlap with LCA and L1 shows that the method is selecting a diverse set of neurons captured by a variety of methods. Similar to the discussion on compatibility score, here we observe a substantial overlap drop in the last layer and this highlights the potential vulnerability of the method to certain representations.

6.2 Concept Datasets

We consider concepts from three linguistic tasks: parts of speech tags (POS, Marcus et al., 1993), semantic tags (SEM, Abzianidze et al., 2017) and syntactic chunking (Chunking) using CoNLL 2000 shared task dataset (Tjong Kim Sang & Buchholz, 2000). For the POS dataset, we used 20 concepts

Table 5: *Task: POS, Model: BERT,* Average NeuronVote score of MeanSelect using 10–50 neurons

	BERT			RoBERTa			XLMR		
Layers	1	6	12	1	6	12	1	6	12
Random	0.019	0.021	0.023	0.020	0.019	0.017	0.019	0.017	0.022
MeanSelect	0.464	0.476	0.402	0.392	0.451	0.328	0.368	0.408	0.253

Table 6: *Task: Semantic tagging,* Average NeuronVote compatibility scores across Semantic tagging concepts when selecting the top 10, 30, and 50 neurons from layers 1, 6 and 12. Bold numbers, underline numbers, and dashed numbers show the first, second, and third best scores respectively

	BERT			RoBERTa			XLMR		
Layers	1	6	12	1	6	12	1	6	12
Random	0.018	0.013	0.017	0.011	0.026	0.017	0.026	0.014	0.026
Gaussian	0.257	0.256	0.222	0.237	0.282	0.245	0.176	0.256	0.195
LCA	**0.474**	**0.541**	0.385	0.488	0.493	0.347	0.309	0.455	0.367
Lasso	0.407	0.492	0.322	0.391	0.460	0.328	0.294	0.396	0.358
Ridge	0.316	0.343	0.361	0.468	0.492	0.575	0.372	0.430	**0.473**
Probeless	0.450	0.501	**0.476**	**0.547**	**0.586**	**0.640**	**0.495**	**0.571**	0.464
IoU	0.344	0.332	0.380	0.288	0.287	0.242	0.277	0.262	0.282

which have a total dataset size of 40137. These concepts include VBG (777), VBZ (908), NNPS (204), DT (4015), TO (1177), CD (1935), JJ (2836), PRP (801), MD (463), RB (1348), VBP (534), VB (1244), NNS (3021), VBN (1082), POS (433), IN (5039), NN (6660), CC (1220), NNP (4698), and VBD (1742).

For the SEM dataset, we used three concepts which have a total dataset size of 120941. These concepts include IST (72240), NOW (24137) and EXS (24564). We used 10 concepts from Chunking which have a total dataset size of 220606. These concepts include B-ADJP (2493), B-ADVP (5081), B-NP (67285), B-PP (26005), B-VP (26078), I-ADJP (805), I-ADVP (532), I-NP (77368) I-PP (339) and I-VP (14620). For all the datasets used in the experiments, we use train/valid/test split 70%, 15% and 15%.

6.3 Results

6.3.1 Semantic Tagging Concepts

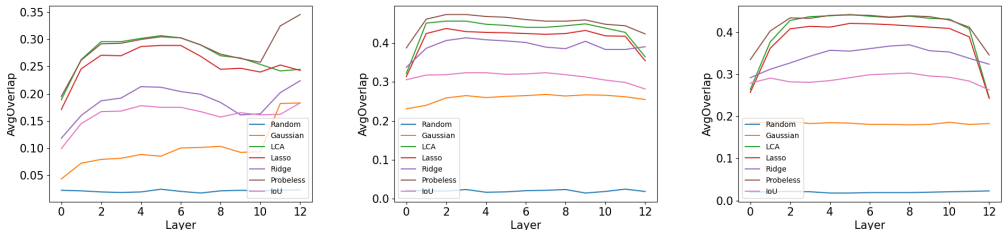
For the SEM dataset, we sample 20000 sentences for experimental validation with train/valid/test split 70%, 15% and 15%. We select three tags: IST (intersective), NOW (present tense) and EXS (untensed simple event). Table 6 presents the average NeuronVote scores across three models. We observed identical trends to that of POS i.e. Probeless is the most consistent method, LCA and Lasso are second best methods but they suffer on the last layers.

6.4 Chunking Concepts

Figures 3 and 4 show layer-wise results for the two voting methods proposed in the paper, across the three understudied models (BERT, RoBERTa and XLM-R). The results show that voting methods consistently rank the *Probeless* method as the most compatible in terms of neuron rankings across the layers. We are including detailed results in Tables 5–10 give detailed results with exact numbers.

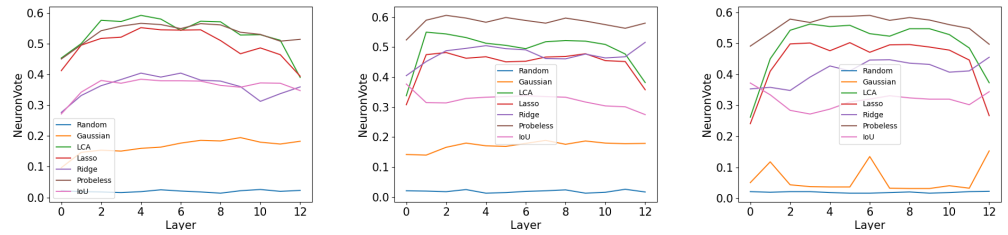
Table 7: *Task: Chunking,* Average `NeuronVote` compatibility scores across Chunking concepts when selecting the top 10, 30, and 50 neurons from layers 1, 6 and 12. Bold numbers, underline numbers, and dashed numbers show the first, second, and third best scores respectively

	BERT			RoBERTa			XLMR		
Layers	1	6	12	1	6	12	1	6	12
Random	0.018	0.017	0.023	0.022	0.018	0.022	0.023	0.019	0.014
Gaussian	0.122	0.181	0.174	0.111	0.163	0.164	0.110	0.121	0.143
LCA	0.395	0.469	0.300	0.440	0.422	0.328	0.336	0.447	0.388
Lasso	0.396	0.472	0.301	0.399	0.395	0.322	0.366	0.425	0.395
Ridge	0.235	0.255	0.256	0.303	0.330	0.386	0.254	0.289	0.410
Probeless	**0.465**	**0.506**	**0.422**	**0.502**	**0.500**	**0.514**	**0.499**	**0.507**	**0.463**
IoU	0.346	0.361	0.321	0.285	0.298	0.244	0.319	0.283	0.352



(a) BERT (b) RoBERTa (c) XLMR

Figure 3: `AvgOverlap` score across different Layers in different models. All methods perform better than the baseline, Random. Probeless is the most consistent method across all models, concepts and across all layers. It is among the top methods with LCA and Lasso. However, LCA and Lasso show low score on last layers. The performance of Gaussian deteriorates significantly and is closer to Random when applied on XLMR.



(a) BERT (b) RoBERTa (c) XLMR

Figure 4: `NeuronVote` score across different Layers in different models. All methods perform better than the baseline, Random. Probeless is the most consistent method across all models, concepts and across all layers. It is among the top methods with LCA and Lasso. However, LCA and Lasso show low score on last layers.

Table 8: This is an extension of Table.4. Average `AvgOverlap` compatibility scores across concepts when selecting the top 10, 30, and 50 neurons. Bold numbers show the best scores. Probeless, LCA are the top performing methods. However, Probeless is most consistent performing method. LCA drops substantially for the last layers.

	BERT												
Layers	0	1	2	3	4	5	6	7	8	9	10	11	12
Random	0.022	0.021	0.019	0.018	0.019	0.024	0.020	0.017	0.021	0.022	0.021	0.022	0.023
Gaussian	0.043	0.072	0.079	0.081	0.088	0.085	0.100	0.101	0.103	0.092	0.093	0.182	0.183
LCA	0.189	**0.263**	**0.296**	**0.296**	**0.302**	**0.307**	**0.303**	**0.290**	0.27	**0.266**	0.254	0.242	0.245
Lasso	0.171	0.246	0.271	0.270	0.287	0.289	0.289	0.269	0.245	0.247	0.24	0.253	0.243
Ridge	0.118	0.16	0.187	0.192	0.213	0.212	0.204	0.199	0.184	0.161	0.163	0.202	0.224
Probeless	**0.195**	0.262	0.292	0.293	0.3	0.305	**0.303**	**0.29**	**0.273**	**0.265**	**0.258**	**0.325**	**0.346**
IoU	0.099	0.145	0.167	0.168	0.178	0.175	0.175	0.167	0.157	0.165	0.161	0.162	0.183

Table 9: This is an extension of Table.4. Average `NeuronVote` compatibility scores across concepts when selecting the top 10, 30, and 50 neurons.

	BERT												
Layers	0	1	2	3	4	5	6	7	8	9	10	11	12
Random	0.022	0.019	0.018	0.016	0.019	0.025	0.021	0.018	0.014	0.022	0.026	0.020	0.023
Gaussian	0.097	0.147	0.154	0.151	0.160	0.164	0.177	0.186	0.184	0.195	0.180	0.174	0.183
LCA	**0.454**	**0.501**	**0.577**	**0.573**	**0.593**	**0.581**	0.544	**0.574**	**0.572**	0.529	**0.530**	**0.512**	0.391
Lasso	0.413	0.496	0.518	0.522	0.553	0.546	0.545	0.546	0.511	0.468	0.487	0.465	0.395
Ridge	0.277	0.332	0.364	0.384	0.405	0.392	0.405	0.382	0.379	0.361	0.313	0.338	0.360
Probeless	0.451	0.497	0.543	0.558	0.567	0.563	**0.550**	0.566	0.562	**0.538**	**0.531**	0.509	0.515
IoU	0.272	0.343	0.381	0.372	0.385	0.380	0.380	0.379	0.365	0.359	0.373	0.372	0.348

Table 10: This is an extension of Table.4. Average `AvgOverlap` compatibility scores across concepts when selecting the top 10, 30, and 50 neurons.

	RoBERTa												
Layers	0	1	2	3	4	5	6	7	8	9	10	11	12
Random	0.021	0.020	0.019	0.023	0.016	0.017	0.020	0.021	0.023	0.014	0.018	0.024	0.018
Gaussian	0.231	0.240	0.259	0.265	0.260	0.263	0.265	0.268	0.264	0.267	0.266	0.262	0.255
LCA	0.322	0.452	0.457	0.457	0.449	0.446	0.441	0.441	0.445	0.450	0.439	0.428	0.365
Lasso	0.314	0.425	0.438	0.430	0.428	0.427	0.425	0.423	0.425	0.433	0.419	0.418	0.355
Ridge	0.338	0.387	0.407	0.414	0.409	0.406	0.402	0.390	0.386	0.405	0.384	0.384	0.391
Probeless	**0.388**	**0.462**	**0.474**	**0.474**	**0.469**	**0.467**	**0.461**	**0.457**	**0.457**	**0.460**	**0.449**	**0.445**	**0.424**
IoU	0.306	0.318	0.319	0.324	0.324	0.320	0.321	0.324	0.319	0.313	0.305	0.299	0.282

Table 11: This is an extension of Table.4. Average `NeuronVote` compatibility scores across concepts when selecting the top 10, 30, and 50 neurons.

	RoBERTa												
Layers	0	1	2	3	4	5	6	7	8	9	10	11	12
Random	0.021	0.020	0.018	0.025	0.013	0.015	0.019	0.021	0.024	0.013	0.016	0.026	0.017
Gaussian	0.142	0.140	0.166	0.180	0.171	0.169	0.179	0.189	0.176	0.187	0.180	0.178	0.179
LCA	0.338	0.550	0.544	0.532	0.513	0.506	0.495	0.518	0.522	0.520	0.509	0.477	0.382
Lasso	0.308	0.475	0.482	0.463	0.468	0.451	0.453	0.467	0.469	0.478	0.455	0.452	0.358
Ridge	0.405	0.452	0.488	0.496	0.505	0.495	0.491	0.462	0.461	0.477	0.464	0.468	0.516
Probeless	**0.524**	**0.590**	**0.606**	**0.597**	**0.583**	**0.599**	**0.589**	**0.580**	**0.597**	**0.587**	**0.575**	**0.563**	**0.580**
IoU	0.377	0.315	0.314	0.329	0.333	0.335	0.338	0.335	0.333	0.317	0.304	0.301	0.275

Table 12: This is an extension of Table.4. Average `AvgOverlap` compatibility scores across concepts when selecting the top 10, 30, and 50 neurons.

	XLMR												
Layers	0	1	2	3	4	5	6	7	8	9	10	11	12
Random	0.020	0.021	0.022	0.021	0.018	0.018	0.019	0.019	0.019	0.020	0.021	0.022	0.023
Gaussian	0.185	0.186	0.188	0.183	0.185	0.184	0.181	0.181	0.180	0.181	0.186	0.181	0.183
LCA	0.264	0.377	0.428	**0.437**	**0.439**	**0.442**	0.438	**0.435**	**0.438**	0.433	**0.431**	0.408	0.245
Lasso	0.257	0.362	0.408	0.414	0.412	0.421	0.420	0.418	0.415	0.412	0.409	0.389	0.243
Ridge	0.292	0.312	0.327	0.343	0.357	0.355	0.361	0.367	0.370	0.356	0.353	0.338	0.324
Probeless	**0.335**	**0.403**	**0.434**	0.433	**0.440**	**0.441**	**0.440**	**0.436**	**0.439**	**0.437**	0.429	**0.412**	**0.346**
IoU	0.279	0.291	0.282	0.281	0.285	0.292	0.299	0.301	0.303	0.296	0.293	0.284	0.263

Table 13: This is an extension of Table.4. Average `NeuronVote` compatibility scores across concepts when selecting the top 10, 30, and 50 neurons.

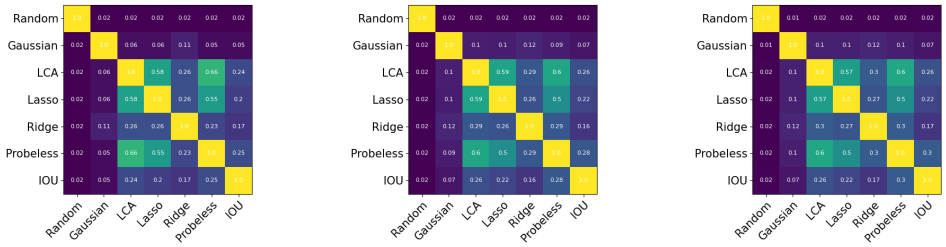
	XLMR												
Layers	0	1	2	3	4	5	6	7	8	9	10	11	12
Random	0.022	0.020	0.022	0.022	0.019	0.017	0.017	0.019	0.021	0.017	0.019	0.022	0.023
Gaussian	0.051	0.118	0.044	0.038	0.037	0.037	0.135	0.033	0.032	0.032	0.041	0.033	0.153
LCA	0.262	0.450	0.542	0.562	0.554	0.558	0.531	0.523	0.547	0.547	0.528	0.485	0.373
Lasso	0.241	0.410	0.498	0.501	0.476	0.502	0.471	0.495	0.496	0.488	0.478	0.446	0.267
Ridge	0.353	0.358	0.348	0.391	0.427	0.411	0.446	0.447	0.436	0.432	0.407	0.411	0.455
Probeless	**0.491**	**0.534**	**0.578**	**0.567**	**0.586**	**0.587**	**0.590**	**0.574**	**0.583**	**0.575**	**0.560**	**0.548**	**0.497**
IoU	0.372	0.334	0.284	0.272	0.288	0.311	0.320	0.331	0.324	0.320	0.320	0.302	0.344

Table 14: This is an extension of Table.5. Average `AvgOverlap` score of MeanSelect when selecting 10, 30, and 50 neurons for all layers

Layers	0	1	2	3	4	5	6	7	8	9	10	11	12
	BERT												
Random	0.022	0.021	0.019	0.018	0.019	0.022	0.020	0.017	0.021	0.022	0.021	0.022	0.023
MeanSelect	0.266	0.351	0.363	0.362	0.365	0.369	0.374	0.370	0.356	0.349	0.343	0.327	0.283
	RoBERTa												
Random	0.021	0.020	0.019	0.023	0.016	0.017	0.020	0.021	0.023	0.014	0.018	0.024	0.018
MeanSelect	0.252	0.294	0.307	0.317	0.326	0.328	0.323	0.320	0.314	0.308	0.283	0.269	0.239
	XLMR												
Random	0.020	0.021	0.022	0.021	0.018	0.018	0.019	0.019	0.019	0.020	0.021	0.022	0.023
MeanSelect	0.233	0.246	0.245	0.247	0.264	0.273	0.294	0.292	0.281	0.261	0.259	0.241	0.159

Table 15: This is an extension of Table.5. Average `NeuronVote` score of MeanSelect when selecting 10, 30, and 50 neurons for all layers

Layers	0	1	2	3	4	5	6	7	8	9	10	11	12
	BERT												
Random	0.022	0.019	0.018	0.016	0.019	0.025	0.021	0.018	0.014	0.022	0.026	0.020	0.023
MeanSelect	0.361	0.476	0.462	0.458	0.470	0.468	0.479	0.488	0.486	0.463	0.457	0.432	0.400
	RoBERTa												
Random	0.021	0.020	0.018	0.025	0.013	0.015	0.019	0.021	0.024	0.013	0.016	0.026	0.017
MeanSelect	0.358	0.388	0.395	0.403	0.426	0.423	0.455	0.438	0.437	0.406	0.383	0.374	0.326
	XLMR												
Random	0.022	0.020	0.022	0.022	0.019	0.017	0.017	0.019	0.021	0.017	0.019	0.022	0.023
MeanSelect	0.376	0.357	0.356	0.342	0.364	0.392	0.414	0.416	0.400	0.372	0.381	0.360	0.243



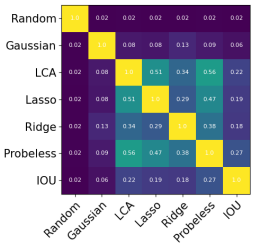
(a) Layer 1 (b) Layer 2 (c) Layer 3

(d) Layer 4 (e) Layer 5 (f) Layer 6

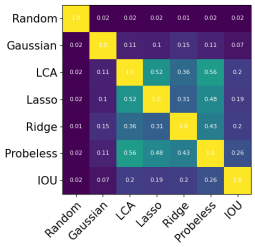
(g) Layer 7 (h) Layer 8 (i) Layer 9

(j) Layer 10 (k) Layer 11 (l) Layer 12

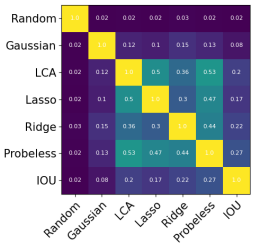
Figure 5: This is an extension of Figure.1. Comparing average overlap of top 10-50 neurons across methods for BERT



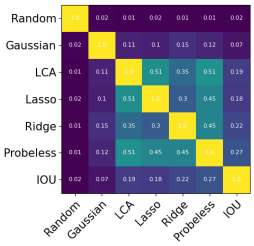
(a) Layer 1



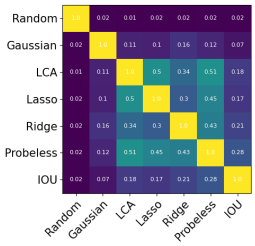
(b) Layer 2



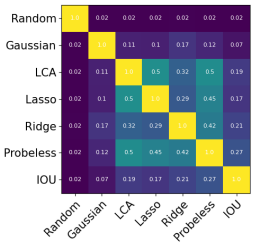
(c) Layer 3



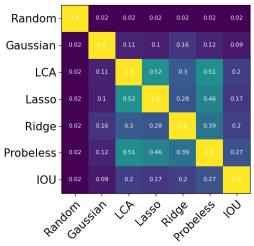
(d) Layer 4



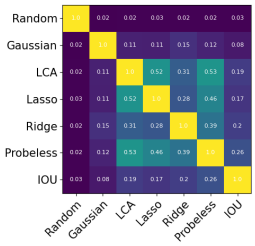
(e) Layer 5



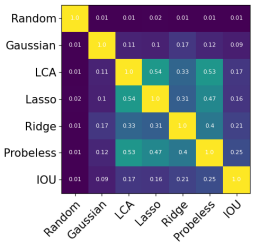
(f) Layer 6



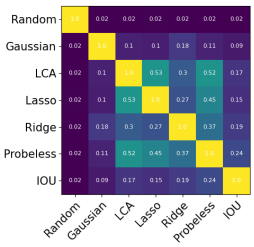
(g) Layer 7



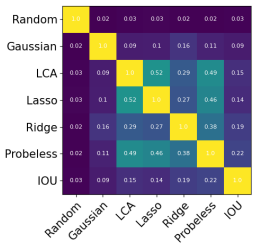
(h) Layer 8



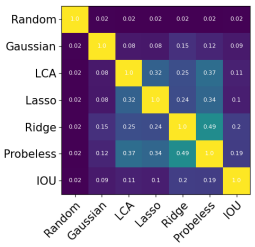
(i) Layer 9



(j) Layer 10

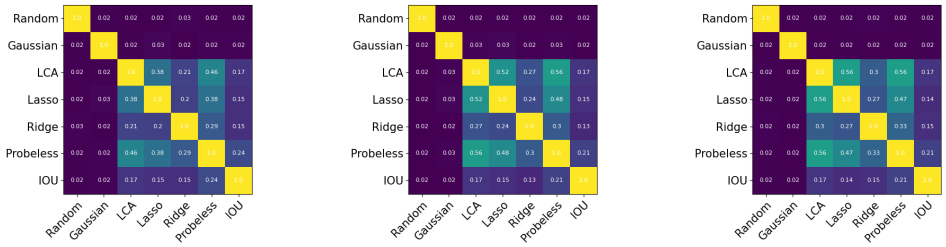


(k) Layer 11



(l) Layer 12

Figure 6: This is an extension of Figure.1. Comparing average overlap of top 10-50 neurons across methods for RoBERTa



(a) Layer 1 (b) Layer 2 (c) Layer 3

(d) Layer 4 (e) Layer 5 (f) Layer 6

(g) Layer 7 (h) Layer 8 (i) Layer 9

(j) Layer 10 (k) Layer 11 (l) Layer 12

Figure 7: This is an extension of Figure.1. Comparing average overlap of top 10-50 neurons across methods for XLMR