Table 8: Detection capabilities for positive and negative directions in the personality big five: *agreeableness* (AGREE), *conscientiousness* (CONSC), *openness* (OPEN), *extraversion* (EXTRA), *neuroticism* (NEURO). Results over ***Llama3-8B-Instruct*** activations in layer 31. Comparing Local Outlier Factor (LOF) method [102], Isolation Forest (IF) and Kmeans to Deep Scan used in this study. $|e|$ is the size of the activation vector defining the clusters, for Deep Scan, $|e| = |O_{S^*}|$

Method	Dimension	$\|e\|$ (↓)	Precision (↑)	Recall (↑)
LOF	AGREE	4096	0.5072 ± 0.0288	0.9904 ± 0.0058
	CONSC	4096	0.5033 ± 0.0359	$\mathbf{1.0 \pm 0.0}$
	OPEN	4096	0.5049 ± 0.0380	$\mathbf{0.9861 \pm 0.0112}$
	EXTRA	4096	0.4867 ± 0.0278	$\mathbf{0.9933 \pm 0.007}$
	NEURO	4096	0.4880 ± 0.0275	$\mathbf{1.0 \pm 0.0}$
IF	AGREE	4096	0.5094 ± 0.0313	$\mathbf{0.9980 \pm 0.0038}$
	CONSC	4096	0.5010 ± 0.0357	0.9865 ± 0.0086
	OPEN	4096	0.4974 ± 0.0390	0.9615 ± 0.0185
	EXTRA	4096	0.4847 ± 0.0292	0.9866 ± 0.0097
	NEURO	4096	0.4880 ± 0.0272	0.9984 ± 0.0034
KMeans	AGREE	4096	0.8333 ± 0.3726	0.8300 ± 0.3712
	CONSC	4096	0.8285 ± 0.3705	0.8333 ± 0.3726
	OPEN	4096	0.8316 ± 0.3719	0.8333 ± 0.3726
	EXTRA	4096	0.7739 ± 0.1568	0.8828 ± 0.1032
	NEURO	4096	0.6260 ± 0.0314	0.6997 ± 0.1394
Deep Scan	AGREE	2210	$\mathbf{0.9971 \pm 0.0113}$	$\mathbf{0.9979 \pm 0.0098}$
	CONSC	2692	$\mathbf{0.9992 \pm 0.0001}$	0.9545 ± 0.0487
	OPEN	2494	$\mathbf{0.9998 \pm 0.0003}$	0.9772 ± 0.0422
	EXTRA	1721	$\mathbf{0.9457 \pm 0.0268}$	0.8901 ± 0.0542
	NEURO	2038	$\mathbf{0.9540 \pm 0.0323}$	0.7565 ± 0.1142