Table 4: **(Q1)**: Validation of usefulness of the set of salient activations O_{S^*} at *Layer 1* and *31* of *Llama3-8B-Instruct* .

Dimension	Layer	$\|O_{S^*}\|$	Precision (↑)	Recall (↑)
AGREE	1	2815	0.5067 ± 0.1432	0.2469 ± 0.0904
	31	1167	$\mathbf{0.9971 \pm 0.0113}$	$\mathbf{0.9979 \pm 0.00984}$
CONSC	1	3240	0.3895 ± 0.1753	0.2237 ± 0.10208
	31	1210	$\mathbf{0.9992 \pm 0.0001}$	$\mathbf{0.95454 \pm 0.04876}$
OPEN	1	3204	0.6048 ± 0.2317	0.2434 ± 0.10216
	31	1177	$\mathbf{0.9998 \pm 0.0003}$	$\mathbf{0.97727 \pm 0.0422}$

A.3 Extended Results

A.3.1 (Q1) Persona Representations Across Layers in Other Models

In Tab. 4, we observe low precision and recall in early layers. This suggests that the activation locations found are not useful to determine MATCHINGBEHAVIOR for dimensions, compared to performance in *Layer 31*. In *Layer 31*, we observe high precision and recall.

This also confirms the quality of the representations that we observe in Fig. 7 and 8, where we plot the PCA embeddings for the *openness* persona in *Granite-7B-Instruct* and the *conscientiousness* persona in *Mistral-7B-Instruct*, respectively. We observe that the separability between MATCHINGBEHAVIOR and NOTMATCHINGBEHAVIOR representations improves in the later layers.

In Tables 5, 6, and 7, we show several clustering metrics to quantify the separation between q^+ (MATCHINGBEHAVIOR) representations and q^- (NOTMATCHINGBEHAVIOR).

A.3.2 (Q2) Unique Locations of Persona Within a Layer

Fig. 9, shows Upset and Venn diagrams plots for intra-persona (Level 2) analysis for personas from all topics, *Personality*, *Ethics*, *Politics*.

In Tab. 8 we report precision and recall regarding MATCHINGBEHAVIOR and NOTMATCHINGBEHAVIOR detection (Level 2) comparing different unsupervised methods. We observe that Deep Scan outperforms in precision while still maintaining a high recall compared to the other unsupervised methods.

In Fig. 10, we show a Venn diagram of the overlap of salient activations at the inter-topic level (Level 0). Between personas from *Ethics*, *Politics*, and *Personality*, we observe very low overlap between salient activations.

[12] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[13] https://huggingface.co/ibm-granite/granite-7b-instruct

[14] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3