Table 2: (**Q1**, **Q2**) Validation of usefulness of salient activations O_{S^*} in detecting sentences X_{S^*} w.r.t. detection hypothesis H_1 at different levels. MATCHINGBEHAVIOR (+) and NOTMATCHINGBEHAVIOR (-) sentences. "all" indicating all other relevant personas, e.g., for *Level 1* CONS$^+$, all= $\{$LIBER$^+ \cup$ IMMI$^+ \cup$ LGBTQ$^+\}$; for *Level 0 Politics*$^+$, all=$\{$*Ethics*$^+ \cup$ *Personality*$^+\}$. Mean ± std over 100 indep. Deep Scan runs, and 200 random test samples. High/low detection power.

Level	H_0	H_1	Precision (↑)	Recall (↑)
Level 2 Intra-Persona	CONSC$^-$	CONSC$^+$	0.8387 ± 0.0399	0.8181 ± 0.0765
	LIBER$^-$	LIBER$^+$	0.8939 ± 0.0507	0.8056 ± 0.0769
	IMMI$^-$	IMMI$^+$	0.8167 ± 0.0507	0.8282 ± 0.0711
	LGBTQ$^-$	LGBTQ$^+$	0.9575 ± 0.0340	0.9365 ± 0.0684
	EXTRA$^-$	EXTRA$^+$	0.9457 ± 0.0268	0.8901 ± 0.0542
	NEURO$^-$	NEURO$^+$	0.9540 ± 0.0323	0.7565 ± 0.1142
	AGREE$^-$	AGREE$^+$	0.9971 ± 0.0113	0.9979 ± 0.0098
	OPEN$^-$	OPEN$^+$	0.9998 ± 0.0003	0.9772 ± 0.0422
	CONSC$^-$	CONSC$^+$	0.9992 ± 0.0001	0.9545 ± 0.0487
	RELAT$^-$	RELAT$^+$	0.8352 ± 0.0629	0.7767 ± 0.0850
	NIHIL$^-$	NIHIL$^+$	0.7777 ± 0.0569	0.7817 ± 0.0831
	UTILI$^-$	UTILI$^+$	0.8316 ± 0.0357	0.7937 ± 0.0548
	VIRTUE$^-$	VIRTUE$^+$	0.8852 ± 0.0386	0.8303 ± 0.0638
	DEONT$^-$	DEONT$^+$	0.7681 ± 0.0800	0.7977 ± 0.1105
Level 1 Inter-Topic	all	CONSC$^+$	0.4739 ± 0.0238	0.7842 ± 0.0810
	all	LIBER$^+$	0.5729 ± 0.0304	0.8953 ± 0.0414
	all	IMMI$^+$	0.7401 ± 0.1462	0.9814 ± 0.0302
	all	LGBTQ$^+$	0.9742 ± 0.0465	0.9030 ± 0.0525
	all	EXTRA$^+$	0.5720 ± 0.1320	0.8573 ± 0.1017
	all	NEURO$^+$	0.9028 ± 0.0843	0.9242 ± 0.0595
	all	AGREE$^+$	0.4193 ± 0.0403	0.7131 ± 0.1078
	all	OPEN$^+$	0.5210 ± 0.0904	0.8943 ± 0.0593
	all	CONSC$^+$	0.4748 ± 0.0315	0.8367 ± 0.1182
	all	RELAT$^+$	0.5051 ± 0.0151	0.9458 ± 0.0512
	all	NIHIL$^+$	0.9615 ± 0.0370	0.7927 ± 0.0860
	all	UTILI$^+$	0.9282 ± 0.1698	0.4997 ± 0.1916
	all	VIRTUE$^+$	0.6278 ± 0.1723	0.8911 ± 0.0471
	all	DEONT$^+$	0.4442 ± 0.1501	0.6616 ± 0.2216
Level 0 Intra-Topic	all	*Politics*$^+$	0.8850 ± 0.2070	0.9511 ± 0.0433
	all	*Ethics*$^+$	0.9958 ± 0.0103	0.8420 ± 0.0541
	all	*Personality*$^+$	0.9799 ± 0.0258	0.8682 ± 0.0701

in final third of layers. Results using Deep Scan suggested that political views have distinctly localized activations in the last layer of *Llama3*, and ethical values show greater polysemantic overlap.

Our analysis is specific to the selected group of datasets and may not generalize to other data sources. The datasets are written in English and primarily reflect WEIRD perspectives[9] [10], and political views largely centered on U.S. politics. Future research should explore a wider range of models and personas, and incorporate beliefs, values, and traits from more diverse cultural contexts. Additionally, we will explore controlled modifications of internal representations—specifically, at the salient activations we identified—which might provide deeper insights into the mechanisms underlying an LLM's encoding of personas.

7 Impact Statement

Our work investigates how personality traits, ethical values, and political beliefs are encoded within LLMs. By analyzing the internal representations of these personas across different LLMs, we provide concrete insights into where these models internalize human values and behaviors. Our findings also offer opportunities for future research on aligning LLM outputs more nuancedly with societal values, such as ensuring a diversity of beliefs and values or enhancing safer user-centric experiences, for example, by improving persona-specific responses.

[9] Western, Educated, Industrialized, Rich, Democratic population.