Table 5: **(Q1)**: Metrics to assess the goodness of *layer 1* and *31* to detect personas (Level 2) in *Llama3-8B-Instruct*. Metrics: Silhouette Score (SH), Calinski-Harabasz Score (CH), Euclidean Distance (ED), Davies-Bouldin Score (DB).

Model	Dimension	layer	SH (↑)	CH (↑)	DB (↓)	ED (↑)
Llama3-8B-Instruct	AGREE	1	0.500 ± 0.0741	340.6 ± 163.9	0.731 ± 0.072	0.403 ± 0.012
		31	0.792 ± 0.0000	3264.5 ± 0.002	0.326 ± 0.000	27.57 ± 0.000
	CONSC	1	0.635 ± 0.0000	718.8 ± 0.026	0.569 ± 0.000	0.370 ± 0.000
		31	0.813 ± 0.0000	4150.4 ± 0.003	0.285 ± 0.000	27.47 ± 0.000
	OPEN	1	0.602 ± 0.0000	570.2 ± 0.0005	0.645 ± 0.000	0.414 ± 0.000
		31	0.795 ± 0.0000	3564.1 ± 0.0125	0.319 ± 0.000	27.60 ± 0.000
	EXTRA	1	0.578 ± 0.0000	527.5 ± 0.001	0.705 ± 0.000	0.382 ± 0.000
		31	0.788 ± 0.0000	3176.5 ± 0.001	0.330 ± 0.000	27.47 ± 0.000
	NEURO	1	0.584 ± 0.0000	615.0 ± 0.065	0.686 ± 0.000	0.378 ± 0.000
		31	0.796 ± 0.0000	3372.4 ± 0.001	0.306 ± 0.000	27.22 ± 0.000
	VIRTUE	1	0.614 ± 0.0000	644.7 ± 0.007	0.639 ± 0.000	0.402 ± 0.000
		31	0.797 ± 0.0000	3471.5 ± 0.000	0.309 ± 0.000	27.24 ± 0.000
	RELAT	1	0.500 ± 0.0000	405.4 ± 0.000	0.935 ± 0.000	0.440 ± 0.000
		31	0.814 ± 0.0000	3960.0 ± 0.002	0.276 ± 0.000	28.06 ± 0.000
	DEONT	1	0.523 ± 0.0230	315.1 ± 104.9	0.758 ± 0.000	0.479 ± 0.054
		31	0.824 ± 0.0000	4297.7 ± 0.004	0.254 ± 0.000	28.17 ± 0.000
	UTILI	1	0.542 ± 0.0405	326.0 ± 160.5	0.645 ± 0.000	0.454 ± 0.062
		31	0.798 ± 0.0000	3587.6 ± 0.001	0.301 ± 0.000	27.59 ± 0.000
	NIHIL	1	0.556 ± 0.0000	472.8 ± 0.001	0.810 ± 0.000	0.421 ± 0.000
		31	0.808 ± 0.0000	3909.8 ± 0.016	0.282 ± 0.000	27.20 ± 0.000
	CONS	1	0.685 ± 0.0001	860.7 ± 0.024	0.475 ± 0.000	0.400 ± 0.000
		31	0.818 ± 0.0000	4529.9 ± 0.007	0.266 ± 0.000	27.17 ± 0.000
	LIBER	1	0.692 ± 0.0001	897.2 ± 0.023	0.469 ± 0.000	0.387 ± 0.000
		31	0.803 ± 0.0000	3809.3 ± 0.004	0.294 ± 0.000	26.80 ± 0.000
	LGBTQ	1	0.634 ± 0.0000	703.1 ± 0.014	0.573 ± 0.000	0.399 ± 0.000
		31	0.774 ± 0.0000	2815.3 ± 0.034	0.352 ± 0.000	27.62 ± 0.000
	IMMI	1	0.570 ± 0.0000	476.5 ± 0.000	0.730 ± 0.000	0.424 ± 0.000
		31	0.806 ± 0.0000	3756.4 ± 0.002	0.288 ± 0.000	27.67 ± 0.000