Table 7: **(Q1)**: Metrics to assess the goodness of *layer 1* and *31* to detect personas (Level 2) in *Granite-7B-Instruct*. Metrics: Silhouette Score (SH), Calinski-Harabasz Score (CH), Euclidean Distance (ED), Davies-Bouldin Score (DB).

Model	Dimension	layer	SH (↑)	CH (↑)	DB (↓)	ED (↑)
Granite-7B-Instruct	AGREE	1	0.420 ± 0.0312	210.7 ± 75.01	0.925 ± 0.225	0.287 ± 0.019
		31	0.611 ± 0.0000	969.0 ± 0.000	0.605 ± 0.000	17.62 ± 0.000
	CONSC	1	0.561 ± 0.0000	578.4 ± 0.003	0.709 ± 0.000	0.250 ± 0.000
		31	0.625 ± 0.0000	1033.5 ± 0.002	0.573 ± 0.000	16.83 ± 0.000
	OPEN	1	0.491 ± 0.0000	312.4 ± 0.000	0.912 ± 0.000	0.265 ± 0.000
		31	0.991 ± 0.0000	37132.7 ± 0.038	0.004 ± 0.000	2073.7 ± 0.000
	EXTRA	1	0.529 ± 0.0000	397.0 ± 0.000	0.820 ± 0.000	0.265 ± 0.000
		31	0.618 ± 0.0000	938.7 ± 0.000	0.595 ± 0.000	18.58 ± 0.000
	NEURO	1	0.491 ± 0.0362	270.3 ± 156.9	0.707 ± 0.157	0.279 ± 0.044
		31	0.637 ± 0.0000	1128.8 ± 0.006	0.549 ± 0.000	18.66 ± 0.000
	VIRTUE	1	0.533 ± 0.0001	350.0 ± 0.002	0.803 ± 0.000	0.242 ± 0.000
		31	0.996 ± 0.0000	164373.4 ± 0.028	0.000 ± 0.000	27948.3 ± 0.000
	RELAT	1	0.491 ± 0.0000	332.1 ± 0.000	0.927 ± 0.000	0.256 ± 0.000
		31	0.614 ± 0.000	828.7 ± 0.000	0.602 ± 0.000	17.32 ± 0.000
	DEONT	1	0.462 ± 0.0023	171.7 ± 65.54	0.678 ± 0.241	0.335 ± 0.048
		31	0.605 ± 0.0000	871.0 ± 0.005	0.622 ± 0.000	17.42 ± 0.000
	UTILI	1	0.481 ± 0.0000	286.0 ± 0.001	0.956 ± 0.000	0.262 ± 0.000
		31	0.994 ± 0.0000	80355.2 ± 0.057	0.002 ± 0.000	3690.19 ± 0.000
	NIHIL	1	0.441 ± 0.0208	184.3 ± 65.67	0.730 ± 0.229	0.300 ± 0.025
		31	0.602 ± 0.0000	292.7 ± 0.000	0.678 ± 0.000	18.44 ± 0.000
	CONS	1	0.511 ± 0.0659	284.2 ± 160.3	0.683 ± 0.126	0.276 ± 0.025
		31	0.993 ± 0.0000	8423.03 ± 0.000	0.001 ± 0.000	21282.9 ± 0.000
	LIBER	1	0.536 ± 0.0529	378.8 ± 236.3	0.624 ± 0.104	0.275 ± 0.039
		31	0.993 ± 0.0000	72452.4 ± 0.017	0.002 ± 0.000	3292.78 ± 0.000
	LGBTQ	1	0.497 ± 0.0509	290.2 ± 167.2	0.682 ± 0.126	0.283 ± 0.039
		31	0.593 ± 0.0000	838.5 ± 0.010	0.643 ± 0.000	15.95 ± 0.000
	IMMI	1	0.510 ± 0.0000	346.1 ± 0.000	0.871 ± 0.000	0.257 ± 0.000
		31	0.617 ± 0.0000	989.7 ± 0.000	0.594 ± 0.000	17.32 ± 0.000