Table 1: **(Q1)** Separation of principal component representations in early (1) vs. late (31) layers (ℓ) of *Llama3* for *Personality* personas. Metrics: Silhouette (Si), Calinski-Harabasz (CH), Euclidean (ED), and Davies-Bouldin (DB). Results are averaged over five seeds (std=0.00, except $\star \approx 0.1$). **Best result** across layers and models. See Appendix Table 5, Table 6, Figure 5, and Table 7 for full results.

Topic	ℓ	SH (↑)	CH (↑)	ED (↑)	DB (↓)
AGREE	1	0.500	340.6*	0.403	0.731
	31	**0.792**	**3264.5**	**27.57**	**0.326**
CONSC	1	0.635	718.8	0.370	0.569
	31	**0.813**	**4150.4**	**27.47**	**0.285**
OPEN	1	0.602	570.2	0.414	0.645
	31	0.795	3564.1	27.60	0.319
EXTRA	1	0.578	527.5	0.382	0.705
	31	**0.788**	**3176.5**	**27.47**	**0.330**
NEURO	1	0.584	615.0	0.378	0.686
	31	**0.796**	**3372.4**	**27.22**	**0.306**

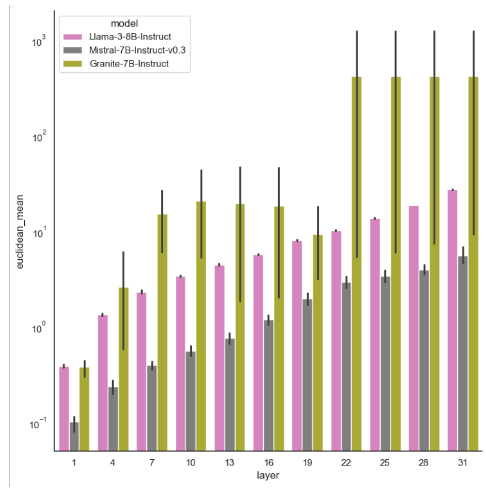

Figure 2: **(Q1)** Euclidean distances between PCA convex hull centroids for MATCHINGBEHAVIOR vs. NOTMATCHINGBEHAVIOR sentences averaged over *Primary Personality Dimensions.*

activations: $S^* = \arg\max_S F(S)$. To efficiently search for this subset Deep Scan uses non-parametric scan statistics (NPSS) [84]. There are three steps to using NPSS on the LLM's activation vectors:

1. **Expectation**: Forming a distribution of "expected" values at each position O_j of the activation vector. We call this expectation our null hypothesis H_0. For instance, we generate the expected distribution over the set of embedding vectors corresponding to NOTMATCHINGBEHAVIOR sentences.
2. **Comparison**: Comparison of embeddings of test set sentences against our expectation H_0. The test set may contain statements from the same distribution as H_0 (e.g., NOTMATCHINGBEHAVIOR) and from the alternative hypothesis H_1 (e.g, MATCHINGBEHAVIOR), which is the hypothesis we are interested in localizing. For each test activation e_{mj}, corresponding to a test sentence X_m and activation position O_j, we compute an empirical p-value. This is defined as the fraction of embeddings from H_0 (Step 1) that exceed the activation value e_{mj}.
3. **Scoring**: We measure the degree of saliency of the resulting test p-values by finding X_S and O_S that maximize the score function F, which estimates how much the observed distribution of p-values from Step 2 deviates from expectation.

Deep Scan uses an iterative ascent procedure that alternates between: 1) identifying the most persona-driven subset of sentences for a fixed subset of activation units, and 2) identifying the most persona-driven subset of activations that maximizes the score for a fixed subset of sentences. For more details on Deep Scan, refer to prior work [51, 53]. This results in the most persona-driven subset $S^* = X_{S^*} \times O_{S^*}$, where O_{S^*} is the localization of a given persona in our study.

Localization Levels. We localize personas at different levels of granularity, corresponding to different hypotheses H_0 and H_1 (see Table 2): At *Level 2* (inter-persona), we identify activations that differentiate MATCHINGBEHAVIOR from NOTMATCHINGBEHAVIOR sentences within the same persona (e.g., CONS$^+$ vs. CONS$^-$); at *Level 1* (intra-topic), we identify activations distinguishing a specific persona from all other personas within the same topic (e.g., CONS$^+$ vs. $\{$LIBER$^+ \cup$ IMMI$^+ \cup$ LGBTQ$^+\}$); at *Level 0* (inter-topic), we identify activations that are common to all personas within a topic and differentiate them from those in other topics (e.g., *Politics*$^+$ vs. $\{$*Ethics*$^+ \cup$ *Personality*$^+\}$).

Precision and Recall of Sentences Subset. To validate the usefulness of the identified salient activations O_{S^*}, we report precision and recall of the corresponding subset of sentences identified X_{S^*} with respect to the identification hypothesis H_1. In our context, precision is the fraction of test sentences in X_{S^*} that truly satisfy H_1 (accuracy of our positive detections), and recall is the fraction of test sentences that satisfy H_1 and are included in X_{S^*} (coverage).

5 Results

We now present and discuss our findings related to our research questions, **(Q1)** and **(Q2)**, as outlined in § 3.4. We denote the first layer (simple input layer) as 0, and the last layer as 31.