4 Localization of Persona Representations

We provide an overview of the methods used to localize persona representations in LLMs. We first describe the methods used to identify and validate the layer in a given model where the embeddings of a specific persona differ most from those of others (§ 4.1), then present the approach for identifying the subset of activations within that layer that play a critical role in encoding a particular persona compared to other personas (§ 4.2).

4.1 Identifying Layers With Strongest Persona Representations

To investigate where persona representations are encoded **(Q1)**, we aim to identify the model layer at which the embeddings for a specific persona (MATCHINGBEHAVIOR) deviate most from those of other personas (NOTMATCHINGBEHAVIOR).[5]

For a given layer, let e^+ represent the set of embedding vectors corresponding to MATCHINGBEHAVIOR sentences, and e^- represent the set of embedding vectors corresponding to NOTMATCHINGBEHAVIOR sentences. Given the high-dimensional nature of these embeddings, we perform dimensionality reduction and compute their principal components (PCs) over the combined set of embeddings ($e^+ \cup e^-$). We denote the embeddings in the PC space as q^+ for MATCHINGBEHAVIOR and q^- for NOTMATCHINGBEHAVIOR.[6] We use several clustering metrics to quantify the differences between these two sets. Thereby, we treat each set, q^+ and q^-, as a cluster and compute the following distance metrics and scores.[7] We report results in § 5 over five independent runs, each using $q^+ = q^- = 100$ randomly sampled data points.

Calinski-Harabasz Score. The score is defined as the ratio of the sum of between-cluster dispersion (BCD) and within-cluster dispersion (WCD) [81]. BCD measures how well clusters are separated from each other. WCD measures the cluster compactness or cohesiveness.

Silhouette Score. The score is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample [82]. Values near 0 would indicate that representations from q^+ and q^- overlap, thus indicating non-sufficient capabilities to capture the given dimension.

Davies-Bouldin Score. The score is defined as the average similarity metric of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances [80, 83]. Thus, farther apart and less dispersed clusters will result in a better score.

Euclidean Distance. We measure the Euclidean distance between centroids C^+ and C^-; where $C^j = \frac{\sum_{p \in Conv(q^j)} p}{|Conv(q^j)|}$, with the convex hull $Conv(q^j)$ as the minimal convex set containing all points p in q^j.

4.2 Identifying a Layer's Activations With Strongest Persona Representations

For our second question **(Q2)**, we examine whether there are consistent activation patterns—distinct groups within sentence embedding vectors—that systematically encode different personas within a given layer. Inspired by previous work [51, 53], we adopt Deep Scan to analyze systematic shifts in neural network activation spaces. For additional related work, see § 2. We now present the method formally.

Let an LLM encode a statement X_m at a layer into an activation vector e_m. For instance, e_m could be the last token embedding vector at the last layer of a *Llama3* model that takes as input the statement X_m : "I believe strongly in family values and traditions", which is a MATCHINGBEHAVIOR for the CONS dimension. Each activation vector e_m consists of J activation units e_{mj}. The positions in this activation vector form the set of $O = \{O_1 \cdots O_J\}$ elements. Thus, J is the dimensionality of the embedding space, e.g., for *Llama3*, $J = 4096$ (see § 3.3). Consider a set of statements from a given persona dataset (e.g., CONSC), denoted as $X = \{X_1, \ldots, X_M\}$. Let $X_S \subseteq X$ and $O_S \subseteq O$, then we define a subset as $S = X_S \times O_S$. We call this a subset of sentences and activations. Our goal is to find the most persona-specific subset. To do this, we use a score function $F(S)$, which quantifies the anomalousness of a subset S. For instance, given the CONS dataset, the scoring function $F(S')$ with $S' = \{X_m\} \times \{O_j\}$ measures how divergent the last token representation of a sentence X_m, is at a given embedding position O_j, compared to the last token representations of all other sentences that are labeled MATCHINGBEHAVIOR. Thus, Deep Scan seeks to find the most salient subset of

[5] See the assumptions of the dataset in § 3.2.

[6] Explained variance ratio across 14 dimensions is 0.657 to 0.898.

[7] We use the scikit-learn implementations [80].