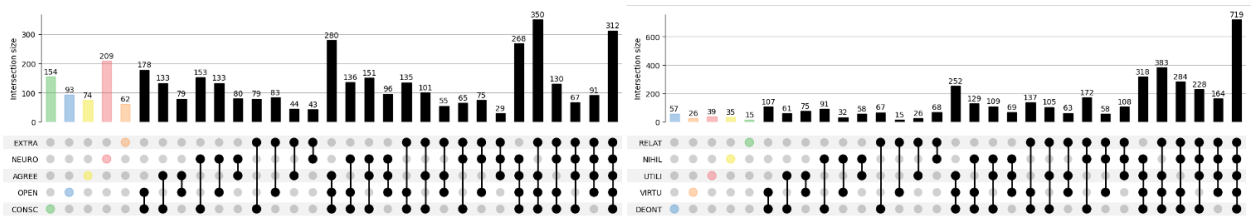

(a) (b)

Figure 3: **(Q2)** Upset plots illustrating the overlap of sets of salient last-layer activations from MATCHINGBEHAVIOR sentences, as identified by Deep Scan, across personas. Each bar represents the number of activations shared by a specific combination of personas.

(Q1) Which Layers and Models Show the Strongest Signal for Persona Representations?

We first study which layers provide the strongest signals for encoding personas for different LLMs. Specifically, we identify the layer that exhibits the greatest divergence between the principal components (PCs) of the last token representations for sentences corresponding to a given persona—comparing q_+ (MATCHINGBEHAVIOR) and q_- (NOTMATCHINGBEHAVIOR) sentences using the methods described in § 4.1. Our findings lay the groundwork for our next step, where we seek to localize sets of activations within a layer encoding persona information.

Results. Fig. 1b shows the first three PC embeddings for the IMMI persona across several layers, comparing q_+ and q_- embeddings. The PC embeddings overlap considerably in the initial layer, while later layers show increasing separation—with the clearest distinction in the final layer of *Llama3*. We find similar trends for other models and personas (see Appendix Fig. 7 and 8).

We use the metrics described in § 4.1 to quantify the separation between the two embedding groups. Fig. 2 shows the Euclidean distances (for all three models) between the centroids of the convex hulls for the two groups of clusters $q_- \cup q_+$, averaged over *Primary Personality Dimensions* personas. See Appendix Figure 6 for all personas. Across the models, the largest distances are found in the later layers (20–31). Tab. 1 reports additional metrics evaluating the separation, overlap, and compactness of the groups q_- and q_+. Most measures indicate that the final layer of *Llama3* achieves the strongest separation. We find, however, that for some personas, certain metrics favor earlier layers or other models. This suggests that while *Llama3* generally provides the best overall separation, for persona-specific applications, evaluating different metrics and models might be beneficial.

Overall, later layers exhibit the greatest separation between q_+ (MATCHINGBEHAVIOR) and q_- (NOTMATCHINGBEHAVIOR) across LLMs, indicating that persona representations become increasingly refined, with final layers encoding the most discriminative features. This aligns with prior work showing that higher layers capture more contextualized, task-specific information [85]. Among the models tested, *Llama3* demonstrated the strongest separation and most cohesive clusters in its final layer, suggesting it most effectively encodes persona-specific information. Consequently, our subsequent analysis focuses exclusively on the last-layer representations of *Llama3*.

(Q2) Are There Unique Locations of Persona Representations Within Layers?

Next, we investigate whether distinct, consistent activation groups within a layer encode different personas. Building on our previous findings, we compare the last token representations from *Llama3* for MATCHINGBEHAVIOR versus NOTMATCHINGBEHAVIOR sentences. We use Deep Scan (§ 4.2) to identify the activation subsets most indicative of persona-specific information O_{S^*}, which we refer to as *salient activations*.

Results. First, we validate the Deep Scan results as described in § 4.2. In Tab. 2 (*Level 2*), we report precision and recall of the corresponding X_{S^*}. We find high precision and recall for all 14 personas, with the precision ranging from 0.778 (NIHIL) to 0.999 (CONSC) and recall from 0.76 (NEURO) to 0.998 (AGREE). This showcases that the found O_{S^*} contains information needed to detect MATCHINGBEHAVIOR of a sentence for a given dimension.

After successful validation, we examine the overlap of salient activation subsets within personas of the same topic, namely *Ethics* (Fig. 3b), *Politics* (Fig. 1c), and *Personality* (Fig. 3a). Recall that the full embedding vector has a dimension of 4096 activations. For *Ethics* personas, only a small fraction of activations are unique—ranging from 0.37% (15 activations) to 1.39% (57)—indicating that few nodes exclusively differentiate each persona. In contrast, we find a substantial overlap among these personas, with 17.55% (719) of the activation vector shared across all. This suggests strong polysemanticity, where the same activation contributes to multiple ethical representations. In comparison, *Politics*