personas display much lower overlap, with only 9.42% (386) shared activations across all. *Personality* personas show a similarly modest overlap at 7.62% (312). *Politics* personas, however, exhibit a larger set of unique activations per persona, ranging from 2.05% (84) to 5.54% (227). Unique activations for *Personality* personas similarly range from 1.51% (62) to 5.10% (209). These findings suggest that individual *Politics*—and to a slightly lesser extent, *Personality*—personas are characterized by more distinct activation patterns. Overall, we find that some personas reveal clearly unique locations within the last-layer last representations of *Llama3*.

What Are the Activation Interactions Between Groups of Personas?

Now, we shift our focus to understanding whether we can differentiate between groups of specific personas (only using MATCHINGBEHAVIOR sentences) based on their embeddings. Specifically, we are interested if we can: (i) distinguish inter-topics, between personas associated with a particular topic (e.g., *Politics*) from other topics, e.g., $\{Ethics \cup Personality\}$, and ii) distinguish intra-topic between a single persona within a topic (e.g., LIBER) and other personas within the same topic, e.g., $\{\text{CONS} \cup \text{LGBTQ} \cup \text{IMMI}\}$. We believe this can provide insights on different levels of granularity that can inform interventions to generate output within a given persona.

Results. We validate our results by reporting the precision and recall of our salient node detection method (see Tab. 2). We achieve high performance at inter-topic *Level 0*. The lowest precision is 0.885 (*Politics*), and the lowest recall is 0.842 (*Ethics*). This suggests that our approach is highly effective at identifying topic-level activation patterns that all personas within a topic share and separating them from personas of other topics.

In contrast, our results are mixed at intra-topic *Level 1*. For 4 of the 14 evaluated personas, we observe high precision (ranging from 0.74 to 0.97) and high recall (ranging from 0.79 to 0.98), indicating reliable detection in these cases. However, for the remaining 9 personas, precision falls (majority ranging from 0.42 to 0.63, with the exception of UTILI at 0.93), and recall is generally lower (ranging from 0.66 to 0.96). This suggests that we can detect broad, inter-topic differences, and patterns are found to be less consistent for intra-topic distinctions—possibly due to overlapping activation patterns or less pronounced differentiating features among some personas.

Given these observations, we focus only on the interplay between salient activations of *Level 0* and *Level 2* in the further analysis. First, at *Level 0*, we find no overlap among salient activations of all three topics—*Ethics*, *Personality*, and *Politics*. In pairwise comparisons, we observe that there is no overlap between *Ethics* and *Personality*, a modest overlap of approximately 7% of activations between *Ethics* and *Politics*, and the largest overlap of roughly 12% between *Politics* and *Personality*. Consequently, the unique nodes attributed to each topic are about 93% for *Ethics*, 88% for *Personality*, and 85% for *Politics*[8]. These findings suggest that distinct activation locations characterize each topic. At the same time, a certain degree of commonality (polysemanticity) remains—particularly between *Politics* and *Personality*—which may reflect shared underlying conceptual features in their representations.

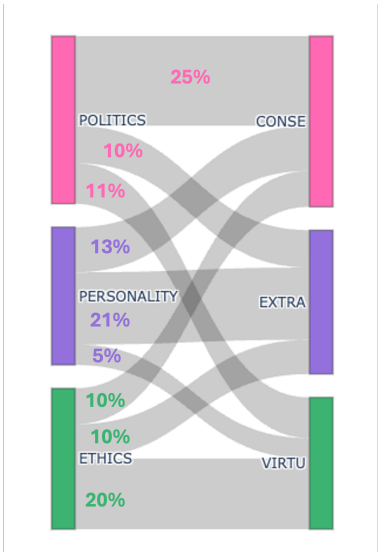

Figure 4: **(Q2)** Overlap of salient activations between topics and sampled persona from each topic.

Lastly, we are interested in understanding how the inter-topic activations (*Level 0*) relate to more detailed inter-persona patterns (*Level 2*). In Fig. 4, we show the overlap of salient activations between these levels for one example persona per topic. We observe an overlap of 25% of the salient activations between *Politics* and political persona CONSC. Similarly, for *Personality* and personality trait EXTRA, we find a 21% overlap, and for *Ethics* and ethical persona VIRTUE, a 20% overlap.

These findings suggest that a significant portion of a persona's encoding includes topic information, while the observed overlaps with other persona topics indicate that some activations are shared across these representational spaces.

6 Summary, Limitations, and Future Work

We investigated where LLMs encode persona-related information within their internal representations, analyzing last token activation vectors from 3 families of decoder-only LLMs using persona-specific statements from 14 datasets across *Politics*, *Ethics*, and *Personality* topics. Our PCA showed the strongest signal in separating persona information

[8]For a visualization, see Appendix Fig. 10.