- *Political Views*, which capture ideological leanings and political affiliations, reflecting individuals' values and preferences for societal order. We consider a subset of four personas:[3] *politically-conservative* (CONS), *politically-liberal* (LIBER), *anti-immigration* (IMMI), and *anti-LGBTQ-rights* (LGBTQ).

For examples of each of the above personas and more explanations, refer to Appendix Tab. 3. By examining these three topics of personas, we can uncover the interplay between these facets.

Following [5], we filter the original dimensions based on a minimum number of sufficiently confident examples. We impose a minimum label confidence of 0.85 and a minimum count of 300 for each direction, MATCHINGBEHAVIOR and NOTMATCHINGBEHAVIOR, of each persona dimension, resulting in a total of 600 examples per persona dataset.

Assumptions of the Dataset. A key assumption underlying the dataset, which is crucial for interpreting our study, is that a NOTMATCHINGBEHAVIOR label does not inherently imply MATCHINGBEHAVIOR for a different persona.[4] For example, a statement from the *politically-liberal* dataset labeled as NOTMATCHINGBEHAVIOR does not automatically mean it aligns with sentences in the *politically-conservative* dataset. Consider the sentence: "Women should be subservient to men." This appears in the *politically-liberal* dataset with a NOTMATCHINGBEHAVIOR label, but this does not imply that a similar sentence exists in the *politically-conservative* dataset (and, in fact, does not). At the same time, we do observe some overlap between persona datasets. For instance, the sentence "I support marriage equality and LGBTQ rights." is labeled as MATCHINGBEHAVIOR in the *politically-liberal* dataset and NOTMATCHINGBEHAVIOR in the *anti-LGBTQ-rights* dataset. It is crucial to understand that the label does not indicate movement along a continuous axis but instead indicates the presence of a behavior.

3.3 Selection of LLMs and Embedding Vectors

We study the internal representations of three models, *Llama3-8B-Instruct* (*Llama3*) [74], *Granite-7B-Instruct* (*Granite*) [75], and *Mistral-7B-Instruct* (*Mistral*) [76]. We focus on instruct models because, unlike base LLMs that rely on a next-word prediction objective, instruct models are fine-tuned specifically for instruction following [22]. They are typically trained using supervised fine-tuning with question-answer pairs annotated by human experts and reinforcement learning with human feedback, allowing them to learn which responses are most useful or relevant to humans [77]. As a result, these models are likely better trained to adhere to persona behaviors. Additionally, instruct models tend to exhibit more predictable behavior than base models [22], making them more reliable for controlled experiments.

We extract the representation vectors at each layer from each model's forward pass when processing MATCHINGBEHAVIOR and NOTMATCHINGBEHAVIOR statements for a given dimension. We only keep the vector corresponding to the last token of each sentence at each layer as it contains relevant and summarized information of the whole sentence [78]. All models considered in this study are decoder-only models with 32 layers, and the activation vector from the last token has a shape of $(1, 4096)$. For a more detailed description of the specific models, see Appendix A.2.

3.4 Research Questions

In this study, we aim to answer two key questions: **(Q1)** Where in the model are persona representations encoded? **(Q2)** How do these representations vary across different personas?

For **(Q1)**, we investigate which layers in LLMs exhibit the strongest signal for encoding persona-specific information. This is important because knowing the layer-wise distribution of persona features can provide better insights into how complex behavioral and human characteristics are encoded in the model. Such insights could drive improvements in model interpretability and enable targeted interventions. Prior work has shown that transformer architectures tend to localize different types of linguistic and semantic information in distinct layers [79], yet the encoding of persona-specific characteristics remains under-explored.

For **(Q2)**, we seek to determine whether there are consistent, unique locations within a given LLM layer where distinct persona representations are encoded. Uncovering such patterns is crucial to understanding whether persona features are confined to specific subspaces within the model. This finding could facilitate more effective methods for controlling and customizing LLM outputs according to desired persona traits. Previous research in neural network interpretability has identified specialized neurons for various linguistic functions [2]. Similar structures regarding persona representations have not yet been studied.

[3] We exclude BELIEVES-IN-GUNRIGHTS and BELIEVES-ABORTION-SHOULD-BE-ILLEGAL.

[4] Prompts asked for statements the persona "would agree with, but others would disagree with," where *others* refers to any persona not aligned with the one under consideration [3].