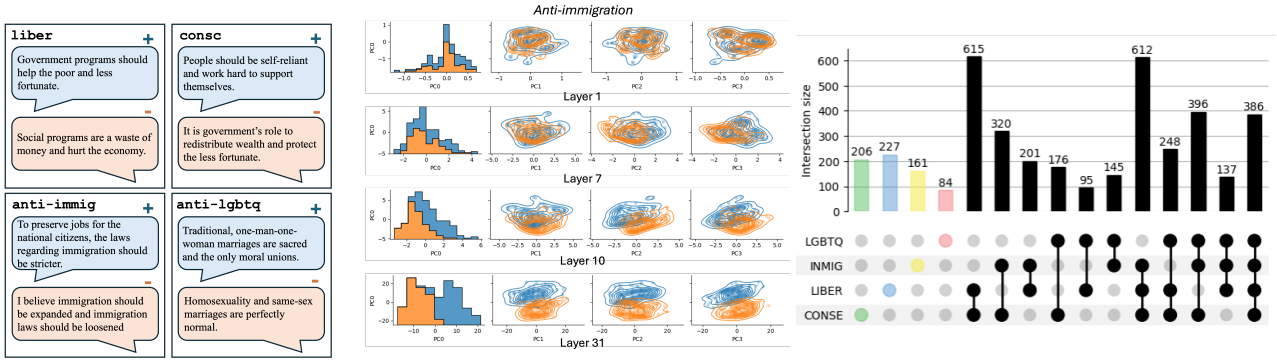

(a) Persona-specific datasets. (b) PCA of activations. (c) Overlap of salient sets of activations.

Figure 1: Overview of our study for *Llama3* on persona topic *Politics*. **(a)** Examples of MATCHINGBEHAVIOR (+) and NOTMATCHINGBEHAVIOR (–) persona statements. **(b)** PCA of representations for + and - sentences for *anti-immigration* **(Q1)**. **(c)** Overlap of sets of salient last-layer activations from + sentences, as identified by Deep Scan, across personas **(Q2)**.

Unlike traditional fair-ML decision-making frameworks, which optimize explicitly defined objectives (e.g., maximizing utility subject to a fairness metric [17, 18]), LLMs learn their decision patterns from massive, largely uncurated text corpora [19, 20], rendering their latent moral or political predispositions opaque. It is possible to refine an LLM's behavior via supervised fine-tuning on small, carefully curated datasets [21, 22]. However, the inherently open-ended nature of linguistic output makes it difficult to anticipate and constrain every downstream use case. If an LLM contains a latent political bias or moral preference that goes undetected, it may systematically privilege certain viewpoints or value systems over others, potentially amplifying polarizing or discriminatory content in high-stakes settings [9, 23, 24, 25]. Research has indeed identified gender biases in LLMs, which favor males' perspectives [26], and political biases favoring liberal viewpoints [9]. These biases are also extended to personality, morality, and stereotypes [23, 27].

Responsible AI governance demands both transparency and explainability ("What does the model encode?") as well as controllability ("How can we steer or constrain its outputs?") [28, 25].Consequently, there has been a growing call to better understand how and where personas are encoded within an LLM's internal representations [29, 30, 31, 32]. Such insights can improve the model's interpretability, transparency, and inform methods to align LLM outputs with human values [32, 33].

In this work, we aim to better understand where personas are encoded in the internal representations of LLMs (see Fig. 1 for an overview). We rely on a publicly available collection of model-generated personas [3] specifically focusing on three categories, spanning human identity and behavior: *Politics*, which include ideological leanings and political affiliations, reflecting individuals' values and societal preferences (e.g., liberal, conservative); *Ethics*, which captures moral reasoning and value-based judgments, central to human decision-making and social interactions (e.g., deontology, utilitarianism); and *Primary Personality Traits* (*Personality*), based on the Big Five model [34, 35], which provides a comprehensive framework for understanding human behavior and interpersonal dynamics (e.g., agreeableness, conscientiousness). These personas span a wide range of values, beliefs, and social preferences, providing a grounded basis for studying how LLMs encode complex human attributes [35].

We feed statements associated with different personas (see Fig. 1a) into various LLMs and extract their internal representations (i.e., activation vectors). We then analyze these representations to address the following two questions:

- **(Q1)** Where in the model are persona representations encoded? Specifically, which layers in the LLM exhibit the strongest signals for encoding persona-specific information (see Fig. 1b)?
- **(Q2)** How do these representations vary across different personas? In particular, are there consistent, uniquely activated locations within a given LLM layer where distinct persona representations are encoded (see Fig. 1c)?

This approach enables us to systematically investigate how LLMs process and differentiate persona-related information.

Our main findings are:

- The final third segment of layers (across *Llama3-8B-Instruct* (*Llama3*), *Granite-7B-Instruct*, and *Mistral-7B-Instruct*) captures the most variance in persona representations, with the last layers exhibiting the strongest separability along principal components. This suggests that higher-level semantic abstractions related to human values are encoded in later layers of the model families. *Llama3* exhibits the largest separation across layers and personas, with the last layer showing the clearest separation.