- For embeddings in *Llama3*'s last layer, we find that personas of different ethical values exhibit the highest activation overlap (17.6% of the embedding vector), while personas with different political beliefs have the most uniquely associated activations (2.1%–5.5%). This suggests that political views are more distinctly localized, while ethical views are more polysemous, sharing activations across multiple concepts.

The remainder of the paper is structured as follows: § 2 reviews related work; § 3 details the study design, dataset, and models; § 4 describes the methods used for our analysis; § 5 presents our empirical results and associated discussion; and § 6 summarizes the paper, identifies limitations, and provides directions for future work.

2 Related Work

Personas. LLM personas have attracted increasing attention for their ability to personalize and steer outputs, particularly in the development of trustworthy models [36, 37]. A persona is a natural language portrayal of an imagined individual belonging to some demographic group or reflecting certain personality traits [38, 39]. Research has found that personas can induce model outputs that are toxic and propagate stereotypes [23, 40, 41], and contain more extreme views with respect to Western vs. non-Western norms [42]. Moreover, personas have been (mis)used to circumvent built-in safety mechanisms by instructing models to adopt specific roles, e.g., that of fictional characters [43]. Understanding how LLMs encode personas is essential for harm mitigation methods [44], aligning models with diverse beliefs [10], and tailoring outputs to users' preferences.

Behavior and Value Encoding in LLMs. Early work argues that representations in BERT [45] can reveal the implicit moral and ethical values embedded in text [46]. These studies focus on quantifying deontological ethics, which determine whether an action is intrinsically right or wrong by analyzing output embeddings. We extend these ideas to a broader spectrum of human values, beliefs, and traits consisting of 14 different personas. Furthermore, we look at internal representations rather than output embeddings.

More recently, understanding an LLM's representations at different layers and token embeddings has gained increased attention [29, 47, 48], aiming to understand how concepts are represented within an LLM's (decoder) neural network, e.g., by generating human-understandable translations of information encoded in hidden representations [47]. In the context of human behavior, prior work has shown how neural activity across all layers can build feature vectors for honest and dishonest behavior detection [29]. In contrast, our work focuses on identifying subsets of the activation vector that are most representative of a given persona.

The work most related to ours [32], performed (parallel to us) a layer-wise analysis of how LLMs encode three Big Five traits by training supervised classifiers on last-token embeddings and using layer-wise perturbations to edit expressed personalities. In contrast, we (i) probe a broader set of moral, personality, and political personas, (ii) identify the minimal subset of embedding dimensions in each layer that drives each persona, and (iii) quantify overlaps between these persona embedding subsets. We link those overlaps to a phenomenon that is often described as polysemanticity, where individual neurons respond to mixtures of seemingly unrelated inputs, which affects interpretability and impacts the generation process [49, 50]. We further show that a classifier using only the activations corresponding to our identified subset of embedding dimensions yields the same test performance as one using the full embedding, thus validating the importance of this subset in encoding a persona. While we do not edit embeddings in this work, these findings open the door for more efficient, fine-grained interventions.

Deep Scan. Deep Scan has been predominantly used to detect anomalous samples in various computer vision, text, and audio tasks by analyzing patterns in neural networks [51, 52, 53]. Recent work has also taken initial steps in exploring which subsets of activations are most responsible for encoding harmful concepts, such as toxicity [51]. We build on this work and extend it by focusing exclusively on measuring the localization consistency and uniqueness of activations that encode human beliefs, values, and traits as personas in LLMs. Our study offers a systematic framework for localizing persona representations and their interactions in LLMs.

3 Study Design

This section outlines the study design, including the socio-technical motivation (§ 3.1), the dataset selection and assumptions (§ 3.2), the models used (§ 3.3), and the motivation for our research questions (§ 3.4).