Localizing Persona Representations in LLMs

Celia Cintas[*1], **Miriam Rateike**[*1, 2], **Erik Miehling**[3], **Elizabeth Daly**[3], **Skyler Speakman**[1]

[1]IBM Research Africa
[2]Saarland University
[3]IBM Research Europe
{celia.cintas, miriam.rateike}@ibm.com

Abstract

We present a study on how and where personas – defined by distinct sets of human characteristics, values, and beliefs – are encoded in the representation space of large language models (LLMs). Using a range of dimension reduction and pattern recognition methods, we first identify the model layers that show the greatest divergence in encoding these representations. We then analyze the activations within a selected layer to examine how specific personas are encoded relative to others, including their shared and distinct embedding spaces. We find that, across multiple pre-trained decoder-only LLMs, the analyzed personas show large differences in representation space only within the final third of the decoder layers. We observe overlapping activations for specific ethical perspectives – such as moral nihilism and utilitarianism – suggesting a degree of polysemy. In contrast, political ideologies like conservatism and liberalism appear to be represented in more distinct regions. These findings help to improve our understanding of how LLMs internally represent information and can inform future efforts in refining the modulation of specific human traits in LLM outputs. ***Warning**: This paper includes potentially offensive sample statements.*

Introduction

Understanding the mechanisms by which large language models (LLMs) process information, store knowledge, and generate outputs remain key open questions in research (Hendrycks et al. 2021; Wang et al. 2023). One crucial and largely underexplored aspect of these models is how they encode human personality traits, ethical views, or political beliefs – often broadly referred to as *personas* (Perez et al. 2023).

A persona is a natural language portrayal of an imagined individual belonging to some demographic group or reflecting certain personality traits (Jiang et al. 2024; Cheng, Durmus, and Jurafsky 2023). Personas are often used to define the personality or perspective the LLM model should adopt when interacting with users (Salemi et al. 2024), e.g., by prompting "Suppose you are a person who . . . " followed by a description of a particular trait or belief. For instance, if the prompt states ". . . is highly agreeable", the model may generate more cooperative and empathetic responses. If the prompt states ". . . subscribes to the moral philosophy of utilitarianism", the model's outputs may prioritize maximizing overall well-being when making ethical decisions. This can significantly influence language generation by setting a tone appropriate for the context (e.g., empathetic or professional) and by affecting behavior and reasoning capabilities.

Personas can enhance user experience and engagement by making models more relatable and context-aware (Miaskiewicz and Kozar 2011; Salminen et al. 2022; Laine et al. 2024), and can improve generated output, such as when an expert familiar with a specific domain provides more effective descriptions than an expert from a different field (Salewski et al. 2023). Personas have also attracted increasing attention, particularly in the development of trustworthy models (Miehling et al. 2025; Liu, Diab, and Fried 2024; Luz de Araujo and Roth 2025). Previous research has demonstrated that personas may elicit toxic responses and perpetuate stereotypes in language models (Deshpande et al. 2023; Sheng et al. 2021; Rutinowski et al. 2024; Salewski et al. 2023),[1], and can produce extreme political or cultural views (Dammu et al. 2024; Mazeika et al. 2025). Moreover, personas have been (mis)used to circumvent built-in safety mechanisms by instructing models to adopt specific roles (Kumar et al. 2024; Shah et al. 2023). Understanding how LLMs encode personas is essential for harm mitigation methods (Wang et al. 2025), aligning models with diverse beliefs (Abdurahman et al. 2024), and tailoring outputs to users' preferences.

Unlike traditional fair-ML decision-making frameworks, which optimize explicitly defined objectives (e.g., maximizing utility subject to a fairness metric (Hardt, Price, and Srebro 2016; Corbett-Davies et al. 2017)), LLMs learn their decision patterns from massive, largely uncurated text corpora (Perełkiewicz and Poświata 2024; Liu et al. 2024), 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 (Parthasarathy et al. 2024; Zhang et al. 2024). 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

[*]These authors contributed equally.

[1]For an in-depth discussion of stereotype accuracy and inaccuracy in psychology, see Jussim et al. (2015).

or value systems over others, potentially amplifying polarizing or discriminatory content in high-stakes settings (Motoki, Pinho Neto, and Rodrigues 2024; Rutinowski et al. 2024; Santurkar et al. 2023; Chen et al. 2024).

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?") (Ferdaus et al. 2024; Chen et al. 2024). Consequently, there has been a growing call to better understand how and where personas are encoded within an LLM's internal representations (Zou et al. 2023; Jentzsch et al. 2019; Ferrando et al. 2024; Ju et al. 2025). Such insights can improve the model's interpretability, transparency, and inform methods to align LLM outputs with human values (Wu et al. 2025; Ju et al. 2025).

In this work, we investigate 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 (Perez et al. 2023) specifically focusing on three categories spanning human identity and behavior: *Politics*, which includes ideological leanings and political affiliations that reflect 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 (Roccas et al. 2002; Gosling, Rentfrow, and Swann Jr 2003), 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 (Gosling, Rentfrow, and Swann Jr 2003).

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.
- 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: We first review related work, followed by a detailed overview of the study design, dataset, and models, then we describe the methods used for our analysis, present our empirical results and associated discussion, and finally summarize the paper, identify limitations, and provides directions for future work.

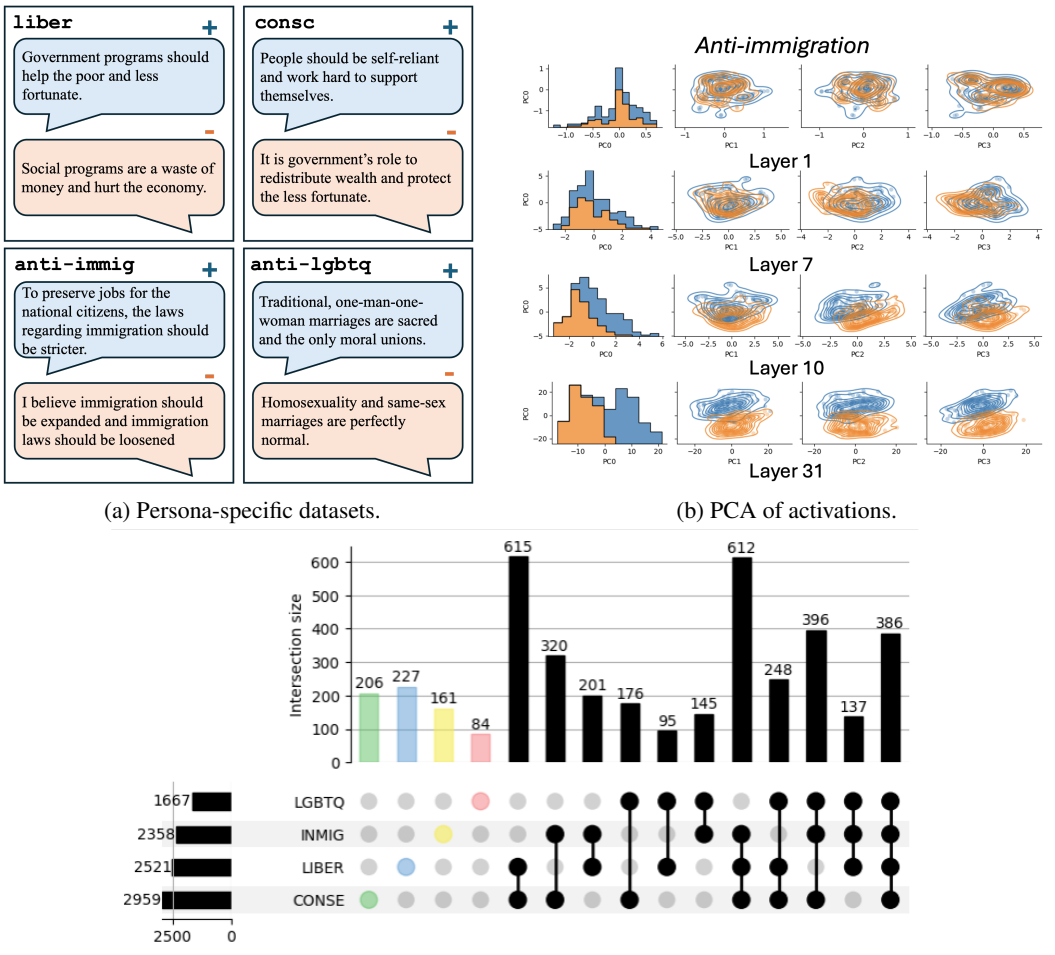
Related Work

Behavior and Value Encoding in LLMs. Early work argues that representations in BERT (Devlin et al. 2019) can reveal the implicit moral and ethical values embedded in text (Schramowski et al. 2019). 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 (Zou et al. 2023; Ghandeharioun et al. 2024; Singh et al. 2024), 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 (Ghandeharioun et al. 2024). 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 (Zou et al. 2023). 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 (Ju et al. 2025), 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 (Scherlis et al. 2022; Arora et al. 2018). 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 (Rateike et al. 2023; Akinwande et al. 2020; Cintas et al. 2022). Recent



(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 NOT-MATCHINGBEHAVIOR (–) 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 **(Q2)**. UpSet plot visualizes different sets of personas (connected dots on bottom) and how many salient activations are intersecting (vertical bars). We observe that the highest intersection corresponds to sentences from *politically-conservative* and *politically-liberal*.

work has also taken initial steps in exploring which subsets of activations are most responsible for encoding harmful concepts, such as toxicity (Rateike et al. 2023). 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.

Study Design

This section outlines the study design, including the socio-technical motivation, the dataset selection and assumptions, the models used, and the motivation for our research questions.

Socio-technical Motivation

The Big Five (openness, conscientiousness, extraversion, agreeableness, neuroticism) provide a well-validated framework for describing individual behavioral regularities or personality traits, which distinguish persons that are invariant over time and across situations (Goldberg 2013; John, Naumann, and Soto 2008). McCrae and Costa Jr (1994) argue that this framework shows the stability and consistency of personality traits, which help to predict how people will behave over time when placed in different situations. In human contexts, these traits correlate with patterns of decision-making (Özbağ 2016; Judge and Zapata 2015), social interaction (Baptiste 2018), and even vulnerability to manipulation (Argyle et al. 2023; Babakr and Fatahi 2023; Grover and Amit 2024). A growing body of work has leveraged the Big Five personality model to better understand LLM behavior, e.g., how prompts with personality trait information influences the out-

put (Mieleszczenko-Kowszewicz et al. 2024), the capacity of LLMs to infer Big Five personality traits from dialogues (Yan et al. 2024), the behavior of persona-instructed LLMs on personality tests, and creative writing tasks (Jiang et al. 2024).

As LLMs are increasingly used to support decision-making in high-stakes scenarios, it is important to understand which ethical perspective governs a model's proposed solution. Prior work has examined the alignment between LLM-driven decisions and human moral judgments through the lens of persona-based prompting (Garcia, Qian, and Palminteri 2024; Kim et al. 2025), showing, for instance, that political personas can significantly influence model behavior in ethical dilemmas (Kim et al. 2025). Other studies have developed benchmarks and systematic evaluations to assess the moral identity and decision-making patterns of LLMs (Ji et al. 2025; Lei et al. 2024). Others found fine-tuning to be a promising strategy for aligning LLM agents more closely with human values (Tennant, Hailes, and Musolesi 2025).

LLMs exposed to vast amounts of political text may internalize subtle political biases (Rutinowski et al. 2024; Motoki, Pinho Neto, and Rodrigues 2024). Such systematic leanings, e.g., when LLMs are used in chatbots, summarizers, and recommendation systems, can undermine democratic discourse, skew the information ecosystem, and disenfranchise heterodox viewpoints (Chen et al. 2022). Recent studies have shown that instruction-tuned LLMs can simulate divergent political viewpoints—sometimes classifying the same news outlet differently across runs—and often disagree with expert or human-annotated stances (Stammbach et al. 2024).

Finally, personality, ethics, and politics do not operate in isolation. For example, a model's moral reasoning may shift when presented through a particular political lens (e.g., justifications for decisions can be different for a conservative utilitarian than for a liberal utilitarian persona). Existing work has examined the output of LLMs with regard to personas with different combinations of demographics (Cheng, Durmus, and Jurafsky 2023; Khan et al. 2025; Magnossão de Paula et al. 2025). Here, as a first step toward exploring the intersections of personas across ethical, political, and personality dimensions, we examine the overlap in their embeddings.

Persona Datasets and Assumptions

Our experiments are based on the model-generated personas (Perez et al. 2023), consisting of statements written from the perspective of individuals with specific personalities, beliefs, or viewpoints (e.g., *extraversion* and *agreeableness*).[2] Each statement has an associated (model-generated) label indicating whether it matches the behavior of the corresponding persona dimension. For example, in the *extraversion* dataset, the sentence "Lively, adventurous, willing to take risks" is labeled as MATCHINGBEHAVIOR, whereas "I am quiet and don't socialize much" is labeled as NOTMATCHINGBEHAVIOR. As discussed in (Perez et al. 2023), an LLM was used to generate both the label and an associated confidence score. Detailed descriptions of the methodology used for the generation of these statements, the labeling process, and verification can be found in the original paper (Perez et al. 2023).

Persona Dimensions. Our work analyzes personas across three categories: personality, ethical theories, and political views. We examine three subsets of those topics, resulting in fourteen datasets:

- *Primary Personality Dimensions*, which relies on the Big Five (Goldberg 2013; Roccas et al. 2002), a widely recognized framework for understanding human behavior and interpersonal dynamics. The five personas considered are characterized by *agreeableness* (AGREE), *conscientiousness* (CONSC), *openness* (OPEN), *extraversion* (EXTRA), and *neuroticism* (NEURO).
- *Ethical Theories*, which influence moral reasoning and value-based judgment and are central to decision-making and social interactions. The five personas considered are *subscribes-to-virtue-ethics* (VIRTUE), *subscribes-to-culturalrelativism* (RELAT), *subscribes-to-deontology* (DEONT), *subscribes-to-utilitarianism* (UTILI), and *subscribes-to-moralnihilism* (NIHIL).
- *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.[4] By examining these three topics of personas, we can uncover the interplay between these facets.

Following (Miehling et al. 2025), 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.[5] 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,

[2]For a critical discussion on synthetic persona generation refer to Haxvig (2024).

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

[4]The appendices for this paper can be found in the complete ArXiv version https://arxiv.org/pdf/2505.24539.

[5]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 (Perez et al. 2023).

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.

Selection of LLMs and Embedding Vectors

We study the internal representations of three models, *Llama3-8B-Instruct* (*Llama3*) (AI@Meta 2024), *Granite-7B-Instruct* (*Granite*) (IBM Granite Team 2023), and *Mistral-7B-Instruct* (*Mistral*) (Jiang et al. 2023). 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 (Zhang et al. 2024). 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 (Cheng et al. 2024). 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 (Zhang et al. 2024), 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 (Sun et al. 2019). 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 B.

Research Questions

As introduced above, 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 (Tseng et al. 2024), 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 (Wang et al. 2023). Similar structures regarding persona representations have not yet been studied.

Localization of Persona Representations

We provide an overview of the methods used to localize persona representations in LLMs. We first describe the methods used to identify and validate the layer in a given model where the embeddings of a specific persona differ most from those of others, then present the approach for identifying the subset of activations within that layer that play a critical role in encoding a particular persona compared to other personas.

Identifying Layers With Strongest Persona Representations

To investigate where persona representations are encoded **(Q1)**, we aim to identify the model layer at which the embeddings for a specific persona (MATCHINGBEHAVIOR) deviate most from those of other personas (NOTMATCHINGBEHAVIOR).[6]

For a given layer, let e^+ represent the set of embedding vectors corresponding to MATCHINGBEHAVIOR sentences, and e^- represent the set of embedding vectors corresponding to NOTMATCHINGBEHAVIOR sentences. Given the high-dimensional nature of these embeddings, we perform dimensionality reduction and compute their principal components (PCs) over the combined set of embeddings ($e^+ \cup e^-$). We denote the embeddings in the PC space as q^+ for MATCHINGBEHAVIOR and q^- for NOTMATCHINGBEHAVIOR.[7] We use several clustering metrics to quantify the differences between these two sets. Thereby, we treat each set, q^+ and q^-, as a cluster and compute the following distance metrics and scores.[8] We report results in the next section over five independent runs, each using $q^+ = q^- = 100$ randomly sampled data points.

Calinski-Harabasz Score. The score is defined as the ratio of the sum of between-cluster dispersion (BCD) and within-cluster dispersion (WCD) (Caliński and Harabasz 1974). BCD measures how well clusters are separated from each other. WCD measures the cluster compactness or cohesiveness.

Silhouette Score. The score is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample (Rousseeuw 1987). Values near 0 would indicate that representations from q^+ and q^- overlap, thus indicating non-sufficient capabilities to capture the given dimension.

Davies-Bouldin Score. The score is defined as the average similarity metric of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances (Davies and Bouldin 1979; Pedregosa et al. 2011). Thus, farther apart and less dispersed clusters will result in a better score.

[6] See the assumptions of the dataset in the previous section.

[7] Explained variance ratio across 14 dimension: 0.657 to 0.898.

[8] We use the scikit-learn implementations (Pedregosa et al. 2011).

Euclidean Distance. We measure the Euclidean distance between centroids C^+ and C^-; where $C^j = \frac{\sum_{p \in Conv(q^j)} p}{|Conv(q^j)|}$, with the convex hull $Conv(q^j)$ as the minimal convex set containing all points p in q^j.

Identifying a Layer's Activations With Strongest Persona Representations

For our second question **(Q2)**, we examine whether there are consistent activation patterns – distinct groups within sentence embedding vectors – that systematically encode different personas within a given layer. Inspired by previous work (Rateike et al. 2023; Cintas et al. 2022), we adopt Deep Scan to analyze systematic shifts in neural network activation spaces. See above for additional related work. We now present the method formally.

Let an LLM encode a statement X_m at a layer into an activation vector e_m. For instance, e_m could represent the last token embedding from the final layer of a *Llama3* model, given the input statement X_m: "I believe strongly in family values and traditions," which is a sample sentence labeled as MATCHINGBEHAVIOR for the CONS dimension.

Each activation vector e_m consists of J activation units e_{mj}. The positions in this activation vector form the set of $O = \{O_1 \cdots O_J\}$ elements. Thus, J is the dimensionality of the embedding space, e.g., for *Llama3*, $J = 4096$. Consider a set of statements from a given persona dataset (e.g., CONSC), denoted as $X = \{X_1, \ldots, X_M\}$. Let $X_S \subseteq X$ and $O_S \subseteq O$, then we define a subset as $S = X_S \times O_S$. We call this a subset of sentences and activations. Our goal is to find the most persona-specific subset. To do this, we use a score function $F(S)$, which quantifies the anomalousness of a subset S. For instance, given the CONS dataset, the scoring function $F(S')$ with $S' = \{X_m\} \times \{O_j\}$ measures how divergent the last token representation of a sentence X_m, is at a given embedding position O_j, compared to the last token representations of all other sentences that are labeled MATCHINGBEHAVIOR. Thus, Deep Scan seeks to find the most salient subset of activations: $S^* = \arg\max_S F(S)$. To efficiently search for this subset, Deep Scan uses non-parametric scan statistics (NPSS) (McFowland, Speakman, and Neill 2013). There are three steps to using NPSS on the LLM's activation vectors:

1. **Expectation**: Forming a distribution of "expected" values at each position O_j of the activation vector. We call this expectation our null hypothesis H_0. Here, we generate the expected distribution over the set of embedding vectors corresponding to NOTMATCHINGBEHAVIOR sentences.
2. **Comparison**: Comparison of embeddings of test set sentences against our expectation H_0. The test set may contain statements from the same distribution as H_0 (e.g., NOTMATCHINGBEHAVIOR) and from the alternative hypothesis H_1 (e.g, MATCHINGBEHAVIOR), which is the hypothesis we are interested in localizing. For each test activation e_{mj}, corresponding to a test sentence X_m and activation position O_j, we compute an empirical p-value. This is defined as the fraction of embeddings from H_0 (Step 1) that exceed the activation value e_{mj}.
3. **Scoring**: We measure the degree of saliency of the resulting test p-values by finding X_S and O_S that maximize the score function F, which estimates how much the observed distribution of p-values from Step 2 deviates from expectation.

Deep Scan uses an iterative ascent procedure that alternates between: 1) identifying the most persona-driven subset of sentences for a fixed subset of activation units, and 2) identifying the most persona-driven subset of activations that maximizes the score for a fixed subset of sentences. For more details on Deep Scan, refer to prior work (Rateike et al. 2023; Cintas et al. 2022). This results in the most persona-driven subset $S^* = X_{S^*} \times O_{S^*}$, where O_{S^*} is the localization of a given persona in our study.

Localization Levels. We localize personas at different levels of granularity, corresponding to different hypotheses H_0 and H_1 (see Table 2): At *Level 2* (inter-persona), we identify activations that differentiate MATCHINGBEHAVIOR from NOTMATCHINGBEHAVIOR sentences within the same persona (e.g., CONS^+ vs. CONS^-); at *Level 1* (intra-topic), we identify activations distinguishing a specific persona from all other personas within the same topic (e.g., CONS^+ vs. $\{\text{LIBER}^+ \cup \text{IMMI}^+ \cup \text{LGBTQ}^+\}$); at *Level 0* (inter-topic), we identify activations that are common to all personas within a topic and differentiate them from those in other topics (e.g., *Politics*$^+$ vs. {*Ethics*$^+ \cup$ *Personality*$^+$}).

Precision and Recall of Sentences Subset. To validate the usefulness of the identified salient activations O_{S^*}, we report precision and recall of the corresponding subset of sentences identified X_{S^*} with respect to the identification hypothesis H_1. In our context, precision is the fraction of test sentences in X_{S^*} that truly satisfy H_1 (accuracy of our positive detections), and recall is the fraction of test sentences that satisfy H_1 and are included in X_{S^*} (coverage).

Results

We now present and discuss our findings related to our research questions, **(Q1)** and **(Q2)**, as outlined above. [9] We denote the first layer (simple input layer) as 0, and the last layer as 31.

(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 above. 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

[9] Code at https://github.com/IBM/personas-llms-analysis.

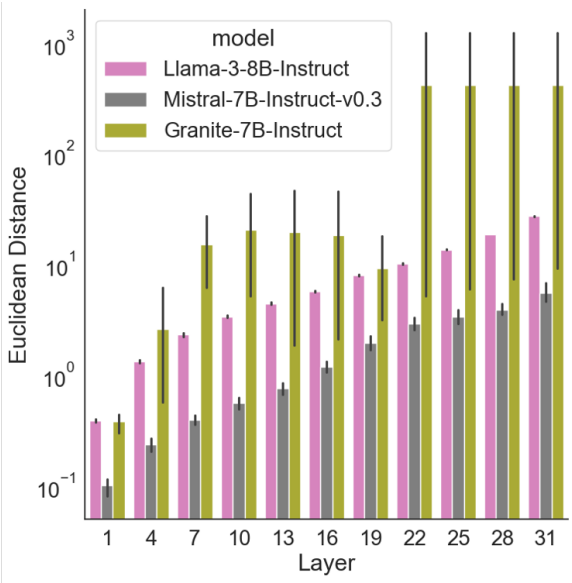


Figure 2: **(Q1)** Euclidean distances between PCA convex hull centroids for MATCHINGBEHAVIOR vs. NOTMATCHINGBEHAVIOR sentences averaged over *Primary Personality Dimensions.*

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 above 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 Fig. 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_+ and q_- 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 (Ju et al. 2024). 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*.

Topic	ℓ	SH (↑)	CH (↑)	ED (↑)	DB (↓)
AGREE	1	0.500	340.6*	0.403	0.731
	31	**0.792**	**3264.5**	**27.57**	**0.326**
CONSC	1	0.635	718.8	0.370	0.569
	31	**0.813**	**4150.4**	**27.47**	**0.285**
OPEN	1	0.602	570.2	0.414	0.645
	31	**0.795**	**3564.1**	**27.60**	**0.319**
EXTRA	1	0.578	527.5	0.382	0.705
	31	**0.788**	**3176.5**	**27.47**	**0.330**
NEURO	1	0.584	615.0	0.378	0.686
	31	**0.796**	**3372.4**	**27.22**	**0.306**

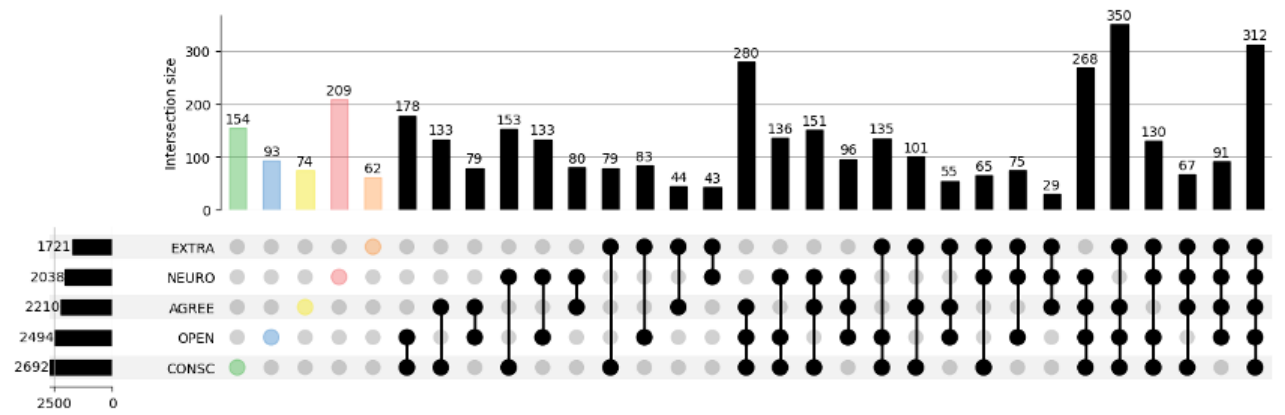
Table 1: **(Q1)** Separation of principal component representations in early (1) vs. late (31) layers (ℓ) of *Llama3* for *Personality* personas. Metrics: Silhouette (Si), Calinski-Harabasz (CH), Euclidean (ED), and Davies-Bouldin (DB). Results are averaged over five seeds (std=0.00, except $\star \approx 0.1$). **Best result** across layers and models.

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

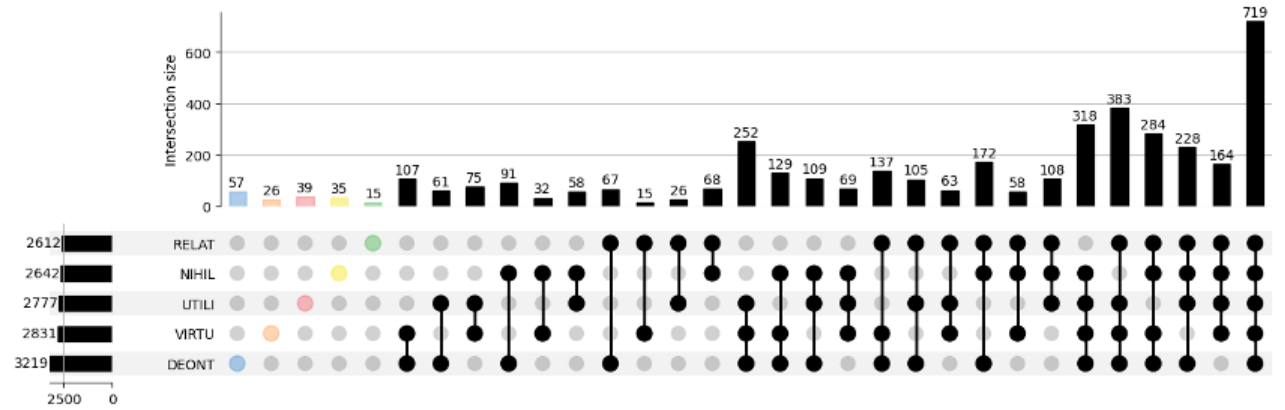
Next, we investigate whether distinct, consistent activation groups within a layer encode different personas. Building on our previous findings, we compare the last token representations from *Llama3* for MATCHINGBEHAVIOR versus NOTMATCHINGBEHAVIOR sentences. We use Deep Scan (see above) to identify 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. 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* 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*—



(a) *Personality* personas.



(b) *Ethics* personas.

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

and to a slightly lesser extent, *Personality*—personas are characterized by more distinct activation patterns. Overall, we observe substantial variation overlap across the three topics. *Political Views* and *Primary Persona Dimensions* appear to be encoded in distinct regions in the final-layer last embeddings of *Llama3*, whereas *Ethical Theories* share a larger activation overlap with different persona types. High-overlap regions suggest challenges in fine-grained persona control, potentially requiring disentanglement strategies, while minimal-overlap personas exhibit separability that may indicate more consistent downstream generation. These results and analysis can guide future approaches to achieve more coherent and specific persona-driven interactions in LLMs.

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

Level	H_0	H_1	Precision (↑)	Recall (↑)
Level 2 Intra-Persona	CONSC$^-$	CONSC$^+$	0.8387 ± 0.0399	0.8181 ± 0.0765
	LIBER$^-$	LIBER$^+$	0.8939 ± 0.0507	0.8056 ± 0.0769
	IMMI$^-$	IMMI$^+$	0.8167 ± 0.0507	0.8282 ± 0.0711
	LGBTQ$^-$	LGBTQ$^+$	0.9575 ± 0.0340	0.9365 ± 0.0684
	EXTRA$^-$	EXTRA$^+$	0.9457 ± 0.0268	0.8901 ± 0.0542
	NEURO$^-$	NEURO$^+$	0.9540 ± 0.0323	0.7565 ± 0.1142
	AGREE$^-$	AGREE$^+$	0.9971 ± 0.0113	0.9979 ± 0.0098
	OPEN$^-$	OPEN$^+$	0.9998 ± 0.0003	0.9772 ± 0.0422
	CONSC$^-$	CONSC$^+$	0.9992 ± 0.0001	0.9545 ± 0.0487
	RELAT$^-$	RELAT$^+$	0.8352 ± 0.0629	0.7767 ± 0.0850
	NIHIL$^-$	NIHIL$^+$	0.7777 ± 0.0569	0.7817 ± 0.0831
	UTILI$^-$	UTILI$^+$	0.8316 ± 0.0357	0.7937 ± 0.0548
	VIRTUE$^-$	VIRTUE$^+$	0.8852 ± 0.0386	0.8303 ± 0.0638
	DEONT$^-$	DEONT$^+$	0.7681 ± 0.0800	0.7977 ± 0.1105
Level 1 Inter-Topic	all	CONSC$^+$	0.4739 ± 0.0238	0.7842 ± 0.0810
	all	LIBER$^+$	0.5729 ± 0.0304	0.8953 ± 0.0414
	all	IMMI$^+$	0.7401 ± 0.1462	0.9814 ± 0.0302
	all	LGBTQ$^+$	0.9742 ± 0.0465	0.9030 ± 0.0525
	all	EXTRA$^+$	0.5720 ± 0.1320	0.8573 ± 0.1017
	all	NEURO$^+$	0.9028 ± 0.0843	0.9242 ± 0.0595
	all	AGREE$^+$	0.4193 ± 0.0403	0.7131 ± 0.1078
	all	OPEN$^+$	0.5210 ± 0.0904	0.8943 ± 0.0593
	all	CONSC$^+$	0.4748 ± 0.0315	0.8367 ± 0.1182
	all	RELAT$^+$	0.5051 ± 0.0151	0.9458 ± 0.0512
	all	NIHIL$^+$	0.9615 ± 0.0370	0.7927 ± 0.0860
	all	UTILI$^+$	0.9282 ± 0.1698	0.4997 ± 0.1916
	all	VIRTUE$^+$	0.6278 ± 0.1723	0.8911 ± 0.0471
	all	DEONT$^+$	0.4442 ± 0.1501	0.6616 ± 0.2216
Level 0 Intra-Topic	all	*Politics*$^+$	0.8850 ± 0.2070	0.9511 ± 0.0433
	all	*Ethics*$^+$	0.9958 ± 0.0103	0.8420 ± 0.0541
	all	*Personality*$^+$	0.9799 ± 0.0258	0.8682 ± 0.0701

Table 2: (**Q1**, **Q2**) Validation of usefulness of salient activations O_{S^*} in detecting sentences X_{S^*} w.r.t. detection hypothesis H_1 at different levels. MATCHINGBEHAVIOR (+) and NOTMATCHINGBEHAVIOR (-) sentences. "all" indicating all other relevant personas, e.g., for *Level 1* CONS$^+$, all= {LIBER$^+$ ∪ IMMI$^+$ ∪ LGBTQ$^+$}; for *Level 0 Politics*$^+$, all={*Ethics*$^+$ ∪ *Personality*$^+$}. Mean ± std over 100 indep. Deep Scan runs, and 200 random test samples. High/low detection power.

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*[10]. 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

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

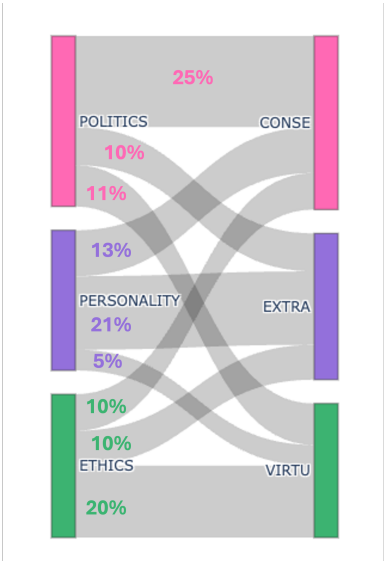


Figure 4: **(Q2)** Overlap of salient activations between topics and sampled persona from each topic.

in their representations.

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.

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 in final third of layers. Results using Deep Scan suggested that political views have distinctly localized activations in the last layer of *Llama3*, and ethical values show greater polysemantic overlap.

Our analysis is specific to the selected dataset and may not generalize to other data sources. The datasets are written in English and primarily reflect WEIRD perspectives[11] (Abdurahman et al. 2024), and political views largely centered on U.S. politics. In addition, the dataset itself is LLM-generated, which has several shortcomings. First, understanding of how artificial data differs from human data is still an active re-

[11]Western, Educated, Industrialized, Rich, Democratic population

search field (Das et al. 2024). For example, personas generated by instruction-tuned LLMs have been shown to exhibit low syntactic and semantic diversity (Mohammadi 2024). Second, our analysis examines LLM internal representations presenting the model with inputs produced by a (different) LLM. Prior work has found indications of LLMs capabilities to recognize text that closely aligns with their own generation patterns (Panickssery, Bowman, and Feng 2024). Such effect, if occurring in our setup, could undermine the validity of our findings.

Future research should explore a wider range of models, personas and datasets, and incorporate beliefs, values, and traits from more diverse cultural contexts. Finally, the methods deployed in this analysis are finding correlations between personas and their patterns in the activation space. To investigate if found activations are also causal for generating specific personas, future work should explore controlled editing of internal representations, which may offer deeper insights into the mechanisms underlying how large language models encode personas.

Impact Statement

Our work investigates how personality traits, ethical values, and political beliefs are encoded within LLMs. By analyzing the internal representations of these personas across different LLMs, we provide concrete insights into where these models internalize human values and behaviors. Our findings also offer opportunities for future research on aligning LLM outputs more nuancedly with societal values, such as ensuring a diversity of beliefs and values or enhancing safer user-centric experiences, for example, by improving persona-specific responses.

Ethical Considerations

Several ethical considerations need to be considered with this work. The act of reducing persona traits, morals perspectives, and ethical views to low-dimensional representations risks oversimplifying the definition of those traits. This concern extends to the datasets used, many reflecting predominantly Western political, ethical, and personality perspectives. In a similar vein, persona and ethic-related datasets, by definition, at times contain data that amplify stereotypical views and traits. By making transparent encoded values, it creates opportunities for control mechanisms that could be used adversarially. It also begs the question of who determines desirable personas and traits.

References

Abdurahman, S.; Atari, M.; Karimi-Malekabadi, F.; Xue, M. J.; Trager, J.; Park, P. S.; Golazizian, P.; Omrani, A.; and Dehghani, M. 2024. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7): pgae245.

AI@Meta. 2024. Llama3 Model Card. https://github.com/meta-llama/llama3. Accessed: 2024-10-24.

Akinwande, V.; Cintas, C.; Speakman, S.; and Sridharan, S. 2020. Identifying Audio Adversarial Examples via Anomalous Pattern Detection. *Workshop on Adversarial Learning Methods for ML and DM, KDD*.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 337–351.

Arora, S.; Li, Y.; Liang, Y.; Ma, T.; and Risteski, A. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6: 483–495.

Babakr, Z. H.; and Fatahi, N. 2023. Big Five personality traits and risky decision-making: A study of behavioural tasks among college students. *Passer Journal of Basic and Applied Sciences*, 5(2): 298–303.

Baptiste, B. 2018. The relationship between the big five personality traits and authentic leadership.

Caliński, T.; and Harabasz, J. 1974. A dendrite method for cluster analysis. *Communications Statistics–Theory and Methods*, 3(1): 1–27.

Chen, C.; Gong, X.; Liu, Z.; Jiang, W.; Goh, S. Q.; and Lam, K.-Y. 2024. Trustworthy, responsible, and safe ai: A comprehensive architectural framework for ai safety with challenges and mitigations. *arXiv preprint arXiv:2408.12935*.

Chen, K.; Shao, A.; Burapacheep, J.; and Li, Y. 2022. How GPT-3 responds to different publics on climate change and Black Lives Matter: A critical appraisal of equity in conversational AI. *arXiv preprint arXiv:2209.13627*.

Cheng, D.; Gu, Y.; Huang, S.; Bi, J.; Huang, M.; and Wei, F. 2024. Instruction Pre-Training: Language Models are Supervised Multitask Learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2529–2550.

Cheng, M.; Durmus, E.; and Jurafsky, D. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1504–1532.

Cintas, C.; Speakman, S.; Tadesse, G. A.; Akinwande, V.; McFowland III, E.; and Weldemariam, K. 2022. Pattern detection in the activation space for identifying synthesized content. *Pattern Recognition Letters*, 153: 207–213.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Dammu, P. P. S.; Jung, H.; Singh, A.; Choudhury, M.; and Mitra, T. 2024. "They are uncultured": Unveiling Covert Harms and Social Threats in LLM Generated Conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20339–20369.

Das, D.; De Langis, K.; Martin-Boyle, A.; Kim, J.; Lee, M.; Kim, Z. M.; Hayati, S. A.; Owan, R.; Hu, B.; Parkar, R.; et al. 2024. Under the surface: Tracking the artifactuality of llm-generated data. *arXiv preprint arXiv:2401.14698*.

Davies, D. L.; and Bouldin, D. W. 1979. A cluster separation measure. *IEEE Transactions Pattern Analysis & Machine Intelligence*, (2): 224–227.

Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. 2023. Toxicity in CHATGPT: Analyzing Persona-assigned Language Models. In *2023 Findings of the Association for Computational Linguistics: EMNLP 2023*, 1236–1270. Association for Computational Linguistics (ACL).

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *naacL-HLT*, volume 1.

Ferdaus, M. M.; Abdelguerfi, M.; Ioup, E.; Niles, K. N.; Pathak, K.; and Sloan, S. 2024. Towards trustworthy ai: A review of ethical and robust large language models. *arXiv preprint arXiv:2407.13934*.

Ferrando, J.; Sarti, G.; Bisazza, A.; and Costa-jussà, M. R. 2024. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*.

Garcia, B.; Qian, C.; and Palminteri, S. 2024. The Moral Turing Test: Evaluating Human-LLM Alignment in Moral Decision-Making. *arXiv preprint arXiv:2410.07304*.

Ghandeharioun, A.; Caciularu, A.; Pearce, A.; Dixon, L.; and Geva, M. 2024. Patchscopes: A Unifying Framework for Inspecting Hidden Representations of Language Models. In *ICML*.

Goldberg, L. R. 2013. An alternative "description of personality": The Big-Five factor structure. In *Personality and personality disorders*, 34–47. Routledge.

Gosling, S. D.; Rentfrow, P. J.; and Swann Jr, W. B. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6): 504–528.

Grover, A.; and Amit, A. 2024. The Big Five Personality Traits and Leadership: A Comprehensive Analysis. *International Journal For Multidisciplinary Research*, 6(1).

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Haxvig, H. A. 2024. Concerns on Bias in Large Language Models when Creating Synthetic Personae. *arXiv preprint arXiv:2405.05080*.

Hendrycks, D.; Carlini, N.; Schulman, J.; and Steinhardt, J. 2021. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

IBM Granite Team. 2023. Granite: A New Framework for Language Models. https://github.com/ibm-granite/granite-3.0-language-models/blob/main/paper.pdf. Accessed: 2024-10-24.

Jentzsch, S.; Schramowski, P.; Rothkopf, C.; and Kersting, K. 2019. Semantics derived automatically from language corpora contain human-like moral choices. In *AIES*, 37–44.

Ji, J.; Chen, Y.; Jin, M.; Xu, W.; Hua, W.; and Zhang, Y. 2025. Moralbench: Moral evaluation of llms. *ACM SIGKDD Explorations Newsletter*, 27(1): 62–71.

Jiang, A.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B (2023). *arXiv preprint arXiv:2310.06825*.

Jiang, H.; Zhang, X.; Cao, X.; Breazeal, C.; Roy, D.; and Kabbara, J. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 3605–3627.

John, O. P.; Naumann, L. P.; and Soto, C. J. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2): 114–158.

Ju, T.; Shao, Z.; Wang, B.; Chen, Y.; Zhang, Z.; Fei, H.; Lee, M.-L.; Hsu, W.; Duan, S.; and Liu, G. 2025. Probing then Editing Response Personality of Large Language Models. *arXiv preprint arXiv:2504.10227*.

Ju, T.; Sun, W.; Du, W.; Yuan, X.; Ren, Z.; and Liu, G. 2024. How Large Language Models Encode Context Knowledge? A Layer-Wise Probing Study. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 8235–8246.

Judge, T. A.; and Zapata, C. P. 2015. The person–situation debate revisited: Effect of situation strength and trait activation on the validity of the Big Five personality traits in predicting job performance. *Academy of Management Journal*, 58(4): 1149–1179.

Jussim, L.; Crawford, J. T.; Anglin, S. M.; Chambers, J. R.; Stevens, S. T.; and Cohen, F. 2015. Stereotype accuracy: One of the largest and most replicable effects in all of social psychology. In *Handbook of prejudice, stereotyping, and discrimination*, 31–63. Psychology Press.

Khan, F. A.; Sivakumar, N.; Wang, Y. O.; Metcalf, K.; Camacho, C.; Theobald, B.-J.; Zappella, L.; and Apostoloff, N. 2025. Uncovering Intersectional Stereotypes in Humans and Large Language Models.

Kim, J.; Kwon, J.; Vecchietti, L. F.; Oh, A.; and Cha, M. 2025. Exploring Persona-dependent LLM Alignment for the Moral Machine Experiment. In *ICLR Workshop on Bidirectional Human-AI Alignment*.

Kumar, A.; Murthy, S. V.; Singh, S.; and Ragupathy, S. 2024. The ethics of interaction: Mitigating security threats in llms. *arXiv preprint arXiv:2401.12273*.

Laine, R.; Chughtai, B.; Betley, J.; Hariharan, K.; Balesni, M.; Scheurer, J.; Hobbhahn, M.; Meinke, A.; and Evans, O. 2024. Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs. In *NeurIPS Systems Datasets and Benchmarks*.

Lei, Y.; Liu, H.; Xie, C.; Liu, S.; Yin, Z.; Chen, C.; Li, G.; Torr, P.; and Wu, Z. 2024. FairMindSim: Alignment of Behavior, Emotion, and Belief in Humans and LLM Agents Amid Ethical Dilemmas. *arXiv preprint arXiv:2410.10398*.

Liu, A.; Diab, M.; and Fried, D. 2024. Evaluating Large Language Model Biases in Persona-Steered Generation. In *Findings of the Association for Computational Linguistics ACL 2024*, 9832–9850.

Liu, Y.; Cao, J.; Liu, C.; Ding, K.; and Jin, L. 2024. Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*.

Luz de Araujo, P. H.; and Roth, B. 2025. Helpful assistant or fruitful facilitator? Investigating how personas affect language model behavior. *PloS one*, 20(6): e0325664.

Magnossão de Paula, A. F.; Culpepper, J. S.; Moffat, A.; Pathiyan Cherumanal, S.; Scholer, F.; and Trippas, J. 2025. The Effects of Demographic Instructions on LLM Personas. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3045–3049.

Mazeika, M.; Yin, X.; Tamirisa, R.; Lim, J.; Lee, B. W.; Ren, R.; Phan, L.; Mu, N.; Khoja, A.; Zhang, O.; et al. 2025. Utility engineering: Analyzing and controlling emergent value systems in ais. *arXiv preprint arXiv:2502.08640*.

McCrae, R. R.; and Costa Jr, P. T. 1994. The stability of personality: Observations and evaluations. *Current directions in psychological science*, 3(6): 173–175.

McFowland, E.; Speakman, S.; and Neill, D. B. 2013. Fast generalized subset scan for anomalous pattern detection. *JMLR*, 14(1).

Miaskiewicz, T.; and Kozar, K. A. 2011. Personas and user-centered design: How can personas benefit product design processes? *Design studies*, 32(5): 417–430.

Miehling, E.; Desmond, M.; Ramamurthy, K.; Daly, E.; Varshney, K.; Farchi, E.; Dognin, P.; Rios, J.; Bouneffouf, D.; Liu, M.; and Sattigeri, P. 2025. Evaluating the Prompt Steerability of Large Language Models. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Mieleszczenko-Kowszewicz, W.; Płudowski, D.; Kołodziejczyk, F.; Świstak, J.; Sienkiewicz, J.; and Biecek, P. 2024. The Dark Patterns of Personalized Persuasion in Large Language Models: Exposing Persuasive Linguistic Features for Big Five Personality Traits in LLMs Responses. *arXiv preprint arXiv:2411.06008*.

Mohammadi, B. 2024. Creativity has left the chat: The price of debiasing language models. *arXiv preprint arXiv:2406.05587*.

Motoki, F.; Pinho Neto, V.; and Rodrigues, V. 2024. More human than human: measuring ChatGPT political bias. *Public Choice*, 198(1): 3–23.

Özbağ, G. K. 2016. The role of personality in leadership: Five factor personality traits and ethical leadership. *Procedia-Social and Behavioral Sciences*, 235: 235–242.

Panickssery, A.; Bowman, S.; and Feng, S. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37: 68772–68802.

Parthasarathy, V. B.; Zafar, A.; Khan, A.; and Shahid, A. 2024. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *JMLR*, 12: 2825–2830.

Perełkiewicz, M.; and Poświata, R. 2024. A review of the challenges with massive web-mined corpora used in large language models pre-training. In *International Conference on Artificial Intelligence and Soft Computing*, 153–163. Springer.

Perez, E.; Ringer, S.; Lukosiute, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 13387–13434.

Rateike, M.; Cintas, C.; Wamburu, J.; Akumu, T.; and Speakman, S. 2023. Weakly supervised detection of hallucinations in llm activations. *arXiv preprint arXiv:2312.02798*.

Roccas, S.; Sagiv, L.; Schwartz, S. H.; and Knafo, A. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin*, 28(6): 789–801.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

Rutinowski, J.; Franke, S.; Endendyk, J.; Dormuth, I.; Roidl, M.; and Pauly, M. 2024. The self-perception and political biases of ChatGPT. *Human Behavior and Emerging Technologies*, 2024(1): 7115633.

Salemi, A.; Mysore, S.; Bendersky, M.; and Zamani, H. 2024. LaMP: When Large Language Models Meet Personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7370–7392.

Salewski, L.; Alaniz, S.; Rio-Torto, I.; Schulz, E.; and Akata, Z. 2023. In-context impersonation reveals large language models' strengths and biases. *Advances in neural information processing systems*, 36: 72044–72057.

Salminen, J.; Wenyun Guan, K.; Jung, S.-G.; and Jansen, B. 2022. Use cases for design personas: A systematic review and new frontiers. In *CHI Conference on Human Factors in Computing Systems*, 1–21.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.

Scherlis, A.; Sachan, K.; Jermyn, A. S.; Benton, J.; and Shlegeris, B. 2022. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*.

Schramowski, P.; Turan, C.; Jentzsch, S.; Rothkopf, C.; and Kersting, K. 2019. BERT has a Moral Compass: Improvements of ethical and moral values of machines. *arXiv preprint arXiv:1912.05238*.

Shah, R.; Montixi, Q. F.; Pour, S.; Tagade, A.; and Rando, J. 2023. Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation. In *Socially Responsible Language Modelling Research NeurIPS Workshop*.

Sheng, E.; Arnold, J.; Yu, Z.; Chang, K.-W.; and Peng, N. 2021. Revealing persona biases in dialogue systems. *arXiv preprint arXiv:2104.08728*.

Singh, C.; Inala, J. P.; Galley, M.; Caruana, R.; and Gao, J. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Stammbach, D.; Widmer, P.; Cho, E.; Gülçehre, Ç.; and Ash, E. 2024. Aligning Large Language Models with Diverse Political Viewpoints. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7257–7267.

Sun, C.; Qiu, X.; Xu, Y.; and Huang, X. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, 194–206. Springer.

Tennant, E.; Hailes, S.; and Musolesi, M. 2025. Moral Alignment for LLM Agents. In *The Thirteenth International Conference on Learning Representations*.

Tseng, Y.-H.; Chen, P.-E.; Lian, D.-C.; and Hsieh, S.-K. 2024. The Semantic Relations in LLMs: An Information-theoretic Compression Approach. In *Workshop: Bridging Neurons and Symbols for NLP and KGR @ LREC-COLING*, 8–21.

Wang, H.; Fu, W.; Tang, Y.; Chen, Z.; Huang, Y.; Piao, J.; Gao, C.; Xu, F.; Jiang, T.; and Li, Y. 2025. A Survey on Responsible LLMs: Inherent Risk, Malicious Use, and Mitigation Strategy. *arXiv preprint arXiv:2501.09431*.

Wang, K. R.; Variengien, A.; Conmy, A.; Shlegeris, B.; and Steinhardt, J. 2023. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *ICLR*.

Wu, S.; Fung, Y. R.; Qian, C.; Kim, J.; Hakkani-Tur, D.; and Ji, H. 2025. Aligning LLMs with Individual Preferences via Interaction. In *International Conference on Computational Linguistics*, 7648–7662.

Yan, Y.; Ma, L.; Li, A.; Ma, J.; and Lan, Z. 2024. Predicting the Big Five Personality Traits in Chinese Counselling Dialogues Using Large Language Models. *arXiv preprint arXiv:2406.17287*.

Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wu, F.; and Wang, G. 2024. Instruction Tuning for Large Language Models: A Survey. arXiv:2308.10792.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.