### 3.1 Socio-technical Motivation

We systematically study of off-the-shelf instruction-tuned models across three persona categories—moral philosophies (e.g., utilitarianism), personality traits, and political ideologies (e.g., liberalism)— to determine if such constructs emerge as patterns in hidden-state activations, and (if so) precisely where they are located within the model's representational space.

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 [54, 55]. McCrae and Costa [56] 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 [57, 58], social interaction [59], and even vulnerability to manipulation [27, 60, 61]. 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 output [62], the capacity of LLMs to infer Big Five personality traits from dialogues [63], the behavior of persona-instructed LLMs on personality tests, and creative writing tasks [38].

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 [64, 65], showing, for instance, that political personas can significantly influence model behavior in ethical dilemmas [65]. Other studies have developed benchmarks and systematic evaluations to assess the moral identity and decision-making patterns of LLMs [66, 67]. Others found fine-tuning to be a promising strategy for aligning LLM agents more closely with human values [68].

LLMs exposed to vast amounts of political text may internalize subtle political biases [9, 23]. 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 minority viewpoints [69]. 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 [70].

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 [39, 71, 72]. Here, as a first step toward exploring the intersections of personas across ethical, political, and personality dimensions, we examine the overlap in their embeddings.

### 3.2 Persona Datasets and Assumptions

Our experiments are based on the model-generated personas [3], 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 [3], 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 [3].

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

- *Primary Personality Dimensions*, which relies on the Big Five [34, 54], 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).

[2] For a critical discussion on synthetic persona generation refer to Haxvig [73].