Localizing Persona Representations in LLMs

Celia Cintas[*]
IBM Research Africa
Nairobi, Kenya
celia.cintas@ibm.com

Miriam Rateike[*]
IBM Research Africa
Nairobi, Kenya
miriam.rateike@ibm.com

Erik Miehling
IBM Research Europe
Dublin, Ireland
erik.miehling@ibm.com

Elizabeth Daly
IBM Research Europe
Dublin, Ireland
elizabeth.daly@ie.ibm.com

Skyler Speakman
IBM Research Africa
Nairobi, Kenya
skyler@ke.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.*

1 Introduction

Understanding the mechanisms by which large language models (LLMs) process information, store knowledge, and generate outputs remain key open questions in research [1, 2]. 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* [3]. Clearly understanding the mechanics of personas is important for a variety of reasons. Personas are often used to define the personality or perspective the LLM model should adopt when interacting with users [4], 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 [5]. Personas have also been leveraged in tasks such as synthetic data generation [6] and decision-making solutions such as LLM-as-a-judge [7]. Personas have been debated in multiple social, political, and behavioral studies as a potential replacement for human participants [8, 9, 10]. While there are interesting applications, the use of LLMs in studies raises key concerns that require careful attention [11]. Personas can enhance user experience and engagement by making models more relatable and context-aware [12, 13, 14], however, they have also been used to bypass safety measures or to trigger unintended consequences rooted in underlying biases [15], raising significant ethical and security concerns [16].

[*] Equal contribution.