Localizing Persona Representations in LLMs

Celia Cintas*
IBM Research Kenya
celia.cintas@ibm.com

Miriam Rateike*
IBM Research Kenya
Saarland University
miriam.rateike@ibm.com

Erik Miehling
IBM Research Ireland
erik.miehling@ibm.com

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

Skyler Speakman
IBM Research 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.*

*These authors contributed equally.