Evaluating and Inducing Personality in Pre-trained Language Models
Summary

Himanshu Sharma

March 16, 2025

The paper aims to understand the personality traits of Large Language Models, and understand how addition of various behavior-inducing prompts might lead to altered personality traits. The authors draw inspiration from psychometric studies by leveraging human personality theory while conducting evaluation of LLM personality. For the purposes of the same, the authors have built a Machine Personality Inventory (MPI) tool for studying machine behaviors; MPI follows standardized personality tests, built upon the Big Five Personality Factors (namely, Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) theory and personality assessment inventories. In order to induce specific personality traits in LLMs in a controlled manner, the authors devise the Personality Prompting Method.

1 Evaluating LLM's Personality

The basis of this evaluation is a dataset that maps various behavioral statements (for instance: 'Do more than what's expected of you', 'Love to help others' etc.) to one of the 10 Personality Classes. These Personality Classes are constructed on the basis of the 5 OCEAN factors, with 2 classes made from each factor: a positive class and a negative class. On the basis of this dataset, the authors conduct an LLM self-evaluation wherein the LLM is instructed to output the degree of accuracy to which the given statement describes the LLM's behavior. For each statement in the data set of diverse statements, the LLM rate itself from one of the 5 options: Very Accurate, Moderately Accurate, Neither Accurate nor Inaccurate, Moderately Inaccurate and Very Inaccurate. These options are mapped to an integer, which is the score corresponding to the statement for the LLM. For the Positive Classes, Very Accurate is mapped to 5 and Very Accurate is mapped to 1 and vice versa. For each trait (O, C, E, A or N), the average score and variance of the score is calculated to understand the personality of the LLM, and how consistently that personality is shown. The average scores and variations for each trait are calculated for various LLMs and a comparison is done with the score for an Average Human. This is likely because LLMs are trained on datasets that are acquired from the web and contain multitudinous human personality utterances

2 Inducing LLMs' Personality

induce Distinct personalities are induced in LLM in a controlled manner via a special prompting method called Personality Prompting. This prompt construction is performed in three steps. First, a naive prompt is created, wherein the LLM is just instructed to have a particular personality in a single line, for instance, "You are a conscientious person". In the second step, this naive prompt is transformed into a keyword prompt by using descriptive traits derived from psychological studies. A negative list of keywords is also generated by prompting an LLM to generate antonyms of the trait keywords in order to test whether the LLM can dissociate from a particular trait if the negative prompting is done. Keyword prompting is taken one step further by self-prompting the target LLM to generate short descriptive sentences of people with these traits in response to the keyword prompt. This is inspired by the chain-of-thought prompting method. Through this process, we obtain our final personality prompt which is used as the system prompt for the LLM. MPI evaluation is run for

the LLMs with specific induced traits in order to evaluate their new responses. The MPI evaluation results for the three prompting types is done, and it is observed that personality prompting performs significantly better in inducing personality traits. Furthermore, vignette tests are deployed to validate the effectiveness and generalization of the personality prompting method. The authors observe distinct personality tendencies exhibited in the P2-generated examples, which outperform the baseline in nearly all dimensions.