Evaluating language model alignment with free associations

Dirk Wulff
Max Planck Institute for Human Development, Berlin, Germany

Zak Hussain
University of Basel, Basel, Switzerland

Samuel Aeschbach
Max Planck Institute for Human Development, Berlin, Germany

Rui Mata
University of Basel, Basel, Switzerland

Abstract

The alignment between large language models and humans' knowledge and preferences is central to such tools' safe and fair deployment. A number of approaches to quantifying alignment exist, but current work is fragmented, preventing an overview across categories of stimuli and demographic groups. We propose that free associations from massive citizen-science projects can advance representational alignment by helping evaluate both content and demographic inclusivity. We assess the representational alignment of GPT-4 Turbo and data from the English Small World of Words Study (ca. 80.000 respondents, 3.7 million responses). Our results indicate that while the language model can capture some procedural signatures of human responses, it shows heterogeneous alignment across stimuli categories, poor representational alignment for controversial topics (e.g., religion, nationality), and differential representation of demographic groups (e.g., males, females). All in all, our work suggests that free association can be used to evaluate the representational alignment of large language models.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).