

## **UC Merced**

# **Proceedings of the Annual Meeting of the Cognitive Science Society**

### **Title**

Evaluating language model alignment with free associations

### **Permalink**

<https://escholarship.org/uc/item/98p056xc>

### **Journal**

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

### **Authors**

Wulff, Dirk U

Hussain, Zak

Aeschbach, Samuel

et al.

### **Publication Date**

2024

Peer reviewed

# 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.