

## **UC Merced**

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

### **Title**

Humanizing Language Models: Exploring behavioral and brain data as language model inputs

### **Permalink**

<https://escholarship.org/uc/item/71v3w2f0>

### **Journal**

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

### **Authors**

Hussain, Zak

Mata, Rui

Newell, Ben

et al.

### **Publication Date**

2024

Peer reviewed

# Humanizing Language Models: Exploring behavioral and brain data as language model inputs

**Zak Hussain**

University of Basel, Basel, Switzerland

**Rui Mata**

University of Basel, Basel, Switzerland

**Ben Newell**

University of New South Wales, Sydney, Australia

**Dirk Wulff**

Max Planck Institute for Human Development, Berlin, Germany

## Abstract

Language models have traditionally been trained on massive digitized text corpora. However, alternative data sources exist that may increase the representational alignment between language models and human knowledge. We contribute to the assessment of the role of data sources on language model representations. Specifically, we present work aimed at understanding differences in the content of language representations ('embeddings') trained from text, behavioral data (e.g., free associations), and brain data (e.g., fMRI). Using a method from neuroscience known as 'representational similarity analysis', we show that embeddings derived from behavioral and brain data encode different information than their text-derived cousins. Furthermore, using an interpretability method that we term 'representational content analysis', we find that, in particular, behavioral embeddings better encode dimensions relating to dominance, valence, and arousal, which are likely critical for the representational alignment of language models.