Human feedback makes Large Language Models more human-like

Pablo Contreras Kallens
Saarland University, Saarbrücken, Germany

Ross Kristensen-McLachlan
Aarhus University, Aarhus, Denmark

Morten Christiansen
Cornell University, Ithaca, New York, United States

Abstract

The most recent generation of Large Language Models owes its success not only to scale, but also a novel step in their training: reinforcement learning from human feedback (RLHF). In this study, we assessed the impact that this training regime has on the fit between model and human behavior in regards to linguistic behavior. We evaluated three versions of OpenAI's GPT-3 davinci – original, instruction-tuned, and RLHF-trained – using psycholinguistic tasks: subject-verb agreement, sentence acceptability, and event knowledge. We then compared their performance to human participants. We found that the RLHF model is significantly more human-like in its answers, including in the errors it commits. Moreover, the uncertainty of the distribution of its output is closely tied with between-subject variation in humans. This suggests that human feedback improves not only the overall quality of LLMs, but also the alignment between their behavior and the linguistic, metalinguistic, and discursive intuitions of humans.

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.*