Wrong for the Right Reason? Using Successes and Failures of Large Language Models to Understand Human Thinking

Zach Studdiford
University of Wisconsin-Madison, Madison, Wisconsin, United States

Gary Lupyan
University of Wisconsin - Madison, Madison, Wisconsin, United States

Abstract

If a person answers a question correctly, how can we tell if the answer reflects an underlying understanding of the phenomenon, or if it is based on merely surface-level associations? Cognitive science has developed multiple tests, such as Winograd Schemas, that ostensibly require a respondent to use some kind of world/situation model rather than just associations. What then are we to make of large language models (LLMs) successes on some of these tasks? We present a series of probes to LLMs and people about everyday situations, finding that models sometimes respond correctly for the wrong reason and in other cases make seemingly 'catastrophic' mistakes by applying the wrong model–often in human-like ways. Our results suggest that probing the basis of LLMs' successes and failures can help inform human problem solving and in some cases call into question our previous tests of human understanding.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*