Representation in Large Language Models

Cameron Yetman
University of Toronto, Toronto, Ontario, Canada

Abstract

Cognitive scientists attribute representations to complex systems in order to explain their behavior. The shocking facility with which Large Language Models (LLMs) perform difficult linguistic and non-linguistic tasks has generated an increasing amount of speculation concerning what sorts of internal representations might underlie this behavior (whether personal, sub-personal, and of which kinds) and what properties such representations might have (for instance, whether they are grounded). This paper aims to elaborate and defend a conservative explanatory methodology, based on analyses of particular LLM behaviors, according to which attribution of sub-personal representations is key to explaining model performance which is robust, systematic, and flexible, especially in zero-shot settings, and that behavioral benchmarking alone is insufficient to resolve questions about representation due to the mutual underdetermination of performance and competence. The resulting view should help frame future explanations of LLM behavior, and provide an empirically grounded alternative to mere a priori speculation.

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.*