The emergence of flexible perspective reasoning in large language models

Pablo Leon Villagra
Brown University, Providence, Rhode Island, United States

Tiana Simovic
University of Toronto, Toronto, Ontario, Canada

Craig Chambers
University of Toronto, Toronto, Ontario, Canada

Abstract

Work on human reference processing has shown that, in sentences like "Mary asked her daughter Sally if she understood the assignment", readers overwhelmingly interpret "she" as co-referring with "Sally". This reflects perspective inference, or reasoning about who possesses at-issue information, and is inconsistent with a statistically-learned bias toward subject antecedent selections. The flexibility of inferencing is evident from the effect of manipulating the object character description ("Mary asked her tutor..."), where readers now prefer Mary as the antecedent. Until recently, these patterns have been largely unaccounted for by large language models (LLMs). Leveraging advancements in LLM interpretability techniques, the present study systematically examines how LLMs fare in relation to human judgments. We determine which layer activations impact these inferences and perturb them to causally link activations to model performance. Finally, we examine performance across training iterations, analyzing the point where subjecthood biases become evident and when more nuanced inferencing emerges.

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