LLMs have ”mental” models: Latent world models in LLM network weights can be inferred from output layer tokens at inference

Cole Robertson
Emory University, Atlanta, Georgia, United States

Phillip Wolff
Emory University, Atlanta, Georgia, United States

Abstract

Do large language models (LLMs) construct and manipulate internal ”mental models” of physical systems, or do they rely solely on statistical associations represented as output layer token probabilities learned from data? We adapt cognitive science methodologies from human mental models research, testing LLMs on pulley system problems using TiKZ-rendered stimuli. Study 1 examines whether LLMs can estimate mechanical advantage (MA) while distinguishing relevant from irrelevant system components, and disregarding distractor elements. We found that contemporaneous state-of-the-art models performed marginally but significantly above chance when exact estimate-label matches were required, and that their estimates correlate significantly with ground-truth MA. Crucially, tested models selectively attended to meaningful variables (e.g., number of ropes and pulleys) while ignoring system features that are irrelevant to MA (rope diameter, pulley diameter, ceiling height). Study 2 extends this by investigating the extent to which LLMs may internally represent gestalt system features, which are crucial to estimating MA: LLMs evaluated a functionally connected pulley system against a ”fake” system comprising unconnected components. Without explicit cues that one system was non-functional, models correctly identified the connected system as having greater MA with an average accuracy of 84%. However, their explanations failed to acknowledge the fundamental distinction between connected and unconnected systems, instead relying on post hoc rationalizations over false premises (e.g., assuming both systems were connected and inferring MA from supporting ropes). This suggests that while LLMs manipulate internal ”world models” analogous to human mental models, these may be conceptually uncoupled from explicit reasoning at the output layer. These findings provide evidence that LLMs may construct latent world models that inform token probabilities, challenging the notion that they are ”only” next-token predictors.

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