Playing with Knowledge: Evaluating Common Sense in LLMs Through Language Games

Ine Gevers, University of Antwerp

Abstract

Large Language Models (LLMs) today achieve strikingly high scores on different benchmarks, designed to test math, language understanding, or coding skills. Yet, these same models can exhibit surprising failures of common sense, like suggesting to use glue on your pizza if the cheese slips off. These mismatches highlight an open question: what do LLMs really understand about the world? Evaluating common sense knowledge in AI systems has been a longstanding challenge in NLP, spanning a wide collection of topics such as implicit language understanding, social or cultural norms, or everyday reasoning strategies.

In this talk, we will explore how language games can reveal different facets of an LLM's knowledge and reasoning than standardized tests do. Often, succeeding in these games requires a mix of implicit world knowledge, estimating opponents' strategic intentions, and of course, understanding the rules of the game. We introduce the game 'Concept', a simple word-guessing game, as such a challenge. Can LLMs guess the intended concept given clues provided by humans? Our results suggest that LLMs struggle with interpreting other players' intentions, and often fail to correct their initial hypotheses.

Bio

Ine Gevers is a PhD researcher in Natural Language Processing at the CLiPS research group of the University of Antwerp, supervised by Prof. Walter Daelemans. She earned her Master's degree in Digital Text Analysis, where she focused on the automatic detection of online hate speech. Afterwards, she collaborated closely with political scientists to develop computational methods to detect claims of representations made by Flemish politicians. For her current PhD topic, Ine researches how to evaluate common sense knowledge in Large Language Models, both by re-examining existing benchmarks, as well as creating new evaluation tasks.