Title: Investigating Large Language Models' Representations Of Plurality Through Probing Interventions

Author: Michael Hanna

Institute: Institute of Formal and Applied Linguistics

Supervisor: RNDr. David Mareček, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Large language models (LLMs) have become ubiquitous in natural language processing, but how exactly they process their input and arrive at good downstream task performance is still poorly understood. While much work has been done using probing to examine LLM internals, or behavioral studies, to determine LLMs' linguistic capabilities, these techniques are too weak to allow us to draw conclusions how LLMs process language. In this paper, I use both probing and causal intervention methods to investigate the question of subject-verb agreement with respect to the subject's plurality. I find that while probing reveals that subject plurality information is distributed throughout a sentence, causal interventions suggest that only information stored in linguistically relevant tokens is used. Probing interventions suggest that some but not all probes capture information in a way that reflects LLMs' usage thereof.

Keywords: Interpretability, Probing, Natural Language Processing, Computational Linguistics