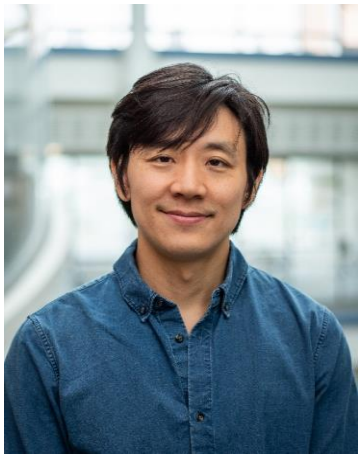
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE



Andrew Lee

Interpretability, Controllability, and Expressivity in the Age of Language Models

Friday, June 14, 2024
3:30pm – 5:30pm
3725 Beyster
Hybrid – Zoom Passcode: UM-CSE

ABSTRACT: Language models have received a lot of recent attention, given their expressivity. However, their expressivity comes at the cost of interpretability and controllability.

Given their large number of parameters, it is infeasible for humans to understand their computational process. This leads to a lack of interpretability, which can lead to undesirable outcomes in downstream applications if deployed irresponsibly.
Meanwhile, language models often generate compelling outputs, but can also generate undesirable outputs, such as biased, toxic, or nonfactual statements, which are inadvertently learned during pre-training. Thus one may be interested in steering a model away from such undesirable behavior. However, this poses challenges – similar to interpretability, complex computations make models more difficult to control.

With that said, my work aims to build systems using language models that are interpretable and/or controllable.

An important component for interpreting language models is to understand the representations that they learn. In the first half of my thesis, I demonstrate that language models are full of human-interpretable representations of concepts. A deep enough understanding of a model's inner representations not only allows us to 1) control what it generates, but also 2) explain what occurs during fine-tuning. As it turns out, these inner representations often have predictive power of the model's outputs. Put differently, they have a causal role for the model's predictions, and thus manipulating such representations can lead to predictable changes in model outputs. Lastly, once we understand the role of these representations, we can observe their changes after being fine-tuned, and explain how such changes lead to a difference in model behavior.

In the second half of my thesis, I demonstrate methods for building interpretable or controllable systems by grounding language models to new domains. By doing so, one can integrate domain expertise into their computations. I demonstrate my methods in the domains of healthcare and games. One example of my methods is a novel framework – micromodels – which is an interpretable and modular design that leverages language models to build domain-level representations for downstream applications to use.

CHAIR: Prof. Rada Mihalcea