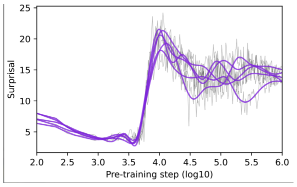
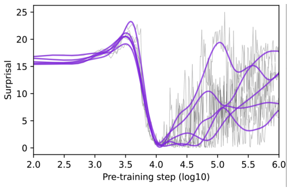
UC San Diego
Cognitive Science

Tyler A. Chang's Dissertation Defense

N-Gram Learning and Pretraining Dynamics in Transformer Language Models



Monday, May 12th, 2025
11:00am – 1:00pm, Public Engagement Building 721
or
https://ucsd.zoom.us/j/97808053907



Abstract:

Transformer language models have received unprecedented attention in recent years due to impressive performance on natural language tasks, but how they acquire these capabilities is largely unknown. During pretraining, language models are trained to predict next tokens (e.g. words) in text, and the field of *pretraining dynamics* aims to demonstrate how this learning objective leads to consistent learning patterns and mechanisms. The primary contributions of this dissertation are to demonstrate relationships between *n*-gram probabilities (next token predictions based only on the previous *n*-1 tokens) and language model learning during pretraining. I first show that language models learn words differently from human children; word frequencies (unigram probabilities) account for the vast majority of variance in words' ages of acquisition in language models, but not in children. I then characterize language model learning as early *n*-gram learning for increasing *n*, then gradual refinement of low probability *n*-gram predictions based on longer contexts and more nuanced capabilities. Mechanistically, subnetworks that make bigram predictions exist in the models long after the full models have moved on from simple bigram predictions. These bigram subnetworks are critical to language modeling performance even in fully trained models, and they recreate several key properties of a language model's activation space across layers. Together, these results demonstrate a close relationship between *n*-gram probabilities and language model learning, suggesting that traditional distributional statistics still play some role in these more complex models.

Committee members:

Dr. Benjamin Bergen (Co-Chair) - UCSD Cognitive Science
Dr. Zhuowen Tu (Co-Chair) - UCSD Cognitive Science
Dr. Leon Bergen - UCSD Linguistics
Dr. Ndapandula Nakashole - UCSD Computer Science and Engineering