Contribution ID: **6** Type: **not specified**

Transformers as token-to-token function learners

Friday 26 September 2025 14:45 (45 minutes)

The ambitious question of understanding a Transformer can be decomposed into understanding the functions it implements: the class of functions one can theoretically approximate using a Transformer, which subclass of them is learnable via gradient descent, which training data distribution is implicitly biased towards which set of functions, how they are implemented across the neural components of the model, and so on. In this talk, I will focus on Transformers implementing language functions. A primer on mechanistic interpretability will be given, followed by certain open problems in this area. Then, I will present an alternate view of the Transformer functions that can potentially solve many of the existing limitations: existence of multiple parallel computation paths, lack of robustness of autoencoder-based replacement models, and how to formalize causal models embedded in training.

Presenter: DUTTA, Subhabrata