ICTS Statistical Physics and Condensed Matter Seminar

Title : A model of errors in transformers

Speaker : Suvrat Raju (ICTS-TIFR, Bengaluru)

Date : Tuesday, 13 January 2026

Time : 2:00 PM (IST)

Abstract : We study the error rate of LLMs on tasks like arithmetic that require a deterministic output, and repetitive processing of tokens drawn from a small set of alternatives. We argue that incorrect predictions arise when small errors in the attention mechanism accumulate to cross a threshold, and use this insight to derive a quantitative two-parameter relationship between the accuracy and the length of the task. The two parameters vary with the prompt and the model; they can be interpreted in terms of an elementary noise rate, and the mean number of erroneous alternatives during next-token prediction. Our analysis is inspired by an ``effective field theory'' perspective: the LLM's fundamental parameters can be organized into a small number of effective parameters for the determinination of the error rate. We perform extensive empirical tests, using Gemini 2.5 Flash, Gemini 2.5 Pro and Deepseek R1, and find excellent agreement between the predicted and observed accuracy for a variety of tasks, although we also identify deviations in some cases. Our model provides an alternative to suggestions that errors made by LLMs on long repetitive tasks indicate the ``collapse of reasoning'', or an inability to express ``compositional'' functions. Finally, we show how to construct prompts to reduce errors.

Venue : Madhava Lecture Hall
Zoom Link: https://icts-res-in.zoom.us/j/92495432478?pwd=RrxbAD0m0qQnfQJJ2qQSD0aFJQDwbU.1
Meeting ID: 924 9543 2478
Passcode: 202030