Title of the Talk

Rethinking Efficiency: Economical, Adaptable and Interpretable Small Language Models

by: Prof. Tanmoy Chakraborty

Abstract: Despite the superprior performance demonstrated by Transformer-based LLMs across numerous applications involving natural languages, their high computational cost, energy consumption, and limited accessibility underscore the need for efficient, interpretable, and adaptable small language models (SLMs). This talk highlights methods to develop economical and interpretable SLMs that rival their larger counterparts in performance without significant computational requirements. Our research emphasizes three key dimensions: economical resource usage, adaptability to diverse and low-resource tasks, and enhanced interpretability. Techniques like competitive knowledge distillation, leveraging student-teacher dynamics, and activation sparsity in manifold-preserving transformers demonstrate significant efficiency gains without compromising performance. We formulate novel decomposer components for LLMs for modularizing problem decomposition and solution generation, allowing smaller models to excel in complex reasoning tasks. We also propose innovative prompt construction and alignment strategies that boosted in-context knowledge adaptation in low resource settings for SLMs. Our findings demonstrate that SLMs can achieve scalability, interpretability, and adaptability, paving the way for broader and sustainable AI accessibility.