Beyond Scaling Laws: From "Large Is Beautiful" to Seeing a World in a Grain of Sand

Abstract:

As multimodal large language models (LLMs) continue to advance, their insatiable appetite for training data and compute has become increasingly unsustainable. This has sparked a wave of innovation focused on improving data and parameter efficiency. Some approaches turn to synthetic data to address data scarcity, while others, such as Microsoft's Phi-4-multimodal, demonstrate how compact small language models (SLMs), using distillation, context compression, and external tools, can achieve strong performance with far fewer resources. Yet these strategies, while promising, may introduce risks of overfitting, brittleness, and model collapse.

To navigate these challenges, we must look beyond brute-force scaling and reimagine AI not as monolithic predictors, but as structured, context-aware, and grounded agents. Insights from disciplines like mathematics, physics, and systems theory—drawing on concepts such as entropy, dynamical systems, and statistical reasoning—provide foundational principles for designing robust, interpretable, and adaptable agentic AI.

Meanwhile, AI is evolving into a collaborative partner in both scientific discovery and real-world productivity. LLMs are increasingly used to generate mathematical conjectures, design algorithms, streamline workflows, support creative processes, and optimize complex systems across domains such as business, healthcare, and education. This growing synergy signals a shift from viewing LLMs as passive tools to recognizing their role as active participants in generating knowledge, insight, and value.

In this talk, we explore a future beyond scaling laws, where intelligence emerges not from scale alone, but from structure, interaction, and contextual grounding. As generative AI becomes interwoven into every aspect of our lives, we may glimpse "heaven in a wild flower," and truly see "a world in a grain of sand."