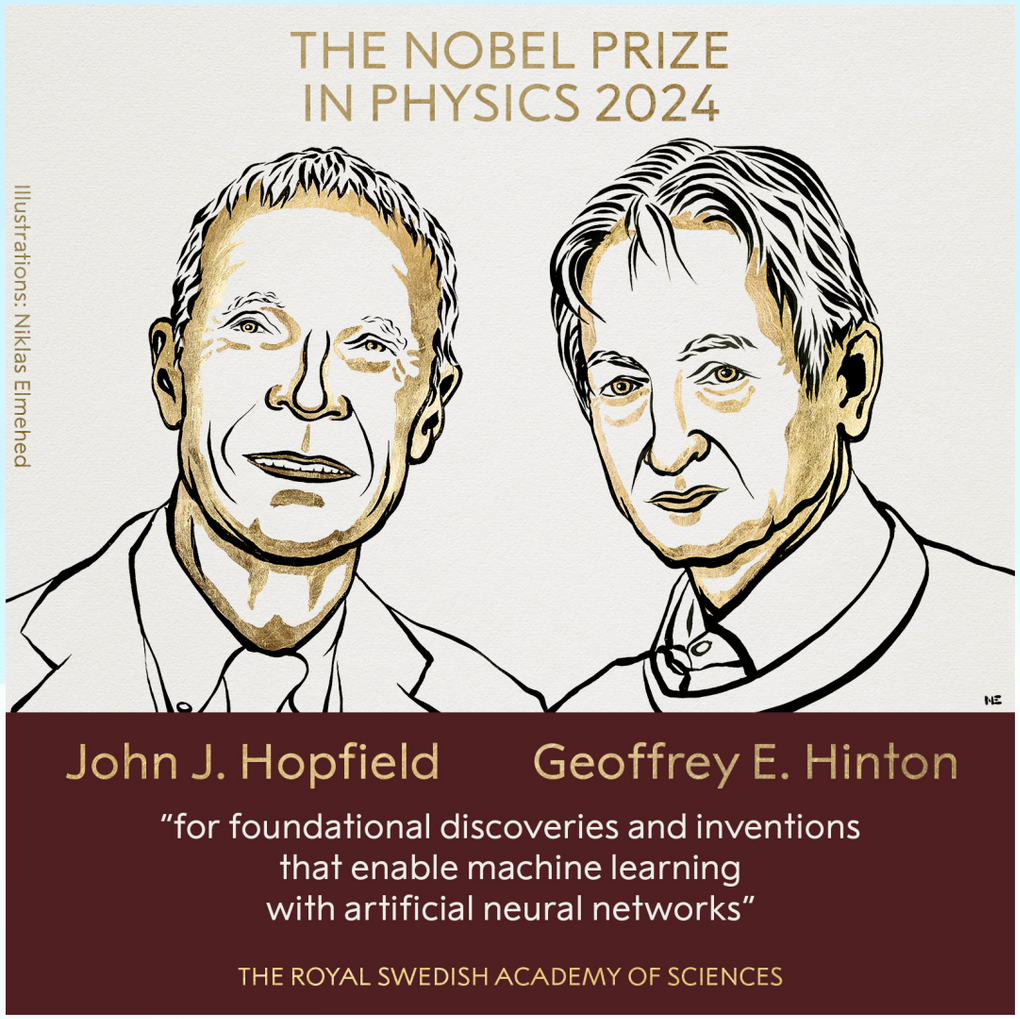
AI4All Institute

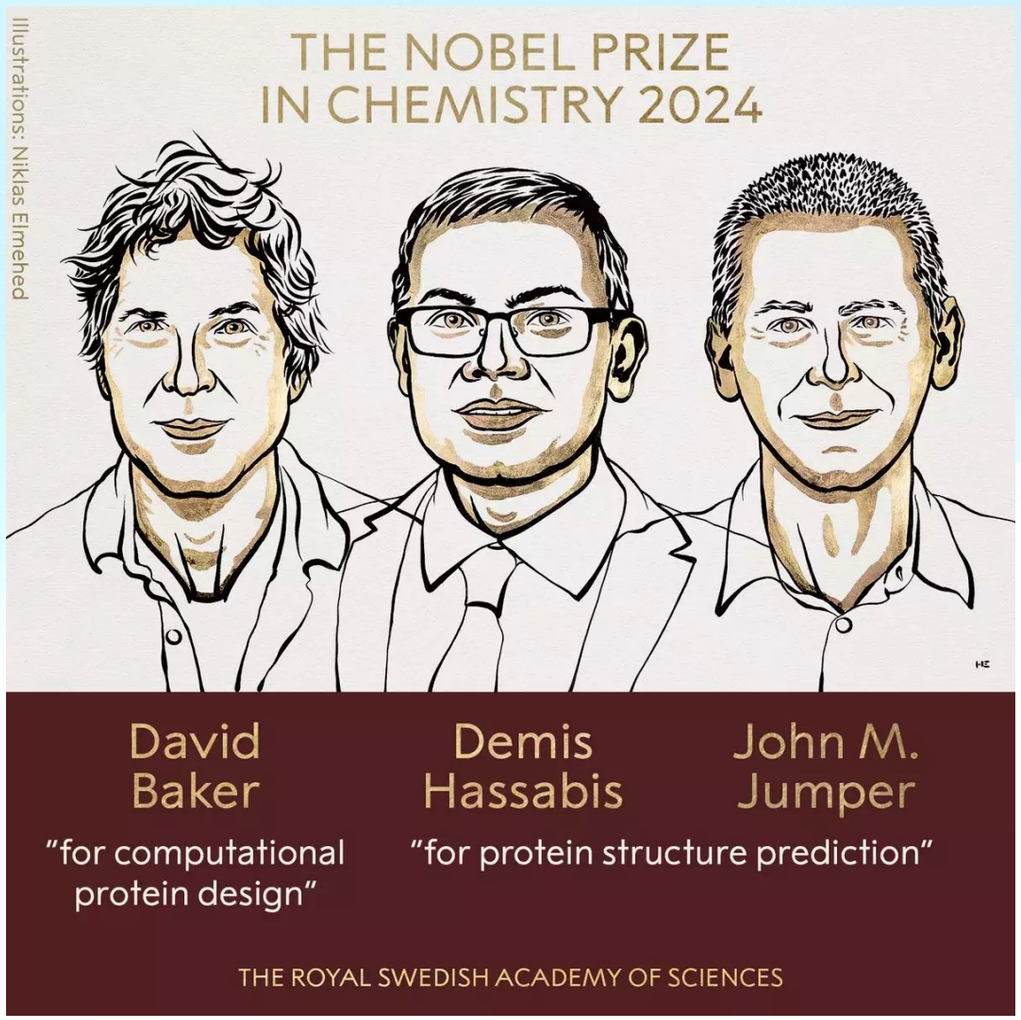
An AI Community of all, by all, for all

AI4AllInstitute.github.io

2024.12.7

AI will be more influential

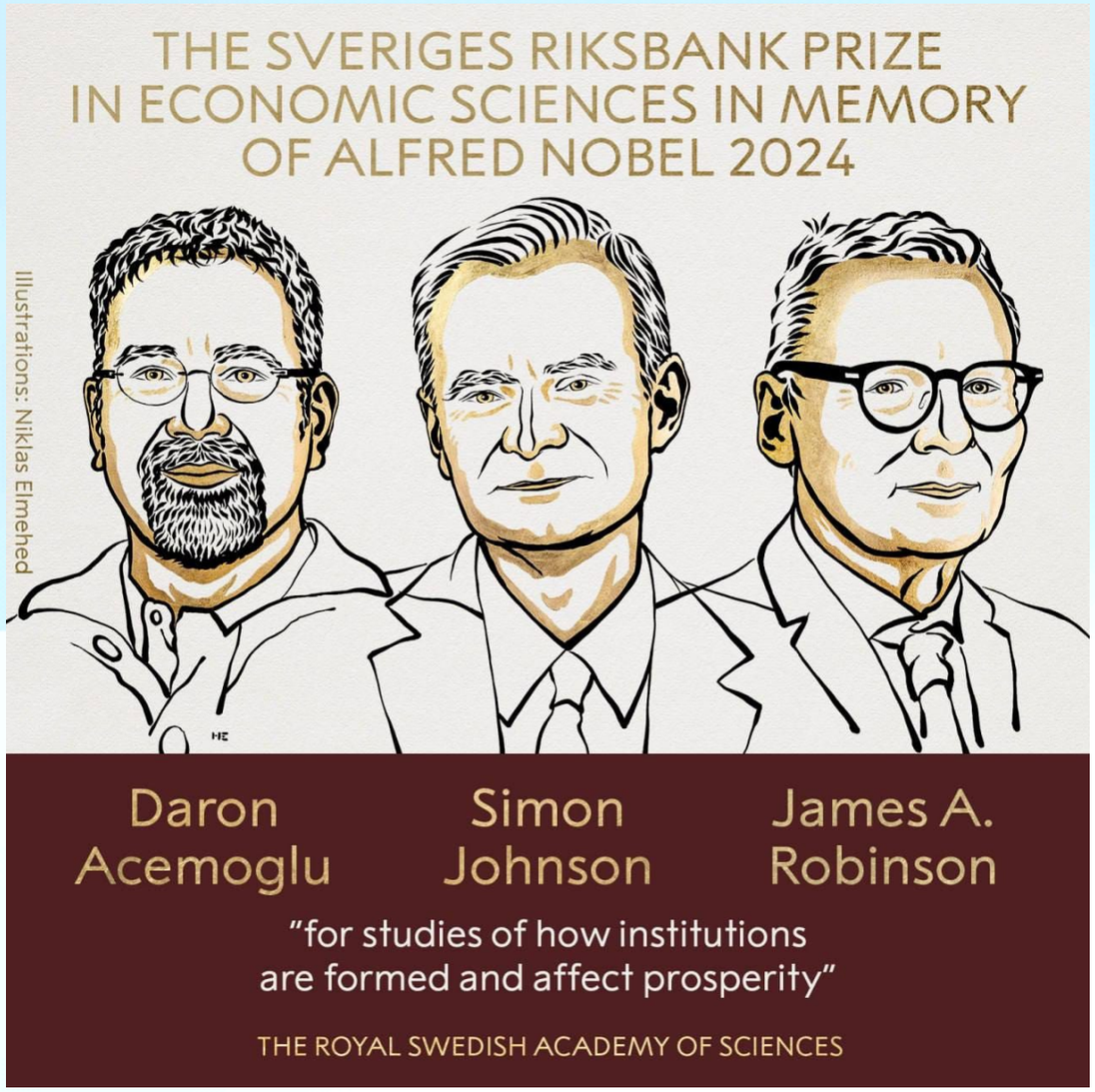




AI's prowess is demonstrated in more and more fields and recognized by more and more people. The 2024 Nobel prizes in physics and chemistry would make such a trend wider, deeper and faster.

AI4All Institute

An AI community of all, by all, for all



Let’s make it conclusive, with respect to who, what and how, for its best prosperity, inspired by the 2024 Nobel prize in economic science.

AI Seminar by AI4All Institute

Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters

Charlie Snell
UC Berkeley

Abstract

Enabling LLMs to improve their outputs by using more test-time computation is a critical step towards building generally self-improving agents that can operate on open-ended natural language. In this paper, we study the scaling of inference-time computation in LLMs, with a focus on answering the question: if an LLM is allowed to use a fixed but non-trivial amount of inference-time compute, how much can it improve its performance on a challenging prompt? Answering this question has implications not only on the achievable performance of LLMs, but also on the future of LLM pretraining and how one should tradeoff inference-time and pre-training compute. Despite its importance, little research attempted to understand the scaling behaviors of various test-time inference methods. Moreover, current work largely provides negative results for a number of these strategies. In this work, we analyze two primary mechanisms to scale test-time computation: (1) searching against dense, process-based verifier reward models; and (2) updating the model's distribution over a response adaptively, given the prompt at test time. We find that in both cases, the effectiveness of different approaches to scaling test-time compute critically varies depending on the difficulty of the prompt. This observation motivates applying a "compute-optimal" scaling strategy, which acts to most effectively allocate test-time compute adaptively per prompt. Using this compute-optimal strategy, we can improve the efficiency of test-time compute scaling by more than 4x compared to a best-of-N baseline. Additionally, in a FLOPs-matched evaluation, we find that on problems where a smaller base model attains somewhat non-trivial success rates, test-time compute can be used to outperform a 14x larger model.

Dec 7, 2024

8:00-9:00pm EST

ID: 568 500 8086
Passcode: 666

ai4allinstitute.github.io

AI Seminar by AI4All Institute

Language Agents: Realizing Generalist Agents with the Power of Natural Language

Yu Gu
OSU

Abstract

Generalist agents have been a long-standing goal in AI research, dating back to its inception. Early efforts in this field, whether rule-based or reinforcement learning (RL)-based, have predominantly focused on specialist agents designed to achieve specific goals. I argue that the key distinction between specialist and generalist agents lies in their ability to handle natural language. Natural language empowers agents to: 1) process diverse goals specified in human language, 2) use language as a unified medium for internal reasoning and decision-making, and 3) ultimately, acquire new knowledge in a way that aligns more naturally with human learning processes. In this talk, I will introduce the research on language-powered agents through two representative testbeds, tracing progress from early efforts built on language models like BERT or T5 to new opportunities unlocked by more recent LLMs. In addition, I will reflect on the power of natural language in intelligence in general and outline a vision for future research on language agents.

Dec 7, 2024

9:30-10:30pm EST

ID: 568 500 8086
Passcode: 666

ai4allinstitute.github.io

A platform for collaboration and communication.

Welcome to test it.