AI roadmap's inherent unpredictability

Ryoji Kashiwagi

22 April 2024

Iakyara vol.384

Nomura Research Institute, Ltd.

Ryoji Kashiwagi
Expert Researcher
Financial Market & Digital Business Research Department

Executive Summary

Since the advent of deep learning, AI has been advancing faster than even experts have predicted. Today, AI faces multiple threats to the sustainability of its recent rate of progress. However, initiatives to neutralize these threats are underway and progressing apace. How rapidly AI functionality will advance going forward is anybody's guess.

Generative AI's progress has surpassed even experts' predictions

OpenAI unveiled its first GPT large language model in 2018. In 2022, a mere four years later, it released ChatGPT, an AI model capable of generating content in a variety of formats, including text, source code, images and speech as directed by users.

Subsequent advancements in generative AI functionality have far surpassed experts' predictions. In May 2023, University of Tokyo professor Yutaka Matsuo predicted that AI video generation would not become reality for at least several and perhaps up to 10 years. Just nine months later, OpenAI released a real-life AI video generator named Sora. Many of Professor Matsuo's fellow experts have likewise persistently underestimated how fast AI would progress. In fact, not one expert has accurately forecasted the pace of AI advancements.

One AI milestone still on the horizon is high-level machine intelligence (HLMI), a state where unaided machines can perform any task better and more cheaply than humans. A survey of 2,778 AI researchers published in January 2024 reported that the researchers are collectively forecasting a 10% probability of HLMI by 2027 and 50% probability of HLMI by 2047[1]. The latter is 13 years earlier than implied by the previous year's responses to the same survey question. The consensus AI roadmap is in extreme flux even among experts.

Threats to AI progress

Meanwhile, the following concerns have been raised about optimistic extrapolations of the recent rapid pace of AI progress.

NOTE
1) "Thousands of AI Authors on the Future of AI" https://arxiv.org/abs/2401.02843

■Energy intensity

The IEA recently forecasted that electricity consumed by digital infrastructure, including data centers, AI and cryptocurrencies, will keep growing long into the future even as overall growth in global electricity consumption slows[2]. It expects power consumption related to AI in particular to grow tenfold between 2023 and 2026.

On a micro level, the IEA estimates that a ChatGPT request requires about 10 times more electric power to process than a Google search query. If generative AI usage grows sharply, the AI industry would face pressure to improve its energy efficiency. In fact, the Biden Administration has already proposed a 30% tax on cryptocurrency miners' electricity consumption in the US[3].

2) "Electricity 2024" - IEA https://iea.blob.core.windows.net/assets/ddd078a8-422b-44a9-a668-52355f24133b/Electricity2024-Analysisandforecastto2026.pdf

3) Cointelegraph, *Biden budget proposes 30% tax on crypto mining electricity usage,* https://cointelegraph.com/news/biden-budget-proposes-30-tax-on-crypto-mining-electricity-usage

■Computing resource availability constraints

Training a generative AI model requires a mind-bogglingly enormous number of computations. The CPUs in today's computers were designed to efficiently execute a wide variety of complex processes. They are not well suited to high-speed mass execution of repetitive, simple operations. GPUs can execute such operations much faster than CPUs. AI trainers consequently favor GPUs over CPUs. Use of GPUs has dramatically shortened AI training times. NVIDIA, a global GPU supplier, has seen its earnings burgeon in recent years and its market capitalization surpass $2tn since February 2024. NVIDIA GPUs optimized for AI training have essentially become the de facto standard for training generative AI models.

On the flipside of NVIDIA's near-monopolistic dominance of the GPU market, some analysts warn that its GPU supply capacity could become a bottleneck that constrains AI's progress.

■Training data scarcity

Training of generative AI models requires huge datasets of various types such as text, image, audio and video. Most such training data are sourced from the Internet. Epoch, an AI research institute, warned in 2022 that the AI industry may soon run out of high-quality training data[4]. It says the existing stock of high-quality textual training data (e.g., Wikipedia entries, news articles, scientific papers) may

4) Will We Run Out of ML Data? Evidence From Projecting Dataset Size Trends - Epoch https://epochai.org/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset

be entirely exhausted by as early as 2026.

The amount of data required to train an AI model is roughly synonymous with the number of parameters the model has. Of OpenAI's large language models released to date, GPT-1 has 110mn parameters, GPT-2 has 1.5bn, GPT-3 has 175bn and GPT-4 has 1tn. In other words, the amount of data used to train OpenAI's latest GPT has grown by a factor of roughly 10,000 in just four years.

Additionally, major news sites have started prohibiting their content from being used in AI models. If availability of high-quality training data is curtailed, AI progress may be adversely affected.

Unpredictable roadmap

We have highlighted a few potential caveats to the rosy narrative that AI functionality will perpetually continue advancing at the same pace or even faster than it has to date. However, researchers are currently working on solving these issues.

To address the energy consumption challenge, one potential solution that has been proposed to is to retool inefficient CPU-based data centers with GPU servers, thereby reducing power consumption and substantially upgrading computational efficiency[5]. Other initiatives are afoot to improve the energy efficiency of GPUs themselves and utilize clean energy to power data centers.

In terms of the computing resource issue, Microsoft has generated some buzz with a new AI training model named BitNet, which uses addition operations in lieu of the matrix multiplication previously required to train generative AI models[6]. If BitNet lives up to its apparent promise, it could drastically reduce the AI training process's computational intensity. If so, computational resource constraints may be avoidable.

Lastly, to mitigate the risk of running out of training data, research is being conducted on usability of low-quality data, defined as data of non-assured quality, such as social media and blog content. Such data is orders of magnitude more plentiful than high-quality data. If low-quality data can be used to train AI models, the AI industry may be able to avoid running out of training data. Another approach now being researched is to have an AI model generate training data to

5) NVIDIA, *Going Green: New Generation of NVIDIA-Powered Systems Shows Way Forward,* https://blogs.nvidia.com/blog/green/

6) Gigazine, *Microsoft releases a 1.58-bit large-scale language model, allowing matrix calculations to be added, dramatically reducing calculation costs* https://gigazine.net/gsc_news/en/20240229-microsoft-1bit-llm/

be used to further train the model. If successful, this approach also may help solve the problem of training data scarcity.

In sum, AI has evolved to date through a series of paradigm shifts. Given the inherent unpredictability of paradigm shifts, please pardon us if we beg off from making AI predictions.

about NRI

Founded in 1965, Nomura Research Institute (NRI) is a leading global provider of system solutions and consulting services with annual sales above $5.1 billion. NRI offers clients holistic support of all aspects of operations from back- to front-office, with NRI's research expertise and innovative solutions as well as understanding of operational challenges faced by financial services firms. The clients include broker-dealers, asset managers, banks and insurance providers. NRI has its offices globally including New York, London, Tokyo, Hong Kong and Singapore, and over 16,500 employees.

For more information, visit https://www.nri.com/en

Inquiries to : Financial Market & Digital Business Research Department
Nomura Research Institute, Ltd.
Otemachi Financial City Grand Cube,
1-9-2 Otemachi, Chiyoda-ku, Tokyo 100-0004, Japan
E-mail : kyara@nri.co.jp

https://www.nri.com/en/knowledge/publication/fis/lakyara/