Marubeni
Marubeni Washington Report

September 4, 2025

Is Scaling the Key to an AI Future?

The "Scaling Hypothesis", Alternative Pathways, and the "Narrative" We Choose to Believe

Marubeni America Corporation Washington Office
So Uehara, Senior Government and International Affairs Manager
uehara-so@marubeni.com

- In recent years, scaling has been critical for AI development, but uncertainty looms. LLMs continue to generate commercial value and drive adoption, yet mounting evidence of diminishing returns has raised real doubts about whether scaling alone can deliver AGI.
- Industry leaders are hedging their bets. Even companies that profit most from the scaling boom, like NVIDIA, are reinvesting in alternative architectures, signaling that insiders see the need for breakthroughs beyond brute-force compute.
- The stakes extend beyond science. Trillions in capital, infrastructure, and national strategy are being wagered on a contested hypothesis, making the narratives we choose to believe as a key determinant for AI's trajectory.

Artificial intelligence is moving faster than ever. According to Stanford HAI's newly released 2025 AI Index Report[1], the past year marked another leap forward. AI systems in 2024 didn't just improve—they accelerated, posting remarkable gains on difficult benchmarks and showing dramatic progress in code-solving tasks.

Right now, much of the spotlight falls on generative AI. Large language models (LLMs) like ChatGPT, Claude, and Gemini have become the most visible face of this transformation, driving adoption across consumers, businesses, and even governments. What once felt experimental is now becoming standard. On the commercial side, AI is no longer just about pilots and prototypes. It's embedding itself into everyday tools and services: Microsoft's Copilot in productivity suites, Netflix's recommendation systems, Zendesk's AI-powered customer service, Adobe's Firefly for creative work, and GitHub Copilot for coding, to name a few. Companies are also leaning into monetization, from subscription models to enterprise-wide deployments.

The Stanford report makes one point especially clear: AI systems are closing the gap with humans on many benchmarked tasks—and in some cases, they're surpassing us. But the picture isn't complete. Humans still have an edge in areas like open-ended reasoning, long-term planning, and, of course, physical interaction with the world. The real question now isn't whether it will change how we work, create, and make decisions; it's how quickly, and in what ways, those changes will reshape our lives.

So where exactly are we going with AI and what's the path to get there? Most roads seem to point toward artificial general intelligence (AGI). But AGI doesn't have one universally accepted definition. Some describe it narrowly, as systems that outperform humans at most economically valuable cognitive tasks. Others take a broader view, imagining human-level adaptability, reasoning, and causal understanding that works across any domain. Which definition you choose matters, because it changes how you think we'll get there.

This is where the scaling hypothesis comes in. Advocates of this hypothesis don't just argue that scaling can reach a narrow version of AGI. They believe scaling existing LLMs with more data and compute could eventually deliver richer forms of intelligence too: reasoning, transfer of knowledge,

[1] Stanford's HAI AI Index Report is an annual benchmark study that tracks global progress in artificial intelligence across research, industry, policy, and societal impact. (link)

and even forms of social and physical adaptability. In other words, they see today's trajectory as sufficient to unlock human-level qualities themselves, not just a cheaper version of human labor.

Critics argue that LLMs are fundamentally limited: they excel at pattern prediction but lack true reasoning, grounding in the physical world, and causal understanding. From this perspective, AGI will require new architectures entirely, not just ever-larger LLMs.

For now, the scaling optimists have the momentum. Capital is flowing into GPU farms, specialized chips, and massive power infrastructure. Investors find the thesis simple and intuitive: whoever has the largest compute wins. That in turn creates a reinforcement loop: money flows toward scaling, companies amplify the scaling story to raise more, and the narrative fuels itself.

Still, the outcome is far from certain. Even if scaling keeps producing breakthroughs (and it already has, powering new business models and scientific advances) the open question remains: does the curve keep climbing all the way to AGI, or does it eventually flatten out? And if it does flatten, what happens to the investors, governments, and societies that placed their bets on scaling as the one true path?

What is "AI" Today?

When most people think of AI today, they picture apps like ChatGPT or Claude. These are large language models (LLMs), built on neural networks. At their core, LLMs are trained on a simple task: predict the next word (or token) given the ones that came before. Through self-supervised learning on massive amounts of text, they've become remarkably good at capturing the statistical patterns of language. Scaled up to billions of parameters and trained on diverse datasets, LLMs generate responses so fluent and context-aware that they appear to "understand" and even "reason."

But LLMs are only one piece of the puzzle. Other techniques (also built on neural networks) power breakthroughs in different domains. Convolutional Neural Networks (CNNs) drive facial recognition, medical imaging, and the vision systems of autonomous cars. Recurrent Neural Networks (RNNs) have long been used for tasks like weather prediction and speech recognition. Reinforcement Learning (RL) underpins milestones like AlphaGo's victory over human champions and advanced robotics. The applications differ, but the foundation is the same: neural networks as powerful engines for making astonishingly accurate predictions.

Underneath it all, modern AI is built on three essential ingredients: data, compute, and algorithms.

- Data: Vast amounts of high-quality text, images, audio, video, and synthetic data are needed to train AI systems.
- Compute: Specialized hardware (mainly GPUs and TPUs)[2] provides the raw processing power to train large neural networks. Scaling compute through ever-larger clusters has been central to AI's progress since the transformer architecture was introduced in 2017.
- Algorithms: The architecture of the model or the actual design of how data and parameters[3]

[2] GPU (Graphics Processing Unit): Originally designed for rendering graphics, GPUs are now widely used in AI because they can process many calculations in parallel, making them highly effective for training large neural networks. TPU (Tensor Processing Unit): A type of application-specific integrated circuit (ASIC) developed by Google, optimized specifically for accelerating machine learning tasks, particularly deep learning using the TensorFlow framework.

[3] Parameters represent the size of the model within a chosen architecture. More parameters mean more "knobs to tune," but how effectively they're used depends on the algorithm and the available compute.

interact. Transformers, CNNs, RNNs, and reinforcement learning all represent different algorithmic choices.

The Scaling Hypothesis and the "LeCun Ceiling"

Today, frontier labs such as OpenAI, Anthropic, Google, DeepSeek, and others are competing to build the most powerful LLMs. Powerful in the sense that they show competence across a range of cognitive skills, in some cases approaching human levels. To achieve this, their dominant strategy follows the "scaling hypothesis"[4]; the assumption that increasing data, model parameters, and compute (via vast data centers powered by GPUs/TPUs) will continue to unlock new capabilities. This is why billions of dollars are being poured into expanding data centers and acquiring cutting-edge processors.

Sam Altman, CEO of OpenAI and Dario Amodei, co-founder and CEO of Anthropic are among the strongest proponents of the scaling hypothesis, expressing confidence that increasing model size, data, and compute will push systems toward artificial general intelligence (AGI), an AI system with general-purpose intelligence comparable to humans. They argue that each leap in scale has consistently unlocked new, often unexpected capabilities, suggesting that further scaling will eventually produce models with broadly human-level performance.

In contrast, Yann LeCun[5] of Meta contends that this approach will hit a ceiling. He argues that while larger models may get better at pattern recognition, they will not achieve genuine reasoning or understanding without fundamentally new architectures designed to build world models and learn from real-world input.

Scaling's Promise and Its Risks

In 2024, OpenAI rolled out o1 and o3 mini, models marketed as having advanced reasoning capabilities. On some benchmarks, they even surpassed average human performance. But peel back the layers and what looks like "reasoning" is still statistical pattern recognition at massive scale. LeCun argues that without grounded world models (a baseline understanding of how the world actually works) these systems remain brittle. Faced with problems outside their training distribution, like rare edge cases or counterfactual scenarios, they often stumble, producing the familiar "hallucinations." LLMs, in other words, is not ready for mission critical decision making.

By late 2024, cracks in the "scaling hypothesis" began to draw sharper scrutiny. In November, reports of diminishing returns on ever-larger models unsettled both researchers and investors, with some warning that scaling was no longer producing breakthroughs at the pace once expected. Even Google's Sundar Pichai acknowledged the issue: "When you start out quickly scaling up, you can throw more compute and you can make a lot of progress, but you definitely are going to need deeper breakthroughs as we go to the next stage," he said. "So you can perceive it as there's a wall, or there's some small barriers."

The unease grew in January 2025, when Chinese upstart DeepSeek released its R1 model, delivering GPT-class performance at a fraction of the cost and compute. The news sent shockwaves through

---

[4] In the media, it is sometimes referred to as the "scaling law".

[5] Yann LeCun is Meta's Chief AI Scientist and a Turing Award–winning computer scientist, best known as a pioneer of deep learning and convolutional neural networks. He is one of the most prominent skeptics of the scaling hypothesis, arguing that new architectures will be needed to achieve true intelligence.

markets, erasing hundreds of billions from Nvidia's market capitalization in a single day[6], as investors suddenly questioned whether future demand for GPUs would be as insatiable as previously assumed. In the days that followed, however, Nvidia CEO Jensen Huang and Wall Street analysts sought to calm fears, stressing that frontier models would still require enormous training and inference compute. Their reassurances helped stabilize the stock and reinforced the argument that DeepSeek's efficiency could actually expand AI adoption rather than curtail demand for hardware.

Today, scaling optimists continue argue that with more data, more compute, and bigger models, AGI could arrive as soon as 2030. Skeptics counter that we may be years or decades away. What's undeniable is that the current AI boom, and the sky-high valuations of OpenAI, Anthropic, Nvidia, and others, rests on one assumption: that scaling will keep delivering breakthrough capabilities. Billions are flowing into GPU purchases, hyperscale data centers, and mega-projects like Stargate under the belief that scaling alone will carry us to AGI.

But what if scaling does not deliver on its promise? Already researchers and companies are exploring alternative approaches, hedging against the possibility that the scaling curve will eventually flatten. As OpenAI's co-founder Ilya Sutskever said, "the 2010s were the age of scaling, now we're back in the age of wonder and discovery once again"[7]. An exploratory path brings higher R&D risk, longer timelines, and potentially weaker returns on the colossal GPU and data-center buildouts now underway. LLMs are expected to remain commercially useful (they already power profitable applications). But the "AGI imminence premium" baked into valuations, particularly because many leading AI companies are "priced for perfection"[8], could quickly unravel. Capital may start shifting from raw compute plays toward startups experimenting with alternative architectures.

And here's the bigger issue: the pioneers of AI, including Turing Award winners, leading linguists, and cognitive scientists can't agree on whether scaling will ever get us to AGI. In most high-investment technology bets (semiconductors, biotech, aerospace) the underlying science is relatively settled; the gamble is on commercialization. AI is different. Here, trillions are being wagered on a hypothesis still under active debate at the very top of the discipline.

That uncertainty isn't just a concern for investors. It has major implications for policymakers and strategists in Washington, who are tying national competitiveness and geopolitical strategy to the trajectory of this technology.

Alternative Pathways

If the LeCun Ceiling turns out to be real, it won't just be a scientific detour. It would force a fundamental redirection of R&D. And in fact, some of those alternative pathways to AGI are already being explored.

At the research frontier, Meta is developing its Joint Embedding Predictive Architecture (JEPA)[9], which

---

[6] Nvidia's stock plunged by 17%, marking the largest one-day loss in U.S. corporate history, with almost $600 billion in market cap wiped out.

[7] November 15, 2024 Reuters article (link).

[8] "Priced for perfection" describes a situation where investors assume flawless execution and maximum growth, leaving little room for error. In such cases, even small disappointments or shifts in expectations can trigger outsized market corrections.

[9] Meta recently unveiled its new V-JEPA2 AI model, a "world model" designed to help AI agents understand the world around them. (link)

aims to build an internal "map" of how the world works. Instead of simply memorizing patterns, JEPA is designed to help AI form a kind of mental model, allowing it to transfer lessons from one context to another, much as people apply knowledge learned in school to work or daily life. Google DeepMind has taken related steps through AlphaZero and MuZero[10]. AlphaZero reached superhuman skill in chess, shogi, and Go through self-play alone, while MuZero went further by mastering games without being told the rules, instead constructing its own internal model of the environment. Another DeepMind project, Dreamer[11], learns by building a simplified model of the world and then running "imagined" scenarios inside it. That ability to daydream about possible futures makes it more efficient and closer to how humans plan and learn.

The industry's largest players are not blind to these possibilities. Even as they profit from the scaling boom, they are channeling record earnings into research that hedges against its potential limits. Most strikingly, NVIDIA, the company that has gained the most from the scaling hypothesis, is itself pursuing alternatives. Beyond its flagship GPUs, it is investing in new chip designs such as data-processing units (DPUs) and neuromorphic research, while expanding software ecosystems like CUDA and Omniverse to support simulation and more efficient architectures. The fact that NVIDIA is preparing for "post-scaling" futures underscores a critical point: the leaders of the industry are not putting all their eggs in the scaling basket but actively positioning for a paradigm shift in how intelligence might ultimately be achieved.

What remains uncertain are the resource implications if these alternatives take hold. If scaling alone reaches its limits, the insatiable demand for ever-larger LLMs (and the GPUs and TPUs that fuel them) could give way to new kinds of hardware, including chips designed to mimic the brain more directly. Neuromorphic processors like Intel's Loihi[12] are one example, modeling aspects of brain architecture in hopes of achieving greater efficiency. This work is still at an early stage, but it suggests a future where the emphasis shifts from quantity (bigger models, more GPUs) to quality (smarter architectures and more efficient compute).

Such a shift would ripple outward. Electricity and infrastructure requirements could look very different. New architectures might prove more efficient and ease the need for mega-scale power planning. Or, if they turn out to be computationally inefficient, they could push electricity demand even higher. For utilities and governments, that uncertainty makes long-term planning more difficult.

Financial flows would also be reshaped. Today, investors assume the winning bet is on scaling: more GPUs, more data centers, and frontier labs racing to train the largest models. But if scaling falters, capital may pivot toward startups and labs pursuing new architectures. Governments, too, may redirect subsidies away from energy-hungry scaling toward more fundamental research in algorithms and brain-inspired hardware.

Artificial intelligence today sits at a crossroads. In the near term, LLMs continue to generate value and justify massive investment in compute. But their very limits are catalyzing research into new

---

[10] A system like MuZero is a powerful reinforcement learning (RL) algorithm that demonstrates progress towards general-purpose AI. While not a true Artificial General Intelligence (AGI), its ability to learn a model and plan in a wide range of scenarios shows a pathway to more adaptable and flexible AI systems. (link)

[11] Google DeepMind programmed Dreamer to play the popular video game Minecraft. It is a RL agent that learns a world model from images. (link)

[12] Neuromorphic chips like Intel's Loihi aim to mimic how the brain processes information, firing only when needed. This design holds the potential for far greater efficiency and lower energy use than today's GPUs, though the technology is still in early stages and not yet ready to replace large-scale AI training hardware. (link)

approaches that may shape the longer-term trajectory. These paths are not necessarily mutually exclusive: LLMs could remain an important subcomponent of a broader AGI technology stack, or they could be eclipsed entirely. The truth is, no one knows which way the field will break.

The Narrative We Choose To Believe

What stands out is not just the science, but the story. The market today appears confident that scaling will carry us forward, even though the pioneers of the field remain deeply divided. That should give any investor pause. And yet, the narrative is self-propagating. The narrators (OpenAI, Nvidia, and even Washington) are themselves stakeholders with enormous bets on scaling. By championing this vision, they channel ever more capital, both political and financial, into GPUs, power infrastructure, and data centers, creating an ecosystem whose financial interests become increasingly intertwined.

History offers a cautionary reminder. The dot-com bubble was also built on a compelling story: one that fueled vast infrastructure buildouts, dazzling valuations, and a sense of inevitability, until it wasn't. If scaling ultimately hits a ceiling, today's AI boom could face a similar reckoning.

And yet, even if scaling does not deliver superintelligence or AGI, the short-term trajectory of LLMs still holds enormous potential. These systems are already reshaping business models, creating new tools, and enabling forms of productivity and creativity that were unthinkable only a few years ago. Whether the value they create will ultimately justify the massive costs of compute, energy, and infrastructure remains to be seen—but the gains to date are undeniable.

In the end, the story of AI is not just about science or engineering; it is about the narratives we choose to believe, the value we are willing to pay for, and the massive bets societies make on those stories.

Marubeni America Corporation Washington Office

1717 Pennsylvania Ave., N.W., Suite 375, Washington, DC 20006
https://www.marubeni.com/jp/research/