SPECIAL COMPETITIVE STUDIES PROJECT

AGI Will Arrive in Three Ways

Version 2. November 2024.

P.J. Maykish
Ylli Bajraktari
Abigail Kukura
Nyah Stewart

As artificial intelligence (AI) gets stronger, we are forced to grapple with the mysteries of the human mind once more. The adult brain has approximately 86 billion neurons.[1] There are trillions of connections between neurons in your mind alone, creating a network far greater than that of the internet,[2] which involves only tens of billions of connected devices. This comparison underscores the brain's unparalleled complexity and processing power, which exceed even our most advanced technology creations. With this complexity comes a lack of consensus over how our minds actually work. Why do we dream, for example? Some think dreams help us process emotions or memories, while others believe they're a way for our brain to practice for real-life situations. The truth is, we're still deciphering the dream code. And what about memory? How can a smell or a song suddenly transport us back to a specific moment from childhood? While we know a lot about the brain structures involved in memory, the precise mechanisms of how memories are encoded, stored, and retrieved remain elusive. Like the journey of

[1] Suzana Herculano-Houzel, The Remarkable, Yet Not Extraordinary, Human Brain as a Scaled-Up Primate Brain and its Associated Cost, Proceedings of the National Academy of Sciences (2012).
[2] Satyajit Sinha, Number of Connected IoT Devices Growing 13% to 18.8 billion Globally, IoT Analytics (2024).

understanding our minds, it is essential to the human experience to step back and imagine how a more powerful form of AI will arrive in leaps, including the next 12-18 months.

By examining three pivotal megatrends, we can imagine the rapid development and deployment of a markedly more powerful form of AI that will lead to artificial general intelligence or AGI.[3] We can also imagine the possibility of these three megatrends collectively producing something greater. This paper builds on a previous newsletter[4] by adding the release of "o1," open-ended machine learning, liquid neural network news, and an updated LLM trajectory graph.

First, generative AI models like GPT-4 will continue to improve. If you were impressed with GPT-4, imagine GPT-7 and its competitors. This phenomenon is generally referred to as the LLM "scaling" hypothesis,[5] as large generative AI models scale their performance by getting bigger, faster, and stronger (see chart below).[6] This path is driven by the few global labs that have the funding, know-how, compute, and power to build these advanced models. The scaling vector of improved AI performance will likely continue unless, or until, scale "breaks,"[7] meaning LLM innovation halts or plateaus due to energy costs, the amount of data available, or the money needed to build them compared to the returns of spending more. More-general AI may be so useful that hyperscale companies may find technological offsets to these scaling barriers. Sam Altman has stated he sees no plateau in sight[8] of the scale vector and stated AGI could arrive as soon as 2025.[9] Dario Amodei has suggested 2026-27[10] but has also argued that it is not productive to fixate on the specific day of "when" AGI will arrive. Instead, the simple phenomena is that these models are getting better and better[11] and may already be better than even the best humans at some things. Dario further notes that models today are being trained at $1B, and the next foundation models – as soon as 2025 – are estimated to cost upwards of $10B to create. Yet even if the scale does "break," LLMs will plateau at some useful cost point and continue to serve as a user interface (UI) between people and AI systems-of-systems.

The newest contribution to LLMs is Open AI's "o1" – the latest model in the progression of GPT 1, 2, 3, and 4. The o1 model outperforms humans in scientific benchmarks. The "o1"

---

[3] Meredith Ringel Morris, et al., Position: Levels of AGI for Operationalizing Progress on the Path to AGI, arXiv (2024).

[4] AGI Will Arrive in Three Ways, Special Competitive Studies Project (2024).

[5] Anson Ho, et al., Algorithmic Progress in Language Models, Epoch AI (2024).

[6] Cade Metz, How 2024 WIll be A.I.'s 'Leap Forward', The New York Times (2024).

[7] Azeem Azhar, AI's $100bn Question: The Scaling Ceiling, Exponential View (2024).

[8] In Conversation with Sam Altman, All-In Podcast (2024).

[9] Noor Al-Sibai, Sam Altman Says the Main Thing He's Excited About Next Year Is Achieving AGI, Futurism (2024).

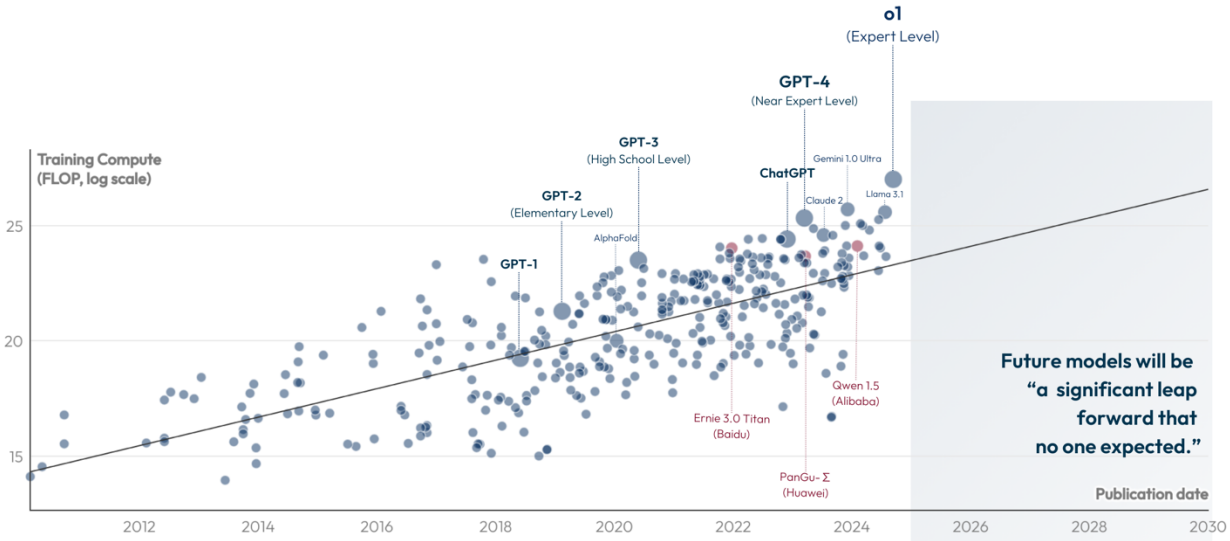
[10] Dario Amodei: Anthropic CEO on Claude, AGI & the Future of AI & Humanity, Lex Fridman Podcast (2024).

[11] Dario Amodei - CEO of Anthropic, In Good Company Podcast, Norges Bank Investment Management (2024).

model advanced AI performance by introducing a new paradigm: **test-time compute**, where the model is given more processing time to reason through problems, leading to increased accuracy. This approach, combined with **chain-of-thought reasoning** (see "Trend 2" below), allows "o1" to break complex tasks down into smaller steps, mimicking human thought processes and achieving higher performance on tasks involving logic, mathematics, and code generation. As a result, "o1" is paving the way for more intelligent and capable AI systems in the future.[12]

AI Progresses Toward Artificial General Intelligence

As AI progresses towards a more general and powerful form, models will grow in scale and capability while demands for their underlying foundations – the AI stack – increase.



Note: Noteworthy models are emphasized by adjusting the size of the data points.
Source: Epoch AI; Newsweek

	Trends Today	Trends Tomorrow
Data	AWS has the most hyperscale data centers worldwide, with 126.	By 2030, AWS will have 186
Hardware	Currently, training clusters have scaled to 100,000 GPUs.	300,000+ GPU clusters in 2025
Energy	Today, electricity dedicated to compute for AI is 3 gigawatts.	Over 56 gigawatts by 2028
Capital	Today, it costs an estimated $1 billion to train AI models.	By 2027, $100 billion in costs

Note: Some leading-edge models lack publicly available data on FLOP and are therefore missing from the graph, such as Gemini 1.5.[13]

[12] Learning to Reason with LLMs, Open AI (2024).
[13] Data on Notable AI Models, Epoch AI (last accessed 2024); Alan D. Thompson, GPT-5, Life Architect (last accessed 2024); MMLU Benchmark (Multi-task Language Understanding), Papers With Code (last accessed 2024); How Bad Will the AI Power Crunch Be?, Special Competitive Studies Project (2024); Sarah Chudleigh, Everything You Should

Second, new approaches to specific AI functions will both *improve and combine* with each other to perform tasks that are more human-like, while simultaneously supporting the scaling megatrend by combining with LLMs. A survey of several of these AI capabilities presents a simple thought experiment about what kind of AI humans will use when these vectors combine. If you combine today's AI with capabilities akin to human functions such as better reasoning, planning, creativity, memory, open-ended self-improvement, and sensing, as well as the ability to orchestrate the inputs from over 700 specialized GPT models, what would you have? The answer is something much more general. The second megatrend has economic consequences, as it opens a dynamic space in which small and medium-sized companies can flourish.

As the global AI community continues to innovate daily, new paradigms and AI functions are constantly emerging. In this rapidly evolving field, the following list stands out as the most significant advancements today:

- **Multimodality.** The ability to fuse data from different kinds of sensors and sources grants an AI system capabilities akin to the human senses. Consider how modern electric vehicles combine cameras, radar, lidar, and ultrasonic sensors into one advanced driver-assistance system. That is a modern example of multimodality and when connected to an AI system, it is like providing an algorithm with a synthetic equivalent to the human senses. Multimodality[14] is a central concept to AI systems, and it bridges the cyber-physical domain to impact life beyond a computer screen and includes subjects ranging from computer vision to digital smell. The sub-tech-vectors in multimodality are vast but include this year's "Segment Anything Model" (SAM),[15] which can accurately identify and isolate any object or region of interest within an image or video, even without specific prior training on those particular objects or scenes.

- **Chain-of-Thought (CoT) and Reasoning.** The ability of an AI application to "reason"[16] refers to the ability of a system to use logic and inference to draw conclusions, make predictions, or solve problems based on available information. It involves understanding relationships, identifying patterns, and evaluating evidence in a manner similar to human thought processes. CoT is a part of this and

---

Know About GPT-5, Botpress (2024); Norges Bank Investment Management, In Good Company Podcast: Dario Amodei - CEO of Anthropic, YouTube at 13 minutes (2024); Dylan Patel, et.al, Multi-Datacenter Training: OpenAI's Ambitious Plan To Beat Google's Infrastructure, SemiAnalysis (2024); Yih-Khai Wong, How Many Data Centers Are There and Where Are They Being Built?, Abi Research (2024); Cade Metz, Robots Learn, Chatbots Visualize: How 2024 Will Be AI's 'Leap Forward', New York Times (2024); Matthias Bastian, Sam Altman says GPT-5 could be a "Significant Leap Forward," But There's Still "A Lot of Work to Do", The Decorder (2024).

[14] Erik P. Blasch, et al., Artificial Intelligence in Use by Multimodal Fusion, 22nd International Conference on Information Fusion (2019).

[15] Alexander Kirillov, et al., Segment Anything, Computer Vision Foundation (2023).

[16] Logic-Based Artificial Intelligence, Stanford Encyclopedia of Philosophy (2024).

involves an algorithm breaking complex problems down into a series of smaller, more manageable steps. Just as humans often use intermediate thoughts and logical deductions to arrive at a conclusion, CoT prompting encourages AI models to generate intermediate reasoning steps[17] before reaching a final answer. Researchers are exploring various pathways to integrate reasoning into model architectures[18] and to control different modes of reasoning. As an example, Meta AI created "Dualformer" to mimic human thinking "fast" vs thinking "slow" from Daniel Khaneman's psychology work and this technique is outperforming "o1" in some areas.[19]

- **Planning/Scaffolding/Framing.** Framing or scaffolding combines planning and reasoning into something akin to strategy. As the engineer and investor Leopold Aschenbrenner states, "Think of CoT++: rather than just asking a model to solve a problem, have one model make a plan of attack, have another propose a bunch of possible solutions, have another critique it, and so on," using a family of AI planning applications.[20] AI models are advancing towards human-like strategy and problem-solving, utilizing a combination of tools and algorithm techniques to generate results greater than the sum of their parts. A new planning technique intersects with agentic AI by a "model-based planning framework" for web-agents that evaluate the outcomes of candidate actions by using an LLM as a world-model before deciding what action to take thus making less dead-ends in the AI planning function.[21]

- **Agentic AI.** This is a class of AI systems designed to act as autonomous agents[22] that can do general-purpose work like completing a task end-to-end using planning and software tool-calling skills. These systems are capable of making decisions, interacting with their environments without requiring constant human intervention, and even interacting agent-to-agent ("multi-agent") via the internet to learn from each other, cooperate, and even develop their own language. Agentic AI systems emphasize goal-oriented behavior and adaptive decision-making,[23] often leveraging advanced algorithms and sensory inputs to execute actions in real time and learn from continuous feedback. The bleeding edge of this

---

[17] Jason Wei, et al., Chain of Thought Prompting Elicits Reasoning in Large Language Models, Google Research, Brain Team, arXiv (2023).

[18] Such as OpenAI's o1.

[19] Dijia Su, et al., Dualformer: Controllable Fast and Slow Thinking by Learning with Randomized Reasoning Traces, Meta FAIR, arXiv (2024).

[20] Leopold Aschenbrenner, Situational Awareness: The Decade Ahead, Situational-Awareness AI (2024).

[21] Yu Gu, et al., Is Your LLM Secretly a World Model of the Internet? Model-Based Planning for Web Agents, arXiv (2024).

[22] Tula Masterman, et al., The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey, arXiv (2024).

[23] Hendrik Leitner & Sally Fletcher, What Is Agentic AI & Is It The Next Big Thing?, SSON (2024).

subject includes automated design of agentic systems (ADAS)[24] where combining the building blocks to automate design is possible. AI Agents have even been created to replicate each person on a software development team[25] in a functioning multi-agent framework. These agents then collaborate using standardized operating procedures and a shared "memory" to complete complex tasks, producing outputs like project requirements, design documents, and even functional code.

- **Specialization/Pruning/Quantization.** Specialization[26] refers to taking pre-trained AI models like LLAMA-2 and tailoring or finetuning them for a specific purpose. GitHub Copilot,[27] for an early example, is a code-completing app drawn from OpenAI's general Codex model that works in various programming languages (think of the way word processing tools can complete a sentence for you). Specialization has emerged as a computer science phenomenon that can amplify the power of large open-source LLMs as it allows others to take expensive, highly trained LLMs and specialize them without incurring such high expenses. Model pruning[28] involves removing unnecessary or redundant parameters (weights) from a trained model to reduce model size and complexity, leading to faster inference and potentially lower memory requirements. Quantization reduces the precision of the model's numerical representations (weights and activations). This typically involves converting 32-bit floating-point numbers (the standard in most training) to lower-precision formats[29] like 16-bit or even 8-bit integers.

- **Liquid Networks.** Otherwise known as liquid neural networks (LNNs), liquid networks[30] are a compute-efficient type of neural network designed for enhanced adaptability and robustness in handling time-series data. Unlike traditional neural networks with fixed architectures, LNNs possess a "liquid" quality, allowing them to dynamically adjust their internal parameters and equations in response to new incoming data, even after the initial training phase is complete. Liquid AI has now released models[31] that have smaller memory footprints and more efficient inference compared to traditional transformer-based architectures.

---

[24] Shengran Hu, et al., Automated Design of Agentic Systems, arXiv (2024).

[25] Sirui Hong, et al., MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework, arXiv (2024).

[26] Brian Lester, et al., The Power of Scale for Parameter-Efficient Prompt Tuning, arXiv (2021).

[27] GitHub Copilot, GitHub (last accessed 2024).

[28] Seul-Ki Yeom, et al., Pruning by Explaining: A Novel Criterion for Deep Neural Network Pruning, Pattern Recognition (2021).

[29] Suryabhan Singh, et al., Pruning and Quantization for Deeper Artificial Intelligence (AI) Model Optimization, Intelligent Control, Robotics, and Industrial Automation (2023).

[30] Ramin Hasani, et al., Liquid Time-Constant Networks, Proceedings of the AAAI Conference on Artificial Intelligence (2021).

[31] Chris McKay, Liquid AI Announces First Generation of Language Liquid Foundation Models, Maginative (2024).

- **Text-to-Action; Embedded AI, Cyber-Physical Advances.** Since 2000, AI has been used to optimize industrial processes[32] to help create zero-downtime systems. Now text-to-action puts that technology in the hands of individuals on the Internet of Things. In AI studies, "text to action" refers to the process of an AI system interpreting and understanding written or spoken language, translating that understanding into specific, executable actions like "set a timer for 10 minutes," and then setting a timer. These capabilities will grow to take on more complex actions.[33] Rather than just setting timers, they will be able to create autonomous cyber attack systems that can execute tailored-access operations like "turning off" the power in Washington, DC. Additionally, AI systems functioning in stand-alone machines – embedded AI – are accelerating robotics[34] and manufacturing[35] as independent trends to follow.

- **Mixture of Experts (MOE) and Composition of Experts (COE).** Specialization of generative AI models takes on a whole new meaning if *hundreds* of them can be coordinated to act as one in a mixture of experts. This AI front has two main paths. The first path is MOE.[36] MOE uses a gating network (or router) that dynamically selects which expert models to activate for a given input, based on its characteristics. The second is COE.[37] In a COE, multiple smaller expert models are chained together to form a larger, more powerful model. Each expert processes the output of the previous one, allowing them to specialize in different levels of abstraction or aspects of the problem.

- **Memory/Context Strings.** A context string[38] or context window refers to the maximum number of tokens (words or subwords) that the model can consider at once when processing a prompt or generating a response. It's analogous to the AI system's "short-term memory" or its ability to retain and utilize information from the immediate conversation or document. As Aschenbrenner[39] notes it is extraordinary that "Gemini 1.5 Pro, with its 1M+ token context, was even able to learn a new language... from scratch."

---

[32] Jay Lee, Industrial AI: Applications with Sustainable Performance, Springer Singapore (2021).

[33] Shunyu Yao, et al., ReAct: Synergizing Reasoning and Acting in Language Models, arXiv (2022).

[34] Rachmad Vidya Wicaksana Putra, et al., Embodied Neuromorphic Artificial Intelligence for Robotics: Perspectives, Challenges, and Research Development Stack, arXiv (2024).

[35] Grace Harmon, Modernizing Manufacturing: Divergent Raises $230M to Aid Growth, Los Angeles Business Journal (2023).

[36] Saeed Masoudnia & Reza Ebrahimpour, Mixture of Experts: A Literature Survey, Springer (2012).

[37] Raghu Prabhakar, et al., SambaNova SN40L: Scaling the AI Memory Wall with Dataflow and Composition of Experts, SambaNova Systems, Inc., arXiv (2024).

[38] Alan Akbik, et al., Contextual String Embeddings for Sequence Labeling, Proceedings of the 27th International Conference on Computational Linguistics (2018).

[39] Leopold Aschenbrenner, Situational Awareness: The Decade Ahead, Situational-Awareness AI (2024).

- **Inference Engines.** Inference engines[40] in AI are like the brains behind the operation, taking in information and using it to make decisions or predictions. They take the knowledge stored within a trained AI model and apply it to new, unseen data, allowing the AI to understand and respond to real-world situations. This is crucial for a wide range of AI applications, from self-driving cars making split-second decisions based on sensor data to chatbots understanding and responding to human language. In essence, inference engines are what enable AI systems to move from theoretical models to practical, real-world problem-solvers.

- **Co-Piloting and Program Synthesis.** The trend of AI systems working on computer code has followed a basic three-step progression. First, it helped a human to complete code (co-piloting), then it refined/tested code (fine-tuning), and, finally, it now autonomously generates its own code to solve problems given by a human (program synthesis). In 2017, program synthesis was described by two researchers as the holy grail for AI: "The grand dream of program synthesis is to make programmers obsolete. The holy grail is to be able to simply state one's intent in some natural form, and have the computer automatically synthesize an efficient program that meets that intent."[41] Seven years later, this front is advancing and you can observe certain freely available models that will pause to build code for whatever task was in your prompt. For example, Claude 3.5 Sonnet improved on a model that was already good at coding: "Claude 3.5 Sonnet can independently write, edit, and execute code with sophisticated reasoning and troubleshooting capabilities."[42]

- **Grounded Search.** This is a technique that aims to reduce hallucinations in model performance (generating incorrect or nonsensical information) by anchoring the model's responses in reliable external knowledge[43] sources such as the internet. Grounded search is one way for a non-living AI system to have an external reference for accuracy once it has been trained. Relatedly, *Retrieval-Augmented Generation (RAG)*[44] involves augmenting an LLM with external knowledge sources like databases or private documents. This enables the model to access up-to-date information and generate more factual and contextually relevant responses.

- **RLHF and Auto-HF.** Reinforcement learning with human feedback (RLHF) is what you do when you "thumbs up/thumbs down" an LLM response to your question. You give the machine feedback, telling it if it succeeded, or if it did not. *Automatic*

---

[40] What is an Inference Engine? Types and Functions, Geeks for Geeks (2024).

[41] Sumit Gulwani, Program Synthesis, Foundations and Trends in Programming Languages (2017).

[42] Introducing Claude 3.5 Sonnet, Anthropic (2024).

[43] Jacob Andreas, Neural Module Networks, Computer Vision Foundation (2016).

[44] Jiawei Chen, et al., Benchmarking Large Language Models in Retrieval-Augmented Generation, Proceedings of the AAAI Conference on Artificial Intelligence (2024).

human feedback (Auto-HF)[45] is the ability for an AI model to autonomously get the human feedback it needs to formulate outputs, such as automatically characterizing and machine learning from human references from videos on the internet. Auto-HF allows a trained model to predict or approximate human feedback on its generated outputs, without requiring direct input from human evaluators for each instance. The "Constitutional AI"[46] technique is an example of auto-HF where the core idea is to use AI feedback itself, rather than relying solely on human labels, to ensure the AI behaves in accordance with a set of values, or a constitution, determined by the model developers. This constitution, often called a reward model, is first trained on a dataset of human preferences or rankings of AI-generated outputs.

- **Open-ended Machine Learning.** Imagine an AI system that could go on learning without human direction. This is called open-ended machine learning and it represents an exciting match to the way humans muse and pursue learning undirected by someone else. Open-ended ML[47] allows for an AI model to continuously generate novel and surprising solutions or ideas (variations), shrink these learning pathways down to the most interesting,[48] and even gather empirical evidence to test ideas.[49] This is a path to an auto-self-improving form of ML that effectively exhibits an unbounded capacity for learning and discovery. This is in contrast to current AI systems that are typically trained for specific tasks and have limited ability to adapt or innovate beyond their initial training data.

- **Compositionality and Creativity.** When an image generator creates artwork based on the input of human words, or an LLM is able to generate a decent new poem, this reflects early signs of AI *compositionality* that is like humans. AI creativity appeared in 2016 when AlphaGo's unexpected move 37 in its victory over Lee Sedol shocked the Go world and would change how humans play the game.[50] Compositionality and creativity are key challenges in AI research on the path toward AGI as these qualities enable systems to go beyond simple pattern recognition and demonstrate a deeper understanding of the underlying structure and meaning of their inputs to co-create with people.

---

[45] Nisan Stiennon, et al., Learning to Summarize From Human Feedback, 34th Conference on Neural Information Processing Systems (2020).

[46] Yuntao Bai, et al., Constitutional AI: Harmlessness from AI Feedback, Anthropic, arXiv (2022).

[47] Edward Hughes, et al., Open-Endedness is Essential for Artificial Superhuman Intelligence. arXiv (2024).

[48] Jenny Zhang, et al., OMNI: Open-endedness via Models of human Notions of Interestingness arXiv (2023).

[49] Tim Rocktäschel, Is Artificial Superintelligence Imminent? TWIML (2024).

[50] Cade Metz, In Two Moves, AlphaGo and Lee Sedol Redefined the Future, Wired (2016).

- **Learning with Less/One-Shot/No-Shot Learning.** In a nutshell, machine learning with less data[51] is about being "compute efficient" with your training resources. One-shot learning[52] is about learning quickly from minimal examples like an AI model seeing one picture of an animal and then recognizing it like a human would. Zero-shot learning[53] is about generalizing knowledge to completely new situations.

- **Reflexion.** Reflexion[54] allows an AI model to generate multiple diverse reasoning chains or "thoughts" before arriving at a final answer. These intermediate thoughts are then evaluated, and the likeliest correct one is selected as the output. This process mimics the human thought process of reflection, where multiple possibilities are considered before making a decision. It is easy to imagine how this computer science front could combine with AI agents to make them more capable of general applications.

- **Attention Modeling.** Attention modeling[55] – the ability to measure (descriptively) and guide (prescriptively) where an AI system turns for data to perform its tasks – is not new but it continues to grow in significance. Attention mechanisms are being used in computer vision tasks[56] like image recognition, object detection, and image captioning. They enable models to focus on the most relevant parts of an image. For advanced scientific discovery, attention-based models are being used to analyze complex scientific data, such as protein sequences[57] and molecular structures, and can identify patterns difficult for humans to detect. In these ways, attention modeling helps AI systems process information more intelligently and effectively, thus contributing to more general and sophisticated AI applications.

A **third** way AGI will arrive is by transforming the fundamental technologies upon which AI depends – such as compute, data, microelectronics, networks, and energy. Leaps in these fundamental parts of the AI stack will take AI performance to another level.

- **Compute.** Different paradigms like quantum, neuromorphic, and reversible computing, plus existing forms of classical supercomputing, will grant both strengths and weaknesses for AI performance. We should expect the potential of

---

[51] Greg Schohn & David Cohn, Less is More: Active learning with Support Vector Machines, Just Research (2000).

[52] Oriol Vinyals, et al., Matching Networks for One Shot Learning, 30th Conference on Neural Information Processing Systems (2016).

[53] Wei Wang, et al., A Survey of Zero-Shot Learning: Settings, Methods, and Applications, ACM Transactions on Intelligent Systems and Technology (2019).

[54] Noah Shinn, et al., Reflexion: Language Agents with Verbal Reinforcement Learning, 37th Conference on Neural Information Processing Systems (2023).

[55] Ashish Vaswani, et al., Attention is All You Need, 31st Conference on Neural Information Processing Systems (2017).

[56] Meng-Hao Guo, Attention Mechanisms in Computer Vision: A Survey, Computational Visual Media (2022).

[57] John Jumper, et al., Highly Accurate Protein Structure Prediction with AlphaFold, Nature (2021).

large QBIT, fault-tolerant quantum computing[58] to arrive before 2030. While quantum computing is not a driver of AI innovation per se, when this new paradigm of compute arrives, AI systems operating on classical computers will have a powerful external source for modeling the real world.[59] Thus, a real frontier lies in integrating the advantages of each kind of computing into one functioning system optimized for AI performance. SCSP has previously called for dominating the hybrid computing space in a way that resembles U.S. leadership in the design of microelectronics currently (i.e. as a future economic position worth achieving as a nation).[60]

- **Microelectronics.** Both the scaling of LLMs and the advancement of novel paradigms previously mentioned are both based in whole or in part on improvements in microelectronics.[61] For example, both the context strings that give an AI system human-like memory capabilities and COE algorithms that can orchestrate pull over 700 specialized GPTs into concert were only made possible by the development of new microelectronics.

- **Data Centers and Science.** Innovation in data management includes advancements in the data center itself, in better organizing and labeling data, and in identifying novel sources of data for AI systems. Hyperscale companies like Microsoft are opening a new data center every three days.[62] Building them better, faster, and cheaper is its own broad field. On organization, labeling data (structuring data to be machine-readable for AI use) was a barrier for machine learning systems in 2021. Now, auto-labeling[63] has arrived: the AI systems can do the labeling autonomously. This means that machine learning models can now be trained much faster and more efficiently. Another profound front in data innovation is access to different categories of data for training and grounding AI models such as open internet-available, synthetic/simulated, multimodal, provided databases, and proprietary "future" data such as that in scientific labs. Innovation across the entire data science vertical could fundamentally improve AI performance and scale.

---

[58] National Action Plan for United States Leadership in Advanced Compute & Microelectronics, Special Competitive Studies Project (2023).

[59] Hsin-Yuan Huang, et al., Quantum Advantage in Learning from Experiments, arXiv (2021).

[60] National Action Plan for United States Leadership in Advanced Compute & Microelectronics, SCSP (2023).

[61] Michael Acton & George Hammond, Chip Challengers Try to Break Nvidia's Grip on AI Market, Financial Times (2024).

[62] AI for Power & Energy with Laurent Boinot, The TWIML AI Podcast (2024).

[63] Shikun Zhang, et al., A Survey on Machine Learning Techniques for Auto Labeling of Video, Audio, and Text Data, arXiv (2021).

- **Networks, IOT, and Edge.** As networks advance, this will free up AI applications to function more like a system of systems.[64] Think of super-fast internet networks like optical networks or 6G as lightning-fast highways for information. With these highways, AI systems connected to the Internet of Things (IoT) can exchange massive amounts of data in the blink of an eye. This means AI systems can react and make decisions almost instantly, like a self-driving car avoiding an accident before you even see the danger. This super-speed also allows AI models to learn from the constant stream of data from IoT devices in real time, making them smarter and more capable every second. Basically, these advanced networks transform AI models from brainy but slow thinkers into super-fast, super-smart decision-makers, revolutionizing how they interact with our connected world.

- **Energy.** Imagine a world where new forms of renewable energy like fusion drive energy costs closer to "zero." Cheaper energy would lower the price point for owning and operating complex AI systems (including data storage, transport, and training). Yet as SCSP has noted, there is much that needs to be done[65] to reach this positive future.

The likelihood is that these three megatrends will develop, interact, intertwine, and combine to create the **fourth** way to AGI. Big questions remain about how the megatrends will combine: when it will happen, who will do it, and which drivers will matter most. Imagine a combination of a future LLM or AI-agent as a user interface, with AI systems that combine better memory, reasoning, creativity, and learning plus advanced modeling on quantum computers integrated with classical computers like the one you are using now. As PsiQuantum has suggested, that combination has the potential to transform whole industries.[66]

The resulting path to AGI will proceed in four simple stages: 1) AI of today; 2) some intermediate stage that is "more-general" as discussed by the NSCAI;[67] 3) AGI that begins to disrupt (positively and negatively) whole verticals of the economy; and 4) something better than humans in some aspects called artificial super-intelligence (ASI). As the confluence of these megatrends accelerates us toward the threshold of AGI, getting positioned and organized for the arrival of this general purpose technology is a defining challenge of our era.

---

[64] AI-Powered 6G Networks Will Reshape Digital Interactions, MIT Technology Review (2023).

[65] How Bad Will the AI Power Crunch Be?, Special Competitive Studies Project (2024).

[66] Applications, PsiQuantum (last accessed 2024).

[67] Final Report, National Security Commission on Artificial Intelligence (2021).