Carnegie Mellon Institute for Strategy & Technology

JULY 2025

WHITE PAPER

AGENTS OF CHANGE

Rapid Shifts in AI Economics Are Redefining How Agentic Systems Are Built, Powered, and Deployed

By Harry Krejsa and Dr. Thomas Şerban von Davier

Carnegie Mellon University

About the Authors

Harry Krejsa is the director of studies at the Carnegie Mellon Institute for Strategy & Technology. Harry joined Carnegie Mellon from the White House's Office of the National Cyber Director, where he led development of the 2023 National Cybersecurity Strategy, established national modern energy security priorities, and represented the U.S. government in technology security consultations with foreign partners and the global private sector. Harry previously worked at the intersection of technology, industrial strategy, and US-China competition for the Department of Defense, the Cyberspace Solarium Commission, and the Center for a New American Security.

Thomas Şerban von Davier, DPhil, is an AI research scientist at the Carnegie Mellon University Software Engineering Institute. Thomas joined SEI from the University of Oxford, where he completed his doctoral work on artificial intelligence (AI) in the Department of Computer Science. He has a growing body of published work on AI engineering and AI safety. Previously he worked as a senior data researcher for a media and tech consultancy group, empowering clients with insights into their data and how it could be operationalized for improved model performance and user satisfaction.

Acknowledgements

The authors are indebted to many friends, colleagues, and mentors who contributed immeasurably to the production of this report. Expert feedback from Tyler Books, Dr. Marissa Connor, Dr. Jodi Forlizzi, Dr. Shannon Gallagher, Keltin Grimes, Dr. Eric Heim, and Carol Smith was invaluable. Support and direction from Dr. Audrey Kurth Cronin of the Carnegie Mellon Institute for Strategy & Technology and Dr. Matt Gaston of the Software Engineering Institute were key to the authors' success. Project guidance and design provided by Jess Regan and Carolyn Just were critical. Editing by Sandra Tolliver and Aleksaundra Handrinos was excellent.

The authors employed artificial intelligence tools to assist with initial research, drafting and editing, and preliminary visual concepts. All AI-generated material was rigorously checked, revised, approved, and integrated by the authors. The views herein are the authors' alone, along with any errors of fact, omission, or interpretation.

This report and its findings are the sole responsibility of the authors. Cover illustration by Jess Regan.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The economics of artificial intelligence are shifting. In a few short months, our understanding of what AI systems can do, who can build them, and how they will be powered and deployed has been transformed.

Early AI models initially focused resources on their *training* phase, building sophisticated but relatively static models of knowledge from vast datasets. Subsequent models began focusing more on their *inference* phase, the costly computational process of applying an AI model's trained knowledge to "reason" over new, real-world tasks. While inference-heavy models were quickly shown to be highly capable, their steep computational demands limited widespread use and created significant economic bottlenecks, especially for the more autonomous "agents" the industry envisioned as the future of AI technology.

Several technology trends have recently converged to overcome these cost and computing barriers, with the Chinese model DeepSeek spotlighting their potential. By harnessing these trends—and adding several of its own innovations—DeepSeek slashed its inference costs to such a small fraction of comparable systems that it could even run locally on a personal computer with minimal loss in performance. Its developers then invited the world to do just that, making their model available to download and tailor as users desired. DeepSeek did not invent freely-customizable models, where Meta has long been an industry leader, but it reinvigorated interest in such models after demonstrating how quickly the market can transform when performance, efficiency, and accessibility align.

These shifts in AI economics are enabling a more dynamic, democratized, and diversified ecosystem than many expected even a short time ago. Cheaper, lightweight, and customizable models are making room not just for more autonomous (or "agentic") use cases, but also for systems of multiple specialized AI systems that interact with one another alongside more powerful cloud-based services. **Early evidence suggests this arrangement—multiple AI agents of various sizes and specialties complementing one another and checking each other's work—produces superior, more accurate outcomes.** If these advantages hold in broader deployment, they may unlock the level of trust and reliability required for agents' broader use by organizations and consumers alike.

This transition toward more flexible, affordable, and distributed AI systems is also reshaping how we have envisioned their deployment, from the way we build data centers and power infrastructure to how we will manage these systems' cybersecurity and data governance. While the coming "agentic era" will surely still require substantial new sources of electricity, the distribution of that

electricity consumption may differ from projections made just a year ago. **As AI systems diversify away from a few cloud-based monoliths into a spectrum of differently sized and customized agents, some computing requirements will move closer to users—from regional inference facilities, to on-premises servers, to individual devices—and our infrastructure requirements will likely evolve accordingly.** This partially-distributed approach also offers potential security benefits. Research suggests that multiple, specialized cybersecurity agents can collaborate to produce more effective intrusion detection and remediation, and locally-executable models will allow organizations to leverage sophisticated AI capabilities while minimizing how frequently sensitive data has to leave their facilities.

Drawing on DeepSeek's breakthroughs as a catalyst for understanding these shifts in AI economics, this brief examines how dramatically-reduced inference costs and open distribution models are combining to democratize access to powerful AI capabilities. We then explore the imminent consequences this democratization holds for our energy infrastructure, computing requirements, and the security and governance of our data.

Dramatically-reduced inference costs and open distribution models are combining to democratize access to powerful AI capabilities.

Looking forward, this brief recommends that policymakers, industry, and civil society adopt four priorities for seizing the potential of these developments while mitigating their risks:

1. Swift and broad deployment of more electricity
2. Strengthening open-source AI development
3. Establishing robust interoperability protocols for agent collaboration
4. Incentivizing business models and equipment that support locally-executable AI

These steps will help ensure that the emerging distributed and multi-agent ecosystem develops in ways that enhance both innovation and resilience—creating an AI future shaped not by a few dominant platforms, but by diverse, specialized, and collaborative intelligence working at every scale.

THE SHIFTING LANDSCAPE OF AI ECONOMICS

Building generative artificial intelligence requires amassing vast quantities of raw information—made up of books, websites, images, and other digital information—to "train" AI models. This training leverages specialized computer chips, immense computational power, significant energy, and sophisticated, self-improving algorithms to drive what the field calls machine learning. The resulting models can identify patterns in their training information and generate what they predict are the most likely (and, consequently, human-like) responses when prompted across a variety of fields, specialties, or contexts.

Many assumed that this initial training phase would be the costliest or most economically consequential part of creating an AI system. That assumption would have allowed AI products to follow business patterns most familiar to investors in traditional software development: large capital expenditures up front (i.e., building a complex piece of software) followed by the near-zero marginal cost of deploying that software for each subsequent use. The reality, however, has proven more complex. The computationally-intensive process of AI models actually applying their training on a case-by-case basis—what the field refers to as "inference" —has emerged as an equally, if not more, demanding challenge.

Inference generalizes an AI model's insights, gleaned from vast but static quantities of knowledge, for application to a novel context prompted by a single user—a technical marvel that is also much costlier and less scalable than the training phase that preceded it.[1] Early products like OpenAI's ChatGPT 3.5 were designed to be minimalists when deploying inference. These models focused all of their training-derived "intelligence" on predicting the next word in a sentence but largely lacked the ability to step back and comprehensively evaluate a complex question, much less pause to examine its response to that question and reevaluate or correct its approach. The result was a fast and prodigious model, but also one prone to "hallucination"—in the AI field, one lay term for generating false or inaccurate information in response to a factual question or request.[2] These inference-light models were frequently powerless to prevent their responses from compounding on an error once an error was made.

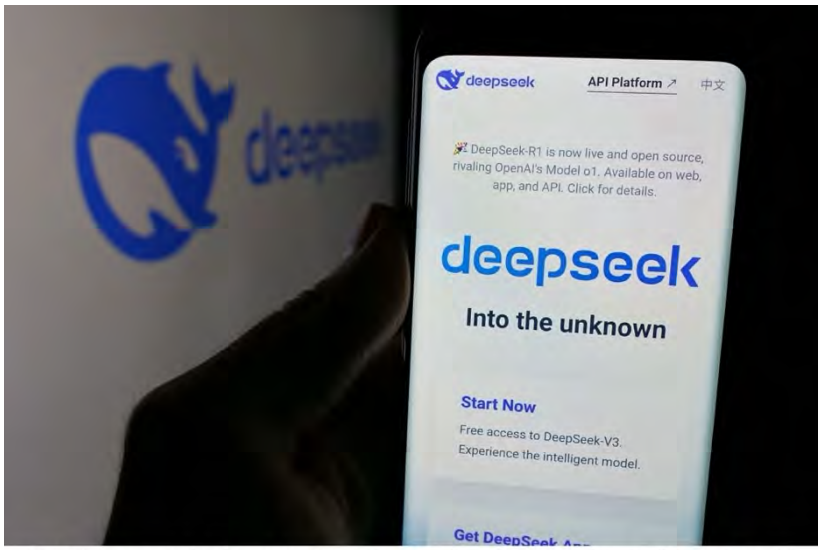
By late 2024, it was clear this maximum-training-minimum-inference approach was primed for change. As researchers encountered diminishing returns from training—where throwing more data and processing power at models no longer yielded proportional improvements in "intelligence"—they discovered that investing more time and computing power into inference in response to an individual user could create superior results.[3] Models instructed to "reason" longer before answering—essentially pausing the generation of their response to reconsider and revise before

proceeding on a more refined path—produced insights that were often higher quality and less prone to hallucination.

This was an important discovery, but one with hurdles to practical application. The improved performance from rebalancing toward inference came at the cost of significantly higher computation and energy requirements, implying that such capabilities would remain centralized among only the largest companies with access to vast computing resources. Further, these computational and economic constraints presented a serious bottleneck for the next stage of AI developers' ambitions: "agents," or models that could plan, make decisions, and take actions on their own. Realizing this agentic future for any appreciable number of users would require significant advances in model efficiency, new paradigms for model distribution and deployment, and a reimagining of energy and compute infrastructure. As it turned out, transformational shifts across all these variables were mere months away.

BREAKING DOWN BARRIERS TO AGENTIC AI

By late 2024, several developments in artificial intelligence development were converging to overcome cost and compute barriers to more autonomous AI, and even potentially democratize AI development writ large. Though many were already underway across the industry, it took the emergence of the Chinese model DeepSeek to spotlight both those trends and the enduring and global competitiveness of the sector. Operating within China's artificial intelligence ecosystem—constrained by US export controls on cutting-edge chips—DeepSeek was born from High-Flyer, a quantitative hedge fund already organized to pursue competitive advantages through computational innovation.[4] This constraint-driven creativity led High-Flyer to adopt a series of bleeding-edge hardware, software, and algorithmic innovations that resulted in a smaller, cheaper, and more efficient model than many of its peers. This was especially so in the case of inference, where High-Flyer's inference-heavy reasoning model, DeepSeek-R1, reportedly operated at 10 percent the cost of comparable OpenAI models.[5]

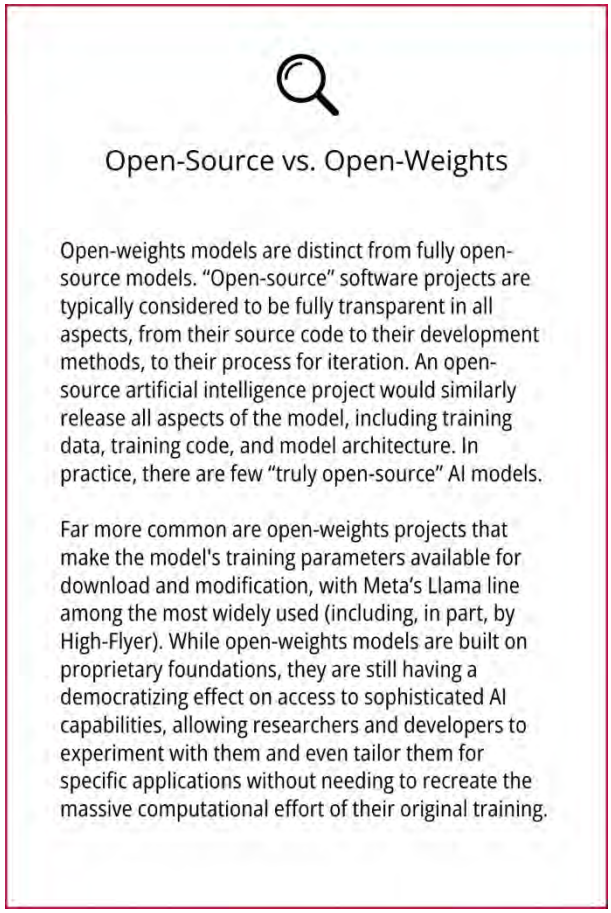


High-Flyer's constraint-driven creativity led to bleeding-edge innovations, making their model smaller, cheaper, and more efficient than competitors.

If DeepSeek creatively implemented a number of trends underway in model *optimization*, it also did the same in model *distribution*. Unlike many of its American contemporaries who monetized their models via subscription offerings, DeepSeek chose to release its model as an "open-weights" product. In the context of AI, *weights* refers to the numerical parameters that define how an AI model processes and interprets information—essentially a quantification of the "brain" that has been shaped through training. Releasing the model as an open-weights product means releasing these numerical parameters publicly, allowing anyone to view the model file and use or customize it directly. Further, **because High-Flyer had been so successful in optimizing DeepSeek's computing requirements, it instantly became one of the most capable AI models that users could download and run on mainstream consumer or enterprise hardware.** DeepSeek had opened up a broader world of AI use cases that could run entirely locally—no cloud-based data flows or subscription services required.

The American AI ecosystem was already making significant strides in open or locally-executable AI models—perhaps most notably by Meta's open-weights Llama line of models. DeepSeek, however, demonstrated that there was not only more room in the marketplace for world-class open-weights competitors, but that you might not need to be a globe-spanning tech giant to be able to field powerful AI products. In the near term, this attracted more mainstream attention to the open-weights sector; shortly after DeepSeek's release, Google revealed the latest, much more capable additions to its Gemma line of fast, open-weights models[6], and even OpenAI, which had famously retreated from its original vision of "open" models before developing ChatGPT, announced its intent to return to the open-weights ecosystem.[7] Over the medium term, the post-DeepSeek jolt of interest in smaller and open models also may have continued mainstreaming the potential for a future AI ecosystem of numerous diverse models interacting and collaborating with one another—rather than one defined by only a few generalist models from a small cohort of tech giants.

Recent moves from both industry and academia provide further evidence that this combination of efficiency gains and open, customizable models is going to be critical as artificial intelligence moves toward more autonomous and capable "agents." Anthropic in late 2024 announced the Model Context Protocol (MCP), an open software development standard (that has been subsequently adopted by other leading AI labs) intended to make data repositories, business tools, and other digital assets interoperable with different AI systems, heading off the risk of proprietary interfaces promoting "lock-in" to a esingle company at this early stage in the field's development.[8] Similarly, Google's Agent2Agent (A2A) protocol, also backed by Microsoft and others, promises to smooth the orderly and collaborative hand-off of tasks between agents designed and operated by different developers and platforms.[9] Early research from academia further suggests these frameworks for multiple collaborative and complementary agents show promise to make AI tools more useful and reliable in everyday life, including by **reducing hallucination, improving accuracy, and buttressing alignment to user intent.**[10]



Open-Source vs. Open-Weights

Open-weights models are distinct from fully open-source models. "Open-source" software projects are typically considered to be fully transparent in all aspects, from their source code to their development methods, to their process for iteration. An open-source artificial intelligence project would similarly release all aspects of the model, including training data, training code, and model architecture. In practice, there are few "truly open-source" AI models.

Far more common are open-weights projects that make the model's training parameters available for download and modification, with Meta's Llama line among the most widely used (including, in part, by High-Flyer). While open-weights models are built on proprietary foundations, they are still having a democratizing effect on access to sophisticated AI capabilities, allowing researchers and developers to experiment with them and even tailor them for specific applications without needing to recreate the massive computational effort of their original training.

> ***The rise of inference and agentic AI is redrawing the map of data center construction, power generation, and secure model deployment.***

Early moves like these show that industry is willing to make "down payments" toward a more open, interoperable ecosystem and embrace the innovative challenge presented by DeepSeek-style shifts in AI economics. Turning these aspirations into reality, however, will next hinge on more complex issues of technical implementation. Cheaper and more diffuse agents mean more intensive inference workloads, more distributed electricity demand, and nuanced security considerations. The next sections explore how these considerations are already redrawing the map of data center construction, power generation, and model deployment.

EVOLVING ENERGY & COMPUTING INFRASTRUCTURE FOR AGENTIC AI

The transition from training-intensive to inference-intensive tools like agents is changing the story of our AI infrastructure expansion from a question of raw gigawatts and chips to a more nuanced one of geography, efficiency, and architecture.

As more AI compute shifts from far-away training centers closer to users (and, with the proliferation of small or open-weights models, even on their personal devices), energy and data infrastructure will similarly need to diffuse outward. This pattern may buy the United States some slack in its coming electricity demand "crunch," but more DeepSeek-style advances in efficiency, and more growth in nationwide power generation—beyond just a selection of data center mega-projects—will still be necessary to ensure that reprieve is not squandered.

Shifting Data Center Priorities: From Energy and Training Toward Latency and Inference

Inference data centers, much like internet-supporting facilities, need substantial and dependable electricity, but also prioritize minimizing latency.

Initially, the boom in AI electricity demand was expected to come from massive, centralized data centers dedicated to the aforementioned training phase for artificial intelligence models. Easy access to power was the primary consideration for locating these facilities, as they were expected to run their fleets of specialized, energy-hungry chips at predictably constant rates around the clock to

complete these vast training tasks, with the (then-believed) more lightweight inference tasks being a less important consideration. This was a partial departure from earlier waves of data centers designed to support our internet infrastructure. While internet-supporting data centers were also energy-hungry facilities requiring robust and reliable access to electricity, they also valued minimizing *latency*, or the delay between a digital question leaving a data center and receiving an answer back from another server or end user.

Tiny milliseconds of latency add up to competitive advantages for major cloud services, so internet-supporting data centers traditionally have sought geographies that balance both access to power *and* proximity to high-throughput "internet backbone" connections. The stereotypically ideal location for an internet-supporting data center was in Northern Virginia, where those backbone connections and relatively robust electricity infrastructure are both plentiful. Training-focused AI data centers, in contrast, were envisioned for more remote parts of the country where they could optimize for energy *without* having to worry about latency. For a training-centric AI market, proximity to end users was less important than cheap electricity markets, permissive regulatory environments, and sometimes even enough land to build a data center's own dedicated solar fields or nuclear reactors. As inference-heavy reasoning and cheaper, specialized agentic systems gain prominence, latency and proximity to users are returning as design and construction imperatives for both AI data centers and power infrastructure.

Reimagining "Where" Proliferating Agents Do Their Work

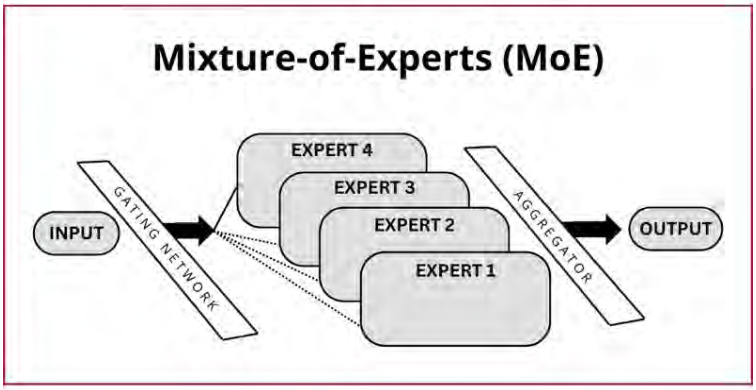
Time-sensitive inference operations for specific user requests—particularly those involving model "thinking" or multi-agent collaboration—struggle with the latency delays that come with a few giant but remote data centers in Texas or Wyoming attempting to serve customers around the country. The market has already responded to this reality, quickly moving AI processes closer to their users via regionally distributed inference-specific computing infrastructure. **While the precise ratio of training to inference data centers is not known, barely a few months into the "reasoning" and "agentic" eras of artificial intelligence, experts agree that a growing majority of all power consumed by AI is almost certainly now for this kind of user-facing computing.**[11]

Moving this type of computing requirement closer to its users could take multiple forms, from regional data centers that serve the cloud-based inference needs for metropolitan areas, to hardware running locally-executable AI models directly in company offices or supporting facilities ("on-premises" or "edge" computing), and even increasingly sophisticated products running directly on phones, laptops, and other individual devices ("on-device AI"). Diffusing out computing like this could end up alleviating the concentrated spikes in new electricity demand that are taxing many parts of the US grid today—or, paradoxically, it could also drive substantially higher energy

consumption systemwide as cheaper AI products spread out onto more numerous but less efficient individual devices. Three hundred million smartphones running all their own local inference processes would have very low latency, but then the grid would have to reckon with a dramatic increase in afternoon battery top-ups from every corner of the country.

Three hundred million smartphones running all their own local inference processes would have very low latency, but then the grid would have to reckon with a dramatic increase in afternoon battery top-ups from every corner of the country.

Toward an Architecture That Works Smarter, Not Harder



In practice, we are likely to see a spectrum of both agent and infrastructure design emerge to balance these increasingly nuanced power and compute requirements. DeepSeek's release signaled that there are likely still dramatic efficiency gains to be had in AI models generally, and inference specifically, that could make our coming energy crunch more manageable while simultaneously making smaller and more numerous AI applications more capable.

One such example is DeepSeek's implementation of an AI model design technique called a Mixture-of-Experts (MoE). Rather than using the entire DeepSeek model for every task, it is able to selectively activate only the specialized parts of its model (or "experts") needed for any given question—using just 9 percent of its total capacity in DeepSeek-V2—allowing it to dramatically reduce its computing requirements while maintaining the same quality of results.[12] Rather than running vast, generalist models for every query, future systems will become increasingly selective, dynamically allocating computing resources based on a task's subject or complexity.

AI labs writ large seem to be following this shift away from "brute-force" approaches, where AI models marshal their maximum capability for every prompt, and instead turning to more precise toolkits, like the Mixture-of-Experts approach or, in the case of reasoning models, a "sliding scale of

intelligence," depending on how difficult the model diagnoses a problem to be.[13] This selective hierarchy of intelligence is so promising that it is likely to extend beyond individual models to entire device ecosystems as well. As more open, locally-executable, and on-device AI capabilities continue to improve, we can expect "triage" systems that process simpler tasks closer to the user while escalating more complex operations to more powerful on-premises computing resources or cloud-based resources as needed.

We can expect "triage" systems that process simpler tasks closer to the user while escalating more complex operations to more powerful on-premises computing resources or cloud-based resources as needed.

"Jevons Paradox" and Efficiency in AI Economics

In the midst of a British coal crunch in 1865, economist William Stanley Jevons observed that as steam engines became more efficient, England's demand for coal did not decline, but instead soared. By squeezing ever more watts of energy out of every lump of coal, technological innovation was making it cheaper for energy-guzzling enterprises to grow and making it easier for newer energy-sipping business ideas to enter the market.

When DeepSeek's efficiency gains first became widely known, tech executives rushed to reassure anxious investors that China's innovations were, in fact, good news for the sector as a whole. Like 1860s energy markets, cheaper AI meant more AI, they argued. Whether you are a tech giant running a massive cloud based subscription model, or a university lab employing a team of open-source, locally-executed ones, the theory goes, a more inclusive AI economy is a rising tide that lifts all boats. Time will tell if the history of efficiency and innovation will again repeat.

Major cloud providers are working to capitalize on these trends by effectively extending versions of their infrastructure into customer premises. Notable examples include tools like Microsoft's Azure Private AI containers,[14] or Google's Anthos edge framework and Gemma on-device AI models, each offering cloud-like AI capabilities while minimizing how often customers' data must leave corporate facilities or even individual devices. Enterprise hardware manufacturers are likewise positioning themselves for this growing interest in more localized AI, with companies from Dell to NVIDIA offering computing platforms and specialized chips to enable on-device artificial intelligence performance that a short time ago could be found only in data centers.[15]

Yet, as these more affordable and diverse implementations of artificial intelligence proliferate, security and reliability implementations will need to advance to match. The next section will discuss how a more open ecosystem of both cloud-based general agents and more numerous tailored or locally-executed ones

will require thoughtful architecture and governance decisions. These decisions will ensure that a more dynamic, affordable, and democratized ecosystem is not undone by declines in trustworthiness or a larger, looser attack surface.

SECURITY & RELIABILITY IN AN ERA OF PROLIFERATING AGENTS

AI products have been dazzling consumers for a few years now, but have been slower to clear the bar for enterprise work in more sensitive applications like clinical settings or national security. DeepSeek-level efficiency gains and a fast-maturing open-weights ecosystem now hint at a turning point, however—one where lower costs and tailored models can deliver outputs reliable enough for critical domains. Realizing that promise, however, still depends on nuanced implementations, hardware tuned for lightweight inference, and an approach to openness and interoperability that lets many agents coordinate without exposing fresh seams for attack.

Reducing Hallucination and Improving Accuracy

Innovations in model efficiency and customization are enabling the new kinds of tools necessary to ensure that AI and coming "agentic" services can be deployed securely and reliably. One such tool slowly making its way from research to deployment is the multi-agent system, where multiple agents interact with one another to handle increasingly complex tasks. As energy and compute requirements for artificial intelligence fall, it is becoming more feasible to implement frameworks where multiple AI agents compete, collaborate, or complement one another to reduce hallucination, improve accuracy, and ensure closer alignment with user intent.[16] These kinds of improvements will be critical for developing next-generation products more

It is becoming more feasible to implement frameworks where multiple AI agents compete, collaborate, or complement one another to reduce hallucination, improve accuracy, and ensure closer alignment with user intent.

appropriate for sensitive applications, be they clinical settings, classified environments, or industrial deployment.

While reducing hallucination remains one of the biggest challenges to these kinds of sensitive AI applications, recent work by researchers like Diego Gosmar and Deborah Dahl suggests that setting up multiple tailored AI agents to review one another's outputs may be part of the solution. In their study, sequential and specialized agents would attempt to respond to a prompt, search that response for unverified claims, incorporate disclaimers or caveats where necessary, and clarify when part of their answer was speculative or uncertain. Their experiment showed that this robotic team-up of a pedant, a skeptic, and nervous novice successfully caught and reduced hallucinations across hundreds of prompts that had been specially designed to tempt AI models into hallucinating.[17] Similarly, Hong Qing Yu and Frank McQuade found in their own recent research that even systems that incorporate "simple" fact-checking agents—armed with (far from simple) continuous knowledge updates against which to validate responses—were able to achieve a 73 percent reduction in hallucinations compared to standalone GPT-4o responses.[18] Anthropic reportedly found similar advantages and detailed how they implemented related multi-agent principles in their most recent Claude research product.[19] But while these multi-agent systems—increasingly enabled by more efficient and tailorable models—show impressive potential for improving performance, reliability, and accuracy, implementing them effectively beyond research functions and in real-world environments will face many of the same model and data security challenges that other AI tools face, but in multiplication.

Multi-Agent Security a Multi-Edged Sword

Cybersecurity is another illustration of how inexpensive, proliferated agents are poised to be a boon, a burden, or both. It is conventional wisdom that one of the field's greatest challenges is its bias toward attackers; hackers need to find just one vulnerability to exploit on their own timeline, while defenders are tasked with protecting against all threats at all times. In a post-DeepSeek world, multi-agent systems now offer the possibility of mitigating that attacker advantage by

Multi-agent systems now offer the possibility of mitigating traditional attacker advantage by economically automating detection, response, and recovery functions at greater speed and scale than the best and most collaborative human teams could hope to achieve.

economically automating detection, response, and recovery functions at greater speed and scale than the best and most collaborative human teams could hope to achieve.

Early research suggests this vision of multiple simultaneous, specialized network security agents could be a feasible one.[20] In one study, simulated agents were able to cooperate across at least three levels of defensibility improvements, from general protections like authentication and cryptography, to more proactive security assessment and monitoring, to (at the most sophisticated level) management and decision-making around which types of specialized agents and actions would provide the best defense against a given threat.[21] While proliferating agents surely also will accelerate hackers' offensive efforts to find vulnerabilities to exploit, prior to just a few short months ago, the energy and computing requirements for a multi-agent defense would have been assumed to be cost prohibitive for all but the most well-resourced enterprises.

SETTING RULES OF THE ROAD TO DELIVER ON THE PROMISE OF A MORE DYNAMIC AGENT ECOSYSTEM

The potential for these kinds of multi-agent systems to benefit cybersecurity, accuracy, and reliability, especially as they become cheaper and more accessible to a wider customer base, will depend on their working effectively together.

In many sensitive applications, AI model security and reliability remain too unpredictable for comfort, a concern that will only compound in complexity as AI agents proliferate and interact. The aforementioned Agent2Agent and Model Context Protocol are important first steps to structuring these interactions, but how multiple agents should collaborate, much less negotiate competing priorities or securely pass sensitive information, remains an intricate and cutting-edge area of research.[22] What degree of autonomy do these agents have, and what form of identification, certification, and authorization should we expect from them when they are acting on behalf of human representatives or corporations? How should these expectations tighten—or loosen—when faced with urgent contingencies?

Questions like these are poised to define the coming era of agentic artificial intelligence, especially as models become cheaper, more accessible, and deployed in multiples—and in March 2025, these

questions were presented as key concerns for the field at the Association for the Advancement of Artificial Intelligence's Presidential Panel on the Future of AI Research.[23] Satisfactory answers will require nuanced collaboration across divergent legal jurisdictions, diverse hardware ecosystems, and complex data governance regimes. High-Flyer's decision to release DeepSeek as an open-weights model spotlighted the promise of more open software development paradigms to navigate this complexity, as well those paradigms' limitations.

Open-weights distribution (even if not, as discussed earlier, entirely open-*source* distribution) enables the local download of a model separate from the company or project that initially produced it. This separation makes more data security considerations possible than would be the case for a purely cloud-based AI service, because all the model's processing of information, inference, and responses to user prompts is conducted locally on the user's (or their organization's) own computing equipment. Open distribution also allows for the customizability of a local model by its user, whose direct control over the model file allows them to more precisely specify the system's behavior for its circumstances or environment. For many sensitive or classified settings, organizations will likely find security advantages in running locally-executable models in certain use cases and frontier cloud-based models in others.

But realizing this vision of a more open and diversified agentic ecosystem could easily require so much specialized knowledge and equipment as to become prohibitive for all but the most sophisticated and well-resourced institutions—threatening to void the democratizing potential of post-DeepSeek efficiency gains. Industry, civil society, and government will need to collaboratively define the rules of the road for this coming era, from security and interoperability protocols, to standardized commodity hardware for openly distributed models, to (perhaps most importantly) striking the right balance of open and proprietary development philosophies to ensure this fast-evolving field remains an inclusive and dynamic one.

Navigating Openness and Control

In practice, like in so many other sectors, the optimal (but difficult) approach likely combines elements of both proprietary and open-source development. In the cybersecurity world, for example, open-source authentication and encryption protocols have proven remarkably effective not despite their transparency, but because of it. When code is openly available for inspection, vulnerabilities can be identified and patched by a global community of experts rather than relying solely on in-house teams.[24] Many technology companies recognize this dynamic and consider maintaining certain open-source tools to be a socially responsible civic duty, either by directly overseeing open projects or libraries, or by incentivizing their employees to contribute to others.

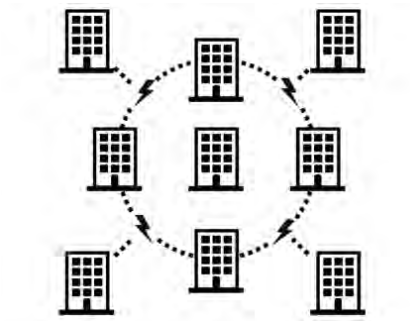
This collaborative approach benefits both the companies themselves and the broader ecosystem by establishing shared standards and distributing maintenance costs.

However, this model can break down when commercial incentives are not properly aligned. Critical open-source tools can languish without sufficient support—a classic "tragedy of the commons" where, even though everyone benefits from a resource, few are motivated to maintain it.

During Microsoft Build 2025, Satya Nadella announced support for competitor-developed interoperability protocols like A2A and MCP.

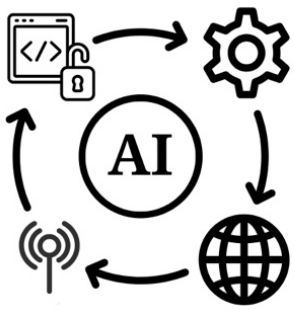
The mixed state of today's artificial intelligence marketplace, young as it may be, demonstrates why the government should partner with industry to ensure that the foundations being laid for our AI-driven future include incentives for **openness, tailorability, and interoperability.** Outside of Meta's suite of Llama models,[25] most leading AI developers paywall or meter access to their best products—which may have contributed to the shock and subsequent tech sector stock sell-offs when it became more widely understood that DeepSeek, an open-weights and locally-executable model, rivaled many of the best American models of the time regardless of whether they were local or cloud-based. Encouragingly, OpenAI subsequently announced a push toward more open-model offerings—perhaps envisioned as lightweight complements to their cutting-edge products, similar to Google's locally-executable Gemma line derived from its larger cloud-based Gemini model.[26] Policymakers should consider what R&D priorities, standards processes, or acquisition policies could ensure that these more open "complements" have the chance to develop into market-wide assets. While research institutions across the country are beginning to examine applications for cheaper, locally-executable, or proliferated agentic systems across the civilian and defense sectors, a broader ecosystem of collaborative development will need national leadership from both government and industry to succeed.

CONSIDERATIONS FOR POLICYMAKERS



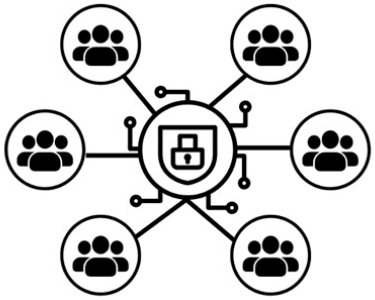
Deploy Energy Infrastructure as Distributed as the Coming Agentic Era

As policymakers continue working to meet historic increases in demand for electricity, they should plan for the many medium-sized, latency-sensitive data centers that will likely dominate an inference-driven infrastructure landscape. This means looking beyond headline-grabbing quick fixes, like "Stargate"-style mega parks or Middle Eastern data center deals, and focusing on scalable policy support for widespread generation, transmission, and storage so that there is abundant electricity closer to where people live and work (and use AI).



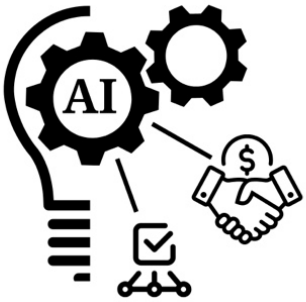
Bolster R&D and the Open-Source AI Ecosystem

Policymakers should prioritize resources and support for research and development initiatives aimed at enhancing the security, reliability, and functionality of open-weights AI models. Establishing collaborative research hubs or public-private partnerships could facilitate the creation and maintenance of tools specifically designed for penetration testing, vulnerability assessment, and continuous validation of open-weights models, as well as experimentation in "truly" open-source ones. Policymakers could further incentivize the participation of private-sector entities in such initiatives through recognition programs, regulatory incentives, or targeted financial support, thereby maintaining an innovative and secure open-weights and open-source AI landscape.



Facilitate Interoperability and Security Standards for Multi-Agent Systems

To support the effective implementation of multi-agent systems, policymakers should encourage the development of industry standards for interoperability, security, and certification. Clear technical standards and best practices can reduce barriers to integration and increase the reliability and security of these systems. Federal procurement of AI systems should support agent interoperability by promoting efforts like the Agent2Agent and Model Context Protocols. This approach could mitigate risks associated with agent malfunctions or mis-prioritizations, enhance public trust, and encourage broader adoption across sensitive sectors such as healthcare, critical infrastructure, and national defense.



Encourage Business Model Innovation for Local AI Systems

Policymakers should explore incentives and regulatory frameworks that encourage the development and adoption of decentralized AI systems, including via locally-executable models and associated computing hardware. Such models can enhance resilience, improve data security, reduce network congestion, and distribute energy intensity. Pilot programs and innovation grants for enterprises pioneering these technologies could help mitigate initial adoption risks and demonstrate the practicality and value of distributed systems, while verifying their potential to support national infrastructure, reduce vulnerabilities inherent in centralized data centers, and democratize access to advanced AI capabilities across sectors and regions.

KEY TERMS

Recent advances in artificial intelligence have filled the lexicon with terminology and buzzwords, including many that sound similar while making important technical distinctions. We have included a short primer in terms that are used in or relevant to this document.

Reinforcement Learning

A traditional form of machine learning that, put simply, involves training a model through trial and error. By establishing a set of rewards or penalties, a developer can provide feedback to the model as it aims to complete a task. The rewards or penalties guide the model toward optimizing for the goal it is trying to accomplish. A common example is when guiding an agent through a maze with a model, every obstacle is a penalty providing feedback that the model made an incorrect decision, while every step of progress toward the end is rewarded with points for the model; over time, the model can be optimized to guide the agent through the maze in under certain specified success criteria.[27]

Edge Computing

In a world filled with devices, sensors, and "smart" technology, a mountain of data is being collected every moment. All this data would clog up network traffic if it was constantly being sent back to one centralized data bank for processing. Therefore, computing at the "edges" of the network has become increasingly desirable, where a device is able to process the bulk of the data locally and respond accordingly with limited need for massive data transfers to a centralized or cloud-based repository[28]. By limiting the data being sent from these devices to only the most important items, or on certain limited time intervals, individuals and organizations can improve their response time and overall efficiency[29]. As AI models become more complex and compute-intensive, moving more of their work to the edge is likely to be an increasingly appealing tool to developers and users.

Agent

The term for a computer program with a degree of autonomy, meaning it can perceive its environment, make decisions, and act without constant human input. While the term isn't new, its meaning has evolved alongside technology. Early examples like ELIZA in the 1960s demonstrated scripted conversation but lacked greater autonomy.[30] Over time, agents grew more sophisticated, capable of planning and adapting in dynamic contexts. Today, LLM-powered agents built on foundation models represent a major shift: They can understand language, reason, and interact with tools to perform complex tasks. This marks a new era in

which agents are flexible, general-purpose systems with the possibility to be specialized and customized for user needs.

Foundation Models

These are models that are trained on a massive amount of raw data without any specifically defined end tasks. They are also called general-purpose AI models. These foundation models are then built upon to create agents or other models with more specific goals, tasks, or capabilities (OpenAI's GPT-4 is a foundation model; the current version of ChatGPT is an agent built on top of the foundation model.) Unlike the agents that are more obviously present in the products and tools people use, the foundation models are generalized. This makes it difficult to identify or measure their direct benefits and it is much more difficult to predict their potential harms.[31]

Multimodal

Refers to a type of AI that can integrate and process multiple types of data (text, images, audio, video, etc.). These added dimensions make the models more resistant to missing data, because the model can look to other sources of data if the primary source is insufficient. However, these models come at a cost of needing even more data for accurate processing and may open themselves up to new types of adversarial attacks.[32]

Mixture-of-Experts (MoE)

Refers to a model architecture that combines expert sub-models with a gating mechanism. The gating mechanism takes the input and decides which sub-model(s) to send the input. The core benefit of this architecture is efficient computational resource management without sacrificing performance. Other benefits include scalability and performance increases: New expert models can be added, and each expert is specifically trained for certain data patterns. One major challenge with this architecture is that all the expert sub-models need to be stored in memory, which can be taxing for a system. Another challenge is that the models might overfit to the data if the experts are too specialized.[33]

Model Distillation

This is a machine learning process where the knowledge and information from a large model is transferred to a smaller model for more efficient deployment into production. Unlike traditional approaches, where foundational models can be built upon for more targeted AI tools and agents, distillation involves condensing these larger agents into smaller, more efficient models that can be easily deployed on a wider variety of machines that might not have the same computing capabilities.[34] There is a risk with model distillation: It does not always perfectly capture the capabilities from the teacher model, and a loss of capability is something that needs to be tested for and addressed when it happens.

Multi-Agent System

This system uses a combination of individual complex agents working collaboratively to achieve a large-scale task for a user. An agent is an AI-informed system that is able not only to make predictions and decisions, but also to act upon the processed information. There are various proposed structures to these systems, depending on the use case, and yet there are few demonstrable use cases (mostly in finance and resource management). There are many challenges in bringing MAS out of research and development and into everyday use, the first of which is making individual agents usable and efficient at scale. Multi-agent systems are also susceptible to agent malfunctions, coordination challenges, unpredictable behaviors, and difficulty with authorizations and certifications.[35]

Open Source

Originated as a way to create software where the source code was freely available for inspection, modification, and enhancement. Over time, it has become more of a philosophy of innovation built on principles of transparency and collaborative improvements.[36] In the context of generative AI, *open source* refers to the release of all aspects of the AI model (weights, training code, training dataset, and data composition). Because this information would give anyone the ability to fully build their own model, most open-source models are released by independent labs and researchers.[37]

Open Weights

A phrase often mistaken for open source. In this case, only the weights of the model are released so that others can download a version of the model and integrate it into their own workflow.[38] These models offer a wider variety of researchers and AI developers access to models without the need to spend time and resources training and establishing parameters.[39] The models work and can be experimented upon and fine-tuned for specific tasks. DeepSeek and Meta both released their models as open-weights, increasing their accessibility without releasing proprietary training data or training code.

[1] "AI Inference vs. Training: What Is AI Inference?," Cloudflare, accessed June 11, 2025, https://www.cloudflare.com/learning/ai/inference-vs-training/.
[2] The authors recognize that "confabulation" is preferred by some researchers to more precisely describe this phenomenon, but employ the term "hallucination" in this document to reflect popular usage.
[3] Kobi Hackenburg, et al., "Scaling Language Model Size Yields Diminishing Returns for Single-Message Political Persuasion," *Proceedings of the National Academy of Sciences* 122, no. 10 (March 7, 2025): e2413443122, https://doi.org/10.1073/pnas.2413443122; Maxwell Zeff, "Current AI Scaling Laws Are Showing Diminishing Returns, Forcing AI Labs to Change Course," TechCrunch, August 20, 2024, https://techcrunch.com/2024/11/20/ai-scaling-laws-are-showing-diminishing-returns-forcing-ai-labs-to-change-course/.
[4] Eduardo Baptista, "High-Flyer, the AI Quant Fund behind China's DeepSeek," Reuters, January 29, 2025, https://www.reuters.com/technology/artificial-intelligence/high-flyer-ai-quant-fund-behind-chinas-deepseek-2025-01-29/.
[5] Christopher Manning, "Explaining DeepSeek and Its Implications," https://www.aixventures.com/explaining-deepseek-and-its-implications-with-chris-manning.
[6] Clement Farabet and Tris Warkentin, "Introducing Gemma 3: The Most Capable Model You Can Run on a Single GPU or TPU," Google, *The Keyword* (blog), March 12, 2025, https://blog.google/technology/developers/gemma-3/.
[7] "OpenAI Plans to Release Open-Weight Language Model in Coming Months," Reuters, April 1, 2025, https://www.reuters.com/technology/artificial-intelligence/openai-plans-release-open-weight-language-model-coming-months-2025-03-31/.
[8] "Introduction," Model Context Protocol, accessed June 11, 2025, https://modelcontextprotocol.io/introduction; "Introducing the Model Context Protocol," Anthropic, November 25, 2024, https://www.anthropic.com/news/model-context-protocol.
[9] Kyle Wiggers, "Microsoft Adopts Google's Standard for Linking up AI Agents," TechCrunch, May 7, 2025, https://techcrunch.com/2025/05/07/microsoft-adopts-googles-standard-for-linking-up-ai-agents/; Rao Surapaneni, et al., "Announcing the Agent2Agent Protocol (A2A)," *Google for Developers* (blog), April 9, 2025, https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/.
[10] Diego Gosmar and Deborah A. Dahl, "Hallucination Mitigation Using Agentic AI Natural Language-Based Frameworks" (arXiv, January 27, 2025), https://doi.org/10.48550/arXiv.2501.13946; Daniel Schwarcz, et al., "AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice" (Social Science Research Network, March 4, 2025), https://doi.org/10.2139/ssrn.5162111.
[11] James O'Donnell and Casey Crownhart, "We Did the Math on AI's Energy Footprint. Here's the Story You Haven't Heard.," MIT Technology Review, May 20, 2025, https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/.
[12] DeepSeek-AI, et al., "DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model" (arXiv, June 19, 2024), https://doi.org/10.48550/arXiv.2405.04434; Ben Thompson, "DeepSeek FAQ," *Stratechery* (blog), January 27, 2025, https://stratechery.com/2025/deepseek-faq/.
[13] Yina Arenas, "Announcing the Availability of the O3-Mini Reasoning Model in Microsoft Azure OpenAI Service," *Microsoft Azure Blog*(blog), January 31, 2025, https://azure.microsoft.com/en-us/blog/announcing-the-availability-of-the-o3-mini-reasoning-model-in-microsoft-azure-openai-service/.
[14] "What Are Azure AI Containers?," Microsoft Learn, March 31, 2025, https://learn.microsoft.com/en-us/azure/ai-services/cognitive-services-container-support.
[15] "AI Recipes from the Dell AI Kitchen" (Dell Technologies), accessed June 11, 2025, https://www.delltechnologies.com/asset/en-us/solutions/business-solutions/briefs-summaries/ai-cookbook-recipes-how-to-run-ai-anywhere-without-internet-connection-brochure.pdf; "NVIDIA Jetson AGX Orin," NVIDIA, accessed June 11, 2025, https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/.
[16] Gosmar and Dahl, "Hallucination Mitigation Using Agentic AI Natural Language-Based Frameworks"; Hong Qing Yu and Frank McQuade, "RAG-KG-IL: A Multi-Agent Hybrid Framework for Reducing Hallucinations and Enhancing LLM Reasoning through RAG and Incremental Knowledge Graph Learning Integration" (arXiv, March 14, 2025), https://doi.org/10.48550/arXiv.2503.13514.
[17] Gosmar and Dahl, "Hallucination Mitigation Using Agentic AI Natural Language-Based Frameworks."
[18] Yu and McQuade, "RAG-KG-IL."
[19] "How we built our multi-agent research system," Engineering at Anthropic, June 13, 2025, https://www.anthropic.com/engineering/built-multi-agent-research-system

[20] Igor Kotenko, "Multi-Agent Modelling and Simulation of Cyber-Attacks and Cyber-Defense for Homeland Security," in *2007 4th IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications* (Dortmund, Germany: IEEE, 2007), 614–19, https://doi.org/10.1109/IDAACS.2007.4488494.
[21] Kotenko.
[22] Ksenija Mandic and Boris Delibašić, "Application Of Multi-Agent Systems In Supply Chain Management," *Management - Journal for Theory and Practice of Management* 17, no. 63 (June 1, 2012): 75–84, https://doi.org/10.7595/management.fon.2012.0014.
[23] "AAAI 2025 Presidential Panel on the Future of AI Research," Association for the Advancement of Artificial Intelligence, March 2025, https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report-Digital-3.7.25.pdf
[24] "Case Studies," *Open Source Security Foundation* (blog), accessed June 11, 2025, https://openssf.org/case-studies/. https://openssf.org/case-studies/
[25] "Llama Models," Meta, accessed June 11, 2025, https://www.llama.com/docs/model-cards-and-prompt-formats/.
[26] Will Knight, "Sam Altman Says OpenAI Will Release an 'Open Weight' AI Model This Summer," *Wired*, March 31, 2025, https://www.wired.com/story/openai-sam-altman-announce-open-source-model/; Farabet and Warkentin, "Introducing Gemma 3."
[27] "Reinforcement Learning," GeeksforGeeks, February 24, 2025, https://www.geeksforgeeks.org/what-is-reinforcement-learning/.
[28] "What Is Edge Computing?," Microsoft Azure, accessed June 11, 2025, https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-edge-computing.
[29] "What Is Edge Computing?," IBM, June 20, 2023, https://www.ibm.com/think/topics/edge-computing.
[30] "Charting the Evolution and Future of Conversational Agents: A Research Agenda Along Five Waves and New Frontiers," Schöbel, S., Schmitt, A., Benner, D., et al., Inf Syst Front 26, 729–754, 2024, https://doi.org/10.1007/s10796-023-10375-9
[31] Elliot Jones, "What Is a Foundation Model?," Ada Lovelace Institute, July 17, 2023, https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.
[32] Cole Stryker, "What Is Multimodal AI?," IBM, July 15, 2024, https://www.ibm.com/think/topics/multimodal-ai.
[33] "Exploring DeepSeek-R1's Mixture-of-Experts Model Architecture," Modular, accessed June 11, 2025, https://www.modular.com/ai-resources/exploring-deepseek-r1-s-mixture-of-experts-model-architecture.
[34] "What Is Model Distillation?," Labelbox, accessed June 11, 2025, https://labelbox-guides.ghost.io/model-distillation/.
[35] Anna Gutowska, "What Is a Multiagent System?," IBM, accessed June 11, 2025, https://www.ibm.com/think/topics/multiagent-system.
[36] "What Is Open Source?," Opensource.com, accessed June 11, 2025, https://opensource.com/resources/what-open-source.
[37] Nathan-Ross Adams, "Openness in Language Models: Open Source, Open Weights & Restricted Weights," ITLawCo, August 8, 2024, https://itlawco.com/openness-in-language-models-open-source-open-weights-restricted-weights/.
[38] Sunil Ramlochan, "Openness in Language Models: Open Source vs Open Weights vs Restricted Weights," Prompt Engineering & AI Institute, December 12, 2023, https://promptengineering.org/llm-open-source-vs-open-weights-vs-restricted-weights/.
[39] Opensource.org, "Open Weights: Not Quite What You've Been Told," Open Source Initiative, accessed June 11, 2025, https://opensource.org/ai/open-weights.

Fig 1: "Person holding web page with logo of artificial intelligence (AI) company DeepSeek on screen in front of logo. Focus on center of phone display." [Stock photo]. Shutterstock Standard Image License. 2025. https://www.shutterstock.com/image-photo/stuttgart-germany-03102025-person-holding-web-2606264485.
Fig 2: "An aerial view of the QTS Data center under construction in Phoenix, Arizona." [Stock photo]. Shutterstock Standard Image License. 2023. https://www.shutterstock.com/image-photo/phoenix-us-nov-09-2023-aerial-2402650457.
Fig 3: "Microsoft corporation logo and company CEO Satya Nadella in the background." [Stock photo]. Shutterstock Standard Image License. 2024. https://www.shutterstock.com/image-photo/microsoft-corporation-logo-company-ceo-satya-2556776997.