Issue Brief

AI System-to-Model Innovation

Transformations from the Shop Floor

Authors
Jonah Schiestle
Andrew Imbrie

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

July 2025

Executive Summary

Recent advances in artificial intelligence (AI), particularly around large language models (LLMs), have made OpenAI's Generative Pre-trained Transformers (GPTs), Google's Gemini, Anthropic's Claude, and Meta's Llama household names. The sequential releases of foundation models from a small number of big players tell a linear story of innovation, but history suggests a more complex narrative. Just as electricity in the Second Industrial Revolution would eventually spur more reliable and distributed power systems, foundation models are paving the way for what some researchers have termed "compound AI systems." These systems encompass a set of distinct components, including at least one AI model and other components that manipulate the data in ways not learned by the model.* The components may aggregate multiple calls to a model, insert context with retrievers, or extend capabilities with tool use. Compound AI systems improve performance on tasks from the base model and can enable the model to perform entirely new tasks.

Consider an autonomous vehicle. The vehicle itself is a complex system of systems, including power, propulsion, sensor systems, and more. The vehicle is not substitutable for a single AI model, but subsystems involve model and compound system design decisions. For example, some manufacturers do not include LiDAR (Light Detection and Ranging) in their self-driving vehicles, opting solely for cameras, while others use a sensor suite, including cameras, LiDAR, and radar. Processing these signals for a task, such as object detection, could take at least two approaches: Train separate networks for each sensor or sensor type and combine the predictions (a compound AI system approach) or train a single model that fuses all sensor inputs (an AI model approach). At the level of object detection, the model and compound system approaches can be direct substitutes. This interchangeability allows lessons from one approach to spur improvements in the other. New model advances could enable compound systems to take advantage of that capability; a compound system approach could add a new sensor network, which might inspire a more efficient single network data fusion model that shares representations between sensors for faster object detection.

System-to-model innovation is an emerging innovation pathway that has driven progress in several prominent areas over the last decade. *System-to-model* advances include DeepMind's incorporation of policy and value calculations in a single network with AlphaGo Zero; chain-of-thought prompting leading to OpenAI's o1 model; the OneGen single pass network and Cohere's Command R family models in retrieval-

* More than one component may be an AI model, learned separately.

augmented generation (RAG); and circuit breakers for AI safety training in language models. *System-to-model* innovations close the feedback loop between model and compound system advances, opening frontier AI breakthroughs to more diverse contributions beyond the top labs.

Instead of one *model-to-model* pathway for progress, recent trends highlight the dynamic interplay between AI model and compound system innovation, where progress along one pathway leads to breakthroughs in the other pathway.

Figure A: AI Model and System Innovation Matrix

1. Model-to-Model Innovation Advances in models to solve a specific problem inspire further model advances for that problem or the use of similar techniques for a different problem. Examples: ImageNet contest winners from 2012 to 2017; the path from early recurrent neural networks to transformers	**3. System-to-Model Innovation** Lessons from compound systems are adapted for model training or new model architectures to accomplish the system task without the additional components. Examples: AlphaGo to AlphaGo Zero; chain-of-thought to OpenAI's o1 model.
2. Model-to-System Innovation A new model with advanced general capabilities inspires compound systems around the model to solve new downstream tasks or improve existing downstream tasks. Examples: Language models to prompt engineering and RAG	**4. System-to-System Innovation** Limitations in compound system performance inspire adapted systems to improve performance under specific conditions. Examples: Vector RAG to Astute RAG and GraphRAG

Policies that foster AI progress are likely to benefit model and compound system developers, but today's debates have focused more on improving model capabilities. If recent *system-to-model* advances are indicative of future trends, then policymakers should also devise tailored solutions for compound system developers to encourage multiple bets on the future of AI. System-level innovations advance with the diffusion of AI and expand the base of contributors to leading-edge progress in the field. While the spread of general-purpose technologies across societies and economies is often more consequential than the innovations themselves, the diffusion of AI will not happen in a geopolitical vacuum.[1] Countries that can identify and harness system-level

innovations faster and more comprehensively will gain crucial economic and military advantages over competitors. This issue brief suggests a three-part framework to navigate the policy implications of *system-to-model* innovation:

Protect smaller companies as incubators of tacit knowledge that could seed future foundation model advances. Governments may need to consider subsidized services, such as safety testing and red team assessments at a system level, for smaller and medium-sized enterprises that lack the resources and expertise to remain compliant with the evolving policy landscape or withstand adversarial attacks. This process should include a detailed mapping of the companies, infrastructure, research organizations, data analytics firms, and technical and nontechnical talent base that comprise U.S. and allied compound AI system innovation networks to determine the actors involved and resources needed for protection.

Diffuse innovation outside of leading foundation model providers by providing models and inference compute access to develop new systems. The National Artificial Intelligence Research Resource (NAIRR) should expand programs that widen access to computational infrastructure not only for training AI models but also for inference. Such an approach would enable researchers using pre-trained foundation models to develop system-level innovations that drive progress in AI models and benefit from the new inference-time compute scaling paradigm. For example, the research community for prompt engineering is large and diverse and frequently produces new techniques that improve model performance on downstream tasks. Policymakers could facilitate the diffusion of AI by supporting work outside of the top research groups to improve future iterations of foundation models.

Anticipate future innovations by monitoring countries diffusing AI models into systems and by adapting tools and methodologies for horizon scanning, technology forecasting, and supply chain risk analysis. A *system-to-model* mindset could encourage a new generation of research and development agreements with U.S. allies and partners. Research communities promoting the diffusion of compound system innovations would also serve as early indicators of future capabilities in frontier models.

As with the rise of electricity in the Second Industrial Revolution, the transformation from market competition to regulated monopolies in power generation represents one possible future for foundation models and their cloud infrastructure providers. Alternatively, foundation models may continue to become increasingly commoditized. How leaders in government, industry, and civil society negotiate complex tradeoffs in innovation, security, and values-based principles will have structural implications for the distribution of resources that make up the AI stack.[2] Industry leaders are already

implementing novel governance frameworks, while governments are pursuing myriad regulatory approaches.[3] Understanding innovation cycles will be critical as policymakers look to shape the trajectory of AI in directions that benefit the public good.

Table of Contents

Introduction

"I can light the entire lower part of New York City," Thomas Edison once mused as he probed the intricacies of electric light and envisioned its broader applications.[4] As one historian noted, what Edison had in mind was not just a light, but a "lighting system" that would transform cities into interlocking thoroughfares of electrified industry and commerce.[5] While Edison would need a bit more time to work through the practical implications of incandescent light, electricity would leave an indelible mark on the factories and industrial wares of the Second Industrial Revolution.

Fully realizing the benefits of electrification required managers and engineers to reimagine dominant modes of organization.[6] Factories had previously centralized their layouts around power sources, such as locating plants near coal mines or plentiful water repositories for steam engines. By contrast, the electric current allowed for more decentralized layouts, feeding machines buzzing with smaller motors located in positions around the factory floor that streamlined production processes. Factories were slow to adapt layouts and modernize assembly lines to harness this new form of power at scale. As manufacturers implemented organizational reforms, however, business cycles spurred improvements in electricity research. Growing demand for reliable power encouraged innovation across electricity generation, distribution, and transmission infrastructure; the adoption of electrified machinery in factories drove advances in new control systems, motor design, and automated technologies; the standardization of voltage and frequency enhanced interoperability and facilitated economies of scale. While innovation and new production methods coexisted alongside massive social upheaval and inequality, the electric lighting system would progress far beyond Edison's early forays into commercial power plants.[7]

Electricity was a marvel, but it also posed new challenges for safety and responsible use. Its seamless integration into our societies took decades, and the battles over direct and alternating currents were fierce.[8] Close observers of this historical period will detect echoes today in the disruptive and transformative potential of artificial intelligence. With the launch of ChatGPT in November 2022, consumers gained a window into advances that had been percolating in the field for years. Models have taken center stage, with OpenAI's GPT family, Google's Gemini, Anthropic's Claude, and Meta's Llama occupying roles that were previously afforded to the latest smartphone and operating system releases. Since then, a rich development ecosystem has matured around these foundation models, including efforts to fine-tune them on specialized data sets and advance retrieval-augmented generation (RAG) techniques for enterprise deployments—in essence, tailoring model outputs for specific users and applications by incorporating external data sources.

Building systems around models can enhance performance on particular tasks, but recent examples also indicate that systems can spur improvements in the models themselves. As with the Second Industrial Revolution, today's AI revolution demands a close look at the system effects of recent advances and their implications for policy. Foundation models are becoming increasingly commoditized as the costs per token decline.[9] While the hyperscalers backing the foundation model companies could end up becoming the new utilities, foundation model providers may also drive market structures with high barriers to entry, high capital costs, and low product differentiation. The industrial organization of AI today invites creative thinking about the different layers of the so-called "AI technology stack."[10] A central policy challenge is to ensure the safe and responsible design, development, and deployment of foundation models, while also providing support for the broader ecosystems around AI to innovate on top of that infrastructure. As with other general-purpose technologies, the diffusion of AI across societies and economies will take years, if not decades.[11] Harnessing the capabilities of system-level innovation may prove to be one of the most consequential pathways for spreading the benefits of powerful AI across societies while also anticipating and mitigating the risks.

Defining Compound AI Systems

To understand interrelationships in model- and system-level developments for AI, it is first necessary to delineate key concepts. In this framing, we focus on input and output structures and training processes in order to separate models from compound systems and illustrate how they influence each other's development. We present definitions for AI algorithms, models, and compound systems in Box 1.

Box 1: Definitions of Key Terms

AI algorithm: A clearly specified set of mathematical rules for creating an AI model.

- This definition for algorithms is adapted from a NIST glossary; however, we remove the second part of NIST's definition, namely, "...a set of rules that, if followed, will give a prescribed result."[12] Stochasticity, or randomness, is an important component of many AI algorithms and successive iterations may not produce the same, prescribed output.
- "AI algorithms" could describe the architecture of a model, such as neural networks, or the method for training a model, such as backpropagation.

AI model: A set of rules learned or developed that take data inputs of a specified form and produce data outputs of a specified form. Rules could take the form of learned parameters in machine learning models or crafted statements in expert systems.

- This AI model definition is broader than IBM's definition, which requires learning from data, thereby precluding AI models outside of the machine learning paradigm.[13]
- In the case of expert systems, the algorithm is the structure of an expert system or method of incorporating new knowledge while the model is the set of rules in the fully built expert system.

Compound AI system: A set of distinct components, including at least one AI model, which take data inputs of a specified form and produce data outputs of a specified form to complete a task. All components of the compound AI system either take inputs of a different form than the system or produce outputs of a different form than the system. Transformations that happen outside of the model component(s), such as combining or formatting data, are not learned within the training routine.

- This definition builds on the original concept from Berkeley Artificial Intelligence Research (BAIR), highlighting data structures and their place within model training.[14] Our added focus on data structures and training regimes helps clarify the boundary between compound AI systems and complex AI models, which use multiple architectural components but whose parameters are learned in a single training routine.
- Compound AI systems are a subset of what previous analysis from CSET has termed "AI systems," which encompass "all components of the overall system—including preprocessing software, physical sensors, logical rules, and hardware devices."[15] The definition of compound AI systems in this report focuses more on input and output structures, and less on hardware or other physical components. For example, in prompt engineering, an initial query is transformed from a basic user question to a form that elicits a stronger response from the language model. This transformation exists outside the model training routine, representing a second component in a compound system, and therefore makes prompt engineering a compound AI system deployment.
- As this definition suggests, a compound AI system could enhance the performance of a single model or multiple complementary models. For example, AlphaGo is a compound system that includes two separate neural networks, a policy network and a value network; GraphRAG is a compound system that combines a graph neural network for retrieval and a large language model for generation. As BAIR notes, compound AI system components include aggregation of repeated model calls, retrievers, and tool use.

Distinguishing between AI models and compound AI systems is important for understanding the complete picture of AI innovation. New compound AI systems can extend the applications of cutting-edge models, while new AI models can improve performance on existing tasks or reach performance thresholds that render novel problems more tractable. Focusing on inputs and outputs highlights the differences in efficiency and flexibility between AI models and compound AI systems. This approach also illuminates a pathway between the two: incorporating features of the compound AI system within the model training process. The distinction in approaches can be

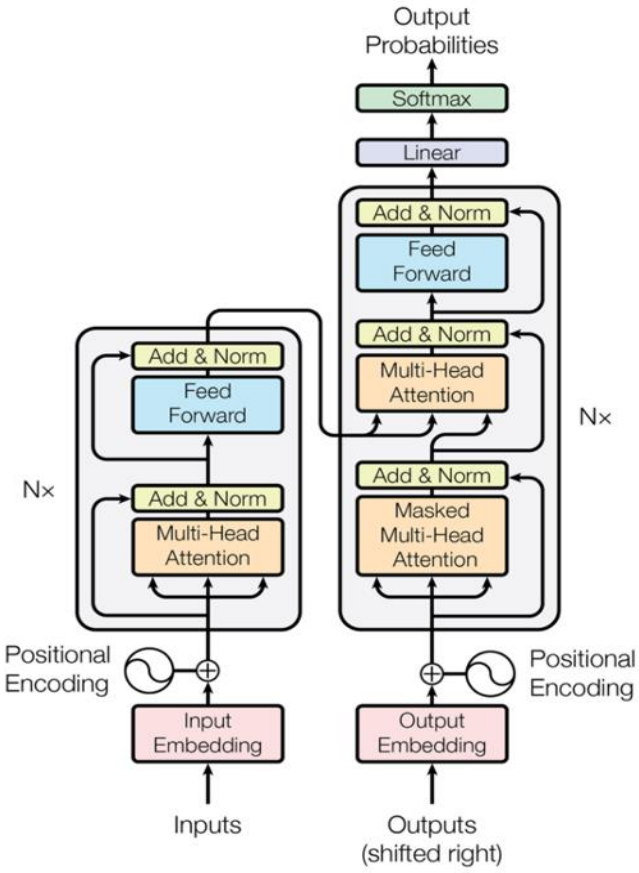
subtle but important; one must carefully consider the operations on data and their inclusion within the training routine.

The example of mixture-of-experts (MoE) models is instructive.[16] Such models match tasks with expert subcomponents to maximize performance. In MoE applications, not all subcomponents generate every prediction. Following the release of OpenAI's GPT-4, observers debated whether this LLM was a single model or a system of eight models.[17] By focusing on AI models and compound AI systems from an input-output perspective, however, the question becomes clearer: Whether an MoE model is a model or a compound AI system depends on the implementation of the gating components. Those components can be trained within the same model where only some parts of the model—the "experts"—are active during a single inference. In this case, the output is generated in a single pass through the trained network, but only one eighth of the network becomes active for a single example. By contrast, a gating network for an MoE can determine independently how to weight the inputs for, or outputs of, each expert and which of the eight separately trained models should be used for a given task. The upshot is that GPT-4 could incorporate MoE techniques and still be a single model, confirming both rumors and OpenAI's claims.

The distinction between AI models and compound AI systems matters for understanding the interplay between AI algorithms and models, on the one hand, and characteristics of performance, on the other. An AI model may use several algorithms in distinct components, but those components take and produce internal representations of the data learned together in a single training process. No external state exists as an input to a separately trained or non-AI component. By visualizing language models, we see they are in fact made up of many such components.

Figure 1, which shows the architecture of the original transformer model, demonstrates this property.[18] The model architecture includes encoding components, multi-headed attention components, and feedforward network components. But these components are connected without interruption by weights learned in a single training process. No intermediate outputs and inputs exist. Capturing tasks within a singular training process of an AI model may produce more efficient outcomes, while building compound systems around models may produce increased extensibility.

Figure 1: The Transformer Model Architecture



Source: Ashish Vaswani, Noam Shazeer, Niki Parmar et al., "Attention Is All You Need."

The construction of an AI model through many subcomponents also highlights an important distinction for assessing AI model improvements; namely, scaling-based performance improvements versus innovative new methods. Many competitions and leaderboards purport to evaluate model improvements based on performance, but performance advances could come from scaling or algorithmic improvements.[19] Scaling has been critical in recent advances, from GPT-4.1 to Google Gemini 2.5 to Claude 3.7 Sonnet. Researchers and companies can also drive massive performance gains by introducing new algorithms, such as the multi-head attention component of transformer architectures. When AlexNet kicked off the deep learning revolution, it took advantage of both model size and algorithmic progress to improve on the image classification error rates by substantial margins.[20]

Many take Richard S. Sutton's The Bitter Lesson to mean scale is all you need, but the more precise lesson in his words is worth underscoring: "general methods *that leverage computation* are ultimately the most effective" (emphasis added).[21]

Advancing from GPT-2 to GPT-3.5 required not only scaling compute but innovating new training methods; namely, fine-tuning and reinforcement learning with human feedback (RLHF).[22] Without fine-tuning, output could lock up in repetition or generate trivial and unhelpful completions. Without RLHF, GPT-3 would not have been aligned with certain human goals and language patterns. To paraphrase The Bitter Lesson more simply, training improvements unlock improved performance through scaling. Understanding where the innovations come from that produce training improvements is critical. Model innovations, compound system innovations, or their interplay can drive progress.

System and Model Innovation Matrix

The dynamic between AI models and compound AI systems that generates innovations and new progress occurs through four potential pathways. Those paths can be pictured as a 2x2 matrix where an innovation in either AI models or compound AI systems leads to a secondary innovation. This matrix and the hypothetical mechanisms to trigger the secondary innovation are in Figure 2.

Figure 2: AI Model and System Innovation Matrix

1. Model-to-Model Innovation Advances in models to solve a specific problem inspire further model advances for that problem or the use of similar techniques for a different problem. Examples: ImageNet contest winners from 2012 to 2017; the path from early recurrent neural networks to transformers	**3. System-to-Model Innovation** Lessons from compound systems are adapted for model training or new model architectures to accomplish the system task without the additional components. Examples: AlphaGo to AlphaGo Zero; chain-of-thought to OpenAI's o1 model
2. Model-to-System Innovation A new model with advanced general capabilities inspires compound systems around the model to solve new downstream tasks or improve existing downstream tasks. Examples: Language models to prompt engineering and RAG	**4. System-to-System Innovation** Limitations in compound system performance inspire adapted systems to improve performance under specific conditions. Examples: Vector RAG to Astute RAG and GraphRAG

First is the *model-to-model* pathway. This route lies at the heart of AI innovation and draws the most attention. *Model-to-model* innovation occurs when a new model kickstarts work to develop more powerful models. Developers could aim to build a model focused on improving performance on the same task, or apply a technique developed for one task to another. *Model-to-model* innovations can be tied to performance on a specific benchmark, such as the series of innovations in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) contest that drove the state-of-the-art in image classification. *Model-to-model* innovations can also spring from research on a concept that is applicable to a broad range of problems, such as the need

for memory mechanisms in neural networks that led from recurrent neural networks (RNNs) to transformers.*

Next are AI model innovations that lead to the creation of new AI systems. *Model-to-system* innovation occurs when models display new capabilities that could be useful in solving a particular problem or class of problems, but additional work is required to perform sufficiently well in a given task. Examples of this innovation pathway abound in recent applications of large language models (LLMs). LLMs displayed a new capability—in-context learning—which allowed for systems to be constructed around the foundation model. Prompt engineering generally, and chain-of-thought prompting in particular, represent one such system innovation. Beyond these examples, RAG is a fruitful system innovation that takes advantage of LLM's ability to learn in-context. RAG systems use an external database of knowledge to augment prompts with additional context. The systems enable users to retrieve information from this database, which is then appended to a prompt to allow LLMs to answer questions not in their training data.

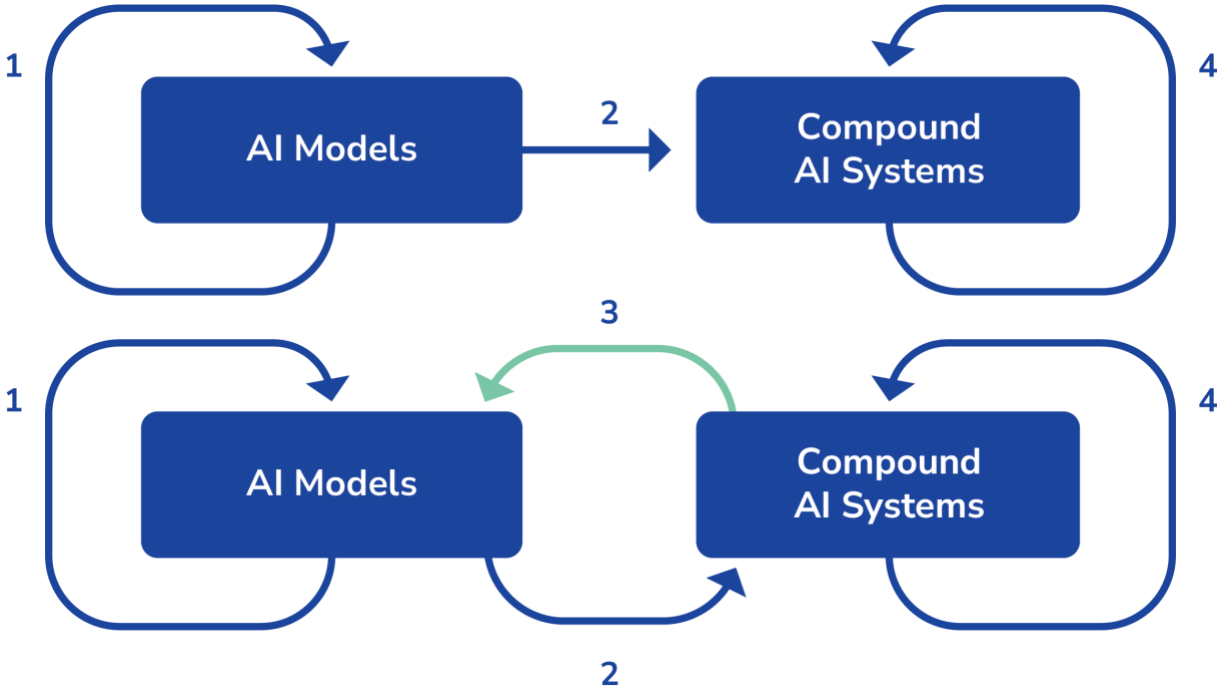
The third pathway is *system-to-system* innovation. *System-to-system* innovations occur when a new system demonstrates extended capabilities on the underlying model but areas for improvement remain in the original system. *System-to-system* innovations can improve performance on an existing system task or enable adaptation to new tasks previously too difficult for the original system. RAG systems provide numerous examples of the *system-to-system* innovation pathway. RAG inspired chunking and other processing improvements in retrieval databases as well as prompting systems to bolster performance and robustness. Innovations in subclasses of RAG, with component substitutions like GraphRAG that replace the retriever sub-system with a knowledge graph, are better at handling highly structured information.

The final pathway, *system-to-model* innovation, is less clear but potentially highly impactful. Can compound AI system innovations lead to AI model innovations? The importance of this pathway becomes clear when looking at the cycles formed in Figure 3. If the prior three pathways exist, then AI model innovations can be mutually reinforcing and lead to new compound AI systems, but system innovations display only self-reinforcing effects. This outcome would imply that direct work on AI models is the key to unlock further progress. But if there is a pathway from compound AI systems to AI model innovations, then there may be a self-reinforcing cycle between model and system advances. This cycle expands the scope of work potentially leading to AI model

* Details on these model-to-model innovations, as well as model-to-system and system-to-system innovations, are in the appendix.

improvements at the bleeding edge, including actors who may not otherwise be involved in such efforts.

Figure 3: AI Model and System Innovation Cycle



Source: Jonah Schiestle and Andrew Imbrie (designed by Jason Ly).

System-to-Model Innovations

As with the prior pathways, *system-to-model* innovation takes multiple forms. Compound AI systems can inspire new model architectures that integrate the system components into the core model, or novel training changes that encourage representations of system components within the model. They can take place within a single organization or between organizations of varying types. These *system-to-model* innovations have myriad implications for future AI innovation and reflect prominent developments in the field over the last decade. We highlight several examples that showcase how *system-to-model* improvements have produced strides in efficient learning in gameplay, accessing Type 2 reasoning in language models, tuning foundation models for specific systems, and training models for safety.

AlphaGo to AlphaGo Zero

The first such example is DeepMind's work in reinforcement learning for gameplay. In 2016, DeepMind announced AlphaGo, a reinforcement learning approach to Go that famously defeated world champion Lee Sedol just two months later, and many years before experts anticipated such a feat.[23] DeepMind achieved this landmark development by creating a novel compound AI system containing many AI model innovations. In games of perfect information like Go, it is possible to define a value for any given board position that corresponds to a player's likelihood of winning that game. Alternatively, one can create policies that take a position and propose a next move and then train to emulate expert human play. With a game the size of Go, these AI models would be combined with Monte Carlo Tree Search (MCTS), a method for looking ahead at certain possible future moves to determine the best current move, as it would be impossible to use value or policy networks to evaluate all possible moves. DeepMind adapted these model innovations, along with advances in reinforcement learning and convolutional neural networks, to create a system with both value and policy networks. The two-network compound AI system used the value network to evaluate positions and the policy network to sample actions. Finally, the networks leveraged MCTS to create a system that could predict the most promising future moves, thereby playing Go at a superhuman level.

AlphaGo contained a limitation, however. One of the key advantages of reinforcement learning is the potential to reduce or eliminate burdensome requirements for human-labeled training data. AlphaGo did not eliminate the need for training data entirely, beginning the policy and value network training with a supervised learning phase using human expert play data. But a long-standing goal is to eliminate this phase entirely, learning *tabula rasa* from self-play. In late 2017, DeepMind achieved

superhuman play with the blank slate model AlphaGo Zero.[24] AlphaGo Zero was so good it never faced human play, beating AlphaGo 100 to 0. AlphaGo Zero was also far more efficient than its predecessors, surpassing their performance after just 72 hours of training using 4 TPUs on a single machine; by comparison, AlphaGo had trained over several months and used 48 TPUs distributed across machines. One of the enabling advancements in AlphaGo Zero was a *system-to-model* architectural innovation. In AlphaGo Zero, the policy and value networks were part of a single deep neural network that provided both the move probabilities and board value. By forming a single network, AlphaGo Zero was able to learn more efficient representations that helped predict both policy and value. The new architecture combined the policy and value networks, thereby collapsing a two-network compound AI system into a single model with improved performance.

AlphaGo Zero has spurred numerous advancements. In gameplay, DeepMind researchers developed the more general AlphaZero within months to achieve *tabula rasa* superhuman play on Go, chess, and shogi.[25] Additionally, DeepMind adapted the system to beat StarCraft in a multi-agent system with AlphaStar and to conquer Atari games using a learned model for rules with MuZero.[26] The innovations from AlphaZero extended beyond gameplay to hard optimization problems in quantum dynamics.[27]

Chain-of-Thought to Reasoning Models

While the AlphaGo Zero *system-to-model* example shows how a compound AI system can be incorporated in a singular, efficient model through architectural changes, the recent release of OpenAI's o1 demonstrates the impact of *system-to-model* innovation on model training.[28] OpenAI incorporated the system-level innovation of chain-of-thought reasoning into their model training process through reinforcement learning to create a model that would elicit step-by-step thinking without the need for prompt engineering. Training CoT behavior into the model by default eliminates the need for the human or templated component of the system to translate a question into a form that induces step-wise reasoning.

In the announcement of o1, OpenAI made several big claims about the performance of the model. The new model family beat out the company's previously best-in-class models in a number of technically rigorous benchmarks, including competition math and coding and PhD-level science questions, despite being at least an order of magnitude smaller than GPT-4. But perhaps the most interesting claim from OpenAI was a new scaling law: test-time (or inference-time) compute. The o1 model uses so-called "time to think" as it generates responses. This hidden chain-of-thought uses additional inference-time compute to produce content useful for in-context learning.

Similar to the observed relationship in train-time compute, OpenAI claimed that as the time allowed for this reasoning increased, there was a similar increase in performance.

At the TEDAI conference, Noam Brown, a research scientist at OpenAI, spoke extensively about this new paradigm for scaling performance.[29] Brown likened the development to Daniel Kahneman's Type 2 thinking, claiming that the memory stored in the model was the Type 1 fast thinking and o1 now opened the door to improved performance through the use of slower Type 2 thinking.[30] Demonstrating the potential of the new paradigm, Brown suggested the need to go from $1 billion to $10 billion in training costs could instead be replaced with a change from $0.01 to $0.15 in inference costs to achieve the same performance on a single query. While this example discounts the fact that the large volume of requests such a model sees may eventually erode any cost advantage saved at training time, it does provide an additional lever to tune performance.

In November 2024, Chinese AI startup DeepSeek announced a new model, R1, which also uses the inference-time compute paradigm and exceeded o1's performance on two mathematical benchmarks.[31] The diffusion of this *system-to-model* innovation to DeepSeek is notable, as it was the first open weight inference-time compute model to be announced. DeepSeek also demonstrates the ability to achieve performance gains without the level of resourcing available to top U.S. AI labs. Former OpenAI policy researcher Miles Brundage commented that the announcement presages "the early stages of a [less compute-intensive] new paradigm."[32] The release of R1 also shows how deferring the cost of frontier-level performance to inference time may allow a larger number of labs to produce threshold-breaking models. A week later, Alibaba released their first reasoning model, QwQ-32B-Preview, becoming the first model available for download in the new paradigm.[33]

In December 2024, Google and OpenAI both announced new reasoning models, demonstrating continuing investment and focus in this *system-to-model* innovation.[34] Both models achieved impressive results on benchmarks, with Gemini 2.0 Flash Thinking Mode achieving a high score on SWE-Bench and o3 achieving the top score on ARC-AGI-PUB and a "step-function increase in AI capabilities," according to François Chollet, co-founder of ARC Prize.[35] The inference-time paradigm is driving even greater optimism among industry players, with OpenAI CEO Sam Altman claiming, "We are now confident we know how to build AGI as we have traditionally understood it."[36]

The new inference-time scaling law could drive similar *system-to-model* innovations in the near future. Reinforcement learning could integrate any of a number of additional

prompt engineering systems into a model in a way that similarly improves model performance through extended inference time. For instance, re-reading has shown additive improvements to chain-of-thought systems, where in addition to reasoning step-by-step the model processes the same question more than once in the prompt.[37] Such a system could be similarly integrated into future versions of the o1 model by adding repetition of the question to the hidden thought through reinforcement learning.

RAG to OneGen

Returning to RAG, another previously described system-level innovation, a recent paper from researchers with Ant Group describes a new architectural advancement for RAG, OneGen.[38] This innovation combines the retriever and generator components of RAG into a single model. OneGen eliminates the need for two different representations of a query—one created by the retriever and the other by the generator. This novel framework instead uses the internal representation of the query within the language model itself to retrieve documents from the knowledge store. In this system, the representations of documents in the knowledge store are also generated by the OneGen model as they are added. The combination of retrieval and generation into a single representation is more efficient and showed competitive performance on a variety of benchmarks for RAG tasks.

While the OneGen *system-to-model* innovation is relatively new and evidence of widespread adoption does not exist yet, OneGen demonstrates the ability of outside groups to contribute to architectural innovations, even for large foundation models. Given the integration of RAG into many of the flagship offerings, such as NotebookLM from Google, or as part of the core product, such as Perplexity or Glean, adoption of new *system-to-model* innovations could quickly spread across the sector. Consider other *system-to-system* RAG innovations: OneGen is designed for vector RAG, but the need for representations similarly applies to GraphRAG systems. Graph neural networks generate embeddings for retrieval from a graph, similar to the process of retrieval from a vector database. GRAG illustrates how a language model can produce embeddings for the retriever component.* Such a retriever could be combined with the

* GRAG is an acronym for Graph Retrieval-Augmented Generation, more traditionally stylized GraphRAG. We use GRAG to refer to the specific implementation cited, which embeds subgraphs for fast retrieval, as opposed to the broader class of GraphRAG. For more information on the original source, please see: Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, Liang Zhao, "GRAG: Graph Retrieval-Augmented Generation," arXiv preprint arXiv:2405.16506 (2024), https://arxiv.org/abs/2405.16506.

OneGen framework to produce a version where the internal representation retrieves relevant data from a graph database instead of a vector store.

While OneGen represents a system-to-model innovation through architectural changes, RAG systems have also inspired training changes. Cohere launched the Command R family of large language models expressly for use in RAG systems. Command R was optimized for RAG with a mixture of supervised and preference fine-tuning.[39] Combined with Cohere's models for the retrieval phase, Embed and Rerank, these models represent a new class of purpose-tuned foundation models. With the subsequent introduction of the larger Command R+ model for more complex and longer context tasks, the model family more closely represents the practice of releasing tiered models for efficient use where smaller models are appropriate. Command R also comes with prompt templates to maximize performance on RAG systems and trigger features like citations in model responses.[40] *System-to-model* innovations like Command R could drive new markets for competition in purpose-focused foundation models. Models tuned for performance in specific compound systems could outcompete larger and more powerful models designed as general purpose.

Circuits to Short Circuits

A final example of *system-to-model* innovation comes from AI safety research on large language models. One area of focus in AI safety research has been the application of circuits research, inspired by prior work in computer vision, to transformer language models.[41] Circuits research looks at the interaction of small subsets of neurons that form circuits and uses these groupings as a level of analysis to describe overall network behavior. Anthropic has focused heavily on this research thread, beginning with their finding on induction heads.[42] Induction heads check the prompt context for previous usage of a token and the token that comes immediately after it. If they find relevant tokens, they predict the next one by suggesting the same pattern; if not, then the circuit leaves the output unchanged.

Researchers at the Center for AI Safety took a different approach by focusing instead on internal representations as the unit of analysis. Calling the approach representation engineering (RepE), the researchers presented unsupervised methods for "reading" and "control."[43] Reading finds the internal representations of high-level concepts in a model, while control modifies those concepts to impact outputs. RepE methods enabled the creation of control vectors, or vectors tied to a representation. For example, the vectors for happiness or honesty could be located and then added or subtracted from model weights at inference time to impact the network's outputs. By

using control vectors, developers can create compound systems that take inputs and conditionally apply control vectors to bias outputs for a desired effect. The control vectors are learned separately from the trained language model. They serve as one component of the compound AI system, and are applied to the language model component at inference time to promote different behaviors than the model exhibits on its own. Control vector systems can be applied statically, meaning they always bias the model in the same way, or dynamically, applying one or more control vectors based on user input.

Building on their RepE work and spinning out a company, Gray Swan, the researchers connected representations and circuits to propose circuit breakers for better aligned models.[44] Circuit breakers apply the system-level innovation of control vectors to the training process of model fine-tuning. Circuit breakers work by locating representations of harmful behavior and adding short circuits to the network to impede the flow to a harmful output generation. The circuit breaker method proved more effective and robust than other prominent safety methods, such as refusal training and adversarial training. Refusal training detects harmful representations and uses that detection to induce a refusal response, thereby leaving the harmful state accessible within the model and vulnerable to jailbreaking. Adversarial training looks to fix these bypasses by training the model to counter attacks, but this technique does not generalize, as it focuses on the attack method and not the harmful representation. Moving from control vectors to circuit breakers demonstrates how studying a single problem in a single model can lead first to a compound system solution that is subsequently integrated into the base model training.

Circuit breakers showed an ability to counter general attacks without producing harmful content, but in safety work this fix often means the model becomes less helpful overall. In the case of circuit breakers, however, the researchers found that a significant reduction in harmfulness corresponded only to a relatively small decrease in helpfulness. Circuit breakers demonstrate a *system-to-model* innovation driven by work in nonprofits and academia, and eventually resulting in a new start-up competing at the frontier of AI safety.

Why System-Level Innovation Matters

These examples underscore the role of *system-to-model* innovation and their impact on key developments in AI, but they leave unanswered a central question: Why should policymakers foster advances at the compound system level? If driving innovation is the goal of AI policy, then it would seem to matter little which pathway the innovation takes. After all, *model-to-model* innovations may be just as likely to drive efficiency and performance gains as *system-to-model* innovations. EfficientNet, the final model to win the ImageNet competition, required 44 times less compute than AlexNet while saturating the ILSVRC benchmarks a mere seven years later.[45]

The properties of system-level innovations offer unique technical advantages. As BAIR researchers note, compound AI systems are easier to update and improve through system design without retraining; add dynamism through connections to tooling and sources; enable increased control and trust through structural components; and extend cost-quality tradeoff options when allowing varying levels of model calls.[46] Moreover, as the *system-to-model* pathway demonstrates, compound systems may inspire new model architectures or training routines that capture these gains.

System-to-model innovation can also shape geopolitical competition in at least two ways. The first is through direct introduction of new AI technologies to the warfighter, combined with novel operational concepts and organizational reforms so impactful they represent a revolution in military affairs.[47] The second is through increased economic growth as AI generates broad productivity gains across the economy, which in turn can enhance long-run national military power.[48]

To generate new military technologies and operational concepts, from drone swarms to dynamic military planning tools, companies across the defense industrial base will have to integrate models from leading AI labs into their platforms.[49] Where model capabilities on their own are insufficient, system-level innovations may fill the breach. For example, military planning tools must be reactive to changing circumstances and will likely require RAG compound AI systems until online learning in language models becomes feasible. As *system-to-model* innovation shows, this pathway may inspire the architectural or training improvements needed to consolidate functionality within the model.

As for economic gains, both the Trump and Biden administrations embraced the idea that economic security is national security.[50] Traditional understandings of the impact of new technologies on geopolitical competition have focused on what scholars have termed “leading sectors,” with countries that generate breakthroughs and monopolize

the profits of strategic industries poised to secure a relative power advantage.[51] The historical record of the first three Industrial Revolutions, however, indicates that diffusion of general-purpose technologies is more likely to lead to productivity gains across the whole economy and determine which countries gain the upper hand geopolitically.[52] By fostering system-level innovations outside of the AI labs, policymakers can promote diffusion of AI across the economy and generate the productivity gains necessary to ensure economic security and national security over the long run.

Some argue that AI will not just be *a* driver of economic and military power, but *the* driver. This mindset is perhaps most explicitly captured in Leopold Aschenbrenner's "Situational Awareness."[53] From this perspective, policymakers should give priority to the top AI labs and ensure security for the model innovations they generate. A shift in focus on AI innovation to the compound system level expands the range of policy options to promote and protect cutting-edge innovations. As the reaction to the inference-time paradigm shows, system-level innovations may be the key that unlocks ever-more powerful AI models. Policy must focus not only on supporting and protecting the work of the AI labs, but also on nurturing relative advantages in applied research, infrastructural support, and education and workforce developments that encourage widespread adoption and responsible use of the technology. Rebalancing policies to promote and protect both model and compound AI system development will help create a more diverse portfolio of bets on AI innovation.

The contributions of system-level advances also have implications for the organization of the AI sector. Apprehension over the concentration of power in Big Tech is a bipartisan issue and the emergence of AI has only exacerbated those concerns.[54] While efforts to limit concentration at the infrastructure and model layers of the AI stack may fall short of displacing the hyperscalers and AI labs, growing competition at the application layer may catalyze *system-to-model* innovations that can widen the base of contributors to the frontier.[55]

The debate on open-source AI underscores the stakes. Proponents argue that openness in AI research is the primary way to democratize the field, identify and remediate vulnerabilities, and spur breakthrough innovations that will unlock new reasoning capabilities and understanding of the physical world.[56] Detractors express concern that openness in AI research will invite catastrophic risks or fuel destabilizing advances in adversarial countries. By promoting system-level innovation, policymakers can bridge these two legitimate perspectives. Where models are relatively open in their access and documentation, *system-to-model* innovations can occur outside of the AI labs, as in circuit breaker innovations; where models are relatively closed, efforts in

the wider community can impact the research direction inside the labs, as in inference-time computing.[57] System-level innovation is not a panacea for addressing the concentration of power in AI innovation, adapting to societal-scale transformations, or managing the complex tradeoffs involved in decisions over model weight releases, but this pathway offers an opportunity for nations to tap into a wider pool of talent for AI development.[58] This talent base is small enough for policymakers to target with appropriate policy interventions, while also being large enough to contribute significantly to future AI advances.

Policy Implications

AI policy to date has placed a greater focus on the continual advance of capabilities in frontier models. From export control regimes that target advanced chips and semiconductor manufacturing equipment to regulations that base enforcement thresholds on training floating point operations (FLOP) or compute costs as a proxy for capabilities, the priority has been to extend the United States' lead in model performance as far as possible.[59] But the declining cost per token and the growing importance of open-weight models should encourage adjustments to policy that include bolstering compound AI systems.[60] To harness the potential of system-level innovation, policymakers will need to tailor solutions for the development ecosystems around AI. The pathways for innovation exist at both the model and system level. U.S. leadership should focus not only on advances at the frontier, but also on the diffusion and adoption of affordable AI models that are safe, secure, and trustworthy. Such an approach will benefit from new frameworks that foster system-level innovations.

System-level innovations open up new apertures for creative policymaking. First, system-level innovations facilitate the diffusion of AI, which will provide crucial gains to the country that can identify and harness them faster and more comprehensively than their geopolitical competitors.[61] Second, system-level innovations encompass a wider base of contributors than model-level development. Countries that can nurture system-level talent pipelines through education and workforce initiatives will reap greater political, military, and economic benefits. In particular, national security leaders and science funding agencies should pursue a calibrated approach to protect smaller developers without stifling their innovative potential; adapt policies, regulations, infrastructure, funding, and workforce development programs to nurture the dynamic interplay between AI models and systems; refine safety tests and red teaming for the system level; and evolve new tools and methodologies for tracking advances in development ecosystems at home and comparative net assessments of progress in other nations, including allies, partners, and competitors.

This brief suggests a three-part framework to navigate the policy implications of *system-to-model* innovation: **protect, diffuse, and anticipate**.

Protect

Researchers and smaller companies developing AI systems are pioneering innovations that will shape the next generation of models. As development ecosystems diversify and the pool of talent contributing to frontier AI broadens, so will the attack surfaces for bad actors and competitors to exploit. Policymakers should consider extending

targeted protections beyond the defense industrial base, the largest technology companies, and computationally intensive cloud configurations. Recent survey data indicates that many companies are poorly equipped to meet basic cybersecurity requirements.[62] In the context of protecting rapidly evolving *system-to-model* advances in AI, governments may need to consider subsidized services, such as safety testing and red team assessments at a system level, for smaller and medium-sized enterprises that lack the resources and expertise to withstand a determined adversarial attack. Policymakers should also conduct a detailed mapping of the smaller companies, infrastructure, research organizations, data analytics firms, and technical and nontechnical talent that comprise U.S. and allied AI innovation networks. System-level innovation highlights a broader pool of contributors, but this talent base and the critical resources they rely on are small enough to be identified for protection measures.

By monitoring and mapping the shared resources and supply chains for the broader development ecosystem, governments will be in a stronger position to target counter-transfer policies and promote information sharing among allies and partners. Foreign investments from countries of concern into the U.S. AI development ecosystem could affect firm-level governance and risk leakage of tacit knowledge.[63] Therefore, policymakers should consider embracing interpretations of relevant AI sectors and developers in the National Security Memorandum (NSM) on AI that include academic researchers and small and medium-sized firms that are driving system-level innovation and, as the example of control vectors to circuit breakers shows, new start-up activity in key fields.[64] Cybersecurity support, research security training for academic institutions, and timely sharing of threat indicators should extend to this broader development ecosystem. The National Science Foundation's SECURE Center could be a natural home for bolstering domestic protections and security-related training for smaller developers without undermining flows of AI talent and investment or stifling competitiveness.[65]

If system innovations drive future improvements of models, there may be reason to consider capability and safety tests at a system level as a leading indicator of model risks and performance. If a compound system today demonstrates risky behavior, developers should incorporate that risk assessment as they train system capabilities into the next generation of models. Evaluations from the U.S. AI Safety Institute and the UK AI Safety Institute of one of Anthropic's leading models underscore the point.[66] The system-level safety tests of the updated Claude 3.5 Sonnet on biological research tasks demonstrated capabilities exceeding human expert baselines while the model-level tests did not. Recent evidence further underscores the brittleness of today's systems to "gray-box" access.[67] Fine-tuning with benign datasets can increase

harmfulness, while document injection attacks to RAG systems can introduce new vulnerabilities.[68] Safety tests at a system level also matter for gauging capability improvements. As researchers have speculated, "Prompting is not going to be that important for that much longer."[69] Larger models appear to be more robust to prompt sensitivity, which suggests that prompt engineering quality could be a leading indicator for model performance in the future.[70]

Diffuse

Policymakers should also tailor approaches for promoting AI development ecosystems. National competitiveness may turn more on which countries can harness diffusion processes for geopolitical advantage. That will require sustained investments not only in a wider base of talent but also in the resources and compute infrastructure necessary to support them.

System-to-model innovation can happen through multiple pathways, including architectural changes or novel training regimes. Architectural changes are generally more resource intensive, while new training methods can build on existing foundation models. The NSM rightly notes that the field is moving too rapidly for the United States to "rely exclusively on a small cohort of large firms." The National Artificial Intelligence Research Resource (NAIRR) should expand its programs to widen access to computational infrastructure for inference with an emphasis on proposals for compound AI system development. Such an approach would enable researchers using pre-trained foundation models to develop system-level innovations that drive further progress in AI models and benefit from the new inference-time compute scaling paradigm. This support could also take the form of federal grants and field-building activities for next-generation inference chips, including field-programmable gate arrays (FPGA) and programs analogous to DARPA's Structured Array Hardware for Automatically Realized Applications (SAHARA) program.[71]

Targeted support would facilitate diffusion of system-level innovations.[72] Funding for applied research and investments in community colleges and nontraditional pathways will encourage system-level innovations among a wider pool of talent. By focusing on *system-to-model* advances, science funding agencies could also lay the groundwork for innovations to come. For example, there may be cases where frontier models are insufficiently advanced for the proposed systems-level innovations. Federal and state funding coupled with infrastructure support for under-resourced academic organizations could sustain efforts to develop systems for model advances that lie just around the corner. Consider the example of combining GRAG with the OneGen framework. This problem is resource intensive, as it requires modifying the architecture

of a large language model. By making shared data sets and computational assets more widely available to smaller companies and academic organizations, federal policy could promote a diverse ecosystem for system-level advances. The AI Grand Challenges Act, as proposed in Senate and House bills of the 118th Congress, would establish grand challenges and prize competitions for AI research. If reconsidered in the 119th Congress and authorized, the NSF should consider using AI Grand Challenges to fund compound AI system projects.[73]

A *system-to-model* mindset could also encourage a new generation of research and development agreements with U.S. allies and partners. The research community for prompt engineering is large and diverse, and frequently produces new techniques. Inference-time scaling and *system-to-model* innovation could represent a pathway for future work outside of the top AI research groups to improve future iterations of foundation models. Countries that lag behind in foundation model advances may have crucial advantages when it comes to promoting these *system-to-model* innovations. Consider fast-growing research clusters for algorithmic advances in the Center for Security and Emerging Technology's (CSET) Map of Science.[74] Researchers across many U.S. allied and partner nations are pursuing work in relevant areas, such as physics-informed neural networks, graph neural networks, Bayesian methods, transfer learning, and optimization techniques.[75]

Anticipate

The rise of *system-to-model* advances also necessitates a shift in tools and methodologies for horizon scanning, technology forecasting, and assessments of supply chain risks to focus on actors driving innovation not just in AI models but also in compound AI systems. The United States boasts many smaller companies and academic organizations that have been crucial to recent *system-to-model* advances. To be a fast follower and collaborator when such advances occur in allied and partner nations, the United States will need to develop its capabilities for spotlighting and adopting system-level advances. The Department of Commerce's SCALE tool could incorporate relevant data and interfaces for assessing supply chain risks among the AI development ecosystem.[76] Similarly, the State Department could leverage the International Technology Security and Innovation (ITSI) Fund to adapt new analytical tools and methods for assessing AI development ecosystems in allied and partner countries for potential R&D partnership opportunities.[77]

Greater focus on *system-to-model* innovations would enhance U.S. technology net assessments. Horizon-scanning methodologies and studies of comparative advantage tend to focus on model-level funding, patenting, investment, and publication activity

over progress at the system level. Assessments of leadership in AI will also vary depending on whether one's metrics highlight frontier model developers and computational resources for training or broader development ecosystems and computational resources for inference. Consider the new inference-time computing paradigm. The trend toward inference scaling may end up strengthening the rationale behind export controls on advanced AI chips, and model-level algorithmic innovations will continue to improve performance. As this brief shows, however, model-driven innovation is not the only pathway for compute efficiency gains. In addition to continuing efforts at model-level tracking, policymakers will also need to devote more time and resources to tracking system-level innovations that diminish the compute required for training and inference.[78] Horizon scanners should pursue novel approaches to identify not only which individual resources to track but also the shared resources and international research networks that would benefit from information sharing, resources, and dedicated services from governments to defend against cyber and physical attacks against U.S. and allied AI development ecosystems.[79]

Structural Transformations in the AI Stack

Edison's lasting accomplishment was not simply to create a more efficient product. He invented a lighting system that lay the groundwork for distributed power through parallel circuits and filaments that required less current.[80] His system-level innovation rivaled the gas companies of the day and took the form in 1882 of the Pearl Street Station, "a full-scale central power station with a system of conductors to distribute electricity to end-users in the high-profile business district in New York City."[81] Market competition defined the early years of this nascent industry, but over time the government legislated a network of utilities that operated regulated monopolies in defined areas.

It is worth considering whether foundation model providers and the hyperscalers that back them will assume similar roles. The answer depends, in part, on the interplay between efficiencies of scale and compute expenditures, on the one hand, and the social costs and national security risks, on the other.[82] The industrial organization of AI is rapidly evolving, with increasing commoditization at the foundation layer and growing concentration among the major cloud service providers. Foundation model providers exhibit some of the characteristics of natural monopolies, such as high barriers to entry, high capital expenditure, and low product differentiation. These trends may change as DeepSeek's R1 and other open-source models drive competition in the foundation model layer. But the medium- to longer-term indicators appear to point in the direction of a new market structure, with hyperscalers such as Google and Microsoft providing the infrastructure (cloud computing); AI labs and well-resourced companies such as OpenAI and Anthropic offering a power source (the model); and governments facilitating permitting and modernizing the grid, while smaller, less capital-intensive companies build on top of that infrastructure.

If system improvements lead to model improvements, then a new dynamic will take shape to bring those innovations into foundation model providers. If national security concerns predominate, then governments may assume a more direct role in securing model weights and other sensitive AI infrastructure.[83] Just as the electric utilities did not end up making the bulk of the profit from the Second Industrial Revolution, today's hyperscalers and foundation model providers may operate a high capital expenditure, tightly regulated business model that enables other innovative firms to sustain thriving AI ecosystems and capture a greater share of the value add.* How government, private

* High capital expenditures and tight regulation could apply to both hyperscalers and foundation model providers, but low profit margins apply largely to the foundation model providers. While hyperscalers' margins may be low relative to the Software-as-a-Service providers that build on their infrastructure,

sector, and civil society leaders influence the design of market structures and diffusion of essential capabilities to enable a wider base of innovation will define the future of AI's industrial organization.[84] These questions will grow in importance as calls for a "Manhattan Project-like program" for AI gather steam, including in the U.S.-China Economic and Security Review Commission's 2024 Report to Congress.[85]

At a minimum, it is time for creative thinking about the institutions that will guide the trajectory of AI.[86] New technologies are not frictionless; they shape and are shaped by socioeconomic conditions and institutional arrangements.[87] Anthropic's Public Benefit Corporation and Long-Term Benefit Trust governance model and debates over the strengths and limitations of industry self-governance suggest that the field is experimenting with new designs and structures for managing a potentially transformative technology.[88] Policymakers and civic leaders must steer these efforts in directions that benefit the public good. Regulations on the hyperscalers or foundation model providers and lighter-touch approaches for development ecosystems could offer the prospect of a middle ground between growing market concentration and the nationalization of AI assets.[89] A patchwork approach to regulation and enforcement in the absence of structural reforms will not create a durable framework for AI innovation.[90] At the same time, putting too much store in systems-level innovation may create new risks and vulnerabilities. Increasing the memory of LLMs, for example, appeared harmless until researchers discovered "many-shot jailbreaking" techniques that exploit longer context windows.[91] Decentralized training runs may lower the costs of compute for researchers, but progress on distributed techniques will create new challenges for policies that rely on centralized computing clusters as an instrument of AI governance.[92] Treating the major cloud service providers and foundation model companies as utilities may also induce complacency in keeping up the base layer, just as perverse incentives caused years of underinvestment in the electrical grid.

Lights do not beget lighting systems on their own. System effects are powerful but also unpredictable.[93] They require social and political institutions to mitigate the risks and diffuse the benefits. A system perspective can help illuminate the path to progress, but policy choices will define it.

---

they remain highly profitable and appear poised to capture a large portion of AI-driven economic growth.

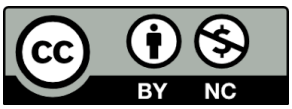
Authors

Jonah Schiestle is the director of engineering at Cymantix and a graduate student in the security studies program at Georgetown University.

Andrew Imbrie is associate professor of the practice in the Gracias chair in security and emerging technology at Georgetown University's School of Foreign Service. He also teaches on the core faculty in the security studies program at Georgetown and is a faculty affiliate at CSET.

Acknowledgments

The authors would like to thank Owen Daniels, Shelton Fitch, Lauren Lassiter, Jason Ly, Igor Mikolic-Torreira, and Susan Moynihan for their feedback and assistance throughout the process. For helpful comments on the underlying issue brief, we thank Jack Corrigan, Andrew Lohn, Sam Manning, and Colin Shea-Blymyer.



Document Identifier: doi: 10.51593/20240056

Appendix: System and Model Innovation Examples

Model-to-Model Innovation

Model-to-model innovations can arise from a series of improvements to the state of the art for a single task. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) contest represents a classic example of the *model-to-model* innovation pathway. In the 2012 contest, AlexNet achieved state-of-the-art results by bringing together a series of model innovations. AlexNet leveraged architectural improvements, including multiple convolution layers with max pooling followed by fully connected layers.[94] It also made training improvements through data augmentation and dropout. To augment training data, the AlexNet team applied transformations that would allow for training of larger networks with the increased volume of data. Dropout removes some neurons randomly during training passes to force the network to learn more robust associations and reduce overfitting. AlexNet kickstarted a chain of model innovations in object recognition. It showed that deep learning could greatly improve performance, and deeper networks could yield further improvements. The 2014 winner, GoogLeNet, introduced a new convolution layer architecture called the Inception module, which used sparsity and dimension reduction techniques to allow deeper networks without overfitting.[95] The 2015 winner, ResNet, introduced another innovation to aid in training deeper networks: the residual learning framework.[96] This new paradigm avoided degradation issues arising in deep networks. ResNet spurred such rapid advances that by 2017, the benchmarks were saturated and the contest was no longer held.

Another type of *model-to-model* innovation incorporates memory into neural networks. Feedforward networks process data in a single pass; they have no memory and make it difficult to account for sequential features in data, such as time series data or natural language. Recurrent neural networks (RNNs) address this issue through loop structures, which require multiple passes to resolve. RNNs have a long history, starting with Frank Rosenblatt's cross-coupled perceptrons in the 1960s.[97] But basic RNNs struggle to retain memory across large gaps in semantic information or sequential data. The effect can be twofold: The mechanism for retaining memory can become vanishingly small over time, losing the desired effect, or the mechanism explodes and becomes sufficiently large to wash out the recent memory.

In 1995, Hochreiter and Schmidhuber introduced Long Short-Term Memory (LSTM) to address this challenge.[98] The LSTM architecture uses a cell state mechanism to pass memory down the line and gating mechanisms to determine how that memory is incorporated. In the mid- to late-2000s, researchers demonstrated the effectiveness of

LSTM architectures across a variety of natural language processing tasks, thereby inspiring more adaptations. In 2014, a team at Google, including Ilya Sutskever, introduced seq2seq.[99] This approach used the newly developed encoder-decoder structure to chain two LSTM networks for machine translation. The encoder subcomponent creates an internal representation of the original language, which the decoder subcomponent then transforms into the output language.

This series of innovations in recurrence still suffered a disadvantage. The recurrence mechanisms for memory render training considerably more difficult than single pass networks. The "Attention Is All You Need" paper references these developments in its motivation but eschews the recurrence mechanism in favor of attention heads, previously used within RNNs, for memory.[100] Attention heads provided a nonsequential way to use data across sequences, which improved parallelizability and reduced training costs while improving performance. The transformer model has since dominated natural language processing tasks. As these examples show, model innovations can spark further model advances across a range of mechanisms inspired by a common need for memory.

Model-to-System Innovation

Model-to-system innovations may take inspiration from emergent capabilities in models that enable new systems. The rise of large language models represents one such case where a new model, GPT-3, demonstrated a level of performance well beyond previous language models. Particularly important was the claim that "Language Models are Few-Shot Learners."[101] The ability to take examples in the model's input and produce effective results beyond the model's training data (referred to as in-context learning) opened up possibilities for adapting foundation language models to myriad downstream tasks without specialized training.

Perhaps the simplest class of compound systems built on in-context learning was the field of prompt engineering.[102] Prompt engineering has the interesting characteristic that the first component in the compound AI system may in fact be the human user. Consider the case of chain-of-thought reasoning. Researchers at Google Brain showed language models performed better on certain tasks when prompting elicited step-by-step reasoning in the response.[103] Chain-of-thought improves answer quality through the in-context learning mechanism. When the model finds a solution through steps, those steps impact the eventual answer because of the autoregressive nature of generative language models. The underlying mechanism of language models, next-token generation, means that the tokens representing the final answer are predicted using the steps generated along the way. This property inspired the development of a

compound system to enhance performance on complex tasks that could be reduced to intermediate steps.

Such a compound system could include the human operator as a component. When asking a question, the user first transforms their question from its original form into one that elicits intermediate reasoning, such as "show your work." The user's goal is not the rehearsal of the steps but the correct final answer. When considering the input-output formulation of AI systems, this alteration to the prompt represents the intermediate output required to qualify as a compound system. This step can also be automated via prompt templating, where prompts are automatically formatted to elicit chain-of-thought, returning to a more traditional formulation of a compound AI system without a human stage.[104]

Another system-level innovation is retrieval-augmented generation (RAG), where a prompt for a language model is first sent to an external database for retrieval of relevant information and then paired with the original question to inform the model's response. RAG techniques draw on the power of in-context learning and expanding context windows, or the ability of a language model to take in longer prompts. Researchers at Facebook, University College London, and New York University introduced RAG systems to accomplish knowledge-intensive tasks. Language models by themselves were limited to knowledge learned during training and stored in parameters, also called parametric memory.[105] By adding a non-parametric memory (knowledge stored outside the model; in this case, a database), a compound system could handle a broader range of tasks by first retrieving relevant information from the database, then providing that data to the model along with the question before finally generating a response. The response goes through the same in-context learning mechanism, which enables the model to extract the knowledge it needs from the retrieved data in the prompt. The original formulation of RAG used the seq2seq model previously discussed, but researchers quickly adapted the technique to the latest LLMs.

RAG systems expand the potential use cases for powerful LLMs by addressing a number of their shortfalls. The most apparent improvement is that of the original proposal: solving knowledge-intensive tasks. RAG alleviates the need to retrain LLMs constantly by incorporating new or niche information in the retrieval component of the compound AI system. But RAG systems can also address some less obvious problems with out-of-the-box LLM usage. One of the most famous challenges in using LLMs is the tendency to produce hallucinations—also described as "stochastic parroting." One driver of hallucinations is when a prompt is outside the distribution of the training data.[106] RAG systems can address this problem by introducing retrieved data in the

areas where the model did not have sufficient data. RAG has reduced hallucinations, though it cannot eliminate the issue entirely.[107]

RAG systems can also help address privacy concerns and the risk of data leakage.[108] Either through failure to sanitize outputs or through adversarial attacks, LLMs could expose their training data in response to certain prompts. Instead, sensitive data can be excluded from training data and introduced through the retrieval component. While this approach does not totally eliminate potential leakage, it represents an easier problem than preventing data leakage mechanisms of the LLM itself, which are not well understood. Additionally, some evidence suggests implementing RAG also lessens the underlying data leakage concern, though the mechanism by which this leakage occurs is unclear.[109]

Both prompt engineering and RAG represent the types of compound AI systems that make possible the foundation model concept.[110] These compound systems built off the innovative work that produced powerful models with new capabilities, and introduced the missing piece that allowed for usage in myriad downstream tasks.

System-to-System Innovation

As with *model-to-model* innovations, *system-to-system* innovations often involve modifications to compound AI systems for a specific performance metric. Returning to the example of RAG systems, researchers have made significant improvements to the way knowledge is stored in and used from the retrieval component. Long context language models have a tendency to forget the context in the middle, such as intermediate documents in multi-document question answering.[111] Given that RAG systems often involve precisely this task, loss of extended context represents a crucial problem for performance. Performance could be highly sensitive to the number of documents added to the question context as well as the order in which the documents are added. Researchers added document chunking to RAG systems to address this problem.[112] Given that the answer to a question likely comes from a specific part of a document and not the whole document, documents can be broken down into smaller chunks and stored. Only the relevant chunks of the document are retrieved and used to augment generation, reducing the "forgetting in the middle" dilemma of long context.

But chunking introduces its own challenges. Chunking documents is a difficult task, with a variety of methods ranging from naive chunks of fixed size to complex, semantic chunking that uses specialized models—which can greatly impact the quality of the chunked database. Additionally, documents may not always contain all relevant information in close proximity so chunks may lose vital context. Anthropic recently

developed a chunking mechanism to combat this challenge, known as contextual retrieval.[113] Contextual retrieval breaks a document down into chunks and then provides document-chunk pairs to a language model prompting the model for context on the chunk in the document. Researchers then take these two-part chunks, direct text from the source and a generative contextual explanation, and add them to the retrieval component to allow for shorter, information dense RAG prompts.

Another challenge for RAG performance in existing tasks is what to do when faced with conflicting information, either due to noisy data or adversarial attack. A number of *system-to-system* innovations address this challenge. RobustRAG addresses the adversarial retrieval problem by implementing RAG as a series of single document question-answering tasks and then aggregating to a final response.[114] By eliciting responses from individual documents, robust aggregation techniques ensure the system can handle adversarial retrieval manipulation. InstructRAG uses in-context learning, inspired by CoT prompting, to address noise in retrieved documents by synthesizing the documents before answering the question.[115] But both methods do not account for the other source of RAG systems: the model's training data. Astute RAG addresses this pitfall by adding a stage to the system that creates a synthetic document representing the model's trained knowledge.[116] With this new context, the compound system then uses similar techniques to synthesize a response and even presents instances of conflict to the user. This *system-to-system* innovation raises the performance floor of RAG systems, allowing them to perform almost as well as direct usage of the model even in the worst-case scenario where the retriever component yields wrong or useless information.

More substantial *system-to-system* innovations adapt compound AI systems to address new problems. GraphRAG represents one such innovation for RAG systems.[117] The prior methods all use the same vector retrieval methods proposed in the original RAG formulation, but there are other retriever systems, such as knowledge graphs.[118] Knowledge graphs capture global relationships across documents and better handle highly structured information—both of which are challenges for traditional RAG methods. Researchers have used GraphRAG in domains where citation linkages are important, as well as for multi-document summarization where the knowledge graph provides efficient compression of the documents to avoid the challenges of long context in language model performance. Building on the GraphRAG approach, HybridRAG introduces a more complicated system, which integrates the advantages of both systems to produce better answers than either system can on its own.[119] The VectorRAG to GraphRAG to HybridRAG *system-to-system* innovation demonstrates the expansion of tasks a system is equipped to handle.

Endnotes

[1] Jeffrey Ding, *Technology and the Rise of Great Powers: How Diffusion Shapes Economic Competition* (Princeton: Princeton University Press, 2024); Nicholas Crafts, "Artificial Intelligence as a General-Purpose Technology: An Historical Perspective, *Oxford Review of Economic Policy* 37, Issue 3 (Autumn 2021): 521-536), https://academic.oup.com/oxrep/article/37/3/521/6374675?login=false.

[2] "OECD AI Principles Overview," OECD, https://oecd.ai/en/ai-principles.

[3] "Responsible Scaling Policy: Version 2.1", Anthropic, effective March 31, 2025, https://www-cdn.anthropic.com/17310f6d70ae5627f55313ed067afc1a762a4068.pdf; "Consultation Paper on AI Regulation Emerging Approaches Across the World," UNESCO, 2024, https://unesdoc.unesco.org/ark:/48223/pf0000390979.

[4] Jill Jonnes, "Let There Be Light", *TIME*, June 23, 2010, https://content.time.com/time/specials/packages/article/0,28804,1999143_1999202_1999203,00.html; Ernest Freeberg, *The Age of Edison: Electric Light and the Invention of Modern America* (New York: Penguin Publishing Group, 2013).

[5] Freeberg, *The Age of Edison*.

[6] Neil Perkin, "What the Electrification of Manufacturing Reveals About Digital Transformation," *Building The Agile Business* (blog), August 24, 2017, https://medium.com/building-the-agile-business/what-the-electrification-of-manufacturing-reveals-about-digital-transformation-f6520a9f171f.

[7] "The History of Electrification: New York City Historic Power Plants," Edison Tech Center, accessed February 10, 2025, https://edisontechcenter.org/NYC.html.

[8] Tim Harford, "Why didn't electricity immediately change manufacturing?," *BBC*, August 20, 2017, https://www.bbc.com/news/business-40673694; "The War of the Currents: AC vs. DC Power," *Energy.gov* (blog), November 18, 2014, https://www.energy.gov/articles/war-currents-ac-vs-dc-power; Calestous Juma, *Innovation and Its Enemies: Why People Resist New Technologies* (New York: Oxford University Press, 2016).

[9] Andrew Ng, "Issue 286," *The Batch* (blog) on DeepLearning.ai, January 29, 2025, https://www.deeplearning.ai/the-batch/issue-286/; Ben Cottier, Ben Snodin, David Owen, and Tom Adamczewski, "LLM Inference Prices Have Fallen Rapidly but Unequally Across Tasks," Epoch AI, March 12, 2025, https://epoch.ai/data-insights/llm-inference-price-trends.

[10] Tejas Narechania and Ganesh Sitaraman, "An Antimonopoly Approach to Governing Artificial Intelligence," *Yale Law and Policy Review* 43, no. 95 (January 2024), https://ssrn.com/abstract=4597080.

[11] Nicholas Crafts, "Artificial intelligence as a general-purpose technology: an historical perspective", *Oxford Review of Economic Policy* 37, no. 3 (Autumn 2021): 521–536, https://doi.org/10.1093/oxrep/grab012.

[12] "Algorithm," s.v., Glossary, Computer Security Resource Center, National Institute of Standards and Technology, accessed February 10, 2025, https://csrc.nist.gov/glossary/term/algorithm.

[13] "What is an AI model?," Think, IBM, accessed February 10, 2025, https://www.ibm.com/think/topics/ai-model.

[14] Matei Zaharia, Omar Khattab, Lingjiao Chen et al., "The Shift from Models to Compound AI Systems," *The BAIR Blog*, February 18, 2024, https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/.

[15] Micah Musser, Andrew Lohn, James X. Dempsey et al., "Adversarial Machine Learning and Cybersecurity" (Center for Security and Emerging Technology, April 2023), https://doi.org/10.51593/2022CA003.

[16] Omar Sanseviero, Lewis Tunstall, Philipp Schmid et al., "Mixture of Experts Explained," Hugging Face (blog), December 11, 2023, https://huggingface.co/blog/moe.

[17] Mandar Karhade, "GPT-4: 8 Models in One; The Secret is Out," *Towards AI* (blog) on Medium, June 24, 2023, https://pub.towardsai.net/gpt-4-8-models-in-one-the-secret-is-out-e3d16fd1eee0.

[18] Brendan Bycroft, "LLM Visualization," accessed February 10, 2025, https://bbycroft.net/llm; Ashish Vaswani, Noam Shazeer, Niki Parmar et al., "Attention Is All You Need," arXiv preprint arXiv:1706.03762 (2017), https://arxiv.org/abs/1706.03762.

[19] "ImageNet Large Scale Visual Recognition Challenge (ILSVRC)," ImageNet, accessed February 10, 2025, https://www.image-net.org/challenges/LSVRC/; "Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots," Hugging Face, accessed February 10, 2025, https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard; Jared Kaplan, Sam McCandlish, Tom Henighan et al., "Scaling Laws for Neural Language Models," arXiv preprint arXiv:2001.08361 (2020), https://arxiv.org/abs/2001.08361; Danny Hernandez and Tom B. Brown, "Measuring the Algorithmic Efficiency of Neural Networks," arXiv preprint arXiv:2005.04305 (2020), https://arxiv.org/abs/2005.04305.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *NeurIPS Proceedings* 25 (2012), https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[21] Richard Sutton, "The Bitter Lesson," Incomplete Ideas (blog), March 13, 2019, http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

[22] Thomas Woodside and Helen Toner, "How Developers Steer Language Model Outputs: Large Language Models Explained, Part 2," *Center for Security and Emerging Technology* (blog), March 8,

2024, https://cset.georgetown.edu/article/how-developers-steer-language-model-outputs-large-language-models-explained-part-2/; Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida et al., "Training language models to follow instructions with human feedback," arXiv preprint arXiv:2203.02155 (2022), https://arxiv.org/abs/2203.02155.

[23] David Silver, Aja Huang, Chris J. Maddison et al., "Mastering the game of Go with deep neural networks and tree search," *Nature* 529 (2016): 484-489, https://doi.org/10.1038/nature16961; Alan Levinovitz, "The Mystery of Go, the Ancient Game that Computers Still Can't Win," *Wired*, May 12, 2014, https://www.wired.com/2014/05/the-world-of-computer-go/.

[24] David Silver, Julian Schrittwieser, Karen Simonyan et al., "Mastering the game of Go without human knowledge," *Nature* 550 (2017): 354–359, https://doi.org/10.1038/nature24270.

[25] David Silver, Thomas Hubert, Julian Schrittwieser et al., "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Science* 362 (2018): 1140–1144, https://doi.org/10.1126/science.aar6404.

[26] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature* 575 (2019): 350–354, https://doi.org/10.1038/s41586-019-1724-z; Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert et al., "Mastering Atari, Go, chess and shogi by planning with a learned model," *Nature* 588 (2020): 604–609, https://doi.org/10.1038/s41586-020-03051-4.

[27] Mogens Dalgaard, Felix Motzoi, Jens Jakob Sørensen, and Jacob Sherson, "Global optimization of quantum dynamics with AlphaZero deep exploration," *npj Quantum Inf* 6 (2020), https://doi.org/10.1038/s41534-019-0241-0.

[28] "Learning to reason with LLMs," *OpenAI*, September 12, 2024, https://openai.com/index/learning-to-reason-with-llms/.

[29] Noam Brown, "AI Won't Plateau—If We Give it Time to Think," *TEDAI*, October 22, 2024, https://tedai-sanfrancisco.ted.com/speakers/noam-brown/.

[30] Daniel Kahneman, "Of 2 Minds: How Fast and Slow Thinking Shape Perception and Choice [Excerpt]," *Scientific American*, June 15, 2012, https://www.scientificamerican.com/article/kahneman-excerpt-thinking-fast-and-slow/.

[31] Carl Franzen, "DeepSeek's first reasoning model R1-Lite-Preview turns heads, beating OpenAI o1 performance," *VentureBeat*, November 20, 2024, https://venturebeat.com/ai/deepseeks-first-reasoning-model-r1-lite-preview-turns-heads-beating-openai-o1-performance/.

[32] Miles Brundage (@Miles_Brundage), "Unsurprising that DeepSeek announced capabilities that are likely beyond what all but ~1-3 frontier developers in the US have. We're in the early stages of a new paradigm, which is, for now, less compute-intensive. And DeepSeek is very competent.," X, November 20, 2024, https://x.com/Miles_Brundage/status/1859325404798124218.

[33] Kyle Wiggers, "Alibaba releases an 'open' challenger to OpenAI's o1 reasoning model," *TechCrunch*, November 27, 2024, https://techcrunch.com/2024/11/27/alibaba-releases-an-open-challenger-to-openais-o1-reasoning-model/.

[34] Will Knight, "OpenAI Upgrades Its Smartest AI Model With Improved Reasoning Skills," *Wired*, December 20, 2024, https://www.wired.com/story/openai-o3-reasoning-model-google-gemini.

[35] François Chollet, "OpenAI o3 Breakthrough High Score on ARC-AGI-PUB," ARC Prize (blog), December 20, 2024, https://arcprize.org/blog/oai-o3-pub-breakthrough.

[36] Sam Altman, "Reflections," Sam Altman (blog), January 5, 2025, https://blog.samaltman.com/reflections.

[37] Xiaohan Xu, Chongyang Tao, Tao Shen et al., "Re-Reading Improves Reasoning in Large Language Models," arXiv preprint arXiv:2309.06275 (2023), https://arxiv.org/abs/2309.06275.

[38] Jintian Zhang, Cheng Peng, Mengshu Sun et al., "OneGen: Efficient One-Pass Unified Generation and Retrieval for LLMs," arXiv preprint arXiv:2409.05152 (2024), https://arxiv.org/abs/2409.05152.

[39] Aidan Gomez, "Command R: Retrieval-Augmented Generation at Production Scale," Cohere (blog), https://cohere.com/blog/command-r; Model Card for C4AI Command-R, Hugging Face, https://huggingface.co/CohereForAI/c4ai-command-r-v01.

[40] "Prompting Command R and R+," Cohere docs, https://docs.cohere.com/docs/prompting-command-r.

[41] "Thread: Circuits," *Distill*, March 10, 2020, https://distill.pub/2020/circuits/.

[42] Nelson Elhage, Neel Nanda, Catherine Olsson et al., "A Mathematical Framework for Transformer Circuits," Transformer Circuits Thread, Anthropic, December 22, 2021, https://transformer-circuits.pub/2021/framework/index.html.

[43] Andy Zou, Long Phan, Sarah Chen et al., "Representation Engineering: A Top-Down Approach to AI Transparency," arXiv preprint arXiv:2310.01405 (2023), https://arxiv.org/abs/2310.01405.

[44] "Enterprise Security for AI-Powered Applications," Gray Swan, https://www.grayswan.ai/; Andy Zou, Long Phan, Justin Wang et al., "Improving Alignment and Robustness with Circuit Breakers," arXiv preprint arXiv:2406.04313 (2024), https://arxiv.org/pdf/2406.04313.

[45] Danny Hernandez and Tom B. Brown, "Measuring the Algorithmic Efficiency of Neural Networks," OpenAI, https://cdn.openai.com/papers/ai_and_efficiency.pdf.

[46] Zaharia et al. "Models to Compound AI Systems."

[47] Owen J. Daniels, "The AI 'Revolution in Military Affairs': What Would it Really Look Like?" *Lawfare*, December 21, 2022, https://www.lawfaremedia.org/article/ai-revolution-military-affairs-what-would-it-really-look.

[48] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," arXiv preprint arXiv:2303.10130 (2023), https://arxiv.org/abs/2303.10130; Paul Kennedy, *The Rise and Fall of the Great Powers* (New York: Vintage, 1989).

[49] Benjamin Jensen and Dan Tadross, "How Large-Language Models Can Revolutionize Military Planning," *War on the Rocks*, April 12, 2023, https://warontherocks.com/2023/04/how-large-language-models-can-revolutionize-military-planning/.

[50] Kathleen Claussen and Timothy Meyer, "How 'Economic Security' is Reshaping Presidential Power," Just Security, July 16, 2024, https://www.justsecurity.org/97760/economic-security-presidential-power/.

[51] Ding, *Technology and the Rise of Great Powers*.

[52] Ding, *Technology and the Rise of Great Powers*.

[53] Leopold Aschenbrenner, "Situational Awareness: The Decade Ahead," https://situational-awareness.ai/.

[54] "A Bipartisan Majority of Voters Are Concerned About the Impact Big Tech Companies Have on the U.S. Economy and on Economic Competition," Data for Progress, June 14, 2022 https://www.dataforprogress.org/blog/2022/6/14/a-bipartisan-majority-of-voters-are-concerned-about-the-impact-big-tech-companies-have-on-the-us-economy-and-on-economic-competition; Amba Kak and Sarah Myers West, "Confronting Tech Power," AI Now Institute, April 11, 2023, https://ainowinstitute.org/wp-content/uploads/2023/04/AI-Now-2023-Landscape-Report-FINAL.pdf.

[55] "Democratize AI? How the Proposed National AI Research Resource Falls Short," AI Now Institute and Data & Society Research Institute, October 5, 2021, https://ainowinstitute.org/publications/policy-brief/democratize-ai-how-the-proposed-national-ai-research-resource-falls-short.

[56] "Debating Technology," World Economic Forum Annual Meeting 2025, https://www.youtube.com/watch?v=MohMBV3cTbg.

[57] Andreas Liesenfeld and Mark Dingemanse, "Rethinking open source generative AI: open-washing and the EU AI Act," in the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil, https://pure.mpg.de/rest/items/item_3588217_2/component/file_3588218/content; see also: Jon Bateman et al., "Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation AI Model Governance," (Carnegie Endowment for International Peace, July 23, 2024), https://carnegieendowment.org/research/2024/07/beyond-open-vs-closed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance?lang=en.

[58] Arvind Narayanan and Sayaash Kapoor, "AI as a Normal Technology: An Alternative to the Vision of AI as a Potential Superintelligence," (Knight First Amendment Institute, Columbia University),

https://kfai-documents.s3.amazonaws.com/documents/c3cac5a2a7/AI-as-Normal-Technology---Narayanan---Kapoor.pdf.

[59] Remarks by National Security Advisor Jake Sullivan at the Special Competitive Studies Project, Global Emerging Technologies Summit, September 16, 2022, https://bidenwhitehouse.archives.gov/briefing-room/speeches-remarks/2022/09/16/remarks-by-national-security-advisor-jake-sullivan-at-the-special-competitive-studies-project-global-emerging-technologies-summit/.

[60] Anonymous, "Ban the H20: Competing in the Inference Age," *ChinaTalk*, March 7, 2025, https://www.chinatalk.media/p/ban-the-h20-competing-in-the-inference; Florence G'sell, Ashok Ayar, and Zeke Gillman, "California's SB1047 vs. EU AI Act: A Comparative Analysis of AI Regulation," *SciencesPo*, October 28, 2024, https://www.sciencespo.fr/public/chaire-numerique/en/2024/10/28/californias-sb1047-vs-eu-ai-act-a-comparative-analysis-of-ai-regulation.

[61] AI diffusion will likely prove uneven and subject to myriad societal and economic constraints that are not amenable to quick technical fixes. For more on the varying timescales of AI development and diffusion, see Narayanan and Sayaash, "AI as a Normal Technology," https://kfai-documents.s3.amazonaws.com/documents/c3cac5a2a7/AI-as-Normal-Technology---Narayanan---Kapoor.pdf.

[62] Carley Welch, "Survey shows very few DoD contractors 'fully' ready for CMMC 2.0 ahead of 2025 rollout," *Breaking Defense*, https://breakingdefense.com/2024/10/survey-shows-very-few-dod-contractors-fully-ready-for-cmmc-2-0-ahead-of-2025-rollout/.

[63] Emily Kilcrease, "The Role of Investment Security in Addressing China's Pursuit of Defense Technologies," Testimony Before the U.S.-China Economic and Security Review Commission, April 13, 2023, https://www.uscc.gov/sites/default/files/2023-04/Emily_Kilcrease_Testimony.pdf.

[64] Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence, White House, October 24, 2024, https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/.

[65] "NSF-backed SECURE Center will support research security, international collaboration," July 24, 2024, https://www.nsf.gov/news/nsf-backed-secure-center-will-support-research.

[66] NIST, "Pre-Deployment Evaluation of Anthropic's Upgraded Claude 3.5 Sonnet," Washington, DC: U.S. Department of Commerce, November 19, 2024, https://www.nist.gov/news-events/news/2024/11/pre-deployment-evaluation-anthropics-upgraded-claude-35-sonnet; note the UK's AI Safety Institute is now "The AI Security Institute" and the U.S. AI Safety Institute is now the "Center for AI Standards and Innovation"; for more details, see: https://www.gov.uk/government/news/tackling-ai-security-risks-to-unleash-growth-and-deliver-plan-

for-change and https://www.commerce.gov/news/press-releases/2025/06/statement-us-secretary-commerce-howard-lutnick-transforming-us-ai.

[67] Kellin Pelrine, Mohammad Taufeeque, Michał Zając, Euan McLean, Adam Gleave, "Exploiting Novel GPT-4 APIs," arXiv preprint arXiv:2312.14302 (2024), https://arxiv.org/abs/2312.14302.

[68] Pelrine et al., "Exploiting Novel GPT-4 APIs."

[69] Ethan Mollick, "A guide to prompting AI (for what it is worth)," One Useful Thing (blog), April 26, 2023, https://www.oneusefulthing.org/p/a-guide-to-prompting-ai-for-what.

[70] Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen, "ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs," arXiv preprint arXiv:2410.12405 (2024), https://arxiv.org/abs/2410.12405.

[71] "Expanding Domestic Manufacturing of Secure, Custom Chips for Defense Needs," DARPA, March 18, 2021, https://www.darpa.mil/news/2021/domestic-manufacturing-secure-chips.

[72] Ding, *Technology and the Rise of Great Powers*.

[73] S.4236 - AI Grand Challenges Act of 2024, Congress.gov, 118th Congress, https://www.congress.gov/bill/118th-congress/senate-bill/4236/text; H. R. 9475, 118th Congress, 2nd Session, https://www.congress.gov/118/bills/hr9475/BILLS-118hr9475ih.pdf.

[74] "Map of Science," Center for Security and Emerging Technology.

[75] "Map of Science," Center for Security and Emerging Technology; Husanjot Chahal, Helen Toner, and Ilya Rahkovsky, "Small Data's Big AI Potential" (Center for Security and Emerging Technology, September 2021), https://cset.georgetown.edu/publication/small-datas-big-ai-potential/#:~:text=Conventional%20wisdom%20suggests%20that%20cutting,not%20require%20massive%20labeled%20datasets.

[76] "Fact Sheet: Department of Commerce Announces New Actions on Supply Chain Resilience," Washington, DC: U.S. Department of Commerce, September 10, 2024, https://www.commerce.gov/news/fact-sheets/2024/09/fact-sheet-department-commerce-announces-new-actions-supply-chain.

[77] "The U.S. Department of State International Technology Security and Innovation Fund," U.S. Department of State, https://www.state.gov/the-u-s-department-of-state-international-technology-security-and-innovation-fund/.

[78] Jordan Schneider and Lily Ottinger, "AI Geopolitics in the Age of Test-Time Compute with Chris Miller + Lennart Heim," *ChinaTalk*, January 6, 2025, https://www.chinatalk.media/p/ai-geopolitics-in-the-age-of-test.

[79] Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence, White House, October 24, 2024, https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/; Andrew Lohn, "Poison in the Well: Securing the Shared Resources of Machine Learning" (Center for Security and Emerging Technology, June 2021), https://cset.georgetown.edu/publication/poison-in-the-well/.

[80] Freeberg, *The Age of Edison*.

[81] "History of Power: The Evolution of the Electric Generation Industry," *Power*, October 1, 2022, https://www.powermag.com/history-of-power-the-evolution-of-the-electric-generation-industry/.

[82] Jon Schmid, Tobias Sytsma, and Anton Shenk, "Evaluating Natural Monopoly Conditions in the AI Foundation Model Market," RAND, September 12, 2024, https://www.rand.org/pubs/research_reports/RRA3415-1.html.

[83] Aschenbrenner, "Situational Awareness;" Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott, "Securing AI Model Weights," RAND, May 30, 2024, https://www.rand.org/pubs/research_reports/RRA2849-1.html.

[84] Ketan Ahuja, "Innovating Antitrust Law: How Innovation Really Happens and How Antitrust Law Should Adapt," Roosevelt Institute, October 19, 2022, https://rooseveltinstitute.org/publications/innovating-antitrust-law/.

[85] Filip Timotija, "Google CEO eyes AI 'Manhattan Project' after Trump inauguration," *The Hill*, December 13, 2024, https://thehill.com/policy/technology/5039230-sundar-pichai-donald-trump-artificial-intelligence-manhattan-project/; 2024 Report Congress, U.S.-China Economic and Security Review Commission, https://www.uscc.gov/sites/default/files/2024-11/2024_Executive_Summary.pdf.

[86] "The Collective Intelligence Project," https://static1.squarespace.com/static/631d02b2dfa9482a32db47ec/t/63e01fcbf73bb003a722792f/1675632588509/CIP+Whitepaper.pdf; Daron Acemoglu, "Institutions, Technology and Prosperity," NBER Working Paper Series, Working Paper 33442, https://www.nber.org/system/files/working_papers/w33442/w33442.pdf.

[87] Calestous Juma, *Innovation and Its Enemies: Why People Resist New Technologies* (Oxford: Oxford University Press, 2016), https://academic.oup.com/book/25649; Daron Acemoglu and Simon Johnson, *Power and Progress: Our 1000-year Struggle Over Technology and Prosperity* (New York: Public Affairs, 2024), https://www.hachettebookgroup.com/titles/daron-acemoglu/power-and-progress/9781541702547/?lens=publicaffairs.

[88] "The Long-Term Benefit Trust," Anthropic, September 19, 2023, https://www.anthropic.com/news/the-long-term-benefit-trust.

[89] Anton Korinek and Jai Vipra, "Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence" (National Bureau of Economic Research, November 2024), https://www.nber.org/papers/w33139. For a different perspective, see Matt Chessen and Craig Martell, "Beyond a Manhattan Project for Artificial General Intelligence," Lawfare, Tuesday, April 22, 2025, https://www.lawfaremedia.org/article/beyond-a-manhattan-project-for-artificial-general-intelligence.

[90] Martin Watzinger, Thomas A. Fackler, Markus Nagler, and Monika Schnitzer, "How Antitrust Enforcement Can Spur Innovation: Bell Labs and the 1956 Consent Decree," January 9, 2017, https://economics.yale.edu/sites/default/files/how_antitrust_enforcement.pdf.

[91] Cem Anil, Esin Durmus, Mrinank Sharma et al., "Many-shot Jailbreaking," Anthropic, April 2, 2024, https://www.anthropic.com/research/many-shot-jailbreaking.

[92] Jack Clark, Import AI 387: Overfitting vs reasoning; distributed training runs; and Facebook's new video models," Import AI, https://jack-clark.net/2024/10/14/import-ai-387-overfitting-vs-reasoning-distributed-training-runs-and-facebooks-new-video-models/.

[93] Robert Jervis, *Systems Effects: Complexity in Political and Social Life* (Princeton: Princeton University Press, 1997), https://press.princeton.edu/books/paperback/9780691005300/system-effects?srsltid=AfmBOoooav5mLp4SG4hUWSvYQ_Bx8OtWZcJmQlyp7X2d3IAjDxUsvllU.

[94] Krizhevsky, Sutskever, and Hinton, "ImageNet Classification."

[95] Christian Szegedy, Wei Liu, Yangqing Jia et al., "Going Deeper with Convolutions," arXiv preprint arXiv:1409.4842 (2014), https://arxiv.org/abs/1409.4842.

[96] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," arXiv preprint arXiv:1512.03385 (2015), https://arxiv.org/abs/1512.03385.

[97] H.D. Block, Bruce Knight, and Frank Rosenblatt, "Analysis of a Four-Layer Series-Coupled Perceptron. II" Digital Commons @ RU, 1962, https://digitalcommons.rockefeller.edu/cgi/viewcontent.cgi?article=1013&context=knight_laboratory.

[98] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation* 9, no. 8 (November 1997): 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[99] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, "Sequence to Sequence Learning with Neural Networks," arXiv preprint arXiv:1409.3215 (2014), https://arxiv.org/abs/1409.3215.

[100] Vaswani et al., "Attention Is All You Need."

[101] Tom B. Brown, Benjamin Mann, Nick Ryder et al., "Language Models are Few-Shot Learners," arXiv preprint arXiv:2005.14165 (2020), https://arxiv.org/abs/2005.14165.

[102] "Text Generation and Prompting," OpenAI Platform, OpenAI, accessed February 11, 2025, https://platform.openai.com/docs/guides/prompt-engineering.

[103] Jason Wei, Xuezhi Wang, Dale Schuurmans et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv preprint arXiv:2201.11903 (2022), https://arxiv.org/abs/2201.11903.

[104] "How to Use Templates for ChatGPT Prompts (With Examples)," CareerCatalyst, Arizona State University, January 10, 2024, https://careercatalyst.asu.edu/newsroom/workforce-education/foundations-of-prompt-template-development/.

[105] Patrick Lewis, Ethan Perez, Aleksandra Piktus et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv preprint arXiv:2005.11401 (2020), https://arxiv.org/abs/2005.11401.

[106] "What are AI hallucinations?," Google Cloud, accessed February 11, 2025, https://cloud.google.com/discover/what-are-ai-hallucinations?hl=en#how-do-ai-hallucinations-occur.

[107] Patrice Béchard and Orlando Marquez Ayala, "Reducing hallucination in structured outputs via Retrieval-Augmented Generation," arXiv preprint arXiv:2404.08189 (2024), https://arxiv.org/abs/2404.08189.

[108] "LLM02:2023 - Data Leakage," OWASP Top 10 for Large Language Model Applications, Open Web Application Security Project, accessed February 11, 2025, https://owasp.org/www-project-top-10-for-large-language-model-applications/Archive/0_1_vulns/Data_Leakage.html.

[109] Shenglai Zeng, Jiankun Zhang, Pengfei He et al., "The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG)," arXiv preprint arXiv:2402.16893 (2024), https://arxiv.org/abs/2402.16893.

[110] Jeanina Matias, "What is a Foundation Model? An Explainer for Non-Experts" (Institute for Human-Centered AI, Stanford University, May 10, 2023), https://hai.stanford.edu/news/what-foundation-model-explainer-non-experts.

[111] Nelson F. Liu, Kevin Lin, John Hewitt et al., "Lost in the Middle: How Language Models Use Long Contexts," arXiv preprint arXiv:2307.03172 (2023), https://arxiv.org/abs/2307.03172.

[112] Roie Schwaber-Cohen, "Chunking Strategies for LLM Applications," Pinecone, June 30, 2023, https://www.pinecone.io/learn/chunking-strategies/.

[113] "Introducing Contextual Retrieval," Anthropic, September 19, 2024, https://www.anthropic.com/news/contextual-retrieval.

[114] Chong Xiang, Tong Wu, Zexuan Zhong et al., "Certifiably Robust RAG against Retrieval Corruption," arXiv preprint arXiv:2405.15556 (2024), https://arxiv.org/abs/2405.15556.

[115] Zhepei Wei, Wei-Lin Chen, and Yu Meng, "InstructRAG: Instructing Retrieval-Augmented Generation via Self-Synthesized Rationales," arXiv preprint arXiv:2406.13629 (2024), https://arxiv.org/abs/2406.13629.

[116] Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arık, "Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models," arXiv preprint arXiv:2410.07176 (2024), https://arxiv.org/abs/2410.07176.

[117] Boci Peng, Yun Zhu, Yongchao Liu et al., "Graph Retrieval-Augmented Generation: A Survey," arXiv preprint arXiv:2408.08921 (2024), https://arxiv.org/abs/2408.08921.

[118] Aidan Hogan, Eva Blomqvist, Michael Cochez et al., "Knowledge Graphs," arXiv preprint arXiv:2003.02320 (2020), https://arxiv.org/abs/2003.02320.

[119] Bhaskarjit Sarmah, Benika Hall, Rohan Rao et al., "HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction," arXiv preprint arXiv:2408.04948 (2024), https://arxiv.org/abs/2408.04948.